North Dakota State Assessment for English Language Arts/Literacy and Mathematics

2020–2021

Volume 4 Evidence of Reliability and Validity

ACKNOWLEDGMENTS

This technical report was produced on behalf of the North Dakota Department of Public Instruction. Requests for additional information concerning this technical report or the associated appendices should be directed to the North Dakota Department of Public Instruction at seschauer@nd.gov.

Major contributors to this technical report from Cambium Assessment, Inc., include the following staff: Trevor Bartlett, Youngmi Cho, Kevin Clayton, Aida Diaz, Mattie MacDonald, Seyfullah Tingir, Pamela Trantham, and Ahmet Turhan. Major contributors from the North Dakota Department of Public Instruction include Stanley Schauer and Bonnie Weisz.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

1. Introduction and Overview of Reliability and Validity Evidence

North Dakota implemented a new assessment program for operational use during the 2017–2018 school year. This new program, named the North Dakota State Assessment (NDSA), replaced the Smarter Balanced Assessment Consortium (SBAC) in English language arts/literacy (ELA/L) and mathematics. The NDSA was previously delivered as an online, fixed-form assessment but starting in the 2020–2021 school year, the State began delivering it as an online, adaptive assessment. The accommodated versions were available to students whose Individualized Education Plans (IEPs) or Section 504 Plans indicated such a need. Table 1 displays the complete list of test administration methods for the 2020–2021 school year.

Table 1: Test Administration

Subject (language/format)	Administration Mode	Grades
ELA (English/adaptive)	Online	3–8, 10
ELA (English/fixed-braille)	Paper	3–8, 10
Mathematics (English/adaptive)	Online	3–8, 10
Mathematics (Spanish/adaptive)	Online	3–8, 10
Mathematics (English/fixed-braille)	Paper	3–8, 10

* *Accommodated versions, including braille and print-on-demand, are delivered on paper. Full descriptions of available accommodations are listed in Volume 5 Section 1.2. The number of students who were provided with accommodations is presented in Volume 1 Section 2.2.*

Given the intended uses of these tests, both reliability evidence and validity evidence are necessary to support appropriate inferences of student academic achievement from the NDSA scores. The purpose of this volume is to provide empirical evidence to support a validity argument regarding the uses and inferences for the NDSA. This volume addresses the following:

- **Reliability.** The reliability of the NDSA adaptive test forms is estimated using marginal reliability in the item response theory (IRT) framework. The reliability estimates are presented by grade and subject and demographic subgroup. This discussion also includes conditional standard error of measurement (CSEM), the reliability of performance classifications, and inter-rater reliability (IRR) of ELA writing scores provided by Cambium Assessment, Inc.'s (CAI) AutoScoring Model.
- **Content validity.** Evidence is provided to show that the test forms constructed to measure North Dakota's educational standards contained a sufficient number of items targeting each area of the blueprint.
- **Internal structure validity**. Evidence is provided regarding the internal relationships among the subscale scores to support their use and how they support using the IRT measurement model. This type of evidence includes observed and disattenuated Pearson correlations among reporting categories per grade. Confirmatory factor analysis (CFA) has also been performed using the second-order factor model. Additionally, local item

independence, an assumption of unidimensional IRT, was evaluated using the Q_3 statistic in spring 2019. The CFA and Q_3 statistics were kept as a reference in this document.

- **Relationship of Test Scores to External Variables.** Evidence of convergent and discriminant validity is provided using observed and disattenuated subscore correlations both within and across subjects. The correlations between the spring 2019 and spring 2021 NDSA summative assessments in ELA/L and mathematics are also presented.

- **Test fairness.** Specialists use content alignment reviews and differential item functioning (DIF) to statistically analyze fairness.

1.1 RELIABILITY

Reliability refers to consistency in test scores. Reliability can be defined as the degree to which an individual's deviation scores remain relatively consistent over repeated administrations of the same test or alternate test forms (Crocker & Algina, 1986). For example, if a student takes the same or parallel tests repeatedly, he or she should receive consistent results. The reliability coefficient refers to the ratio of true score variance to observed score variance:

$$\rho_{XX\prime} = \frac{\sigma_T^2}{\sigma_X^2}.$$

Another way to view reliability is to consider its relationship with the standard error of measurement (SEM): the smaller the standard error, the higher the precision of the test scores. For example, classical test theory (CTT) assumes that an observed score (X) of an individual can be expressed as a true score (T) plus some error (E), $X = T + E$. The variance of X can be shown to be the sum of two orthogonal variance components:

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2$$

Returning to the definition of reliability as the ratio of true score variance to observed score variance, we can arrive at the following theorem:

$$\rho_{XX\prime} = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_x^2 - \sigma_E^2}{\sigma_X^2} = 1 - \frac{\sigma_E^2}{\sigma_X^2}$$

As the fraction of error variance to observed score variance tends to zero, the reliability then tends to 1. The CTT SEM, which assumes a homoscedastic error, is derived from the classical notion expressed above as $\sigma_X\sqrt{1 - \rho_{XX'}}$, where σ_X is the standard deviation of the scaled score, and $\rho_{XX'}$ is a reliability coefficient. Based on the definition of reliability, this formula can be derived as follows:

$$\rho_{XX'} = 1 - \frac{\sigma_E^2}{\sigma_X^2},$$

$$\frac{\sigma_E^2}{\sigma_X^2} = 1 - \rho_{XX'},$$

$$\sigma_E^2 = \sigma_X^2(1 - \rho_{XX'}),$$

$$\sigma_E = \sigma_X\sqrt{(1 - \rho_{XX'})}.$$

In general, the SEM is relatively constant across samples, as the group dependent term, σ_X, can be shown to cancel out:

$$\sigma_E = \sigma_X\sqrt{(1 - \rho_{XX'})} = \sigma_X\sqrt{(1 - (1 - \frac{\sigma_E^2}{\sigma_X^2}))} = \sigma_X\sqrt{\frac{\sigma_E^2}{\sigma_X^2}} = \sigma_X \cdot \frac{\sigma_E}{\sigma_X} = \sigma_E.$$

This equation shows that the SEM in the CTT is assumed to be a homoscedastic error, irrespective of the standard deviation of a group.

In contrast, the SEM in IRT varies over the ability continuum. These heterogeneous errors are a function of a test information function (TIF) that provides different information about test takers depending on their estimated abilities. Often, the TIF is maximized over an important performance cut, such as the proficiency cut score.

Because the TIF indicates the amount of information provided by the test at different points along the ability scale, its inverse indicates the lack of information at different points along the ability scale. This lack of information is the uncertainty, or the measurement error, of the score at various score points. Conventionally, fixed-form tests are maximized near the middle of the score distribution or near an important classification cut and have less information at the tails of the score distribution. Refer to Section 3.2, Test Information Curves and Standard Error of Measurement, of this volume for the derivation of heterogeneous errors in IRT.

1.2 VALIDITY

Validity refers to the degree to which "evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME], 2014). Messick (1989) defines validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores and other modes of assessment." These definitions emphasize the evidence and theory that support the inferences and interpretations of test scores. *The Standards* (AERA, APA, & NCME, 2014) suggest five sources of validity evidence that can be used in evaluating a proposed interpretation of test scores. When validating test scores, these sources of evidence should be carefully considered.

The first source of validity evidence is the relationship between the test content and the intended test construct (refer to Section 4, Evidence of Content Validity, for details). For test score inferences to support a validity claim, the items should be representative of the content domain, and the content domain should be relevant to the proposed interpretation of the test scores. To determine content representativeness, diverse panels of content experts conduct alignment studies. During these studies, experts review individual items and rate them based on how well they match the test specifications or cognitive skills required for a particular construct (refer to Volume 2: Test Development of this technical report for details). Test scores can be used to support an intended validity claim when they contain minimal construct-irrelevant variance. For example, a

mathematics item targeting a specific mathematics skill that requires advanced reading proficiency and vocabulary has a high level of construct-irrelevant variance. Thus, the intended construct of measurement is confounded, which impedes the validity of the test scores.

Statistical analyses, such as factor analysis or multi-dimensional scaling, are also used to evaluate content relevance. The results from factor analysis for the fixed-form spring 2018 NDSA for ELA and mathematics are presented in Section 5.2, Confirmatory Factor Analysis. Factor analysis was not possible for spring 2021 due to the switch to computer-adaptive testing. Evidence based on test content is a crucial component of validity because construct underrepresentation or irrelevancy can result in unfair advantages or disadvantages to one or more groups of test takers.

In addition, technology-enhanced items should be examined to ensure that no construct-irrelevant variance was introduced. If any aspect of the technology impedes, or creates an advantage for a student in his or her responses to items, this could affect item responses and inferences regarding that student's abilities on the measured construct (refer to Volume 2 Section 5.1 of this technical report for details).

The second source of validity evidence is based on "the fit between the construct and the detailed nature of performance or response actually engaged in by examinees" (AERA, APA, & NCME, 2014). This evidence is collected by surveying test takers about their performance strategies or responses to particular items. Because items are developed to measure particular constructs and intellectual processes, evidence that test takers have engaged in relevant performance strategies to answer the items correctly supports the validity of the test scores.

The third source of validity evidence is based on internal structure: the degree to which the relationships among test items and test components relate to the construct on which the proposed test scores are interpreted. DIF, which determines whether particular items may function differently for subgroups of test takers, is one method for analyzing the internal structure of tests (refer to Volume 1, Section 5.2, of this technical report for details). Other possible analyses to examine internal structure are dimensionality assessment, goodness-of-fit model to data, and reliability analysis (refer to Sections 3 and 5 of this volume for details).

The fourth source of validity evidence is the relationship of test scores to external variables. *The Standards* (AERA, APA, & NCME, 2014) divides this source of evidence into three parts: convergent and discriminant evidence, test-criterion relationships, and validity generalization. Convergent evidence supports the relationship between the test and other measures intended to assess similar constructs. Conversely, discriminant evidence delineates the test from other measures designed to assess different constructs. A multi-trait-multi-method matrix can be used to analyze both convergent and discriminant evidence (refer to Section 5.4, Convergent and Discriminant Validity, for details).

Additionally, test-criterion relationships indicate how accurately test scores predict criterion performance. The degree of accuracy mainly depends upon the test's purpose, such as classification, diagnosis, or selection. Test-criterion evidence is also used to investigate predictions of favoring different groups. Due to construct underrepresentation or construct-irrelevant components, the relation of test scores to a relevant criterion may differ from one group to another. Furthermore, validity generalization is related to whether the evidence is situation-specific or can be generalized across different settings and times. For example, sampling errors or range

restrictions may need to be considered to determine whether the conclusions of a test can be assumed for the larger population.

The fifth source of validity evidence should include the intended and unintended consequences of test use in the test-validation process. Determining the validity of the test should depend upon evidence directly related to the test and should not be influenced by external factors. For example, if an employer administers a test to determine the hiring rates for different groups of people, an unequal distribution of skills related to the measurement construct does not necessarily imply a lack of validity for the test. However, if the unequal distribution of scores is due to an unintended, confounding aspect of the test, that aspect would interfere with the test's validity. As described in Volume 1 and in this volume of the technical report, test use should align with the test's intended purpose.

Supporting a validity argument requires multiple sources of validity evidence. Multiple sources of validity evidence allow for an evaluation of whether sufficient evidence has been presented to support the test scores' intended uses and interpretations. Thus, determining test validity requires an explicit statement regarding the intended uses of the test scores first, and subsequently, evidence that the scores can be used to support these inferences. Ultimately, judgments about validity are tied to use and context and are a matter of degree; that is, validity is not a dichotomous condition and is not an immutable characteristic of the assessment.

2. PURPOSE OF THE NORTH DAKOTA STATE ASSESSMENTS

The primary purpose of the NDSA is to yield test scores at the student level and other levels of aggregation that reflect student achievement relative to the North Dakota Content Standards. NDSA supports instruction and student learning by measuring growth in student achievement and providing immediate feedback to educators and parents that can be used to form instructional strategies to remediate or enrich instruction. Assessments can be used as an indicator to determine whether students in North Dakota have the knowledge and skills essential for college and career readiness.

North Dakota's educational assessments also provide evidence of the requirements for state and federal accountability systems. Test scores can be employed to evaluate students' learning progress and help teachers improve their instruction, which has a positive effect on student learning over time.

The tests are constructed to measure student proficiency on the North Dakota Content Standards in ELA/L and mathematics. The test was developed using principles of evidence-centered design and adherence to the principles of universal design to ensure that all students have access to the test content. Volume 2, Test Development, of this technical report describes the North Dakota Content Standards and test blueprints in more detail. This volume provides evidence of content validity in Section 4. The NDSA test scores are useful indicators for understanding individual students' academic achievement of the North Dakota Content Standards and whether students' performance is progressing over time. Additionally, individual test scores can be used to measure test reliability, which can be found in Section 3, Reliability.

The NDSA is a criterion-referenced test designed to measure student performance on the North Dakota Content Standards in ELA/L and mathematics. As a comparison, norm-referenced tests are designed to compare or rank all students to one another.

The scale score and relative strengths and weaknesses at the reporting category (domain) level were provided for each student to indicate student strengths and weaknesses in various content areas of the test relative to other areas and to the district and state. These scores serve as useful feedback for teachers to tailor their instruction, provided that they are viewed with the usual caution that accompanies using reporting category scores. Thus, we must examine the reliability coefficients for these test scores and the validity of the test scores to support practical use across the state. Volume 6 of this technical report is the score interpretation guide and provides details on all scores generated and their appropriate uses and limitations.

3. RELIABILITY

3.1 RELIABILITY FOR ELA AND MATHEMATICS

The NDSA ELA/L and mathematics testing administrations are computer-adaptive tests (CATs). Because there is no set form in adaptive testing, marginal reliability was computed for the scale scores, taking into account the varying measurement errors across the ability range. Marginal reliability is a measure of the overall reliability of an assessment based on the average conditional

standard error of measurement (CSEM), estimated at different points on the ability scale for all students.

Marginal reliability ($\overline{\rho}$) is defined as

$$\overline{\rho} = [\sigma^2 - \left(\frac{\sum_{i=1}^{N} CSEM_i^2}{N}\right)]/\sigma^2,$$

where N is the number of students; $CSEM_i$ is the CSEM of the theta score for student *i*, and σ^2 is the variance of the theta score. The higher the reliability coefficient, the greater the precision of the test. Table 2 presents the marginal reliability coefficients for all students. The reliability coefficients for all subjects and grades range from 0.86–0.92. Appendix A: Reliability Coefficients provides a further breakdown, including reliability coefficients for demographic subgroups and reporting categories.

Table 2: Marginal Reliability Coefficients: ELA and Mathematics

Subject	Grade	Reliability	Subject	Grade	Reliability
ELA	3	0.88	Mathematics	3	0.92
	4	0.88		4	0.92
	5	0.90		5	0.89
	6	0.89		6	0.89
	7	0.88		7	0.88
	8	0.89		8	0.90
	10	0.89		10	0.86

3.2 Test Information Curves and Standard Error of Measurement

Within the item response theory (IRT) framework, measurement error varies across the range of ability as a result of the test, providing varied information across the range of ability as displayed by the test information function (TIF). The TIF describes the amount of information provided by the test at each score point along the ability continuum. The inverse of the TIF is characterized as the conditional measurement error at each score point. For instance, if the measurement error is large, then less information is being provided by the assessment at the specific ability level.

