North Dakota State Assessment for Science

2021–2022

Volume 4: Evidence of Reliability and Validity

NORTH DAKOTA DEPARTMENT OF
PUBLIC INSTRUCTION

TABLE OF CONTENTS

1. INTRODUCTION AND OVERVIEW OF RELIABILITY AND VALIDITY EVIDENCE ...1
1.1 Reliability ...2
1.2 Validity ...3
2. PURPOSE OF THE NORTH DAKOTA STATE ASSESSMENT FOR SCIENCE ...6
3. RELIABILITY ...7
3.1 Standard Error of Measurement ...8
3.2 Reliability of Achievement Classification ...10
3.2.1 Classification Accuracy ...10
3.2.2 Classification Consistency ...11
3.3 Precision at Cut Scores ...12
4. EVIDENCE OF CONTENT VALIDITY ...13
4.1 Content Standards ...13
4.2 Independent Alignment Study ...13
5. EVIDENCE OF INTERNAL-EXTERNAL STRUCTURE ...14
5.1 Correlations Among Discipline Scores ...14
5.2 Convergent and Discriminant Validity ...15
5.3 Cluster Effects ...19
5.4 Confirmatory Factor Analysis ...22
5.4.1 Results ...26
5.4.2 Conclusion ...30
6. FAIRNESS IN CONTENT ...31
6.1 Cognitive Laboratory Studies ...31
6.2 Statistical Fairness in Item Statistics ...32
7. SUMMARY ...33
8. REFERENCES ...34

LIST OF TABLES

Table 1. 2021–2022 Operational Assessment Modes 1
Table 2. Marginal Reliability Coefficients 7
Table 3. Classification Accuracy Index 11
Table 4. Classification Consistency Index 12
Table 5. Achievement Levels and Associated Conditional Standard Error of Measurement 12
Table 6. Number of Items for Each Discipline 13
Table 7. Correlations Among Disciplines 15
Table 8. Correlations Across Subjects, Grade 4 16
Table 9. Correlations Across Subjects, Grade 8 17
Table 10. Correlations Across Subjects, Grade 10 18
Table 11. Correlations Across Spring 2022 English Language Arts, Mathematics, and Science Scores 19
Table 12. Number of Forms, Clusters per Discipline (Range Across Forms), Number of Assertions per Form (Range Across Forms), and Number of Students per Form (Range Across Forms) 23
Table 13. Guidelines for Evaluating Goodness-of-Fit* 26
Table 14. Fit Measures per Model and Form, Grade 6 27
Table 15. Fit Measures per Model and Form, Grade 7 27
Table 16. Fit Measures per Model and Form, Grade 8 28
Table 17. Fit Measures per Model and Form—Grade 6—One Cluster Removed 29
Table 18. Model Implied Correlations per Form for the Disciplines in Model 4 29

LIST OF FIGURES

Figure 1. Conditional Standard Errors of Measurement 8
Figure 2. Cluster Variance Proportion for Operational Items in Elementary School 20
Figure 3. Cluster Variance Proportion for Operational Items in Middle School 21
Figure 4. Cluster Variance Proportion for Operational Items in High School 21
Figure 5. One-Factor Structural Model (Assertions-Overall Science): "Model 1" 24
Figure 6. Second-Order Structural Model (Assertions-Disciplines-Overall Science): "Model 2"24
Figure 7. Second-Order Structural Model (Assertions-Clusters-Overall Science): "Model 3".... 25
Figure 8. Third-Order Structural Model (Assertions-Clusters-Disciplines-Overall Science): "Model 4" 25

LIST OF APPENDICES

Appendix A. Student Demographics and Reliability Coefficients for NDSA for Science
Appendix B. Conditional Standard Error of Measurement for NDSA for Science
Appendix C. Classification Accuracy and Consistency Indices by Subgroups for NDSA for Science
Appendix D. Science Clusters Cognitive Lab Report
Appendix E. Braille Cognitive Lab Report
Appendix F. Independent Alignment Study Report

1. Introduction and Overview of Reliability and Validity Evidence

The new North Dakota Science Content Standards were adopted by the North Dakota Department of Public Instruction (NDDPI) in February 2019. Based on the Standards, the North Dakota State Assessment (NDSA) for Science was developed and first administered operationally in grades 4, 8, and 10 during the 2020–2021 school year. For the 2021–2022 school year, the NDSA for Science continued to be administered online using an adaptive test design. Accommodated braille versions and designated-support Spanish-language versions were available for each grade. For the 2021–2022 school year, remote testing forms were constructed to allow for assessing science among students taking the test remotely; however, before the testing window opened, NDDPI decided not to use these remote forms. For form construction procedures for the unused remote test forms, refer to Volume 2, Appendix L, Remote Testing Forms. Table 1 shows the complete list of summative tests for the operational test administrations.

Table 1. 2021–2022 Operational Assessment Modes

Language/Format	Assessment Mode	Grade
English	Online	4, 8, and 10
Spanish	Online	4, 8, and 10
Braille	Paper	4, 8, and 10

Given the intended uses of these tests, evidence for both reliability and validity is necessary to support appropriate inferences of student academic achievement based on the NDSA for Science scores. The analyses to support reliability and validity evidence that are reported in this volume were conducted based on test results for students whose scores were reported, including those taking the standard and the accommodated versions of the NDSA for Science.

The purpose of this report is to provide empirical evidence that supports a validity argument for the uses of and inferences from the NDSA for Science. This volume addresses the following five topics:

1. ***Reliability.*** The reliability estimates are presented by grade and demographic subgroups. This section also includes conditional standard error of measurement (CSEM), classification accuracy (CA), and classification consistency (CC) results by grade.
2. ***Content Validity.*** This section presents evidence showing that test forms were constructed to measure the three-dimensional 2019 North Dakota Science Content Standards with a sufficient number of items targeting each area of the test blueprint.
3. ***Internal Structure Validity.*** This section provides evidence regarding the internal relationships among the subscale scores to support their use and to justify the item response theory (IRT) measurement model. This type of evidence includes observed and disattenuated Pearson correlations among discipline scores per grade. The IRT model is a multi-dimensional model with an overall dimension representing proficiency in science and nuisance dimensions that consider within-item local dependencies

among scoring assertions (refer to Section 5.1 of Volume 1, Annual Technical Report). In this volume, evidence is provided for the presence of item cluster effects. Additionally, confirmatory factor analysis (CFA) is used to evaluate the fit of the IRT model and to compare it to alternative models, including models with a simpler internal structure (i.e., unidimensional models) and models with a more elaborate internal structure.

4. ***Relationship of Test Scores to External Variables.*** This section provides evidence of convergent and discriminant validity using observed and disattenuated subscore correlations both within and across subjects.
5. ***Test fairness.*** This section details how fairness is an explicit concern during item development. Items are developed following the principles of universal design (UD), which removes barriers to enable access for the widest possible range of students. Test fairness is further monitored statistically through the use of differential item functioning (DIF) analysis in tandem with content reviews by specialists.

1.1 RELIABILITY

Reliability refers to consistency in test scores and can be defined as the degree to which individuals' deviation scores remain relatively consistent over repeated administrations of the same test or alternate test forms (Crocker & Algina, 1986). For example, if a person takes the same or parallel tests repeatedly, they 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, the 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 SEM under the assumption of CTT, 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:

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

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

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

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

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

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

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

In contrast, the SEMs in IRT vary over the ability continuum. These heterogeneous errors are a function of a test information function (TIF) that provides different information about examinees depending on their estimated abilities.

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 cut score at various cut score points. Refer to Section 3.3, Precision at Cut Scores, for the derivation of heterogeneous measurement errors in IRT, and how these errors are aggregated over the cut score distribution to obtain a single, marginal, IRT-based reliability coefficient.