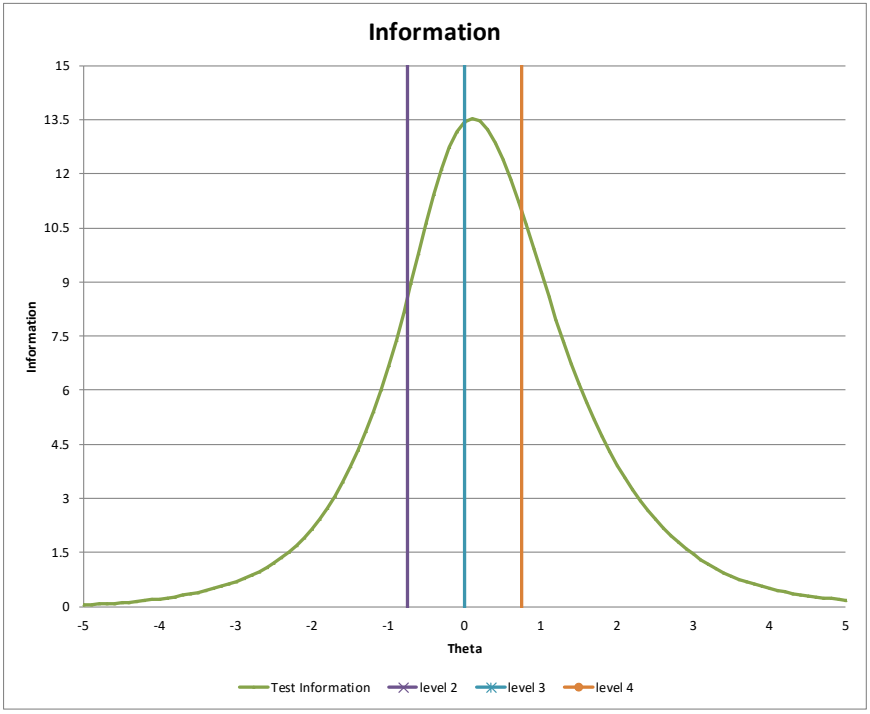
Figure 1 displays a sample TIF with three vertical lines indicating the performance cuts. The graphic shows that this test information is maximized in the middle of the score distribution, meaning it provides the most precise scores in this range. The curve is lower at the tails indicating that the test provides less information about test takers at the tails relative to the center.

Computing these TIFs is useful to evaluate where the test is maximally informative. In IRT, the TIF is based on the estimates of the item parameters in the test, and the formula used for the NDSA is calculated as:

$$TIF(\theta_s) = \sum_{i=1}^{N_{GPCM}} D^2 a_i^2 \left(\frac{\sum_{h=1}^{m_i} h^2 exp\left(\sum_{l=1}^{h} Da_i(\theta_s - b_{il})\right)}{1 + \sum_{h=1}^{m_i} exp\left(\sum_{l=1}^{h} Da_i(\theta_s - b_{il})\right)} - \left(\frac{\sum_{h=1}^{m_i} h exp\left(\sum_{l=1}^{h} Da_i(\theta_s - b_{il})\right)}{1 + \sum_{h=1}^{m_i} exp\left(\sum_{l=1}^{h} Da_i(\theta_s - b_{il})\right)} \right)^2 \right) + \sum_{i=1}^{N_{3PL}} D^2 a_i^2 \left(\frac{q_i}{p_i} \left[\frac{p_i - c_i}{1 - c_i} \right]^2 \right),$$

where N_{GPCM} is the number of items scored using generalized partial credit model (GPCM) items, N_{3PL} is the number of items scored using 3PL or 2PL model, i indicates item i ($i \in \{1,2,\ldots,N\}$), m_i is the maximum possible score of the item, s indicates student s, and θ_s is the ability of student s.

Figure 1: Sample Test Information Function

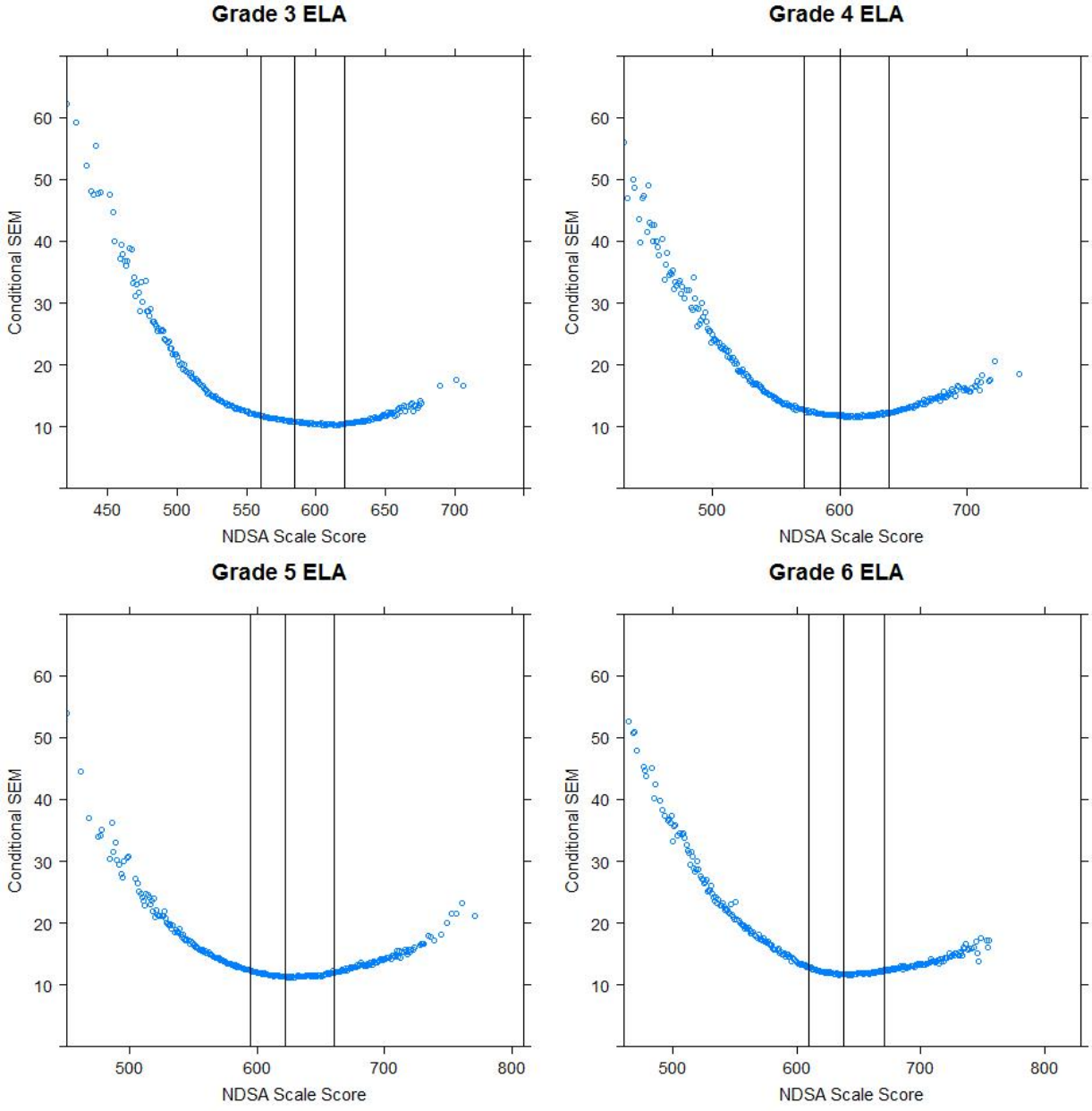


The standard error for estimated student ability (theta score) is the square root of the reciprocal of the TIF:

$$se(\theta_s) = \frac{1}{\sqrt{TIF(\theta_s)}}.$$

It is typically more useful to consider the inverse of the TIF rather than the TIF itself, as the standard errors are more useful for score interpretation. For this reason, standard error plots are presented in Figure 2 and Figure 3 respectively, instead of the TIFs for ELA and mathematics. These plots are based on the scaled scores reported in 2021. Vertical lines represent the three achievement category cut scores.

Figure 2: Conditional Standard Errors of Measurement (ELA)



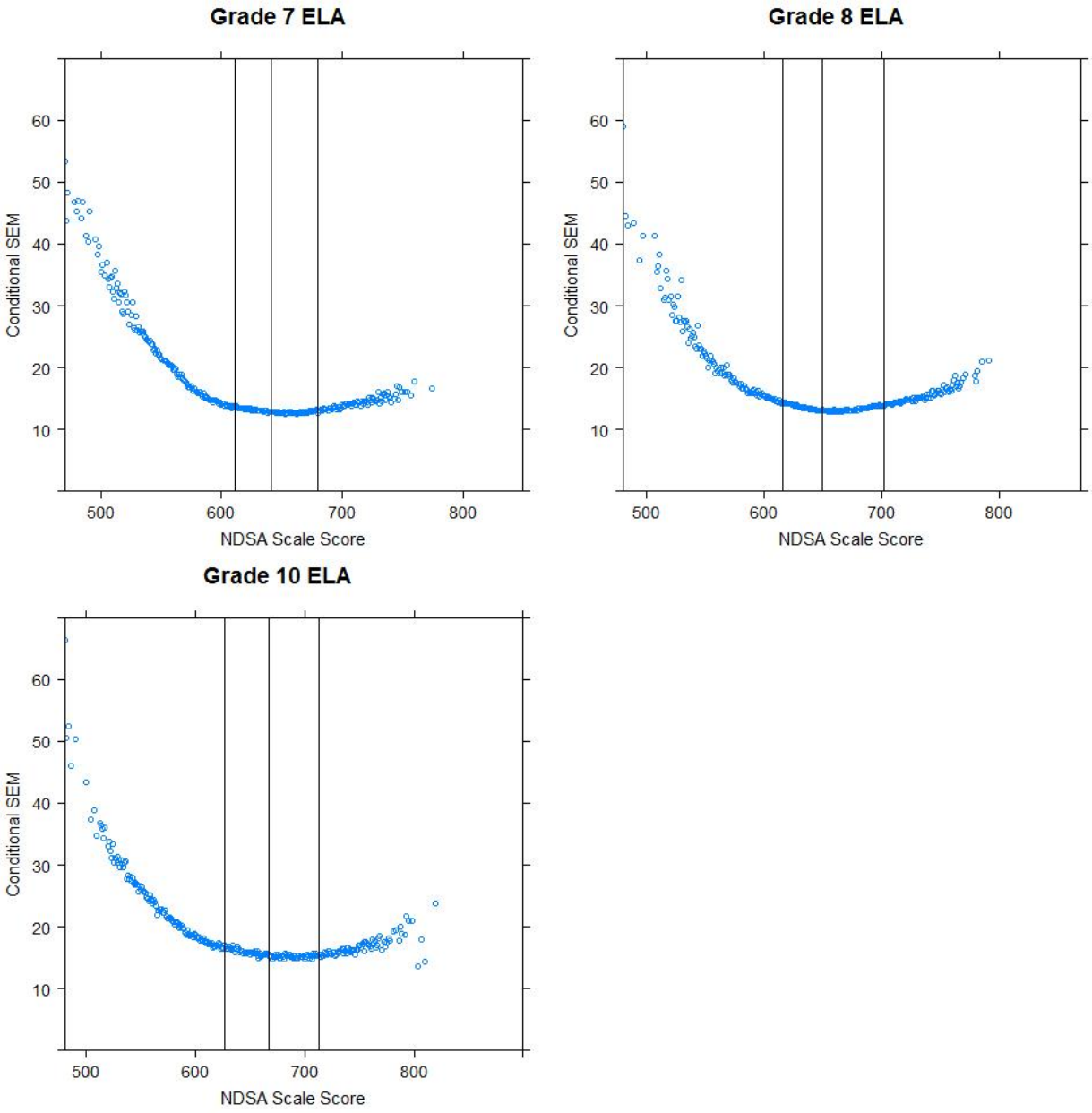
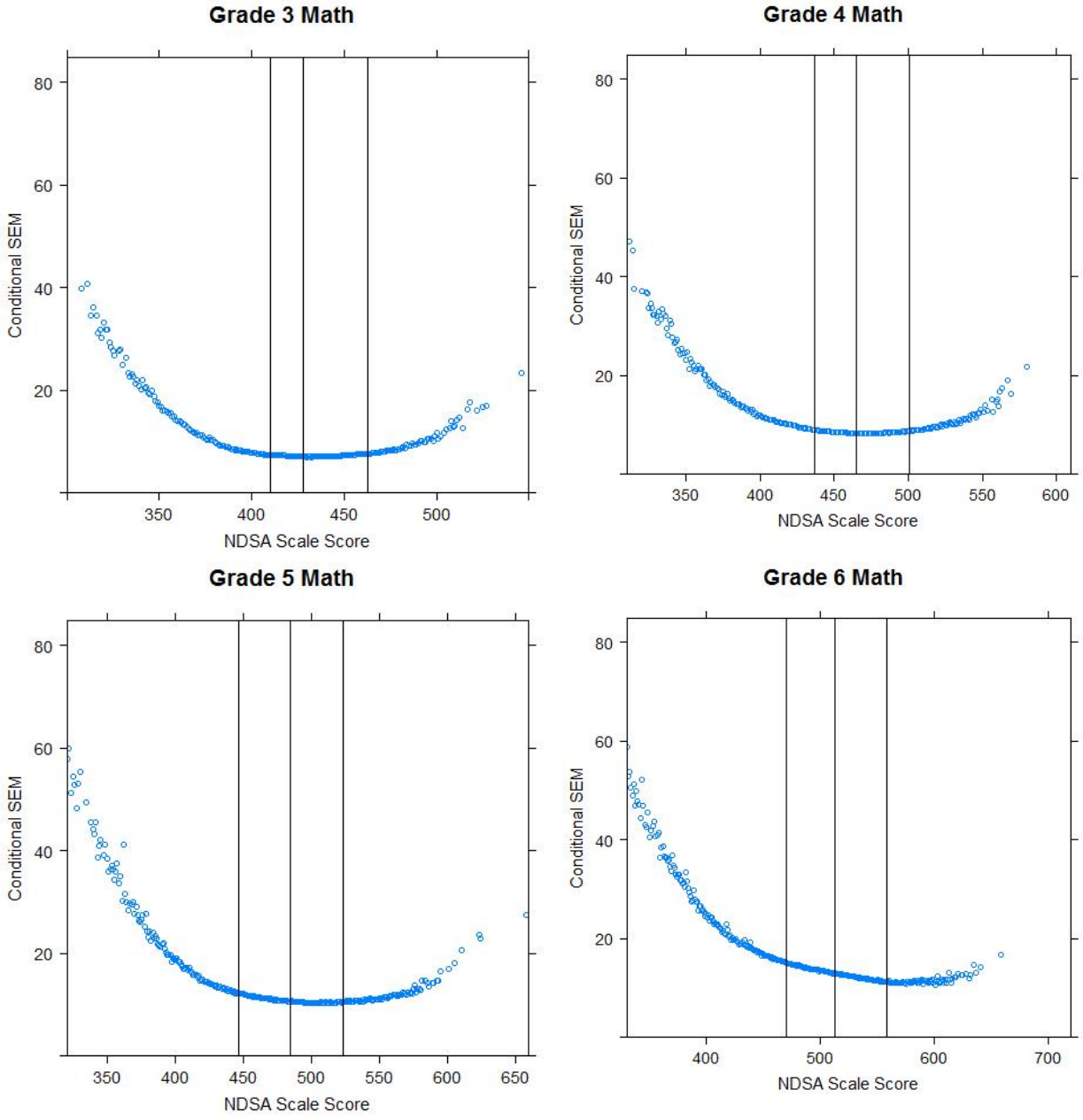
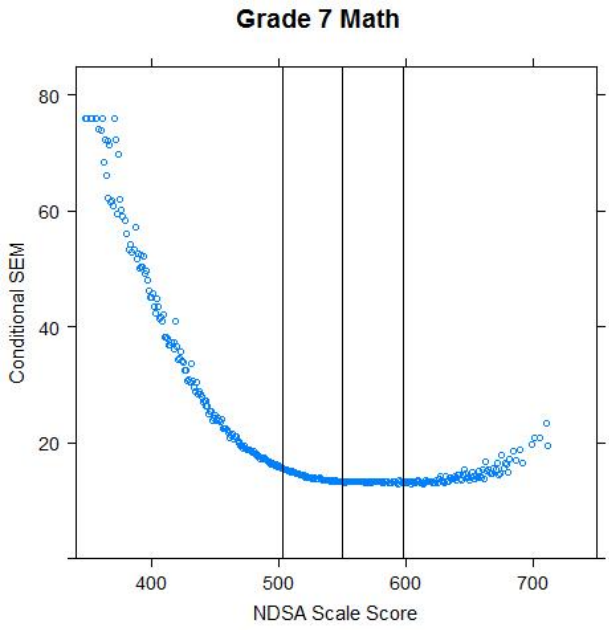
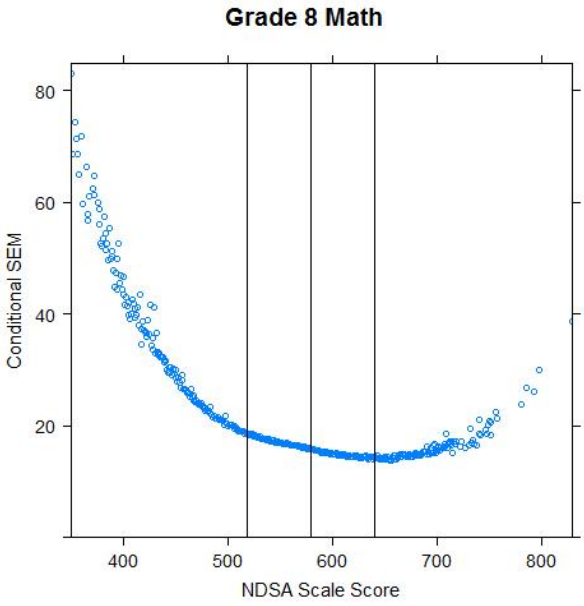
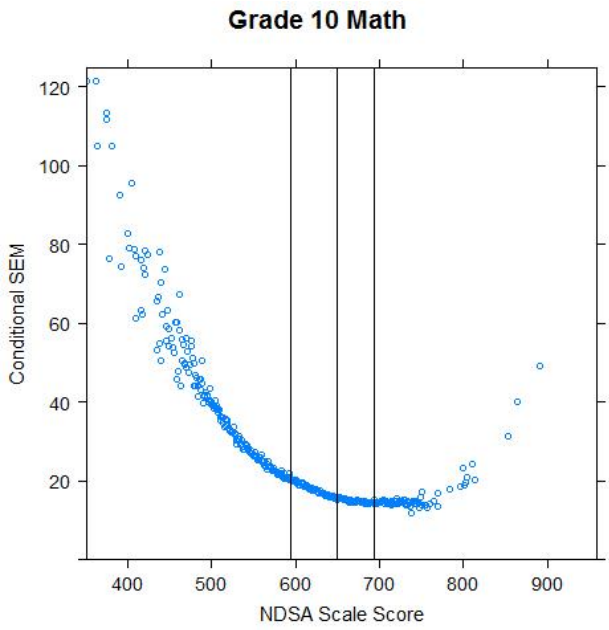

Grade 7 ELA
Conditional SEM
NDSA Scale Score
Grade 8 ELA
Conditional SEM
NDSA Scale Score
Grade 10 ELA
Conditional SEM
NDSA Scale Score

Figure 3: Conditional Standard Errors of Measurement (Mathematics)









The CSEM curves follow the typical expected trends, with the smallest values observed near the middle of the score scale. Desirably, the lowest SEMs are observed at the proficiency cut (the middle vertical line between Partially Meets Standard and Meets Standard score ranges) for most tests.

Reliability coefficients and SEM for each reporting category are also presented in Appendix A, and Appendix B includes scale score by scale score CSEM and the corresponding achievement levels for each scale score.

3.3 RELIABILITY OF ACHIEVEMENT CLASSIFICATION

When students complete the NDSA, they are placed into one of four achievement levels based on their observed scaled score. The reliability of classifying students into a specific level can be computed in terms of the likelihood of accurate and consistent classification as specified in Standard 2.16 in *The Standards for Educational and Psychological Testing* (AERA, APA, NCME, 2014).

The reliability of achievement classification can be examined in terms of classification accuracy and classification consistency. Classification accuracy refers to the degree to which a student's true score and observed score would fall within the same achievement level (Rudner, 2001). Classification consistency refers to the degree to which test takers are classified into the same achievement level, assuming the test is administered twice independently (Lee, Hanson, and Brennan, 2002)—that is, the percentages of students consistently classified in the same achievement levels on two equivalent test forms. In reality, the true ability is unknown, and students do not take an alternate, equivalent form; therefore, classification consistency is estimated based on students' item scores, the item parameters, and the assumed underlying latent ability distribution.

For student *j*, the student's estimated ability is $\hat{\theta}_j$ with SEM of $se(\hat{\theta}_j)$, and the estimated ability is distributed as $\hat{\theta}_j \sim N\left(\theta_j, se^2(\hat{\theta}_j)\right)$, assuming a normal distribution, where θ_j is the unknown true ability of student *j*. The probability of the true score at performance level l $(l = 1, \cdots, L)$ is estimated as

$$p_{jl} = p(c_{Ll} \le \theta_i < c_{Ul}) = p\left(\frac{c_{Ll} - \hat{\theta}_j}{se(\hat{\theta}_j)} \le \frac{\theta_j - \hat{\theta}_j}{se(\hat{\theta}_j)} < \frac{c_{Ul} - \hat{\theta}_j}{se(\hat{\theta}_j)}\right)$$
$$= p\left(\frac{\hat{\theta}_j - c_{Ul}}{se(\hat{\theta}_j)} < \frac{\hat{\theta}_j - \theta_j}{se(\hat{\theta}_j)} \le \frac{\hat{\theta}_j - c_{Ll}}{se(\hat{\theta}_j)}\right) = \Phi\left(\frac{\hat{\theta}_j - c_{Ll}}{se(\hat{\theta}_j)}\right) - \Phi\left(\frac{\hat{\theta}_j - c_{Ul}}{se(\hat{\theta}_j)}\right),$$

where c_{Ll} and c_{Ul} denote the score corresponding to the lower and upper limits of the performance level l, respectively.

Classification accuracy and consistency by achievement level for all students and subgroups are shown side by side for comparison in Appendix C.

3.3.1 Classification Accuracy

Using p_{jl}, the expected number of students at level l based on students from observed level *k* can be expressed as

$$E_{Akl} = \sum_{pl_j \in k} p_{jl},$$

where pl_j is the *j*th student's performance level, the values of E_{Akl} are the elements used to populate the matrix E_A, an $L \times L$ matrix of conditionally expected numbers of students to score

within each performance level based on their true scores. The classification accuracy (*CA*) at level l is estimated by

$$CA_l = \frac{E_{Akl}}{N_k},$$

where N_k is the observed number of students scoring in performance level k.

The classification accuracy for the pth cut (*CAC*) is estimated by forming square partitioned blocks of the matrix E_A and taking the summation over all elements within the block as follows:

$$CAC = \left(\sum_{k=1}^{p} \sum_{l=1}^{p} E_{Akl} + \sum_{k=p+1}^{L} \sum_{l=p+1}^{L} E_{Akl} \right) \Big/ N,$$

where N is the total number of students.

The overall classification accuracy is estimated from the diagonal elements of the matrix:

$$CA = \frac{tr(\boldsymbol{E_A})}{N}.$$

Table 3 and Table 4 provide the overall classification accuracy and the classification accuracy for the individual cuts for ELA and mathematics, respectively. The overall classification accuracy of the test ranges from 77.14%–78.86% for ELA and from 79.85% to around 82.43% for mathematics. The cut accuracy rates are high across all grades and subjects, with a minimum value of 90.96% for ELA and 90.87% for mathematics. This denotes that the degree to which we can reliably differentiate students between adjacent performance levels is typically above 90%.

Table 3: Classification Accuracy Index (ELA)

Grade	Overall Accuracy (%)	Cut Accuracy (%)		
		Cut 1	Cut 2	Cut 3
3	78.62	91.15	91.66	95.69
4	78.48	91.10	91.66	95.62
5	78.29	92.10	91.48	94.61
6	77.14	92.33	90.96	93.72
7	77.42	91.33	90.96	95.01
8	78.86	92.69	91.16	94.97
10	78.31	92.11	91.90	94.24

Table 4: Classification Accuracy Index (Mathematics)

Grade	Overall Accuracy (%)	Cut Accuracy (%)		
		Cut 1	Cut 2	Cut 3
3	82.15	93.46	92.42	96.23

Grade	Overall Accuracy (%)	Cut Accuracy (%)		
		Cut 1	Cut 2	Cut 3
4	82.43	92.71	92.89	96.83
5	79.85	92.95	91.03	95.86
6	79.87	92.31	90.87	96.69
7	80.56	92.77	91.40	96.38
8	81.77	92.98	91.65	97.13
10	80.07	91.37	92.30	96.36

3.3.2 Classification Consistency

Assuming the test is independently administered twice to the same students, similar to accuracy, an $L \times L$ matrix $\boldsymbol{E_C}$ can be constructed. The element of $\boldsymbol{E_C}$ is populated by

$$E_{Ckl} = \sum_{j=1}^{N} p_{jl} p_{jk},$$

where p_{jl} is the probability of the true score at performance level l in Test 1 and p_{jk} is the probability of the true score at performance level k in Test 2, for the jth student. The classification consistency index for the cuts (*CCC*) and overall classification consistency (*CC*) were estimated similarly to CAC and CA.

$$CCC = \left(\sum_{k=1}^{p} \sum_{l=1}^{p} E_{Ckl} + \sum_{k=p+1}^{L} \sum_{l=p+1}^{L} E_{Ckl} \right) \Big/ N,$$

and

$$CC = \frac{tr(\boldsymbol{E_C})}{N}.$$

Table 5 and Table 6 provide the classification consistency for the overall and individual cuts for ELA and mathematics, respectively. The overall classification consistency of the test ranges from 68.55%– 70.53% for ELA and from 71.67%–75.41% for mathematics.

The individual cut consistency rates are high across all grades and subjects. The minimum values for each subject are 87.27% for ELA and 87.17% for mathematics. For all achievement levels, classification accuracy is slightly higher than classification consistency. Classification consistency rates can be lower than classification accuracy; the consistency is based on two tests with measurement errors, but the accuracy is based on one test with a measurement error and the true score.

Table 5. Classification Consistency Index (ELA)

Grade	Overall Consistency (%)	Cut Consistency (%)		
		Cut 1	Cut 2	Cut 3
3	70.48	87.50	88.25	93.84
4	70.08	87.35	88.14	93.77
5	70.03	88.86	87.97	92.44
6	68.55	89.11	87.29	91.18
7	68.87	87.72	87.27	92.94
8	70.53	89.66	87.48	92.88
10	69.97	88.88	88.52	91.93

Table 6. Classification Consistency Index (Mathematics)

Grade	Overall Consistency (%)	Cut Consistency (%)		
		Cut 1	Cut 2	Cut 3
3	75.11	90.75	89.24	94.69
4	75.41	89.79	89.96	95.55
5	71.75	90.09	87.37	94.15
6	71.67	89.07	87.17	95.21
7	72.59	89.72	87.88	94.83
8	74.20	90.01	88.22	95.89
10	72.23	87.85	89.16	94.86

3.4 PRECISION AT CUT SCORES

Table 7 and Table 8 present mean CSEM at each achievement level by grade and subject. These tables also include achievement-level cut scores and the associated CSEM. The NDSA test scores are somewhat more precise for test scores near the middle of the scale, especially around the *Proficient* performance standard cut. The tables also show that test scores remain precise even for students in the lowest and highest achievement levels.

Table 7: Achievement Levels and Associated CSEM (ELA)

Grade	Achievement Level	Mean CSEM	Cut Score (Scale Score)	CSEM at Cut Score
3	1	14.94	-	-
3	2	11.27	560	11.77
3	3	10.51	585	10.82
3	4	11.16	621	10.44
4	1	16.37	-	-

Grade	Achievement Level	Mean CSEM	Cut Score (Scale Score)	CSEM at Cut Score
4	2	12.16	572	12.64
4	3	11.80	600	11.88
4	4	13.37	639	12.22
5	1	14.98	-	-
5	2	11.73	595	12.32
5	3	11.49	622	11.46
5	4	13.11	661	12.04
6	1	17.06	-	-
6	2	12.11	610	12.90
6	3	11.89	638	11.69
6	4	13.03	671	12.47
7	1	17.21	-	-
7	2	13.22	611	13.56
7	3	12.79	641	12.94
7	4	13.72	680	12.78
8	1	17.32	-	-
8	2	13.63	616	14.35
8	3	13.27	650	13.18
8	4	14.74	702	13.92
10	1	21.23	-	-
10	2	15.98	627	17.03
10	3	15.24	667	15.38
10	4	16.18	713	15.44

Table 8: Achievement Levels and Associated CSEM (Mathematics)

Grade	Achievement Level	Mean CSEM	Cut Score (Scale Score)	CSEM at Cut Score
3	1	10.02	-	-
3	2	7.28	410	7.47
3	3	7.25	428	7.10
3	4	8.59	463	7.76
4	1	12.26	-	-
4	2	8.63	437	9.05
4	3	8.46	465	8.31
4	4	9.57	501	8.82

Grade	Achievement Level	Mean CSEM	Cut Score (Scale Score)	CSEM at Cut Score
5	1	18.09	-	-
5	2	11.36	446	12.31
5	3	10.52	484	10.71
5	4	11.2	523	10.58
6	1	20.78	-	-
6	2	13.97	470	15.11
6	3	12.23	513	13.09
6	4	11.28	558	11.25
7	1	25.91	-	-
7	2	14.11	503	15.57
7	3	13.31	550	13.36
7	4	13.61	598	13.19
8	1	28.26	-	-
8	2	17.09	519	18.64
8	3	15.03	580	15.87
8	4	14.97	640	14.48
10	1	35.07	-	-
10	2	17.85	594	20.53
10	3	15.09	650	15.19
10	4	15.04	693	15.19

3.5 WRITING PROMPTS INTER-RATER RELIABILITY

The 2020–2021 writing responses were scored using a combination of CAI's automated scoring engine, Autoscore, and handscoring. This section describes the engine, how the engine scores are combined with handscores, and the engine's performance on a held-out validation sample and during live scoring.

3.5.1 Automated Scoring Engine

CAI's automated scoring engine, Autoscore, uses a statistical process to evaluate writing prompts. It evaluates student essays against the same rubric used by human raters, uses a statistical process to analyze each essay, and assigns a score for each of the three dimensions. Autoscore's training and calibration process creates prompt-specific scoring models used for scoring responses for each prompt.

As previously noted, Autoscore analyzes response characteristics and human-provided scores and predicts what a human rater would do. The response characteristics are collected using features, which are then used to predict scores. Autoscore uses features associated with writing quality and response meaning. Writing quality features include measures of syntax, grammatical and

mechanical correctness, spelling correctness, text complexity, paragraphing quality, and sentence variation and quality. Measures of response meaning include using latent semantic analysis (LSA) and deep learning methods that consider not just the pattern of word frequencies in a response, but also the order of words in the response. LSA ignores word order but identifies key topics associated with the sets of words in a response. Deep learning methods use word order and sets of localized word patterns related to scores humans have assigned. Finally, in Autoscore, two models are built in parallel, and the outputs of these models are optimally combined to predict the response score. This approach allows for a more stable score estimate, similar to using two or more handscorers.

CAI uses approximately 2,000 responses to train and validate Autoscore performance. These responses are divided into three samples: train, ensemble, and held-out validation. The training sample is used to train competing models and to pick the best-performing model. The ensemble sample is used to estimate parameters of a categorical logistic regression (one-vs.-rest) using the probabilities from a model comprised of LSA features, writing features, and the logits from a deep learning model as inputs. Once the ensembling model parameters are estimated, the held-out validation data are scored, and the engine's performance is examined on these data. The engine is trained on the best-available score (the final, resolved score) coming out of the handscoring process described next.

The 2,000 responses were selected using stratified random sampling and scored by two human raters. Essay responses to the grades 3–7 writing prompts were sent to Measurement, Inc., and responses in grades 8 and higher were sent to Data Recognition Corporation for human scoring. Human raters were trained to score writing responses using anchor papers selected by content experts and finalized rubrics (Table 9) at a rangefinding meeting. Raters revisited anchor papers and rubrics to refamiliarize themselves with scoring, including a range of sample responses and scores.

Raters were assigned to groups. Training the raters occurred as the leader of each group read student responses aloud to raters; the raters independently referred back to the anchors and rubrics and shared what they thought the score for the particular response should be. If the decision among raters was unanimous, there was a brief discussion, and they moved to the next response. If the decision was not unanimous, the raters discussed the anchors and rubrics to reach a consensus.

Two trained raters scored each writing item response. When scores from Reader 1 and Reader 2 were not in exact agreement, the response was sent for resolution scoring by a team leader or scoring director. The final item score was based on the resolution score, when present, or on the initial read.