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], & 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 or other modes of assessment" (p. 13). Both definitions emphasize evidence and theory to support inferences and interpretations of test scores. The *Standards for Educational and Psychological Testing* (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 evidence for validity is the relationship between the test content and the intended test construct (refer to Section 4, Evidence of Content Validity). 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 test scores. To determine content representativeness, diverse panels of content experts conduct alignment studies in which experts review individual items and rate them based on how well they match the test specifications or cognitive skills required for a construct (refer to Section 4.2, Independent Alignment Study, for the results of an independent alignment study; and Volume 2, Test Development, for details on the

item development process). Technology-enhanced items should be examined to ensure that no construct-irrelevant variance is introduced; if some aspect of the technology impedes or advantages a student in their responses to items, this could affect item responses and inferences regarding abilities on the measured construct (refer to Volume 2, Test Development).

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, p. 12). This evidence is collected by surveying test takers about their performance strategies or responses to specific items. Because items are developed to measure specific constructs and intellectual processes, evidence that examinees have engaged in relevant performance strategies to correctly answer the items supports the validity of the test scores.

The third source of evidence for validity is based on *internal structure*, which is the degree to which the relationships among test items and test components relate to the construct on which the proposed test scores are interpreted. Dimensionality assessment, goodness-of-model-fit to data, and reliability analysis are possible analyses to examine internal structure (refer to Section 3, Reliability, and Section 5, Evidence of Internal-External Structure). It is important to assess to which degree the statistical relation between items and test components is invariant across groups. DIF analysis can be used to assess whether specific items function differently for subgroups of test takers (refer to Section 4.4 of Volume 1, Annual Technical Report).

The fourth source of evidence for validity is the relationship of test scores to external variables. The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) divides this source of evidence into three parts: (1) convergent and discriminant evidence; (2) test-criterion relationships; and (3) 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 intended to assess different constructs. To analyze both convergent and discriminant evidence, a multitrait-multimethod matrix can be used. Additionally, test-criterion relationships indicate how accurately test scores predict criterion performance. The degree of accuracy depends on the test's purpose, such as classification, diagnosis, or selection. Test-criterion evidence is 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 is the intended and unintended consequences of test use, which should be included in the test-validation process. Determining the validity of the test should depend on evidence directly related to the test; this process should not be influenced by external factors. For example, if an employer administers a test to determine 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 in fact due to an unintended, confounding aspect of the test, this would interfere with the test's validity. Test use should align with the intended purpose of the test.

A successful validity argument requires multiple sources of validity evidence that enable one to evaluate whether sufficient evidence exists to support the intended uses and interpretations of the test scores. Thus, determining the validity of a test first requires an explicit statement regarding the intended uses of the test scores and, second, evidence that the scores can be used to support these inferences.

2. PURPOSE OF THE NORTH DAKOTA STATE ASSESSMENT FOR SCIENCE

The primary purpose of the NDSA for Science is to yield accurate information on students' achievement of the North Dakota's Science Content Standards. The NDSA for Science measures the science knowledge and skills of North Dakota students in grades 4, 8, and 10. The North Dakota Department of Public Instruction (NDDPI) provides an overview of the NDSA for Science at: www.nd.gov/dpi/districtsschools/assessment/ndsa. Information about the 2019 North Dakota Science Content Standards is available at: www.nd.gov/dpi/districtsschools/k-12-education-content-standards.

The NDSA for Science supports instruction and student learning by measuring growth in student achievement. Assessments can be used as indicators to determine whether students in North Dakota are ready with the knowledge and skills that are essential for college and career readiness.

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

The assessments are constructed to measure student proficiency in accordance with best practices as described in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014). The NDSA for Science was developed in adherence to the principles of universal design (UD) to ensure that all students have access to the test content (refer to Volume 2, Test Development, for a description of the NDSA for Science standards and test blueprints in more detail; refer to Section 4, Evidence of Content Validity, for additional evidence of content validity). The NDSA for Science test scores are useful indicators for understanding individual students' academic achievement of the 2019 North Dakota Science Content Standards and for evaluating whether a student is progressing in performance over time. Additionally, both individual and aggregated scores can be used to for measuring the reliability of the test (refer to Section 3, Reliability, for the reliability of the test scores).

The NDSA for Science is a standard-referenced test that is designed to measure student performance on the three-dimensional science standards in North Dakota schools. As a comparison, norm-referenced tests are designed to rank or compare all students with one another (refer to Volume 2, Test Development, for the NDSA for Science standards and test blueprints).

The scale score and relative strengths and weaknesses at the discipline level are provided for each student to indicate student strengths and weaknesses in different content areas of the test, relative to the other areas and to the district and state. These scores serve as useful feedback that teachers can use to tailor their instruction. To support their practical use across the state, we examine the reliability coefficients for and the validity of these test scores.

3. RELIABILITY

Classical test theory (CTT)-based reliability indices are not appropriate for science assessments for two reasons. First, in spring 2022, the NDSA for Science was administered using an adaptive test design. Each student could potentially get a unique set of items, whereas CTT-based reliability indices require that the same set of items be administered to a large group of students. Second, since item response theory (IRT) methods are used for calibration and scoring, the measurement error of ability estimates is not constant across the ability range, even for the same set of items. The reliability of science is computed as follows:

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

where N is the number of students; $CSEM_i$ is the conditional standard error of measurement (CSEM) of the overall ability estimate for student i; and σ^2 is the variance of the overall ability estimates. The higher the reliability coefficient, the greater the precision of the test.

The marginal reliability of science for the overall sample is reported by grade in Table 2. The overall reliability ranges from 0.84 to 0.85. Due to the new structure of the test, Cambium Assessment, Inc. (CAI) has explored the relationships between reliability and other important factors, such as the effect of nuisance dimensions (refer to Section 5 of Volume 1, Annual Technical Report). It was found that if the local dependencies among assertions pertaining to the same item were ignored, the marginal reliability would increase. Ignoring local dependencies can be achieved either by computing the maximum likelihood estimates (MLE) ability estimates under the unidimensional Rasch model or by setting the variance parameters to zero for all item clusters when computing the marginal maximum likelihood estimation (MMLE) ability estimates under the one-parameter logistic (1PL) bifactor model (refer to Section 6 of Volume 1, Annual Technical Report); thus, by ignoring the local dependencies, which are substantial for many item clusters, the reliability coefficient overestimates the test's true reliability. Note, however, that local dependencies are also present to some degree in traditional assessments that use item groups (e.g., a set of items relating to the same reading passage). Local dependencies are typically not accounted for by traditional assessments, and hence the reported reliability coefficients may overestimate, to some degree, the true reliability of these tests. The reliability coefficients are also reported for demographics subgroups and reporting categories in Appendix A, Student Demographics and Reliability Coefficients for the NDSA for Science.

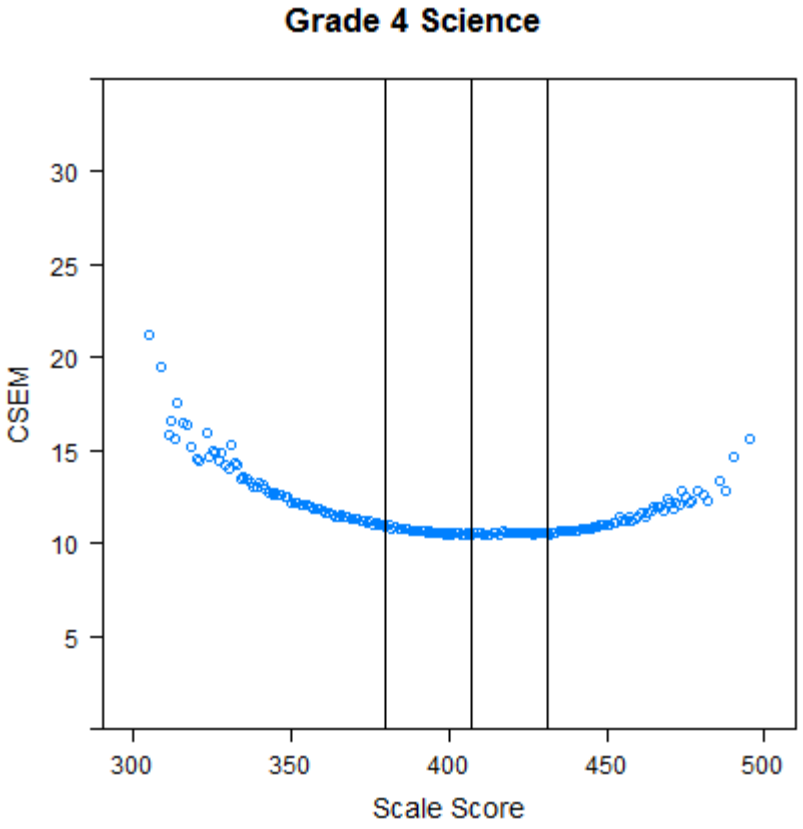
Table 2. Marginal Reliability Coefficients