Table 9: Writing Rubrics

Dimension	Rubric	Score Points
Conventions	The response demonstrates an adequate command of basic conventions. The response may include the following: • Some minor errors in usage but no patterns of errors • Adequate use of punctuation, capitalization, sentence formation, and spelling	0,1,2

Evidence & Elaboration	The response provides thorough and convincing support, citing evidence for the controlling idea or main idea that includes the effective use of sources, facts, and details. The response includes most of the following: • Smoothly integrated, thorough, and relevant evidence, including precise references to sources • Effective use of a variety of elaborative techniques (including but not limited to definitions, quotations, and examples), demonstrating an understanding of the topic and text • Clear and effective expression of ideas, using precise language • Academic and domain-specific vocabulary clearly appropriate for the audience and purpose • Varied sentence structure, demonstrating language facility	1,2,3,4
Purpose, Focus, & Organization	The response is fully sustained and consistently focused within the purpose, audience, and task; and it has a clear controlling idea and effective organizational structure creating coherence and completeness. The response includes most of the following: • Strongly maintained controlling idea with little or no loosely related material • Skillful use of a variety of transitional strategies to clarify the relationships between and among ideas • Logical progression of ideas from beginning to end with a satisfying introduction and conclusion • Appropriate style and objective tone established and maintained	1,2,3,4

The statistics used to examine human-human agreement and Autoscore-human agreement were percentage exact agreement and quadratic weighted kappa (QWK). The percentage exact agreement is the total number of responses in which scores from both scorers are equal, divided by the number of responses scored twice. In addition to the percentage agreement rates, the QWK values were computed for the training sample and the validation sample for the writing prompts.

Cohen's kappa (Cohen, 1968) is an index of inter-rater agreement that accounts for the agreement that could be expected due to chance. This statistic can be computed as

$$K = \frac{P_o - P_c}{1 - P_c},$$

where P_o is the proportion of observed agreement, and P_c indicates the proportion of agreement by chance. Cohen's kappa treats all disagreement values with equal weights. QWK coefficients (QWK: Cohen, 1968), however, allow unequal weights, which can be used as a measure of validity. QWK coefficients were calculated using the following formula:

$$K_w = \frac{P'_o - P'_c}{1 - P'_c},$$

where

$$P'_o = \frac{\sum w_{ij} p_{oij}}{w_{max}},$$

$$P'_c = \frac{\sum w_{ij} p_{cij}}{w_{max}},$$

where p_{oij} is the proportion of the judgments observed in the ijth cell, p_{cij} is the proportion in the ijth cell expected by chance, and w_{ij} is the disagreement weight. QWK ranges from 0–1, where values of 0 indicate no agreement and values of 1 indicate perfect agreement.

Autoscore-human agreement was generally higher than human-human agreement for percentage exact agreement and QWK on the held-out validation sample (see *Table 10*). The agreement metrics were computed between the two human raters and between Autoscore and the final, resolved score. Because Autoscore is trained on and evaluated against a more reliable score (the final, resolved score), the agreement between Autoscore and the final, resolved score should generally be higher than that of two human raters. This result held true for both exact agreement and QWK. Using Williamson, Xi, & Breyer (2012) recommendations, we expect almost all item traits will be such that the Autoscore-HS QWK is no lower than .1 than the H1-H2 QWK. Although not an industry recommendation, almost all item traits will be such that the Autoscore-HS exact agreement rate is no lower than 5.25% than the H1-H2 exact agreement rate.

- Human-human exact agreement rates averaged between 67–70% for each dimension, with minimum values ranging from 56–60% to maximum values ranging from 76–81%. The average Autoscore-human exact agreement ranged between 73–77% for each dimension, with minimum values ranging from 68–70% to maximum values ranging from 77–84%. Autoscore-final score exact agreement was the same or higher than human-human exact agreement for 36 item traits out of the 42 item traits. For two item traits (3054 and 3059 Elaboration), the agreement rate was lower than the human-human agreement (7% and 5.3%, respectively) threshold of 5.25%; these values are underlined in the table.

- Human-human QWK agreement averaged between .62–.64 for each dimension, with minimum values ranging from .50–.53 to maximum values ranging from .71–.80. Autoscore-human QWK agreement averaged between .67–.68 for each dimension, with minimum values ranging from .56–.59 to maximum values ranging from .75–.79. Autoscore-final score QWK agreement was the same or higher than human-human QWK agreement for 34 of the 42 item traits. No items and traits had QWK agreements .1 lower than the human-human QWK.

Table 10: Item Trait-Level Agreement of Autoscore with Human Raters on the Held-Out Validation Sample

Grade	Item ID	Dimension	Number of responses	Exact Agreement			Quadratic Weighted Kappa		
				H1-H2	HS-AS	Diff	H1-H2	HS-AS	Diff
3	7407	Convention	276	69%	74%	5%	0.60	0.66	0.06
		Elaboration	276	60%	68%	8%	0.63	0.66	0.04
		Organization	276	64%	69%	5%	0.67	0.65	-0.02
	7423	Convention	273	70%	80%	10%	0.64	0.79	0.15
		Elaboration	273	62%	72%	10%	0.64	0.70	0.06
		Organization	273	61%	71%	10%	0.65	0.68	0.03
4	3084	Convention	227	63%	71%	8%	0.62	0.71	0.09
		Elaboration	227	74%	79%	5%	0.50	0.59	0.09
		Organization	227	73%	77%	4%	0.54	0.63	0.08
	6517	Convention	297	64%	72%	8%	0.62	0.71	0.08

Grade	Item ID	Dimension	Number of responses	Exact Agreement			Quadratic Weighted Kappa		
				H1-H2	HS-AS	Diff	H1-H2	HS-AS	Diff
		Elaboration	297	75%	81%	5%	0.61	0.60	-0.01
		Organization	297	72%	72%	1%	0.61	0.56	-0.05
5	4283	Convention	302	71%	80%	9%	0.63	0.72	0.09
		Elaboration	302	65%	78%	14%	0.56	0.71	0.15
		Organization	302	64%	75%	11%	0.53	0.63	0.10
	5513	Convention	293	71%	74%	3%	0.59	0.62	0.03
		Elaboration	293	70%	76%	6%	0.58	0.62	0.04
		Organization	293	71%	68%	-3%	0.67	0.59	-0.08
6	3138	Convention	290	68%	76%	8%	0.59	0.65	0.06
		Elaboration	290	60%	72%	12%	0.57	0.67	0.10
		Organization	290	56%	71%	15%	0.59	0.71	0.13
	4291	Convention	297	60%	70%	10%	0.56	0.60	0.05
		Elaboration	297	74%	74%	0%	0.57	0.59	0.02
		Organization	297	65%	70%	4%	0.59	0.63	0.03
7	3037	Convention	279	71%	77%	6%	0.67	0.73	0.06
		Elaboration	279	65%	78%	13%	0.56	0.64	0.08
		Organization	279	68%	75%	7%	0.60	0.65	0.04
	3883	Convention	306	75%	84%	9%	0.57	0.73	0.16
		Elaboration	306	71%	81%	10%	0.64	0.71	0.07
		Organization	306	64%	74%	9%	0.60	0.65	0.05
8	3054	Convention	336	76%	73%	-3%	0.71	0.69	-0.02
		Elaboration	336	80%	73%	-7%	0.80	0.70	-0.097
		Organization	336	74%	76%	3%	0.77	0.77	0.00
	3059	Convention	361	81%	80%	-1%	0.71	0.71	0.00
		Elaboration	360	78%	73%	-5.3%	0.78	0.73	-0.04
		Organization	360	76%	75%	-1%	0.77	0.76	-0.01
10	3888	Convention	331	70%	82%	12%	0.52	0.56	0.05
		Elaboration	329	74%	76%	2%	0.70	0.75	0.05
		Organization	329	70%	77%	7%	0.70	0.79	0.08
	4640	Convention	357	71%	78%	6%	0.59	0.62	0.03
		Elaboration	355	64%	73%	9%	0.65	0.73	0.08
		Organization	355	66%	70%	5%	0.70	0.73	0.04
Average		Convention		70%	77%	6%	0.62	0.68	0.06
		Elaboration		69%	75%	6%	0.63	0.67	0.05
		Organization		67%	73%	6%	0.64	0.67	0.03
Minimum		Convention		60%	70%	-3%	52%	56%	-2%
		Elaboration		60%	68%	-7%	50%	59%	-10%
		Organization		56%	68%	-3%	53%	56%	-8%
Maximum		Convention		81%	84%	12%	0.71	0.79	0.16
		Elaboration		80%	81%	14%	0.80	0.75	0.15
		Organization		76%	77%	15%	0.77	0.79	0.13

**Essays that were given a condition code by Autoscore or human raters were excluded.*

Autoscore generally produced similar mean scores and standard deviations as the final, resolved score on the held-out validation sample. Table 11 presents the mean and standard deviation of the scores produced by the final, resolved score (Human) and by Autoscore. The standardized mean difference (SMD) is calculated from these values to examine the mean differences in standard deviation units, which are then interpretable across items and traits. For this calculation, the Autoscore mean is subtracted from the Human mean and divided by the square root of the average of the two variances. SMD values under .15 are considered adequate (Willimanson, Xi, & Breyer, 2012). Forty item traits met this criterion, and two item traits (7407 and 3138 Conventions) failed this criterion and are underlined in the table.

Because automated scoring engines are statistical techniques, they tend to regress to the mean, particularly for unbalanced score point distributions. The Standard Deviation (SD) Ratio, which is the ratio of the Autoscore SD divided by the Human SD, measures the degree of the regression. While there are no common industry standards to interpret SD Ratio, it is preferable that this ratio is near 1, indicating that the Autoscore and Human Scoring standard deviations are similar. This ratio is often slightly below 1, because of the tendency of statistical methods to regress particularly in the presence of unbalanced data. The SD Ratio values below .85 are underlined in the table. Forty item traits met this criterion, and two item traits (7423 and 3037 Elaboration) failed this criterion and are underlined in the table.

Table 11: Item Trait-Level Autoscore with Human Rater Mean Scores on the Held-Out Validation Sample

Grade	Item ID	Dimension	Number of responses	Human		Autoscore		SMD	SD Ratio
				Mean	SD	Mean	SD		
3	7407	Convention	276	1.49	0.66	1.63	0.59	<u>-0.22</u>	0.89
		Elaboration	276	2.05	0.75	2.10	0.72	-0.06	0.96
		Organization	276	2.16	0.75	2.19	0.66	-0.05	0.88
	7423	Convention	273	1.45	0.72	1.51	0.70	-0.09	0.98
		Elaboration	273	1.99	0.77	2.00	0.66	-0.01	<u>0.85</u>
		Organization	273	2.01	0.77	2.03	0.72	-0.02	0.93
4	3084	Convention	227	1.19	0.73	1.19	0.73	0.01	1.00
		Elaboration	227	1.39	0.52	1.36	0.49	0.06	0.95
		Organization	227	1.49	0.55	1.52	0.57	-0.04	1.03
	6517	Convention	297	1.21	0.71	1.22	0.67	-0.01	0.95
		Elaboration	297	1.31	0.51	1.31	0.49	0.00	0.96
		Organization	297	1.52	0.60	1.55	0.61	-0.04	1.02
5	4283	Convention	302	1.49	0.59	1.49	0.60	0.00	1.01
		Elaboration	302	1.89	0.63	1.90	0.61	-0.01	0.96
		Organization	302	2.02	0.64	1.98	0.61	0.06	0.95
	5513	Convention	293	1.48	0.61	1.55	0.58	-0.11	0.95
		Elaboration	293	1.63	0.60	1.55	0.54	0.14	0.90
		Organization	293	1.85	0.66	1.89	0.61	-0.05	0.92
6	3138	Convention	290	1.52	0.65	1.61	0.61	<u>-0.15</u>	0.93

Grade	Item ID	Dimension	Number of responses	Human		Autoscore		SMD	SD Ratio
				Mean	SD	Mean	SD		
7	4291	Elaboration	290	1.75	0.67	1.80	0.63	-0.08	0.94
		Organization	290	1.89	0.74	1.92	0.72	-0.04	0.97
		Convention	297	1.47	0.69	1.49	0.63	-0.04	0.91
		Elaboration	297	1.48	0.65	1.40	0.61	0.12	0.94
		Organization	297	1.68	0.70	1.65	0.68	0.04	0.96
	3037	Convention	279	1.44	0.64	1.49	0.66	-0.09	1.02
		Elaboration	279	1.89	0.59	1.82	0.50	0.12	0.84
		Organization	279	1.96	0.63	1.90	0.58	0.10	0.92
	3883	Convention	306	1.61	0.55	1.63	0.53	-0.05	0.97
		Elaboration	306	1.60	0.62	1.59	0.61	0.02	0.98
		Organization	306	1.77	0.65	1.79	0.60	-0.02	0.93
8	3054	Convention	336	1.46	0.67	1.49	0.68	-0.04	1.02
		Elaboration	336	1.95	0.70	1.99	0.69	-0.06	0.98
		Organization	336	2.06	0.74	2.12	0.75	-0.09	1.01
	3059	Convention	361	1.57	0.59	1.62	0.57	-0.09	0.96
		Elaboration	360	2.09	0.73	2.05	0.71	0.06	0.97
		Organization	360	2.23	0.75	2.18	0.71	0.06	0.95
10	3888	Convention	331	1.73	0.49	1.78	0.42	-0.10	0.85
		Elaboration	329	2.02	0.72	2.01	0.66	0.02	0.91
		Organization	329	2.18	0.76	2.17	0.73	0.02	0.97
	4640	Convention	357	1.61	0.55	1.67	0.53	-0.10	0.95
		Elaboration	355	2.06	0.74	2.03	0.71	0.04	0.95
		Organization	355	2.17	0.77	2.12	0.73	0.07	0.95

**Essays that were given a condition code by Autoscore or human raters were excluded.*

Aside from rubric-based scores outlined in Table 9, Autoscore can generate condition codes—that is, conditions indicating that the response provided by the student is considered invalid and therefore incorrect. The machine-generated condition codes are as follows:

- NO_RESPONSE: Only blank characters are detected in the response.
- NOT_ENOUGH_DATA: Student response has less than the minimum number of words configured in the rubric (currently set to 11 words).
- PROMPT_COPY_MATCH: Student response is copied from the passage or item prompt (currently flagged when a 70% match is found, but this parameter is configurable).
- DUPLICATE_TEXT: Student response is repeated text copied over and over (currently flagged when a 43% match is found, but this parameter is configurable).
- OUT_OF_VOCAB: Student response is comprised mostly of words that do not overlap with those in the training set vocabulary (currently set to 50%).
- NONSPECIFIC: The essay scoring engine predicts the assignment of a condition code. However, even after training the system to generate codes, there can be responses that do not fall into the pre-set categories. The system will generate a NONSPECIFIC condition code for those responses.

Additionally, Autoscore produces a confidence index for a response, indicating how confident the engine is that its score is correct. This index is on a percentile scale and is computed in a two-stage process. In the first stage, for each item, a confidence level is estimated on each dimension using the held-out validation sample; this level can be interpreted as the probability that a dimension score is accurately produced by the engine and is influenced by whether a response has a borderline score or has unusual characteristics. An overall item confidence level can be interpreted as an average of the confidence levels of each dimension. Then, a sample of approximately 5,000 responses gathered from an operational administration and unseen by the engine is scored by Autoscore, and percentile tables are computed based on the overall confidence level.

CAI uses a hybrid, human/machine scoring approach during live testing that flags low-confidence responses or other unusual responses for handscoring. Responses that received confidence percentiles less than 15 and any responses that received a condition code of NONSPECIFIC, OUT OF VOCAB, or DUPLICATE TEXT were routed for human verification. Because the confidence percentile is estimated based upon samples, there will be variation across items in the actual percentage of responses receiving a "low confidence" score.

Human verification was conducted using the following process:

- If the first verification reader-assigned scores in all dimensions that matched the machine-assigned scores, the machine-assigned scores were accepted as the final dimension scores.
- If the first verification reader did not assign the same dimension scores as the machine-assigned scores, the essay was sent to the second verification reader, who then assigned scores in all dimensions. If the second reader's dimension scores matched with either the machine's or the first reader's score, the matching scores were accepted as the final score.
- If the second verification reader's dimension scores did not match the machine's or first reader's scores, the essay was sent to the scoring supervisor to assign the final scores in all dimensions.
- If a backreader's score was available, their score was accepted as the final score regardless of all other assigned scores.

Finally, in addition to the essays sent for human verification due to the low confidence flag or condition codes, the first 500 essays that did not receive a NO RESPONSE, NOT ENOUGH DATA, or PROMPT COPY MATCH were routed for human scoring. The purpose of handscoring the first 500 essays was to ensure that the human scoring and the engine scoring were performing as expected, recognizing the inherent complexities in the dynamics of human scoring. While the first 500 essays cannot be thought to be representative of the tested population, they should be reasonably indicative of the performance of the essay scoring system for responses encountered after the first 500.

Table 12 presents the number and percentage of responses routed for human verification, overall and by routing condition. As expected, 500 responses were routed as part of the first 500 routing conditions, and these percentages were 11–12% of the tested population for grades 3–8 and 34–35% of the tested population for grade 10. The percentage of responses routed for condition codes ranged between .1%–4.1%, with the higher grades having more routing than lower grades. The percentage of responses routed due to low confidence ranged from 3%–16%. In future administrations, CAI will take steps to limit the magnitude of variation to better ensure that

approximately 15% (+/-5%) are routed due to low confidence. For grades 3–8, the total percentage of responses routed ranged from 16%–26%. For grade 10, the total percentage of responses routed ranged from 40%–42%.

Table 12: Number and Percentage of Responses Routed for Human Verification, by Routing Condition and Item

Grade	Item ID	Total Tested	First 500		Condition Code		Low Confidence		Total Routed	
			%	N	%	N	%	N	%	N
3	7407	4329	12%	500	0.3%	13	6%	248	18%	761
	7423	4439	11%	500	0.5%	19	9%	342	19%	861
4	3084	4239	12%	500	0.3%	11	14%	535	25%	1046
	6517	4236	12%	500	0.2%	8	16%	605	26%	1113
5	4283	4186	12%	500	0.1%	4	11%	390	21%	894
	5513	4311	12%	500	0.1%	4	14%	538	24%	1042
6	3138	4223	12%	500	0.2%	6	10%	364	21%	870
	4291	4237	12%	500	0.2%	8	10%	359	20%	867
7	3037	4306	12%	500	0.2%	6	13%	491	23%	997
	3883	4152	12%	500	0.1%	5	4%	149	16%	654
8	3054	4113	12%	500	1.2%	44	3%	116	16%	660
	3059	4173	12%	500	4.1%	150	4%	154	19%	804
10	3888	1469	34%	500	3.1%	30	9%	90	42%	620
	4640	1440	35%	500	2.2%	21	6%	59	40%	580

**Data do not include responses receiving the NO RESPONSE, NOT ENOUGH DATA, or PROMPT COPY MATCH condition codes.*

The performance of the scoring on the first 500 students can be examined using the handscoring agreements of the held-out validation sample as a benchmark. Currently, there are no standards in the industry for examining live scoring, partly because handscoring is a dynamic and complex process and because the process used during handscoring benchmarks may not match those used during live scoring. CAI used thresholds of 10% and .2 for exact agreement and QWK, respectively, to identify agreements on the first 500 samples that lie below these thresholds as a way to monitor the scoring process. We use larger thresholds for monitoring the operational data (compared to the thresholds for monitoring the held-out validation data) because of the complexities surrounding live scoring situations. This method is necessary because the engine scores were compared to less reliable scores (i.e., non-expert scores) and because the scoring occurs early in the window when human raters are still cementing their understanding and application of the rubric.

Table 13 presents the exact agreement and QWK between the first human rater and Autoscore (H1-AS), the first human rater and the expert scorer (H1-ES), and Autoscore and the expert scorer (AS-ES) for the responses that were routed for final resolution. Looking at the agreement between the first human rater and Autoscore, there were 11 item traits with exact agreement rates more than 10% below the human-human agreement rates on the held-out validation sample and 6 item traits with QWK agreement rates more than .2 below the human-human agreement rates on the held-out validation sample.

Lower agreement rates may be due to various factors, including human scoring error or drift, Autoscore error, or changes to the tested population in terms of ability or writing (such that the held-out validation results no longer serve as an appropriate benchmark). Changes to the population are difficult to discern. However, we can compare the agreement of the first human scorer to the expert scorer and the agreement of Autoscore to the expert scorer for routed responses to better understand how the two scorers perform relative to an expert. Recall that responses are routed to the expert scorer when the set of dimension scores do not exactly match Autoscore, the first human scorer, or a second human scorer.

When comparing the 11 flagged exact agreement rates in this manner, eight of the H1-ES exact agreement rates were lower than the AS-ES, one was the same, and two of the AS-HS exact agreement rates were lower than the H1-ES. When comparing the six QWK agreement values in this manner, three of the H1-ES QWK agreement values were lower than the AS-ES and three of the AS-HS QWK agreement values were lower than the H1-ES. These results suggest that while a few lower agreements have been identified, Autoscore—as the primary source of scores for NDSA—shows adequate agreement rates with either the first human scorer or the expert scorer for almost all items and traits.

Table 13: Item Trait-Level Agreement of Autoscore with Human Raters on the First 500 Samples

Grade	Item ID	Dimension	Number of responses		Exact Agreement				Quadratic Weighted Kappa			
			H1	ES	Bench mark	H1-AS	H1-ES	AS-ES	Bench mark	H1-AS	H1-ES	AS-ES
3	7407	Convention	499	153	69%	70%	68%	75%	0.60	0.64	0.59	0.65
		Elaboration	499	153	60%	64%	65%	69%	0.63	0.64	0.70	0.59
		Organization	499	153	64%	65%	54%	76%	0.67	0.67	0.62	0.73
	7423	Convention	495	124	70%	56%	59%	71%	0.64	0.51	0.35	0.63
		Elaboration	490	122	62%	72%	63%	84%	0.64	0.65	0.60	0.83
		Organization	490	122	61%	72%	71%	75%	0.65	0.66	0.71	0.73
4	3084	Convention	499	167	63%	61%	65%	68%	0.62	0.55	0.39	0.58
		Elaboration	497	166	74%	61%	51%	72%	0.50	0.36	0.36	0.46
		Organization	497	166	73%	58%	51%	63%	0.54	0.44	0.36	0.43
	6517	Convention	500	217	64%	62%	61%	73%	0.62	0.59	0.53	0.62
		Elaboration	500	217	75%	43%	57%	47%	0.61	0.22	0.53	0.27
		Organization	500	217	72%	48%	53%	65%	0.61	0.39	0.52	0.57
5	4283	Convention	500	172	71%	69%	62%	81%	0.63	0.62	0.46	0.67
		Elaboration	499	172	65%	63%	63%	67%	0.56	0.53	0.67	0.61
		Organization	499	172	64%	62%	65%	70%	0.53	0.55	0.70	0.60
	5513	Convention	498	225	71%	71%	68%	81%	0.59	0.59	0.43	0.69
		Elaboration	491	223	70%	46%	38%	74%	0.58	0.33	0.32	0.62
		Organization	491	223	71%	57%	62%	69%	0.67	0.48	0.53	0.52
6	3138	Convention	500	152	68%	74%	64%	74%	0.59	0.57	0.37	0.48
		Elaboration	500	152	60%	75%	69%	66%	0.57	0.64	0.66	0.59

Grade	Item ID	Dimension	Number of responses		Exact Agreement				Quadratic Weighted Kappa			
			H1	ES	Bench mark	H1-AS	H1-ES	AS-ES	Bench mark	H1-AS	H1-ES	AS-ES
7	4291	Organization	500	152	56%	64%	61%	68%	0.59	0.56	0.53	0.65
		Convention	500	153	60%	73%	65%	60%	0.56	0.60	0.42	0.17
		Elaboration	499	152	74%	73%	64%	71%	0.57	0.48	0.57	0.40
		Organization	499	152	65%	62%	41%	72%	0.59	0.51	0.44	0.62
	3037	Convention	499	183	71%	64%	58%	66%	0.67	0.53	0.25	0.40
		Elaboration	499	183	65%	62%	52%	76%	0.56	0.43	0.43	0.63
		Organization	499	183	68%	62%	69%	62%	0.60	0.40	0.58	0.39
	3883	Convention	500	130	75%	66%	63%	68%	0.57	0.49	0.44	0.41
		Elaboration	498	129	71%	79%	64%	72%	0.64	0.55	0.37	0.39
		Organization	498	129	64%	72%	50%	69%	0.60	0.56	0.40	0.54
8	3054	Convention	493	167	76%	55%	44%	68%	0.71	0.46	0.31	0.54
		Elaboration	492	166	80%	68%	61%	61%	0.80	0.54	0.63	0.50
		Organization	492	166	74%	69%	74%	60%	0.77	0.63	0.76	0.52
	3059	Convention	477	121	81%	74%	76%	74%	0.71	0.69	0.65	0.58
		Elaboration	477	120	78%	64%	65%	55%	0.78	0.49	0.55	0.48
		Organization	477	120	76%	68%	62%	73%	0.77	0.60	0.52	0.64
10	3888	Convention	485	132	70%	86%	76%	84%	0.52	0.63	0.49	0.58
		Elaboration	484	132	74%	62%	48%	71%	0.70	0.57	0.49	0.66
		Organization	484	132	70%	62%	43%	73%	0.70	0.59	0.37	0.78
	4640	Convention	492	125	71%	74%	66%	78%	0.59	0.57	0.45	0.47
		Elaboration	492	125	64%	63%	56%	49%	0.65	0.64	0.50	0.44
		Organization	492	125	66%	65%	51%	65%	0.70	0.68	0.56	0.56

**Essays that were given a condition code by Autoscore or human raters were excluded.*

The mean scores and standard deviations can also be compared for the first human score, Autoscore, and the expert score. Again, there are no industry standards around how best to monitor automated scoring performance. Table 14 presents the score means and standard deviations for the first human rater (H1) and Autoscore (AS) on the first 500 samples and the same metrics on the responses routed for the expert scorer (ES). The item traits for which the H1-AS SMD magnitude exceeds .3 are underlined in the table. Again, we used a larger threshold for the operational data than for the held-out validation in consideration of the complexities inherent to live scoring. Fifteen item dimensions using this threshold are underlined in the table.

We can examine the agreement between the first human rater and Autoscore with the final resolved score from the entire hybrid scoring process. While the final resolved score is not independent of these scores, it still represents the best score for a given response. If the two scorers are performing similarly, they should have similar agreement with the final resolved score. To examine this, we can identify SMD magnitudes for each scorer that exceed .15. Of the 15 item traits exceeding .3 between H1 and AS, 9 were such that both the H1-Final and AS-Final SMDs exceeded .15. For 2 of the 15 item traits, the AS-Final SMD exceeded the .15 value, but the H1-Final SMD did not. For 4 of the 15 item traits, the H1-Final SMD exceeded the .15 value, but the AS-Final did not.

Additionally, there were six item traits where the AS-Final SMD magnitude exceeded .15, but the H1-Final did not. And, there were five item traits where the H1-Final SMD magnitude exceeded .15, but the AS-Final did not.

These results indicate that Autoscore and the first human rater differ in their score assignment for many items and that these differences are addressed somewhat by the resolution process. In the 2021-2022 school year, CAI will review the rubrics and scoring processes to ensure that the handscorers and engine scores are better aligned in the score assignment.

Table 14: Item Trait-Level Autoscore and Human Rater Means and Standard Deviations on the First 500 Samples

Grade	Item ID	Dimension	H1		AS		Final		SMD		
			Mean	SD	Mean	SD	Mean	SD	H1-AS	H1-Final	AS-Final
3	7407	Convention	1.32	0.70	1.44	0.66	1.32	0.69	0.17	0.01	-0.17
		Elaboration	1.85	0.83	1.71	0.62	1.77	0.73	-0.19	-0.11	0.09
		Organization	2.01	0.82	1.81	0.65	1.86	0.74	-0.27	-0.19	0.07
	7423	Convention	1.13	0.60	1.44	0.73	1.26	0.68	0.47	0.21	-0.25
		Elaboration	1.62	0.62	1.72	0.65	1.66	0.62	0.15	0.06	-0.10
		Organization	1.67	0.65	1.75	0.66	1.69	0.64	0.12	0.03	-0.09
4	3084	Convention	1.29	0.62	1.05	0.69	1.19	0.61	-0.37	-0.16	0.22
		Elaboration	1.58	0.68	1.23	0.44	1.36	0.52	-0.62	-0.36	0.28
		Organization	1.73	0.75	1.33	0.49	1.50	0.56	-0.64	-0.36	0.32
	6517	Convention	1.31	0.73	1.13	0.62	1.30	0.68	-0.27	0.00	0.27
		Elaboration	1.84	0.75	1.17	0.38	1.59	0.67	-1.13	-0.35	0.77
		Organization	1.91	0.77	1.35	0.50	1.63	0.69	-0.86	-0.37	0.47
5	4283	Convention	1.26	0.69	1.39	0.63	1.36	0.64	0.21	0.16	-0.05
		Elaboration	1.74	0.77	1.64	0.60	1.71	0.69	-0.14	-0.05	0.10
		Organization	1.86	0.80	1.70	0.58	1.79	0.72	-0.23	-0.09	0.14
	5513	Convention	1.48	0.61	1.43	0.60	1.46	0.61	-0.08	-0.03	0.05
		Elaboration	2.03	0.87	1.43	0.52	1.60	0.65	-0.84	-0.57	0.28
		Organization	2.01	0.83	1.66	0.57	1.79	0.66	-0.49	-0.29	0.21
6	3138	Convention	1.53	0.58	1.62	0.56	1.61	0.55	0.16	0.13	-0.03
		Elaboration	1.79	0.66	1.68	0.59	1.76	0.63	-0.19	-0.04	0.14
		Organization	1.87	0.69	1.76	0.64	1.82	0.63	-0.18	-0.08	0.10
	4291	Convention	1.49	0.64	1.54	0.59	1.44	0.62	0.07	-0.09	-0.17
		Elaboration	1.41	0.64	1.25	0.46	1.34	0.53	-0.298	-0.13	0.19
		Organization	1.67	0.71	1.43	0.56	1.51	0.61	-0.37	-0.24	0.13
7	3037	Convention	1.57	0.61	1.37	0.63	1.45	0.63	-0.33	-0.20	0.13
		Elaboration	1.64	0.66	1.57	0.51	1.57	0.56	-0.12	-0.11	0.00
		Organization	1.86	0.62	1.66	0.51	1.71	0.54	-0.36	-0.25	0.11
	3883	Convention	1.38	0.70	1.62	0.56	1.47	0.63	0.38	0.13	-0.25
		Elaboration	1.32	0.55	1.26	0.44	1.30	0.49	-0.11	-0.03	0.09
		Organization	1.52	0.64	1.42	0.50	1.48	0.56	-0.16	-0.06	0.10

Grade	Item ID	Dimension	H1		AS		Final		SMD		
			Mean	SD	Mean	SD	Mean	SD	H1-AS	H1-Final	AS-Final
8	3054	Convention	1.10	0.81	1.40	0.71	1.34	0.73	0.40	0.31	-0.08
		Elaboration	1.46	0.68	1.49	0.54	1.58	0.65	0.05	0.18	0.147
		Organization	1.73	0.72	1.60	0.60	1.71	0.67	-0.20	-0.03	0.18
	3059	Convention	1.44	0.68	1.51	0.62	1.44	0.66	0.12	0.01	-0.11
		Elaboration	1.42	0.63	1.59	0.60	1.45	0.60	0.28	0.04	-0.24
		Organization	1.70	0.68	1.62	0.62	1.64	0.63	-0.11	-0.09	0.02
10	3888	Convention	1.77	0.48	1.80	0.42	1.78	0.43	0.06	0.03	-0.03
		Elaboration	1.89	0.70	1.73	0.66	1.76	0.69	-0.24	-0.19	0.05
		Organization	2.03	0.63	1.86	0.73	1.87	0.68	-0.25	-0.24	0.01
	4640	Convention	1.54	0.64	1.73	0.49	1.64	0.57	0.32	0.154	-0.17
		Elaboration	1.60	0.77	1.83	0.71	1.70	0.73	0.31	0.12	-0.19
		Organization	1.83	0.81	1.96	0.73	1.88	0.75	0.17	0.07	-0.10

**Essays that were given a condition code by Autoscore or human raters were excluded.*

4. Evidence of Content Validity

This section demonstrates that the knowledge and skills assessed by the NDSA were representative of the content standards of the larger knowledge domain. We describe the content standards for the NDSA and discuss the test development process, mapping the NDSA tests to the standards. A complete description of the test development process can be found in Volume 2, Test Development, of this technical report.