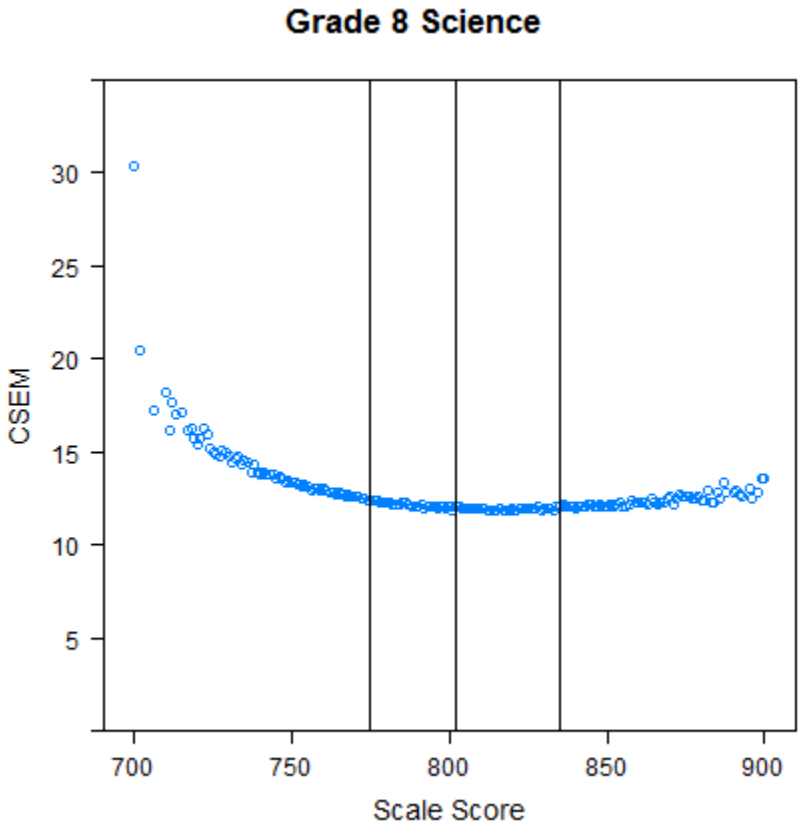
Grade	*N*	Reliability
4	9,131	0.85
8	8,736	0.85
10	7,873	0.84

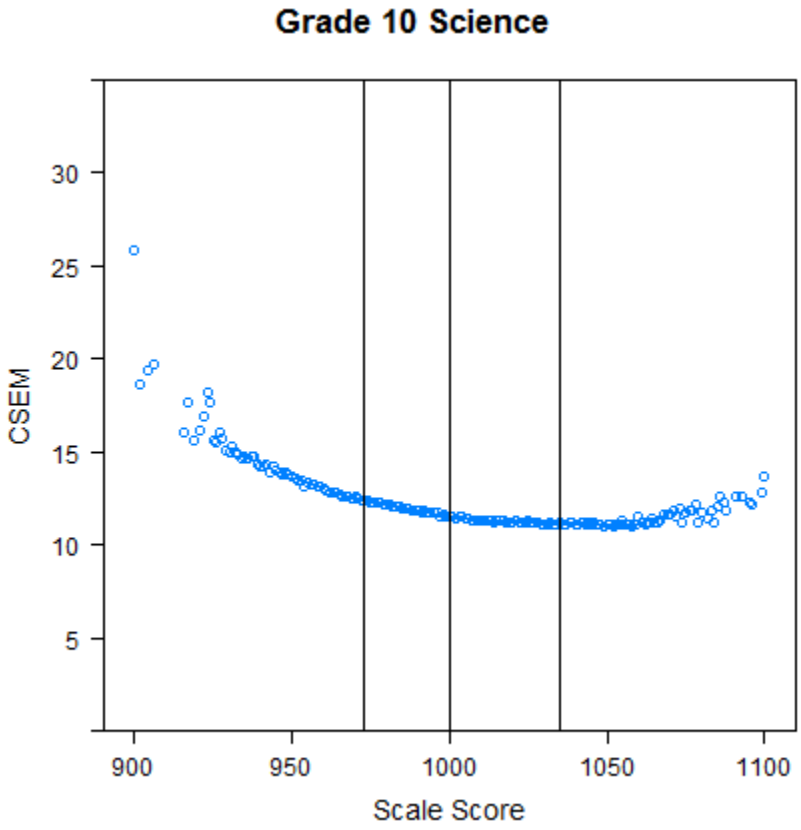
3.1 Standard Error of Measurement

The computation method of CSEM is described in Section 6.4 of Volume 1, Annual Technical Report. Figure 1 presents the average CSEM for each scale score. The lowest standard errors are observed near the proficiency cut score (the middle vertical line) for grades 4 and 8, which is a desirable test property. The CSEM at each scale score is reported in Appendix B, Conditional Standard Error of Measurement for the NDSA for Science.

Figure 1. Conditional Standard Errors of Measurement




Grade 8 Science
CSEM
30
25
20
15
10
5
700
750
800
850
900
Scale Score


Grade 10 Science
CSEM
30
25
20
15
10
5
900
950
1000
1050
1100
Scale Score

3.2 RELIABILITY OF ACHIEVEMENT CLASSIFICATION

When student achievement is reported in terms of achievement levels, 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 the classification accuracy (CA) and classification consistency (CC). CA refers to the agreement between the classifications based on the test taken and the classifications that would be made based on the students' true scores (if they could be obtained). CC refers to the agreement between the classifications based on the form taken and the classifications that would be made based on an alternate, equivalently constructed test form.

In reality, students' true scores (i.e., true abilities) are unknown, and students are not administered an alternate, equivalent form. Therefore, CA and CC are estimated on the basis of students' item scores, the item parameters, and the assumed latent ability distribution, as described in the following sections. The true score is an expected value of the test score with measurement error.

For student j, the student's estimated ability is $\hat{\theta}_j$ with a standard error of measurement (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 score of student j. The probability of the true score at achievement 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 achievement level l, respectively.

3.2.1 Classification Accuracy

Using p_{jl}, an $L \times L$ matrix $\boldsymbol{E_A}$ can be calculated. Each element E_{Akl} of matrix $\boldsymbol{E_A}$ represents the expected number of students at level l (based on their true scores) given students from observed level k, and can be calculated as

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

where pl_j is the jth student's observed achievement level. The 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 achievement level k.

The classification for the pth cut score is estimated by forming square partitioned blocks of the matrix $\boldsymbol{E_A}$ and taking the summation over all elements within the block as follows:

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

where N is the total number of students.

The overall CA is estimated from the diagonal elements of the matrix:

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

Table 3 provides the CA for the individual cut scores. The overall CA of the test ranges from 73.43% to 73.76%. The individual cut score accuracy rates are high across all grades and forms, with the minimum value being 88.62% for grade 8. It denotes that more than 88% of the time, we can accurately differentiate students between adjacent achievement levels in the 2021−2022 NDSA for Science. The CA for demographic subgroups is presented in Appendix C, Classification Accuracy and Consistency Indices by Subgroups for NDSA for Science.

Table 3. Classification Accuracy Index

Grade	Overall Accuracy (%)	Classification Accuracy (%)		
		Level 2 Cut Score	*Level 3 Cut Score*	*Level 4 Cut Score*
4	73.63	89.92	89.85	93.72
8	73.76	91.11	88.62	93.87
10	73.43	90.42	88.80	94.09

3.2.2 Classification Consistency

Assuming the test is administered twice independently to the same group of students, similarly to accuracy, a $L \times L$ 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 achievement level l in the first administration, and p_{jk} is the probability of the true score at achievement level k in the second administration for the jth student. The classification consistency index for the cuts (CCC) and overall CC were estimated in a way similar to the classification accuracy for the cuts and CA.