4.1 Content Standards

The NDSA was aligned to the English language arts (ELA) and mathematics standards adopted in April 2017. The ELA and mathematics standards are available for review at https://www.nd.gov/dpi/sites/www/files/documents/Academic%20Support/ELA-Literacy%20Standards-2017%20Final-Revised%2009-21-2020.pdf. Blueprints were developed to ensure that the test and the items were aligned to the prioritized standards that they were intended to measure. A complete description of the blueprint and test form construction process can be found in Volume 2, Section 2, of the NDSA technical reports.

Table 15 and Table 16 present the reporting categories by grade and test and the number of items measuring each category.

Table 15: Number of Items for Each ELA Reporting Category

Reporting Category	Grade						
	3	4	5	6	7	8	10
Reading Standards for Informational/Nonfiction Text	188	199	170	243	231	229	141
Reading Standards for Literature/Fiction	134	141	139	175	182	148	77

Reporting Category	Grade						
	3	4	5	6	7	8	10
Writing and Language Standards	90	116	105	106	109	106	79

Table 16: Number of Items for Each Mathematics Reporting Category

Grade	Reporting Category	Number of Items
3	Measurement, Data, and Geometry	161
	Number and Operations in Base Ten	108
	Number and Operations - Fractions	159
	Operations and Algebraic Thinking	181
4	Measurement, Data, and Geometry	161
	Number and Operations in Base Ten	183
	Number and Operations - Fractions	190
	Operations and Algebraic Thinking	108
5	Measurement, Data, and Geometry	132
	Number and Operations in Base Ten	148
	Number and Operations - Fractions	158
	Operations and Algebraic Thinking	93
6	Expressions and Equations	196
	Geometry	71
	Ratios and Proportional Relationships and Number Systems	321
	Statistics and Probability	63
7	Expressions and Equations	82
	Geometry	95
	Ratios and Proportional Relationships and Number Systems	180
	Statistics and Probability	94
8	Expressions and Equations and Number Systems	213
	Functions	106
	Geometry	132
	Statistics and Probability	69
10	Algebra	204
	Functions	241
	Geometry	169
	Statistics, Probability, and the Number System	61

4.2 Alignment of ICCR Test Forms to the Content Standards and Benchmarks

Refer to the third-party, independent alignment study in Volume 7 of this technical report for details.

5. Evidence on Internal-External Structure

This section explores the internal structure of the assessment using the scores provided at the reporting category level. The relationship of the subscores is just one indicator of the test dimensionality.

There are three reporting categories in ELA for grades 3, 4, 6, and 10: Reading Standards for Informational/Nonfiction Text, Reading Standards for Literature/Fiction, and Writing and Language Standards. In mathematics, reporting categories differ in each grade or course (refer to

Table 16 for reporting category information).

Scale scores and relative strengths and weaknesses based on each reporting category were provided to students. Evidence is needed to verify that each reporting category's scale scores and relative strengths and weaknesses provide both different and useful information for student achievement.

It may not be reasonable to expect that the reporting category scores are completely orthogonal—this would suggest that there are no relationships among reporting category scores and would make justification of a unidimensional item response theory (IRT) model difficult. However, we could then easily justify reporting these separate scores. On the contrary, if the reporting categories were perfectly correlated, we could justify a unidimensional model, but we could not justify reporting separate scores.

One pathway to explore the internal structure of the test is via a second-order factor model, assuming a general mathematics construct (first factor) with reporting categories (second factor) and that the items load onto the reporting category they intend to measure. If the first-order factors are highly correlated and the model fits data well for the second-order model, this provides evidence of unidimensionality and reporting subscores.

Another pathway is to explore observed correlations between the subscores. However, as each reporting category is measured with a small number of items, the standard errors of the observed scores within each reporting category are typically larger than the standard error of the total test score. Disattenuating for measurement error could offer some insight into the theoretical true score correlations. Both observed correlations and disattenuated correlations are provided in the following section.

5.1 Correlations Among Reporting Category Scores

The correlations between reporting category scores, both observed (below diagonal) and corrected for attenuation (above diagonal) are presented in Table 17 and Table 18. On the diagonal, the reliability coefficient of the reporting category is shown. In ELA, the observed correlations

between the reporting categories range from 0.49–0.60. For mathematics, the observed correlations were between 0.42–0.69. Disattenuated correlations were between 0.67–0.84 for ELA and 0.68–0.91 for mathematics.

In some instances, these correlations were lower than might be expected. However, as previously noted, the correlations were subject to a large amount of measurement error at the reporting category level, given the limited number of items from which the scores were derived. Consequently, over-interpretation of these correlations as either high or low should be made cautiously. Furthermore, somewhat lower correlations support using separate reporting categories because they are shown to be distinct measures.

Table 17: Correlations Among Reporting Categories (ELA)

Grade	Reporting Category	Mean # of Items per Student	RI	RL	WL
3	Reading Standards for Informational/Nonfiction Text (RI)	14.9	0.69	0.70	0.69
	Reading Standards for Literature/Fiction (RL)	18.4	0.50	0.73	0.68
	Writing and Language Standards (WL)	14.0	0.51	0.52	0.80
4	Reading Standards for Informational/Nonfiction Text (RI)	13.7	0.69	0.78	0.67
	Reading Standards for Literature/Fiction (RL)	18.5	0.56	0.75	0.72
	Writing and Language Standards (WL)	13.8	0.49	0.55	0.78
5	Reading Standards for Informational/Nonfiction Text (RI)	14.9	0.70	0.80	0.73
	Reading Standards for Literature/Fiction (RL)	18.1	0.58	0.75	0.74
	Writing and Language Standards (WL)	13.7	0.55	0.58	0.81
6	Reading Standards for Informational/Nonfiction Text (RI)	16.2	0.74	0.77	0.69
	Reading Standards for Literature/Fiction (RL)	15.8	0.57	0.74	0.75
	Writing and Language Standards (WL)	13.3	0.53	0.58	0.8
7	Reading Standards for Informational/Nonfiction Text (RI)	15.5	0.72	0.83	0.76
	Reading Standards for Literature/Fiction (RL)	16.8	0.59	0.71	0.76
	Writing and Language Standards (WL)	13.7	0.56	0.56	0.76
8	Reading Standards for Informational/Nonfiction Text (RI)	16.3	0.74	0.79	0.71
	Reading Standards for Literature/Fiction (RL)	17.1	0.58	0.73	0.70
	Writing and Language Standards (WL)	13.8	0.55	0.54	0.81
10	Reading Standards for Informational/Nonfiction Text (RI)	15.1	0.73	0.84	0.76
	Reading Standards for Literature/Fiction (RL)	14.6	0.60	0.7	0.74
	Writing and Language Standards (WL)	15.9	0.59	0.56	0.82

**Diagonal value represents the reliability coefficient of the reporting category. Observed correlations are below the diagonal, and disattenuated are above.*

Table 18: Correlations Among Reporting Categories (Mathematics)

Grade	Reporting Category	Mean # of Items per Student	MDG	NBT	NF	OA
3	Measurement, Data, and Geometry (MDG)	9.9	0.72	0.84	0.81	0.86
	Number and Operations in Base Ten (NBT)	8.0	0.63	0.79	0.71	0.83
	Number and Operations – Fractions (NF)	8.0	0.59	0.54	0.74	0.74
	Operations and Algebraic Thinking (OA)	11.3	0.65	0.66	0.57	0.80
4	Measurement, Data, and Geometry (MDG)	10.4	0.76	0.85	0.84	0.86
	Number and Operations in Base Ten (NBT)	8.1	0.64	0.75	0.84	0.91
	Number and Operations – Fractions (NF)	9.4	0.65	0.64	0.78	0.85
	Operations and Algebraic Thinking (OA)	9.3	0.66	0.69	0.66	0.77
5	Measurement, Data, and Geometry (MDG)	10.4	0.70	0.85	0.83	0.86
	Number and Operations in Base Ten (NBT)	8.0	0.61	0.73	0.80	0.86
	Number and Operations – Fractions (NF)	9.6	0.60	0.59	0.75	0.79
	Operations and Algebraic Thinking (OA)	9.2	0.61	0.62	0.58	0.72
			EE	**G**	**RPNS**	**SP**
6	Expressions and Equations (EE)	9.5	0.72	0.73	0.91	0.72
	Geometry (G)	8.3	0.49	0.63	0.77	0.68
	Ratios and Proportional Relationships and Number Systems (RPNS)	12.0	0.67	0.53	0.76	0.76
	Statistics and Probability (SP)	8.2	0.48	0.42	0.52	0.61
7	Expressions and Equations (EE)	9.3	0.73	0.77	0.83	0.78
	Geometry (G)	9.0	0.54	0.68	0.86	0.81
	Ratios and Proportional Relationships and Number Systems (RPNS)	9.0	0.63	0.63	0.79	0.87
	Statistics and Probability (SP)	9.6	0.54	0.54	0.63	0.66
			EENS	**F**	**G**	**SP**
8	Expressions and Equations and Number Systems (EENS)	11.2	0.76	0.88	0.81	0.91
	Functions (F)	9.3	0.62	0.65	0.72	0.84
	Geometry (G)	8.3	0.60	0.49	0.72	0.73
	Statistics and Probability (SP)	8.3	0.68	0.58	0.53	0.74
			A	**F**	**G**	**S**
10	Algebra (A)	12.8	0.77	0.77	0.78	0.71
	Functions (F)	11.3	0.55	0.67	0.75	0.68
	Geometry (G)	9.3	0.60	0.54	0.77	0.72
	Statistics and Probability (S)	10.6	0.54	0.48	0.55	0.75

**Diagonal value represents the reliability coefficient of the reporting category. Observed correlations are below the diagonal, and disattenuated are above.*

5.2 CONFIRMATORY FACTOR ANALYSIS

In the 2020–2021 school year, the NDSA was administered as an CAT. Unlike the fixed-form tests administered in the 2018–2019 school year, the number of students who took each item was not always sufficient for conducting confirmatory factor analysis. Due to this restriction, the internal structural validity evidence supported by 2018–2019 NDSA student data is summarized in this section. The 2018–2019 NDSA and 2020–2021 NDSA were constructed using the same content standards and similar test blueprints. The internal structure of the two assessments is expected to be equivalent, with some degree of variability in model coefficients.

The NDSA had test items designed to measure different standards and higher-level reporting categories. Test scores were reported as an overall performance measure. Additionally, scores on the various reporting categories were also provided as indices of strand-specific performance. The strand scores were reported in a fashion that aligned with the theoretical structure of the test derived from the test blueprint.

The results in this section are intended to provide evidence that the methods for reporting the NDSA strand scores align with the test's underlying structure and the appropriateness of the selected IRT models. This section is based on a second-order confirmatory factor analysis, in which the first-order factors load onto a common underlying factor. The first-order factors represent the dimensions of the test blueprint, and items load onto the factors they are intended to measure. The underlying structure of the ELA and mathematics tests was generally common across all grades, which is useful for comparing the results of our analyses across the grades.

While the test consisted of items targeting different standards, all items within a grade and subject were calibrated concurrently using the various IRT models described in this technical report. This implies the pivotal IRT assumption of local independence (Lord, 1980). Formally stated, this assumption posits that the probability of the outcome on item *i* depends only on the student's ability and the item's characteristics. Beyond that, the score of item *i* is independent of the outcome of all other items. From this assumption, the joint density (i.e., the likelihood) is viewed as the product of the individual densities. Thus, the maximum likelihood estimation of person and item parameters in traditional IRT is derived on the basis of this theory.

The measurement model and the score reporting method assume a single underlying factor, with separate factors representing each reporting category. Consequently, it is important to collect validity evidence on the internal structure of the assessment to determine the rationality of conducting concurrent calibrations and using these scoring and reporting methods.

5.2.1 Factor Analytic Methods

A series of confirmatory factor analyses (CFA) were conducted using the statistical program Mplus [version 7.31] (Muthén & Muthén, 2012) for each grade and subject assessment. Mplus is commonly used for collecting validity evidence on the internal structure of assessments. The estimation method, weighted least squares means, and variance adjusted (WLSMV) were employed because it is less sensitive to the sample size and model and is shown to perform well with categorical variables (Muthén, du Toit, & Spisic, 1997).

As previously stated, the reporting scores method used for North Dakota implies separate factors connected by a single underlying factor for each reporting category. This model is subsequently referred to as the implied model. In factor analytic terms, this suggests that test items load onto separate first-order factors, with the first-order factors connected to a single underlying second-order factor. The use of the CFA in this section establishes some validity evidence for the degree to which the implied model is reasonable.

A chi-square difference test is often applied to assess model fit. However, this test is sensitive to sample size, almost always rejecting the null hypothesis when the sample size is large. Therefore, instead of conducting a chi-square difference test, other goodness-of-fit indices were used to evaluate the implied model for the NDSA.

If the internal structure of the test was strictly unidimensional, then the overall person ability measure, theta (θ), would be the single underlying common factor, and the correlation matrix among test items would suggest no discernable pattern among factors. There would be no empirical or logical basis to report scores for the separate performance categories. In factor analytic terms, a strictly unidimensional test structure implies a single-order factor model, in which all test items load onto a single underlying factor. The development below expands the first-order model to a generalized second-order parameterization to show the relationship between the models.

The factor analysis models are based on the matrix $\boldsymbol{S}$ of tetrachoric and polychoric sample correlations among the item scores (Olsson, 1979), and the matrix $\boldsymbol{W}$ of asymptotic covariances among these sample correlations (Jöreskog, 1994) is employed as a weight matrix in a weighted least squares estimation approach (Browne, 1984; Muthén, 1984) to minimize the fit function:

$$F_{WLS} = \text{vech}(\boldsymbol{S} - \widehat{\boldsymbol{\Sigma}})'\boldsymbol{W}^{-1}\text{vech}(\boldsymbol{S} - \widehat{\boldsymbol{\Sigma}}).$$

In the preceding equation, $\widehat{\boldsymbol{\Sigma}}$ is the implied correlation matrix, given the estimated factor model, and the function vech vectorizes a symmetric matrix. That is, the vech stacks each column of the matrix to form a vector. Note that the WLSMV approach (Muthén, du Toit, & Spisic, 1997) employs a weight matrix of asymptotic variances (i.e., the diagonal of the weight matrix) instead of the full asymptotic covariances.

We posit a first-order factor analysis where all test items load onto a single underlying common factor as the base model. The first-order model can be mathematically represented as:

$$\widehat{\boldsymbol{\Sigma}} = \boldsymbol{\Lambda\Phi\Lambda}' + \boldsymbol{\Theta},$$

where $\boldsymbol{\Lambda}$ is the matrix of item factor loadings (with $\boldsymbol{\Lambda}'$ representing its transpose), and $\boldsymbol{\Theta}$ is the uniqueness or measurement error. The matrix $\boldsymbol{\Phi}$ is the correlation among the separate factors. For the base model, items are thought to load onto a single underlying factor only. Hence $\boldsymbol{\Lambda}$ is a *p x 1* vector, where *p* is the number of test items and $\boldsymbol{\Phi}$ is a scalar equal to 1. Therefore, it is possible to drop the matrix $\boldsymbol{\Phi}$ from the general notation. However, this notation is retained to facilitate comparisons to the implied model more easily, such that it can subsequently be viewed as a special case of the second-order factor analysis.

For the implied model, we posit a second-order factor analysis in which test items are coerced to load onto the reporting categories they are designed to target, and all reporting categories share a

common underlying factor. The second-order factor analysis can be mathematically represented as:

$$\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Lambda}(\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}' + \boldsymbol{\Psi})\boldsymbol{\Lambda}' + \boldsymbol{\Theta},$$

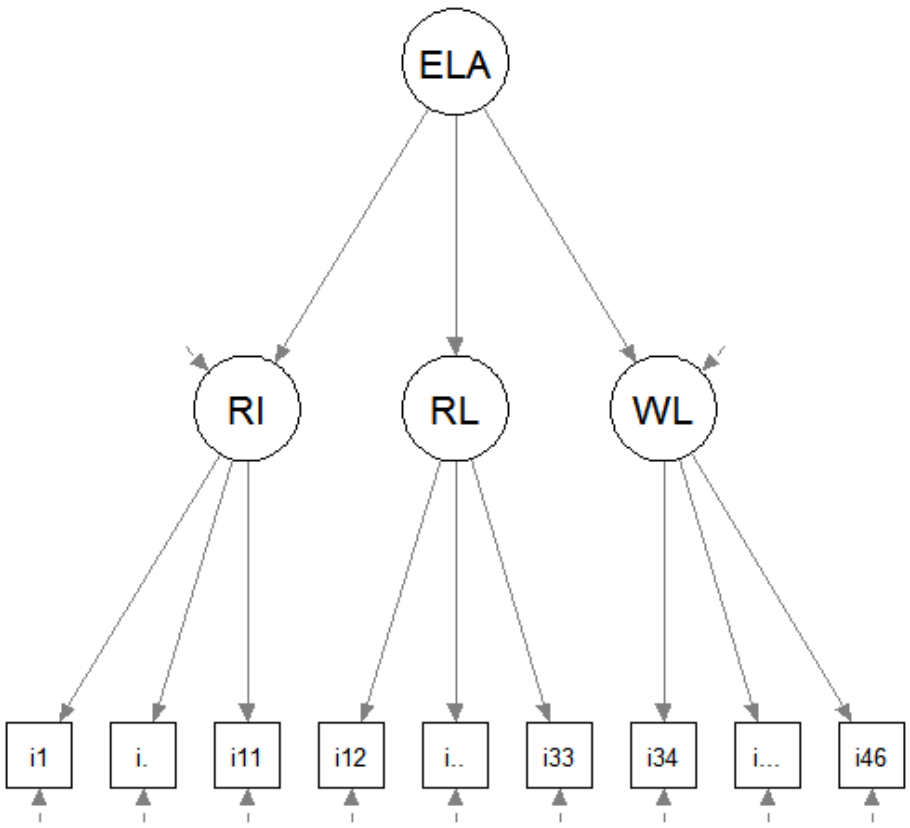
where $\hat{\boldsymbol{\Sigma}}$ is the implied correlation matrix among test items, $\boldsymbol{\Lambda}$ is the *p x k* matrix of the first-order factor loadings relating item scores to first-order factors, $\boldsymbol{\Gamma}$is the *k x 1* matrix of the second-order factor loadings relating the first-order factors to the second-order factor with *k* denoting the number of factors, $\boldsymbol{\Phi}$ is the correlation matrix of the second-order factors, and $\boldsymbol{\Psi}$ is the matrix of the first-order factor residuals. All other notations are the same as in the first-order model. Note that the second-order model expands the first-order model such that $\boldsymbol{\Phi} \rightarrow \boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}' + \boldsymbol{\Psi}$. Therefore, the first-order model is said to be nested within the second-order model.

There is a separate factor for each of three ELA and four mathematics reporting categories (refer to Table 15 and Table 16 for reporting category information). Therefore, the number of rows in $\boldsymbol{\Gamma}$ (*k)* differs between subjects, but the general structure of the factor analysis is consistent across ELA and mathematics.

The second-order factor model can also be represented graphically, and a sample of the generalized approaches is provided on the following page. The general structure of the second-order factor analysis for ELA is illustrated in Figure 4. This figure is generally representative of the factor analyses performed for all grades and subjects, understanding that the number of items within each reporting category could vary across the grades.

The purpose of conducting a CFA for the NDSA was to provide evidence that each assessment in the NDSA implied a second-order factor model: a single underlying second-order factor with the first-order factors defining each of the reporting categories.

Figure 4: Second-Order Factor Model (ELA)



5.2.2 Results

Several goodness-of-fit statistics from each of the analyses are presented in Table 19, which shows the summary results obtained from the CFA. Three goodness-of-fit indices were used to evaluate the model fit of the item parameters to how students responded to the items. The root mean square error of approximation (RMSEA) is referred to as a badness-of-fit index so that a value closer to 0 implies better fit and a value of 0 implies best fit. In general, an RMSEA below 0.05 is considered good fit and an RMSEA over 0.1 suggests poor fit (Browne & Cudeck, 1993). The Tucker-Lewis index (TLI) and the comparative fit index (CFI) are incremental goodness-of-fit indices. These indices compare the implied model to the baseline model, where no observed variables are correlated (i.e., there are no factors). Values greater than 0.9 are recognized as acceptable, and values greater than 0.95 are considered good fit (Hu & Bentler, 1999). As Hu and Bentler (1999) suggest, the selected cut-off values of the fit index should not be overgeneralized and should be interpreted with caution.

The model showed good fit across content domains based on the fit indices produced and the established criteria for evaluating fit. The RMSEA was below 0.05 for all tests, and the CFI and

TLI were equal to or greater than 0.95 except for grade 3 ELA, which had a CIF of 0.931 and TLI of 0.927.

Table 19: Goodness-of-Fit Second-Order CFA

ELA					
Grade	*df*	*RMSEA*	*CFI*	*TLI*	*Convergence*
3*	987	0.039	0.931	0.927	Yes
4	986	0.035	0.951	0.949	Yes
5*	987	0.041	0.960	0.958	Yes
6	899	0.026	0.965	0.963	Yes
7	986	0.028	0.959	0.957	Yes
8	986	0.026	0.964	0.963	Yes
10	942	0.026	0.971	0.969	Yes
Mathematics					
Grade	*df*	*RMSEA*	*CFI*	*TLI*	*Convergence*
3	815	0.026	0.975	0.973	Yes
4	815	0.023	0.982	0.981	Yes
5	815	0.027	0.973	0.971	Yes
6	815	0.025	0.974	0.972	Yes
7	815	0.034	0.967	0.965	Yes
8	815	0.035	0.949	0.946	Yes
10	815	0.025	0.972	0.971	Yes

**For these tests, the second-order model was run by constraining the residual variance of a certain factor to zero due to non-significant negative residual variance.*

The second-order factor model converged for all tests. However, the residual variance for one factor fell slightly below the boundary of 0 for grades 3 and 5 ELA when using the M-Plus software package. This negative residual variance may be related to the computational implementation of the optimization approach in M-Plus, it may be a flag related to model misspecification, or it may be related to other causes (Van Driel, 1978; Chen, Bollen, Paxton, Curran & Kirby, 2001). The residual variance was constrained to 0 for these tests. This is equivalent to treating the parameter as fixed, which does not necessarily conform to our *a-priori* hypothesis.

Table 20 and Table 21 provide the estimated correlations between the reporting categories from the second-order factor model for ELA and mathematics, respectively. In all cases, these correlations are very high. However, the results provide empirical evidence that there is some detectable dimensionality among reporting categories.

Table 20: Correlations Among ELA Factors

Grade	Reporting Category	RI	RL	WL
3*	Reading Standards for Informational/Nonfiction Text (RI)	1.00	-	-
	Reading Standards for Literature/Fiction (RL)	0.93	1.00	-

	Writing and Language Standards (WL)	0.72	0.67	1.00
4	Reading Standards for Informational/Nonfiction Text (RI)	1.00	-	-
	Reading Standards for Literature/Fiction (RL)	0.95	1.00	-
	Writing and Language Standards (WL)	0.70	0.72	1.00
5*	Reading Standards for Informational/Nonfiction Text (RI)	1.00	-	-
	Reading Standards for Literature/Fiction (RL)	0.78	1.00	-
	Writing and Language Standards (WL)	0.89	0.88	1.00
6	Reading Standards for Informational/Nonfiction Text (RI)	1.00	-	-
	Reading Standards for Literature/Fiction (RL)	0.95	1.00	-
	Writing and Language Standards (WL)	0.78	0.76	1.00
7	Reading Standards for Informational/Nonfiction Text (RI)	1.00	-	-
	Reading Standards for Literature/Fiction (RL)	0.98	1.00	-
	Writing and Language Standards (WL)	0.77	0.79	1.00
8	Reading Standards for Informational/Nonfiction Text (RI)	1.00	-	-
	Reading Standards for Literature/Fiction (RL)	0.95	1.00	-
	Writing and Language Standards (WL)	0.82	0.81	1.00
10	Reading Standards for Informational/Nonfiction Text (RI)	1.00	-	-
	Reading Standards for Literature/Fiction (RL)	0.97	1.00	-
	Writing and Language Standards (WL)	0.83	0.82	1.00

**For these tests, the second-order model was run by constraining the residual variance of a certain factor to zero due to non-significant negative residual variance.*

Table 21: Correlations Among Mathematics Factors

Grade	Reporting Category	MDG	NBT	NF	OA
3	Measurement, Data, and Geometry (MDG)	1.00	-	-	-
	Number and Operations in Base Ten (NBT)	0.91	1.00	-	-
	Number and Operations – Fractions (NF)	0.85	0.82	1.00	-
	Operations and Algebraic Thinking (OA)	0.94	0.91	0.85	1.00
4	Measurement, Data, and Geometry (MDG)	1.00	-	-	-
	Number and Operations in Base Ten (NBT)	0.90	1.00	-	-
	Number and Operations – Fractions (NF)	0.92	0.89	1.00	-
	Operations and Algebraic Thinking (OA)	0.95	0.92	0.94	1.00
5	Measurement, Data, and Geometry (MDG)	1.00	-	-	-
	Number and Operations in Base Ten (NBT)	0.94	1.00	-	-
	Number and Operations – Fractions (NF)	0.93	0.92	1.00	-
	Operations and Algebraic Thinking (OA)	0.92	0.91	0.90	1.00
		EE	**G**	**RPNS**	**SP**

6	Expressions and Equations (EE)	1.00	-	-	-
	Geometry (G)	0.85	1.00	-	-
	Ratios and Proportional Relationships and Number Systems (RPNS)	0.94	0.90	1.00	-
	Statistics and Probability (SP)	0.90	0.86	0.94	1.00
7	Expressions and Equations (EE)	1.00	-	-	-
	Geometry (G)	0.88	1.00	-	-
	Ratios and Proportional Relationships and Number Systems (RPNS)	0.92	0.94	1.00	-
	Statistics and Probability (SP)	0.83	0.85	0.89	1.00
		EENS	**F**	**G**	**SP**
8	Expressions and Equations and Number Systems (EENS)	1.00	-	-	-
	Functions (F)	0.87	1.00	-	-
	Geometry (G)	0.89	0.82	1.00	1.00
	Statistics and Probability (SP)	0.93	0.85	0.87	-
		A	**F**	**G**	**S**
10	Algebra (A)	1.00	-	-	-
	Functions (F)	0.97	1.00	-	-
	Geometry (G)	0.96	0.94	1.00	-
	Statistics and Probability (S)	0.97	0.96	0.95	1.00

**For these tests, the second-order model was run by constraining the residual variance of a certain factor to zero due to non-significant negative residual variance.*

5.2.3 Discussion

In all scenarios, the empirical results suggest that the implied model fits the data well. These results indicate that reporting an overall score in addition to separate scores for the individual reporting categories is reasonable, as the intercorrelations among items suggest that there are detectable distinctions among reporting categories.

The correlations among the separate factors are high, which is reasonable. This correlation supports the measurement model, given that the calibration of all items is performed concurrently. If the correlations among factors were exceptionally low, this could possibly suggest that a different IRT model would be needed (e.g., multidimensional IRT) or that the IRT calibration should be performed separately for items measuring different factors. The high correlations among the factors suggest that these alternative methods are unnecessary and that our current approach is preferable.

Overall, these results provide empirical evidence and justifies using our scoring and reporting methods. The results also justify the current IRT model employed.

5.3 LOCAL INDEPENDENCE

The validity of the application of IRT depends greatly on meeting the underlying assumptions of the models. One such assumption is local independence, which means that for a given proficiency estimate, the (marginal) likelihood is maximized, assuming the probability of correct responses is the product of independent probabilities over all items (Chen & Thissen, 1997):

$$\mathrm{L}(\theta) = \int \prod_{i=1}^{I} \Pr(z_i|\theta)\, f(\theta) \mathrm{d}\theta$$

When local independence is not met, there are issues of multidimensionality that are unaccounted for in the modeling of the data (Bejar, 1980). In fact, Lord (1980) noted that "local independence follows automatically from unidimensionality" (as cited in Bejar, 1980, p.5). From a dimensionality perspective, there may be nuisance factors influencing relationships among certain items after accounting for the intended construct of interest. These nuisance factors can be influenced by various testing features, such as speededness, fatigue, item chaining, and item or response formats (Yen, 1993).

Yen's Q_3 statistic (Yen, 1984) was used to measure local independence, which was derived from the correlation between the performances of two items. Simply, the Q_3 statistic is the correlation among IRT residuals and is computed using the following equations:

$$d_{ij} = u_{ij} - T_i\left(\hat{\theta}_j\right).$$

where u_{ij} is the item score of the *j*th test taker for item *i*, $T_i\left(\hat{\theta}_j\right)$ is the estimated true score for item *i* of test taker *j*, which is defined as

$$T_i\left(\hat{\theta}_j\right) = \sum_{l=1}^{m} y_{il} P_{il}\left(\hat{\theta}_j\right)$$

where y_{il} is the weight for response category *l*, *m* is the number of response categories, and $P_{il}(\hat{\theta}_j)$ is the probability of response category *l* to item *i* by test taker *j* with the ability estimate $\hat{\theta}_j$.

The pairwise index of local dependence Q_3 between item *i* and item *i'* is

$$Q_{3ii'} = r(d_i, d_{i'}),$$

where *r* refers to the Pearson product-moment correlation.

When there are n items, n(n-1)/2, Q_3 statistics will be produced. The Q_3 values are expected to be small. Table 22 and Table 23 present summaries of the distributions of Q_3 statistics—minimum, 5th percentile; median, 95th percentile; and maximum values from each grade and subject. The results show that about 90% of the items, between the 5th and 95th percentiles for all grades and subjects, were smaller than the critical value of 0.20 for $|Q_3|$ (Chen & Thissen, 1997).

Table 22: ELA Q_3 Statistic

Grade	Q_3 Distribution					Within Passage Q_3**	
	Minimum	5th Percentile	Median	95th Percentile	Maximum*	Minimum	Maximum
3	-0.193	-0.108	-0.012	0.029	0.87	-0.094	0.177
4	-0.116	-0.072	-0.019	0.030	0.701	-0.06	0.097
5	-0.148	-0.080	-0.018	0.023	0.842	-0.116	0.104
6	-0.154	-0.080	-0.016	0.027	0.581	-0.043	0.118
7	-0.209	-0.101	-0.008	0.022	0.92	-0.040	0.086
8	-0.170	-0.096	-0.014	0.030	0.849	-0.043	0.221
10	-0.210	-0.096	-0.011	0.030	0.811	-0.051	0.082

**Maximum Q_3 values are from elaboration and organization dimensions of the writing prompt.*
***Within Passage Q_3, values are computed for each item pair within a passage.*

Table 23: Mathematics Q_3 Statistic

Grade	Q_3 Distribution				
	Minimum	5th Percentile	Median	95th Percentile	Maximum
3	-0.103	-0.067	-0.025	0.028	0.224
4	-0.130	-0.069	-0.025	0.026	0.192
5	-0.133	-0.075	-0.023	0.026	0.245
6	-0.107	-0.064	-0.023	0.024	0.190
7	-0.106	-0.073	-0.024	0.026	0.374
8	-0.105	-0.076	-0.021	0.028	0.459
10	-0.131	-0.070	-0.017	0.032	0.118

In the 2020–2021 school year, the NDSA was administered as an adaptive test. When calculating the Q_3 statistics, pairwise deletion was used in the fixed-form tests. Therefore, Q_3 provides biased estimates under the CAT administration. Due to this restriction, Q_3 statistics were not calculated for the spring 2021 administration.

5.4 Convergent and Discriminant Validity

According to Standard 1.14 of the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014), it is necessary to provide evidence of convergent and discriminant validity evidence. It is a part of validity evidence demonstrating that assessment scores are related as expected with criterion and other variables for all student groups. However, a second independent test measuring the same constructs as ELA and mathematics in North Dakota, which could easily permit a cross test set of correlations, was not available. Therefore, the correlations between subscores within and across ELA and mathematics were examined alternatively. The *a-priori* expectation is that subscores within the same subject (e.g., mathematics) will correlate more positively than subscore correlations across subjects (e.g., mathematics and ELA). These correlations are based on a small number of items (e.g., typically around 8–18); consequently, the

observed score correlations will be smaller in magnitude due to the very large measurement error at the subscore level. For this reason, the observed score and the disattenuated correlations are provided.