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

and

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

Table 4 provides the CC for the cut scores. The overall CC of the test ranges from 63.56–64.10%. The individual cut score consistency rates are high across all grades and forms, with the minimum value being 84.12% for grade 8. In all achievement levels, CA is slightly higher than CC. The CC rates can be lower than CA; the CC is based on two tests with measurement errors, but the CA is based on one test with a measurement error and the true score. The CC for demographic subgroups is presented in Appendix C, Classification Accuracy and Consistency Indices by Subgroups for the NDSA for Science.

Table 4. Classification Consistency Index

Grade	Overall Consistency (%)	Classification Consistency (%)		
		Level 2 Cut Score	*Level 3 Cut Score*	*Level 4 Cut Score*
4	63.94	85.87	85.68	91.11
8	64.10	87.53	84.12	91.21
10	63.56	86.53	84.27	91.65

3.3 PRECISION AT CUT SCORES

Table 5 presents the mean CSEM at each achievement level by grade and includes achievement-level cut scores and associated CSEM. The CSEM at each scale score is reported in Appendix B, Conditional Standard Error of Measurement for the NDSA for Science.

Table 5. Achievement Levels and Associated Conditional Standard Error of Measurement

Grade	Achievement Level	Mean CSEM	Cut Score (Scale Score)	CSEM at Cut Score
4	Novice	11.68	-	-
	Partially Proficient	10.67	380	10.86
	Proficient	10.55	407	10.52
	Advanced	10.93	431	10.55
8	Novice	13.21	-	-
	Partially Proficient	12.17	775	12.45
	Proficient	11.96	802	12.04
	Advanced	12.22	835	12.06
10	Novice	13.27	-	-
	Partially Proficient	11.97	973	12.46
	Proficient	11.33	1,000	11.59
	Advanced	11.27	1,035	11.22

4. Evidence of Content Validity

The knowledge and skills assessed by the North Dakota State Assessment (NDSA) for Science are representative of the content standards of the larger knowledge domain. We describe the content standards for the NDSA for Science and discuss the test development process and mapping the NDSA for Science tests to the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014). A complete description of the test development process can be found in Volume 2, Test Development.

4.1 Content Standards

The NDSA for Science was aligned to the new three-dimensional North Dakota Science Content Standards, which were adopted by North Dakota in 2019. The 2019 North Dakota Science Content Standards are available for review at: www.nd.gov/dpi/districtsschools/k-12-education-content-standards. Test 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 construction process can be found in Volume 2, Test Development.

Table 6 presents the disciplines by grade, as well as the number of operational items administered to measure each discipline.

Table 6. Number of Items for Each Discipline

Grade	Discipline	Item Clusters	Stand-Alone Items
4	Earth and Space Sciences	8	15
	Life Sciences	12	21
	Physical Sciences	6	18
8	Earth and Space Sciences	9	20
	Life Sciences	8	32
	Physical Sciences	11	19
10	Life Sciences	17	40
	Physical Sciences	7	9

4.2 Independent Alignment Study

While it is critically important to develop and strictly enforce an item development process that works to ensure the alignment of test items to content standards, it is also important to independently verify the alignment of test items to content standards.

EdMetric LLC served as an external evaluator and conducted an alignment study in October 2021. The purpose of the study was to examine the extent to which the NDSA for Science item pool

represented the North Dakota Science Content Standards as represented by the test blueprints in terms of range, complexity, depth, and breadth.

The results of the alignment study are presented in Appendix F, Independent Alignment Study Report.

5. EVIDENCE OF INTERNAL-EXTERNAL STRUCTURE

In this section, the internal structure of the assessment is explored using the scores provided at the discipline level. The relationship between the discipline scores is just one indicator of test dimensionality. The North Dakota State Assessment (NDSA) for Science is modeled with the Rasch testlet model (Wang & Wilson, 2005). The item response theory (IRT) model is a high-dimensional model, incorporating a nuisance dimension for each item cluster (and stand-alone items with four or more assertions), in addition to an overall dimension representing the overall proficiency. This approach is innovative and quite different from the traditional approach of ignoring local dependencies. Validity evidence on the internal structure focuses on the presence of cluster effects and how substantial they are. Additionally, confirmatory factor analysis (CFA) is used to evaluate the fit of the IRT model and to compare the model to alternative models, including models with a simpler internal structure (i.e., unidimensional models without cluster effects) and models with a more elaborate internal structure.

Another pathway explores observed correlations between the discipline scores. However, as each discipline is measured with a small number of items, the standard errors of the observed scores within each discipline 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 subsection.

5.1 CORRELATIONS AMONG DISCIPLINE SCORES

Table 7 presents the observed and disattenuated correlation matrix of the discipline scores. The observed correlations range from 0.62 to 0.66, and disattenuated correlations range from 0.97 to 1.00.

In some instances, the observed correlations were lower than one might expect. However, as previously noted, the correlations were subject to a large amount of measurement error at the discipline 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. After correcting for measurement error, the correlations between the discipline scores become very high. The disattenuated correlations are close to 1, supporting the use of a psychometric model that does not include a separate dimension for each of the three disciplines.

Table 7. Correlations Among Disciplines

Grade	Discipline	Earth and Space Sciences (ESS)	Life Sciences (LS)	Physical Sciences (PS)
4	Earth and Space Sciences	0.63*	0.98	0.97
	Life Sciences	0.64	0.68*	0.97
	Physical Sciences	0.62	0.64	0.65*
8	Earth and Space Sciences	0.63*	0.99	1.00
	Life Sciences	0.66	0.71*	1.00
	Physical Sciences	0.61	0.64	0.57*
10	Life Sciences	NA	0.80*	1.00
	Physical Sciences	NA	0.66	0.54*

Note. *The values for cells shaded on the diagonal are marginal reliabilities for each discipline. Below the cells shaded on the diagonal are the observed correlations, and above the cells shaded on the diagonal are the disattenuated correlations. The disattenuated correlations larger than 1 were truncated to 1.

5.2 Convergent and Discriminant Validity

Collectively, Standard 1.16 through Standard 1.19 of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) emphasize practices to provide evidence of convergent and discriminant validity. It is a part of validity evidence demonstrating that assessment scores are related as expected with criteria and other variables for all student groups. However, a second, independent test measuring the same science construct as the NDSA for Science, which could easily permit for a cross-test set of correlations, was not available. Alternatively, the correlations between subscores were examined. The a priori expectation is that subscores within the same subject (e.g., correlations of science disciplines within science) will correlate more positively than subscores correlations across subjects (e.g., correlation of science disciplines with reporting categories within mathematics). These correlations are based on a small number of items; consequently, the observed score correlations will be smaller in magnitude as a result of the larger measurement error at the subscore level. For this reason, both the observed score and the disattenuated correlations are provided.

Observed and disattenuated subscore correlations were calculated both within and across subjects. The pattern is generally consistent with the a priori expectation that subscores within a test have higher correlations than correlations between tests measuring a different construct. The correlations between reporting categories from science, English language arts (ELA), and mathematics are presented in Table 8 through Table 10. The shaded cells that form a diagonal show the reliability coefficient of the reporting category.

Table 8. Correlations Across Subjects, Grade 4

Subject	Number of Students	Reporting Category	Science			ELA			Mathematics			
			ESS	*LS*	*PS*	*RSI*	*RSL*	*W*	*M*	*NOBT*	*NOF*	*OAT*
Science	9,089	Earth and Space Sciences (ESS)	0.63*	0.98	0.97	0.79	0.79	0.71	0.76	0.74	0.74	0.79
		Life Sciences (LS)	0.64	0.68*	0.97	0.82	0.84	0.74	0.76	0.72	0.74	0.79
		Physical Sciences (PS)	0.62	0.64	0.65*	0.79	0.80	0.69	0.76	0.73	0.75	0.78
ELA		Reading Standards for Informational/Nonfiction Text (RSI)	0.53	0.57	0.53	0.71*	0.81	0.71	0.68	0.65	0.66	0.72
		Reading Standards for Literature/Fiction (RSL)	0.54	0.60	0.55	0.59	0.75*	0.75	0.66	0.64	0.65	0.70
		Writing and Language Standards (W)	0.49	0.54	0.49	0.52	0.56	0.76*	0.66	0.67	0.63	0.71
Mathematics		Measurement, Data, and Geometry (M)	0.53	0.55	0.53	0.50	0.50	0.50	0.76*	0.83	0.85	0.87
		Number and Operations in Base Ten (NOBT)	0.51	0.52	0.51	0.47	0.48	0.50	0.62	0.75*	0.82	0.89
		Number and Operations–Fractions (NOF)	0.52	0.54	0.53	0.49	0.49	0.49	0.65	0.62	0.78*	0.85
		Operations and Algebraic Thinking (OAT)	0.55	0.57	0.55	0.53	0.53	0.54	0.66	0.67	0.65	0.76*

*Cells shaded on the diagonal represent the reliability coefficient of the reporting category. Observed correlations are below the cells shaded on the diagonal; disattenuated correlations are above. The disattenuated correlations larger than 1 were truncated to 1.

Table 9. Correlations Across Subjects, Grade 8

Subject	Number of Students	Reporting Category	Science			ELA			Mathematics			
			ESS	*LS*	*PS*	*RSI*	*RSL*	*W*	*E*	*F*	*G*	*SP*
Science	8,645	Earth and Space Sciences (ESS)	0.63*	0.99	1.00	0.77	0.76	0.73	0.79	0.78	0.69	0.79
		Life Sciences (LS)	0.66	0.71*	1.00	0.78	0.78	0.74	0.79	0.78	0.67	0.79
		Physical Sciences (PS)	0.61	0.64	0.57*	0.79	0.78	0.75	0.81	0.79	0.67	0.81
ELA		Reading Standards for Informational/Nonfiction Text (RSI)	0.51	0.55	0.50	0.70*	0.83	0.78	0.70	0.70	0.59	0.70
		Reading Standards for Literature/Fiction (RSL)	0.50	0.55	0.49	0.58	0.69*	0.77	0.68	0.68	0.57	0.68
		Writing and Language Standards (W)	0.53	0.57	0.52	0.59	0.58	0.82*	0.73	0.71	0.62	0.71
Mathematics		Expressions and Equations and Number Systems (E)	0.54	0.57	0.53	0.50	0.49	0.57	0.75*	0.92	0.84	0.91
		Functions (F)	0.49	0.52	0.48	0.47	0.45	0.52	0.63	0.64*	0.74	0.88
		Geometry (G)	0.45	0.47	0.42	0.41	0.39	0.46	0.60	0.49	0.68*	0.76
		Statistics and Probability (SP)	0.53	0.57	0.52	0.50	0.49	0.55	0.67	0.60	0.53	0.73*

*Cells shaded on the diagonal represent the reliability coefficient of the reporting category. Observed correlations are below the cells shaded on the diagonal; disattenuated correlations are above. The disattenuated correlations larger than 1 were truncated to 1.

Table 10. Correlations Across Subjects, Grade 10

Subject	Number of Students	Reporting Category	Science		ELA			Mathematics			
			LS	*PS*	*RSI*	*RSL*	*W*	*A*	*F*	*G*	*SP*
Science	2,543	Life Sciences (LS)	0.79*	1.00	0.83	0.80	0.76	0.77	0.78	0.78	0.78
		Physical Sciences (PS)	0.65	0.53*	0.80	0.78	0.75	0.79	0.83	0.81	0.80
ELA		Reading Standards for Informational/Nonfiction Text (RSI)	0.62	0.49	0.71*	0.87	0.81	0.70	0.70	0.70	0.68
		Reading Standards for Literature/Fiction (RSL)	0.58	0.47	0.60	0.67*	0.77	0.65	0.64	0.64	0.61
		Writing and Language Standards (W)	0.61	0.50	0.62	0.57	0.83*	0.70	0.70	0.71	0.66
Mathematics		Algebra (A)	0.58	0.49	0.50	0.45	0.55	0.72*	0.81	0.80	0.81
		Functions (F)	0.54	0.47	0.46	0.41	0.50	0.54	0.61*	0.82	0.74
		Geometry (G)	0.60	0.51	0.51	0.44	0.56	0.58	0.54	0.74*	0.83
		Statistics and Probability (SP)	0.57	0.48	0.47	0.41	0.50	0.57	0.48	0.59	0.68*

*Cells shaded on the diagonal represent the reliability coefficient of the reporting category. Observed correlations are below the cells shaded on the diagonal; disattenuated correlations are above. The disattenuated correlations larger than 1 were truncated to 1.

Additionally, the correlation was computed among the overall scores for the three tested subjects: (1) ELA, (2) mathematics, and (3) science, as shown in Table 11.

Table 11. Correlations Across Spring 2022
English Language Arts, Mathematics, and Science Scores

Grade	*N*	ELA and Mathematics	ELA and Science	Mathematics and Science
4	9,089	0.71	0.74	0.72
8	8,645	0.70	0.71	0.70
10	2,543	0.71	0.73	0.72

5.3 CLUSTER EFFECTS

The NDSA for Science is modeled off the Rasch testlet model (Wang & Wilson, 2005). The IRT model is a high-dimensional model that incorporates a nuisance dimension for each item cluster, in addition to an overall dimension representing overall proficiency. Section 5.1, Model Description, of Volume 1, Annual Technical Report, presents a detailed description of the IRT model. The internal (latent) structure of the model is presented in Figure 7 of this volume. The psychometric approach for the assessment is innovative and quite different from the traditional approach of ignoring local dependencies. The validity evidence on the internal structure presented in this section relates to the presence of cluster effects and how substantial they are.

Simulation studies conducted by Rijmen, Jiang, & Turhan (2018) confirmed that both the item difficulty parameters and the cluster variances are recovered well for the Rasch testlet model (Wang & Wilson, 2005) under a variety of conditions. Cluster effects with a range of magnitudes were also recovered well. The results obtained by Rijmen et al. (2018) confirmed earlier findings reported in the literature under conditions that were chosen to closely resemble the assessment (e.g., Bradlow, Wainer, & Wang, 1999). For example, in one of the studies, the item location parameters and cluster variances used to simulate data were based on the results of a pilot study.

CAI examined the distribution of cluster variances obtained from the 2019 IRT calibrations for the entire Independent College and Career Readiness (ICCR) item bank. For elementary school, the estimated value of the cluster variances of all operational, scored items ranged from 0 to 5.13, with a median value of 0.38 and a mean value of 0.78. As a comparison, the estimated variance parameter of the overall dimension for North Dakota elementary school in 2021 was $\widehat{\sigma}^2_{\theta_{ND}} = 0.67$.

For middle school, the estimated value of the cluster variances of all operational, scored items ranged from 0 to 2.47, with a median value of 0.43 and a mean value of 0.57. The estimated variance parameter of the overall dimension for North Dakota middle school in 2021 was $\widehat{\sigma}^2_{\theta_{ND}} = 0.42$.

For high school, the estimated value of the cluster variances of all operational, scored items ranged from 0.07 to 2.58, with a median value of 0.43 and a mean value of 0.52. The estimated variance parameter of the overall dimension for North Dakota high school in 2021 was $\widehat{\sigma}^2_{\theta_{ND}}. = 0.48$.

Figure 2 through Figure 4 present the histograms of the cluster variances expressed as the proportion of the systematic variance due to the cluster variance for each cluster (computed as $\eta_g = \frac{\sigma_g^2}{\sigma_{\theta_{ND}}^2 + \sigma_g^2}$), where $\sigma_{\theta_{ND}}^2$ is the variance estimate of the overall proficiency of North Dakota students.

A wide range of cluster variances was observed in all three grade bands. These results indicate that, for all grades, cluster effects can be substantial and provide evidence for the appropriateness of a psychometric model that explicitly takes local dependencies among the assertions of an item cluster into account.