Observed and disattenuated subscore correlations were calculated both within and across subjects for grades 3–8 and grade 10 ELA and mathematics. Table 24–Table 30 shows the observed and disattenuated score correlations between ELA and mathematics subscores for grades 3–8 and grade 10, where students took both subjects. In general, the pattern is consistent with the *a-priori* expectation that subscores within a test correlate more highly than correlations between tests measuring a different construct with a few small notes on the writing dimensions.

Table 24: Grade 3 Correlations Across Subjects

Subject	Reporting Category	Mathematics				ELA		
		MDG	NBT	NF	OA	RI	RL	WL
Mathematics	Measurement, Data, and Geometry (MDG)	0.72	0.84	0.81	0.86	0.67	0.65	0.70
	Number and Operations in Base Ten (NBT)	0.63	0.79	0.71	0.83	0.60	0.57	0.64
	Number and Operations – Fractions (NF)	0.59	0.54	0.74	0.74	0.59	0.57	0.60
	Operations and Algebraic Thinking (OA)	0.65	0.66	0.57	0.80	0.62	0.59	0.69
ELA	Reading Standards for Informational/Nonfiction Text (RI)	0.47	0.44	0.42	0.46	0.69	0.70	0.69
	Reading Standards for Literature/Fiction (RL)	0.47	0.43	0.42	0.45	0.50	0.73	0.68
	Writing and Language Standards (WL)	0.53	0.51	0.46	0.55	0.51	0.52	0.80

**Diagonal value represents the reliability coefficient of the reporting category. Observed correlations are below the diagonal, and disattenuated are above.*

Table 25: Grade 4 Correlations Across Subjects

Subject	Reporting Category	Mathematics				ELA		
		MDG	NBT	NF	OA	RI	RL	WL
Mathematics	Measurement, Data, and Geometry (MDG)	0.76	0.85	0.84	0.86	0.62	0.66	0.64
	Number and Operations in Base Ten (NBT)	0.64	0.75	0.84	0.91	0.60	0.63	0.65
	Number and Operations – Fractions (NF)	0.65	0.64	0.78	0.85	0.60	0.63	0.64
	Operations and Algebraic Thinking (OA)	0.66	0.69	0.66	0.77	0.64	0.67	0.68
ELA	Reading Standards for Informational/Nonfiction Text (RI)	0.45	0.43	0.44	0.47	0.69	0.76	0.67
	Reading Standards for Literature/Fiction (RL)	0.50	0.47	0.48	0.51	0.55	0.75	0.72
	Writing and Language Standards (WL)	0.49	0.50	0.50	0.53	0.49	0.55	0.78

**Diagonal value represents the reliability coefficient of the reporting category. Observed correlations are below the diagonal, and disattenuated are above.*

Table 26: Grade 5 Correlations Across Subjects

Subject	Reporting Category	Mathematics				ELA		
		MDG	NBT	NF	OA	RI	RL	WL
Mathematics	Measurement, Data, and Geometry (MDG)	0.70	0.85	0.83	0.86	0.66	0.68	0.69
	Number and Operations in Base Ten (NBT)	0.61	0.73	0.80	0.86	0.62	0.61	0.68
	Number and Operations – Fractions (NF)	0.60	0.59	0.75	0.79	0.62	0.63	0.64
	Operations and Algebraic Thinking (OA)	0.61	0.62	0.58	0.72	0.70	0.69	0.75
ELA	Reading Standards for Informational/Nonfiction Text (RI)	0.46	0.44	0.45	0.50	0.70	0.80	0.73
	Reading Standards for Literature/Fiction (RL)	0.49	0.45	0.47	0.51	0.58	0.75	0.74
	Writing and Language Standards (WL)	0.52	0.52	0.50	0.57	0.55	0.58	0.81

**Diagonal value represents the reliability coefficient of the reporting category. Observed correlations are below the diagonal, and disattenuated are above.*

Table 27: Grade 6 Correlations Across Subjects

Subject	Reporting Category	Mathematics				ELA		
		EE	G	RPNS	SP	RI	RL	WL
Mathematics	Expressions and Equations (EE)	0.72	0.73	0.91	0.72	0.63	0.68	0.71
	Geometry (G)	0.49	0.63	0.77	0.68	0.56	0.60	0.61
	Ratios and Proportional Relationships and Number Systems (RPNS)	0.67	0.53	0.76	0.76	0.65	0.68	0.72
	Statistics and Probability (SP)	0.48	0.42	0.52	0.61	0.63	0.67	0.66
ELA	Reading Standards for Informational/Nonfiction Text (RI)	0.46	0.38	0.49	0.42	0.74	0.77	0.69
	Reading Standards for Literature/Fiction (RL)	0.50	0.41	0.51	0.45	0.57	0.74	0.75
	Writing and Language Standards (WL)	0.54	0.43	0.56	0.46	0.53	0.58	0.80

**Diagonal value represents the reliability coefficient of the reporting category. Observed correlations are below the diagonal, and disattenuated are above.*

Table 28: Grade 7 Correlations Across Subjects

Subject	Reporting Category	Mathematics				ELA		
		EE	G	RPNS	SP	RI	RL	WL
Mathematics	Expressions and Equations (EE)	0.73	0.77	0.83	0.78	0.66	0.65	0.64
	Geometry (G)	0.54	0.68	0.86	0.81	0.66	0.66	0.68
	Ratios and Proportional Relationships and Number Systems (RPNS)	0.63	0.63	0.79	0.87	0.69	0.68	0.68
	Statistics and Probability (SP)	0.54	0.54	0.63	0.66	0.70	0.69	0.68
ELA	Reading Standards for Informational/Nonfiction Text (RI)	0.48	0.46	0.52	0.48	0.72	0.83	0.76
	Reading Standards for Literature/Fiction (RL)	0.47	0.46	0.51	0.47	0.59	0.71	0.76
	Writing and Language Standards (WL)	0.48	0.49	0.53	0.48	0.56	0.56	0.76

**Diagonal value represents the reliability coefficient of the reporting category. Observed correlations are below the diagonal, and disattenuated are above.*

Table 29: Grade 8 Correlations Across Subjects

Subject	Reporting Category	Mathematics				ELA		
		EENS	F	G	SP	RI	RL	WL
Mathematics	Expressions and Equations and Number Systems (EENS)	0.76	0.88	0.81	0.91	0.67	0.67	0.71
	Functions (F)	0.62	0.65	0.72	0.84	0.69	0.67	0.69
	Geometry (G)	0.60	0.49	0.72	0.73	0.56	0.55	0.60
	Statistics and Probability (SP)	0.68	0.58	0.53	0.74	0.68	0.67	0.71
ELA	Reading Standards for Informational/Nonfiction Text (RI)	0.50	0.48	0.41	0.50	0.74	0.79	0.71
	Reading Standards for Literature/Fiction (RL)	0.50	0.46	0.40	0.49	0.58	0.73	0.70
	Writing and Language Standards (WL)	0.56	0.50	0.46	0.55	0.55	0.54	0.81

**Diagonal value represents the reliability coefficient of the reporting category. Observed correlations are below the diagonal, and disattenuated are above.*

Table 30: Grade 10 Correlations Across Subjects

Subject	Reporting Category	Mathematics				ELA		
		A	F	G	S	RI	RL	WL
Mathematics	Algebra (A)	0.77	0.77	0.78	0.71	0.68	0.64	0.65
	Functions (F)	0.55	0.67	0.75	0.68	0.66	0.63	0.62
	Geometry (G)	0.60	0.54	0.77	0.72	0.68	0.64	0.67
	Statistics and Probability (S)	0.54	0.48	0.55	0.75	0.62	0.59	0.61
ELA	Reading Standards for Informational/Nonfiction Text (RI)	0.51	0.46	0.51	0.46	0.73	0.84	0.76
	Reading Standards for Literature/Fiction (RL)	0.47	0.43	0.47	0.43	0.60	0.70	0.73
	Writing and Language Standards (WL)	0.52	0.46	0.53	0.48	0.59	0.55	0.82

**Diagonal value represents the reliability coefficient of the reporting category. Observed correlations are below the diagonal, and disattenuated are above.*

Additionally, the correlation was computed among the overall scores for ELA and mathematics. Correlations are presented in Table 31 and are relatively high, between 0.69–0.72. Disattenuated correlations also ranged from 0.78–0.81.

Table 31: Correlations Across Spring 2021 ELA and Mathematics

Grade	N	ELA Reliability	Mathematics Reliability	Correlation	Disattenuated Correlation
3	8842	0.88	0.92	0.70	0.78
4	8493	0.88	0.92	0.69	0.77
5	8506	0.90	0.89	0.70	0.78
6	8475	0.89	0.89	0.72	0.81
7	8474	0.88	0.88	0.70	0.80
8	8284	0.89	0.90	0.72	0.80
10	2911	0.89	0.86	0.70	0.80

5.5 Relationship of Test Scores to External Variables

The relationship of test scores to external variables, measuring the same or related constructs, is an important source of validity evidence. The NDSA was first administered to students during the spring of 2018, replacing the Smarter Balanced Assessment Consortium (SBAC) assessments in ELA and mathematics. Ideally, we would correlate two different tests measuring a common construct administered within a similar time period.

Here, we present correlations between two different tests measuring a common construct but measured one summative test administration apart. We expect the correlations to be high, suggesting that the NDSA has a high relationship with an externally developed measure; the time gap between the two different assessments will be lower than if the two tests were measured within a similar testing window. Table 32 and Table 33 present correlations between the NDSA scores from spring 2019 and spring 2021. Correlations are between 0.71–0.78, which can be considered relatively high compared to industry standards. Additionally, disattenuated correlations are between 0.79–0.86.

Table 32: Correlations Between Spring 2019 Scores and Spring 2021 Scores (ELA)

Spring 2019 Grade	Spring 2021 Grade	N	Spring 2019 Marginal Reliability	Spring 2021 Marginal Reliability	Correlations	Disattenuated Correlations
3	5	7755	0.89	0.90	0.71	0.79
4	6	7751	0.89	0.89	0.73	0.82
5	7	7700	0.88	0.88	0.74	0.84
6	8	7636	0.88	0.89	0.75	0.85
8	10	2696	0.90	0.89	0.75	0.84

Table 33: Correlations Between Spring 2019 Scores and Spring 2021 Scores (Mathematics)

Spring 2019 Grade	Spring 2021 Grade	N	Spring 2019 Marginal Reliability	Spring 2021 Marginal Reliability	Correlations	Disattenuated Correlations
3	5	7803	0.90	0.90	0.75	0.83
4	6	7802	0.90	0.89	0.75	0.84
5	7	7731	0.89	0.88	0.78	0.88
6	8	7675	0.89	0.90	0.77	0.86
8	10	2723	0.88	0.86	0.74	0.85

6. FAIRNESS IN CONTENT

The principles of universal design of assessments provide guidelines for test design to minimize the impact of construct-irrelevant factors in assessing student achievement. Universal design removes barriers to provide access for the widest range of students possible. Seven principles of universal design are applied in the process of test development (Thompson, Johnstone, & Thurlow, 2002). They include the following:

1. Inclusive assessment population
2. Precisely defined constructs
3. Accessible, non-biased items
4. Amenable to accommodations
5. Simple, clear, and intuitive instructions and procedures
6. Maximum readability and comprehensibility
7. Maximum legibility

Content experts have received extensive training on the principles of universal design and apply these principles in the development of all test materials. In the review process, adherence to the principles of universal design is verified by North Dakota leadership.

6.1 STATISTICAL FAIRNESS IN ITEM STATISTICS

Analysis of the content alone is not sufficient to determine the fairness of a test. Rather, it must be accompanied by statistical processes. While various item statistics were reviewed during form building to evaluate the quality of items, one notable statistic used was differential item functioning (DIF). Items were classified into three categories (A, B, or C) for DIF, ranging from no evidence of DIF to severe DIF, according to the DIF classification convention illustrated in Volume 1 of this technical report. Furthermore, items were categorized positively (i.e., +A, +B, or +C), signifying that the item favored the focal group (e.g., African-American/Black, Hispanic, Female), or negatively (i.e., -A, -B, or -C), signifying that the item favored the reference group (e.g., White, Male). Items were flagged if their DIF statistics indicated the "C" category for any group. A DIF classification of "C" indicates that the item shows significant DIF and should be reviewed for potential content bias, differential validity, or other issues that may reduce item fairness. Items were reviewed by the Bias and Sensitivity Committee regardless of whether the DIF statistic favored the focal or the reference group. The details surrounding this review of items for bias is further described in Volume 2 of this technical report.

DIF analyses were conducted to detect potential item bias from a statistical perspective across major ethnic and gender groups. Specifically, DIF analyses were performed for the following groups:

- Male/Female
- White/African-American

- White/Hispanic
- White/Asian, Native Hawaiian, Pacific Islander
- White/Native American
- White/Multiracial

A detailed description of the DIF analysis that was performed is presented in Volume 1, Section 4.2, of this technical report. The DIF statistics for each operational test item are presented in the appendices of Volume 1.

7. SUMMARY

This report is intended to provide a collection of reliability and validity evidence to support appropriate inferences from the observed test scores. The overall results can be summarized as follows:

- **Reliability.** Various measures of reliability are provided at the aggregate and subgroup levels, showing that the reliability of all tests is in line with acceptable industry standards.
- **Content validity.** Evidence is provided to support the assertion that content coverage on each form was consistent with test specifications of the blueprint across testing modes.
- **Internal structural validity.** Evidence is provided to support the selection of the measurement model, the tenability of local independence, and the reporting of an overall score and subscores at the reporting category levels.

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: Author.

Bejar, I. I. (1980). Biased assessment of program impact due to psychometric artifacts. *Psychological Bulletin, 87*(3), 513–524.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3*, 296–322.

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 37*, 62–83.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Chen, F., Kenneth, A., Bollen, P., Paxton, P., Curran, P. J., & Kirby, J. B. (2001). Improper Solutions in Structural Equation Models: Causes, Consequences, and Strategies. *Sociological Methods & Research, 29*, 468–508.

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*(3), 265–289.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*, 213–220.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart, and Winston.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika. 16*, 297–334.

Feldt, L. S., & Brennan, R. (1989). Reliability. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 105–146). New York: Macmillan.

Feldt, L. S., & Qualls, A. L. (1996). Bias in coefficient alpha arising from heterogeneity of test content. *Applied Measurement in Education, 9*, 277–286.

Guo, F. (2006). Expected classification accuracy using the latent distribution. *Practical Assessment, Research & Evaluation, 11*(6).

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55.

Jöreskog, K. G. (1994). On the estimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika, 59*(3), 381–389.

Kuder, G. F., & Richardson, M. W. (1937). The theory of estimation of test reliability. *Psychometrika, 2*(3), 151–160.

Lee, G., & Frisbie, D. A. (1999). Estimating reliability under a generalizability theory model for test scores composed of testlets. *Applied Measurement in Education, 12*, 237–255.

Lee, W., Hanson, B., & Brennan, R. (2002). Estimating consistency and accuracy indices for multiple classifications. *Applied Psychological Measurement, 26(4)*, 412–432.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13–103). New York: Macmillan.

Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*(1), 115–132.

Muthén, B. O., du Toit, S. H. C., & Spisic, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes*. Unpublished manuscript.

Muthén, L. K., & Muthén, B. O. (2012). Mplus user's guide. 7th Edition. Los Angeles, CA: Muthén & Muthén.

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, *44*, 443–460.

Qualls, A. L. (1995). Estimating the reliability of a test containing multiple item formats. *Applied Measurement in Education, 8*, 111–120.

Raju, N. S. (1977). A generalization of coefficient alpha. *Psychometrika*, *42*, 549–565.

Rudner, L. M. (2001). Computing the expected proportions of misclassified examinees. *Practical Assessment, Research & Evaluation, 7*(14).

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3*, 271–295.

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (NCEO Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved October 2002, from http://education.umn.edu/NCEO/OnlinePubs/Synthesis44.html.

van Driel, O, P. (1978). On various causes of improper solutions in maximum likelihood factor analysis. *Psychometrika, 43*, 225–43.

Williamson, D., Xi, X., & Breyer, F.J. (2012). A framework for the evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31(*1*),* 2–13.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125–145.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*(3), 187–213.

Yoon, B., & Young, M. J. (2000). *Estimating the reliability for test scores with mixed item formats: Internal consistency and generalizability.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.