Figure 2. Cluster Variance Proportion for Operational Items in Elementary School

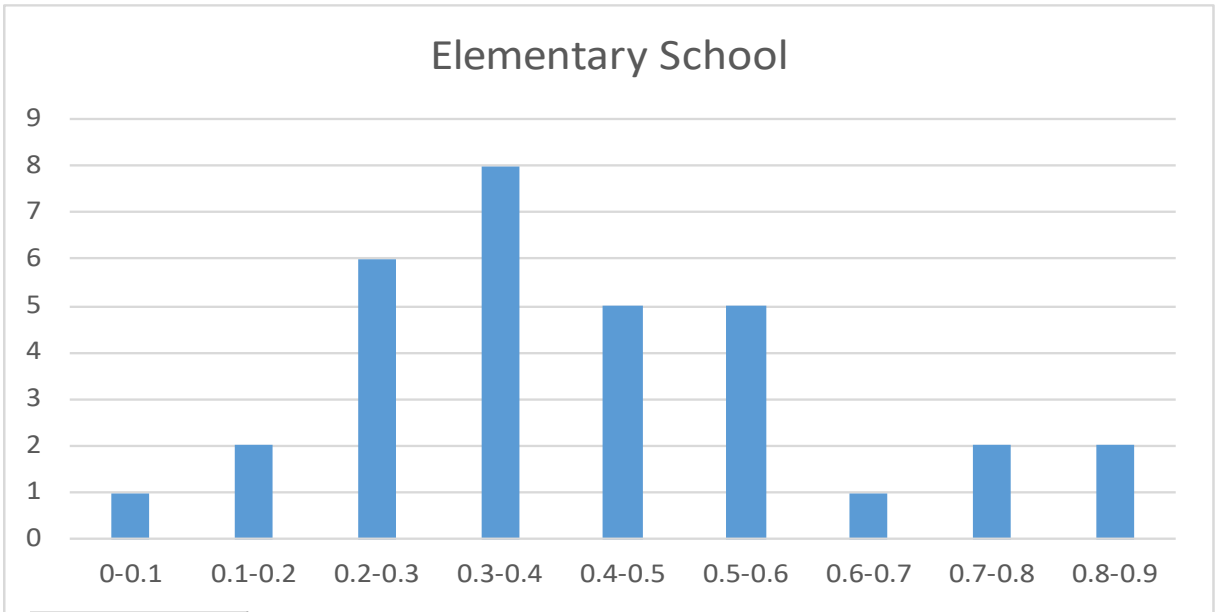


Figure 3. Cluster Variance Proportion for Operational Items in Middle School

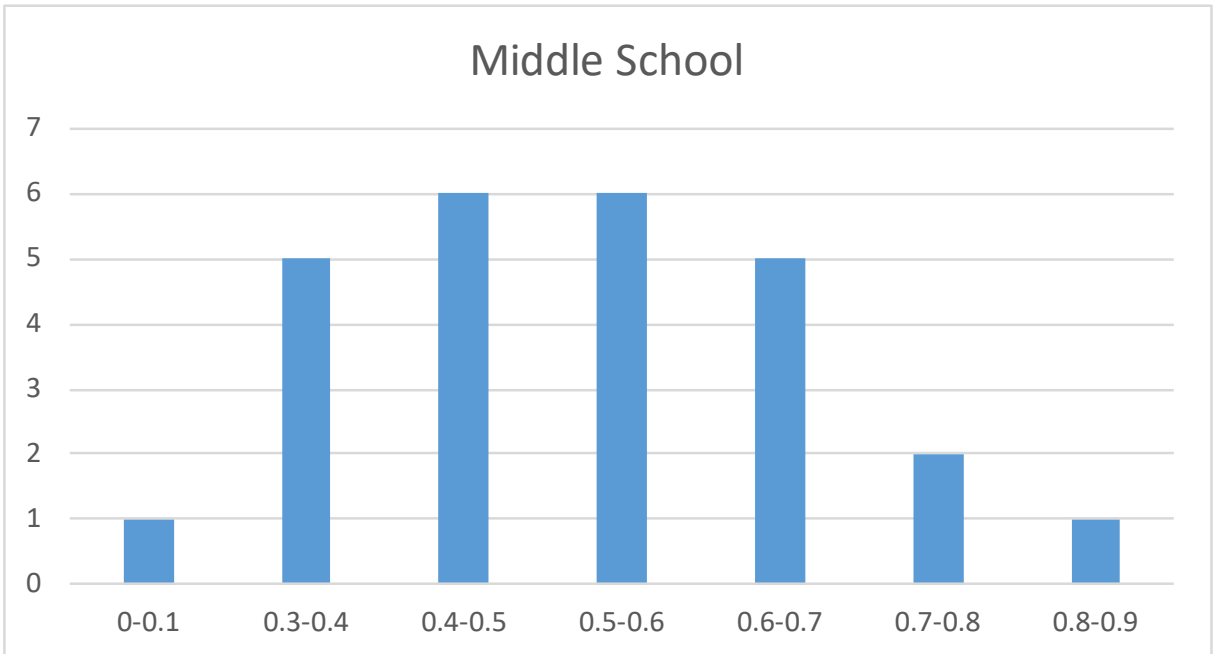
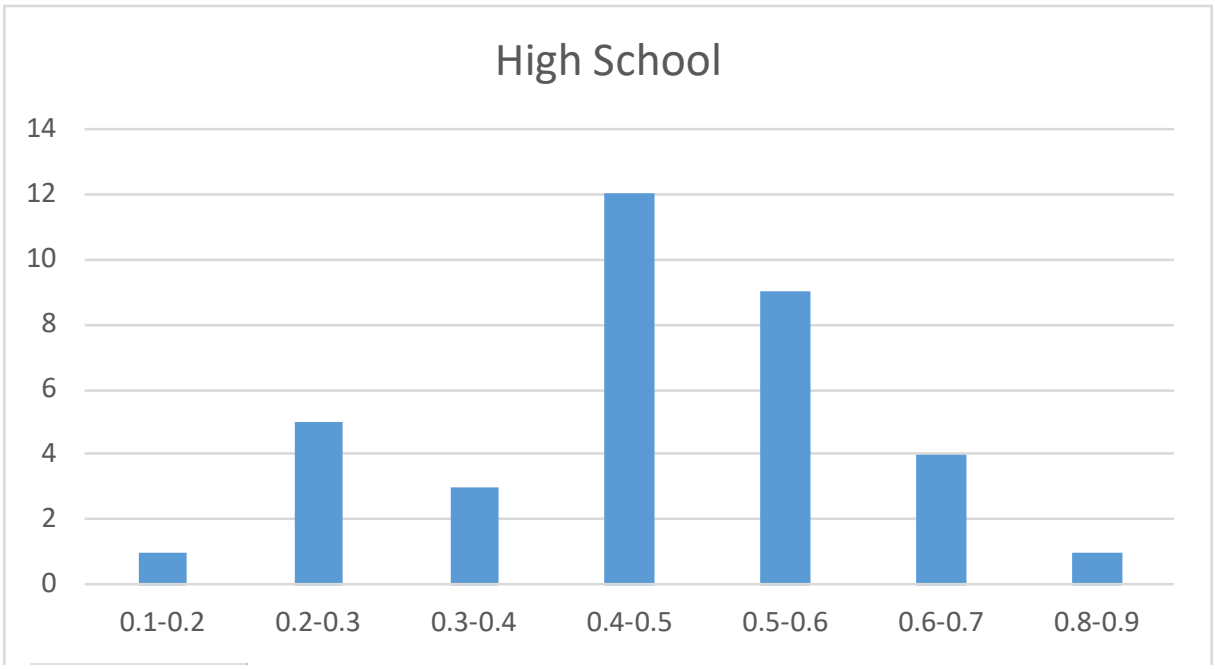


Figure 4. Cluster Variance Proportion for Operational Items in High School



5.4 CONFIRMATORY FACTOR ANALYSIS

In Section 5.3, Cluster Effects, evidence is presented for the existence of substantial cluster effects. In this section (refer to Section 5.4, Confirmatory Factor Analysis), the internal structure of the IRT model used for calibrating the item parameters is further evaluated using CFA. Alternative models are also considered, including models with a simpler internal structure (e.g., unidimensional models) and models with a more elaborate internal structure.

Estimation methods for CFA for discrete observed variables are not well suited for incomplete data collection designs where each case has data only on a subset of the set of observed variables. The linear-on-the-fly (LOFT) test design results in sparse data matrices. Students are responding to only a small number of items relative to the size of the item pool, so data are missing on most of the manifest variables for any given student. In 2018 and 2019, a LOFT test design was used for all operational science assessments inspired by a three-dimensional science framework, except for Utah. As a result, the student responses of these other states are not readily amenable for the application of CFA techniques.

The 2018 Utah science operational field test made use of a set of fixed-form tests for each grade. Therefore, the data for each fixed-form test are complete, and the fixed-form tests are amenable to CFA. The Utah science standards, though the standards are grade-specific for middle school, were developed under a framework similar to the one developed for the Next Generation Science Standards (NGSS), and a crosswalk is available between both sets of standards. Utah is part of the Memorandum of Understanding (MOU), and many of the states participating in the MOU also use the middle school items developed for and owned by Utah. Taken together, analyzing the science fixed-form tests that were administered in Utah in 2018 can provide evidence with respect to the internal structure of the NDSA for Science.

In 2018, Utah's science assessments comprised a set of fixed-form tests for each grade, and all items in these forms were item clusters. The number of fixed-form tests varied by grade, but within each grade, the total number of item clusters was the same across forms. However, some items were rejected during the rubric validation or data review and were removed from this analysis. All students with a "completed" status were included in the CFA. The percentage of students per grade that had a status other than "completed" was less than 0.85%. Table 12 summarizes the number of forms included in this analysis, the number of item clusters per discipline (range across forms), the number of assertions (range across forms), and the number of students (range across forms) for each one of the grades.

Table 12. Number of Forms, Clusters per Discipline (Range Across Forms), Number of Assertions per Form (Range Across Forms), and Number of Students per Form (Range Across Forms)

Grade	Number of Fixed Forms	Number of Item Clusters per Discipline in Each Form			Number of Assertions per Form	Number of Students per Form
		Earth and Space Sciences	*Life Sciences*	*Physical Sciences*		
6	3	2–3	2–3	2	74–83	6,804–6,881
7	6	2	5	2	83–89	3,822–3,890
8	3	2	2	6–7	93–100	5,061–5,104

The factor structure of a testlet model, which is the model used for calibration, is formally equivalent to a second-order model. Specifically, the testlet model is the model obtained after a Schmid–Leiman transformation of the second-order model (Li, Bolt, & Fu, 2006; Rijmen, 2009; Yung, Thissen, & McLeod, 1999). In the corresponding second-order model, the group of assertions related to an item cluster are indicators of the item cluster, and each item cluster is an indicator of overall achievement on the science assessment. Because assertions are not pure indicators of a specific factor, each assertion has a corresponding error component. Similarly, item clusters include an error component indicating that they are not pure indicators of the overall science achievement.

CAI used CFA to evaluate the fit of the second-order model mentioned previously to student data from spring 2018. Three additional structural models were included in the analysis, as well. In the first model, there was only one factor representing overall science achievement. All assertions were indicators of this overall proficiency factor. The first model was a testlet model in which all cluster variances were zero. In the second model, assertions were indicators of the corresponding science discipline, and each discipline was an indicator of the overall science achievement. This was a second-order model with science disciplines rather than item clusters as first-order factors. This model did not take the cluster effects into account. In the last, most general model, assertions were indicators of the corresponding item cluster, and item clusters were indicators of the corresponding science discipline, with disciplines being indicators of the overall science achievement. For the sake of simplicity, the models in the analysis are here referred to as:

- Model 1—Assertions-Overall Science (one-factor model)
- Model 2—Assertions-Disciplines-Overall Science (second-order model)
- Model 3—Assertions-Clusters-Overall Science (second-order model)
- Model 4—Assertions-Clusters-Disciplines-Overall Science (third-order model)

Figure 5 through Figure 8 illustrate these four structural models. Model 1 is nested within Models 2, 3, and 4. Also, Models 2 and 3 are nested within Model 4. The paths from the factors to the assertions represent the first-order factor loadings. Note that all four models include factor loadings for the assertions, which is different from the calibration model for which all the discrimination parameters of the assertions were set to 1.

Figure 5. One-Factor Structural Model (Assertions-Overall Science): "Model 1"

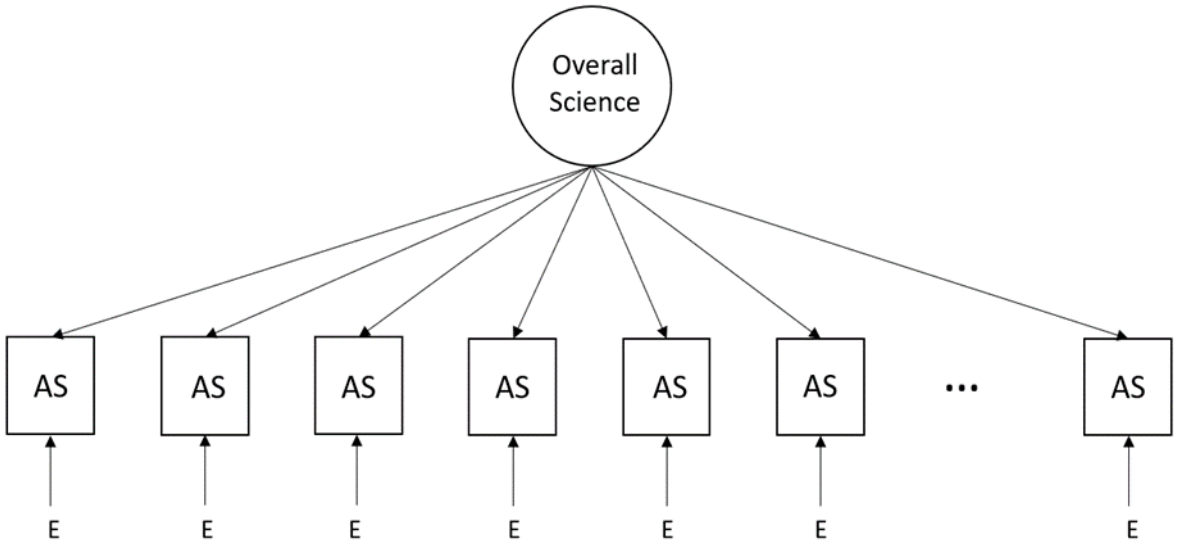


Figure 6. Second-Order Structural Model (Assertions-Disciplines-Overall Science): "Model 2"

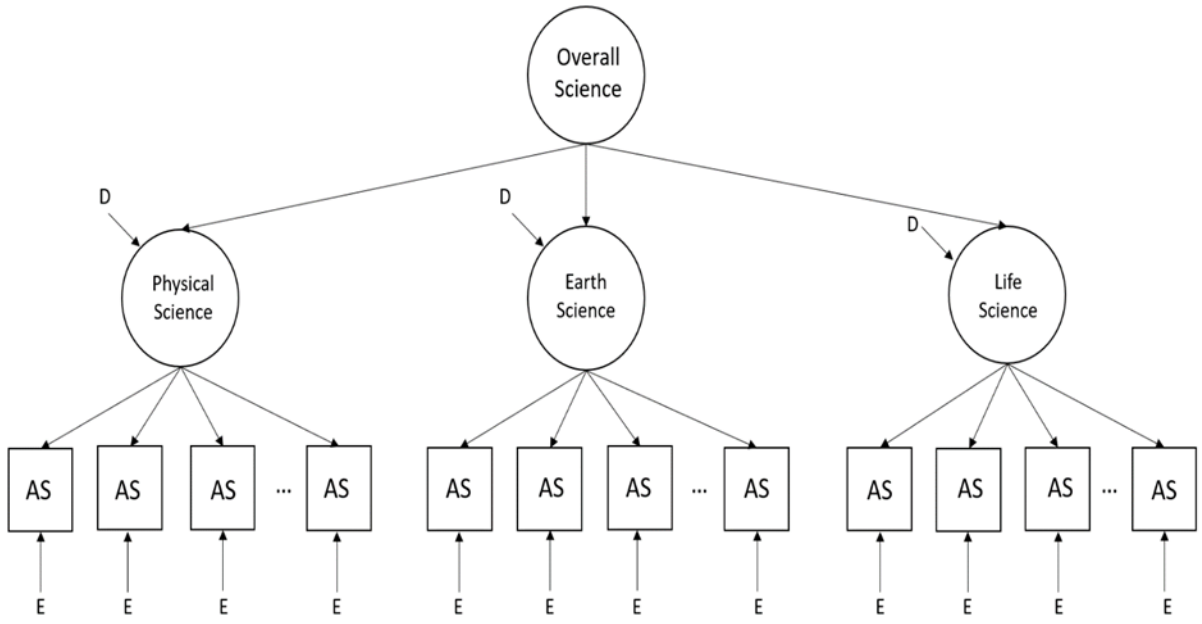


Figure 7. Second-Order Structural Model (Assertions-Clusters-Overall Science): "Model 3"

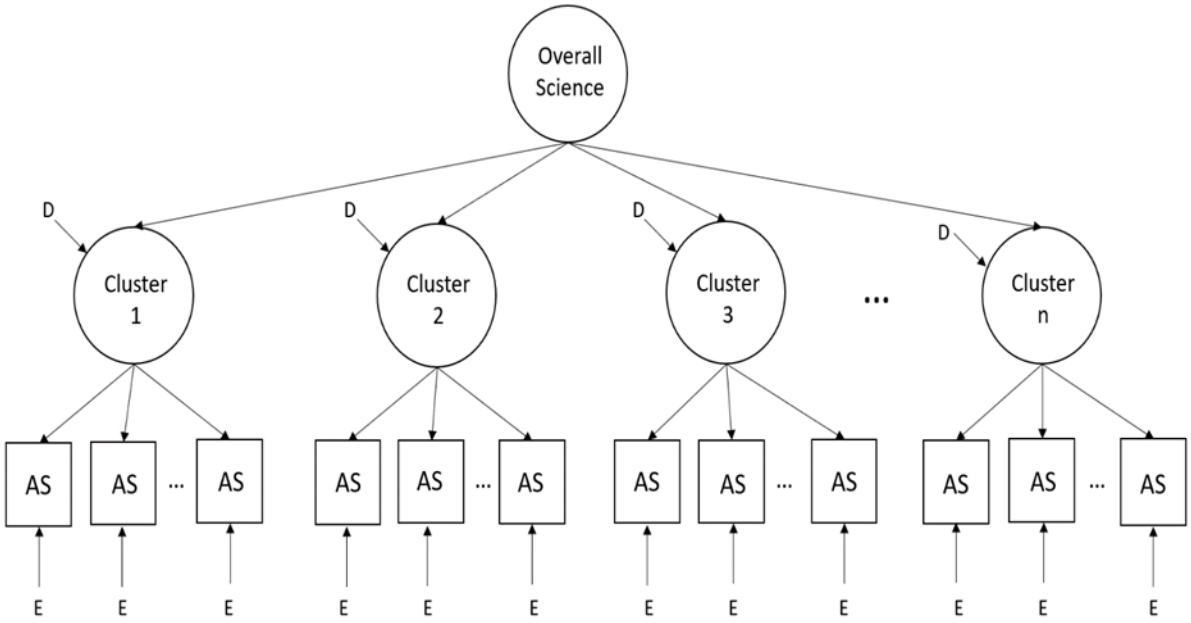
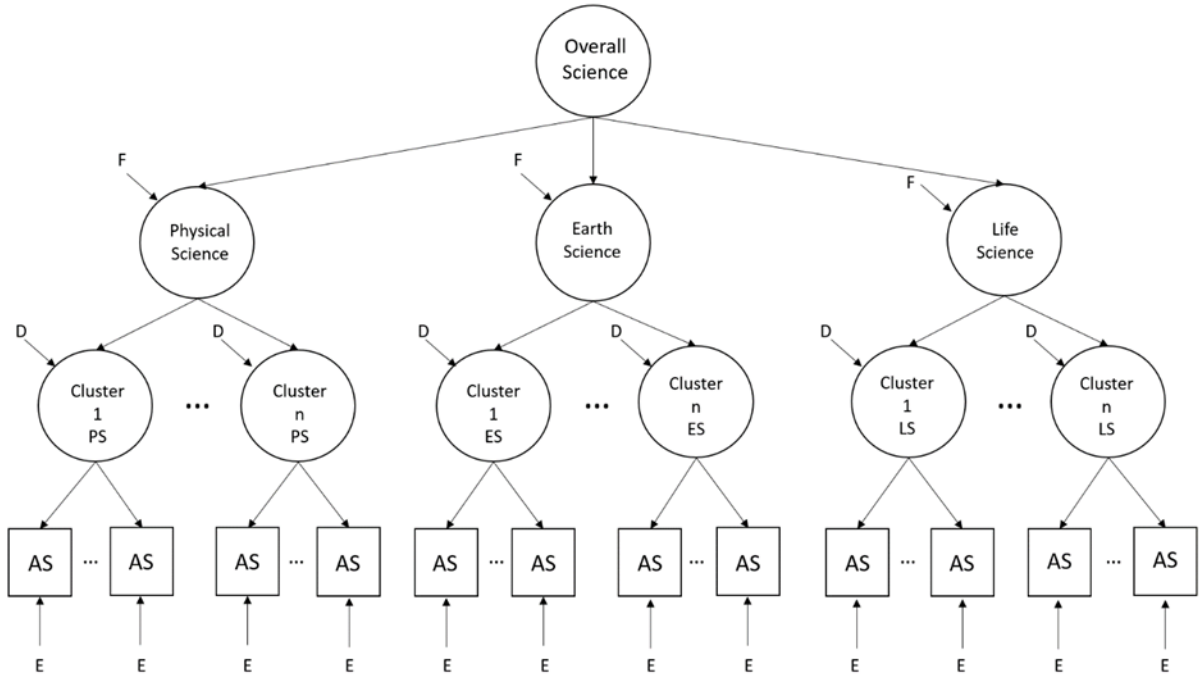


Figure 8. Third-Order Structural Model (Assertions-Clusters-Disciplines-Overall Science): "Model 4"



5.4.1 Results

For each test form, fit measures were computed for each of the four models. The fit measures used to evaluate goodness-of-fit were the comparative fit index (CFI), the Tucker-Lewis index (TLI), the root mean square error of approximation (RMSEA), and the standardized root mean residual (SRMR).

CFI and TLI are relative fit indices, meaning they evaluate model fit by comparing the model of interest to a baseline model. RMSEA and SRMR are indices of absolute fit. Table 13 provides a list of these measures along with the corresponding thresholds indicating a good fit.

*Table 13. Guidelines for Evaluating Goodness-of-Fit**

Goodness-of-Fit Measure	Indication of Good Fit
CFI	≥ 0.95
TLI	≥ 0.95
RMSEA	≤ 0.06
SRMR	≤ 0.08

*Brown, 2015; Hu & Bentler, 1999

Table 14 through Table 16 show the goodness-of-fit statistics for grades 6–8, respectively.[1] Numbers in bold indicate those indices that did not meet the criteria established in Table 13. Across all grades and models, the following conclusions were drawn:

- Model 1 shows the most misfit across grades and forms.
- Relative to Model 1, Model 3 generally showed more improvement in model fit than Model 2 showed across forms (i.e., higher values for CFI and TLI and lower values for RMSEA and SRMR). This means that accounting for the clusters resulted in greater improvement in model fit over a single-factor model than did accounting for disciplines.
- Model 4 did not show improvement in model fit over Model 3. Fit measures remained the same (or had a difference of 0.001 or smaller in very few cases) across forms for Models 3 and 4. Thus, when clusters were taken into account, incorporating disciplines into the model did not improve model fit.
- Overall model fit for Models 3 and 4 decreased with decreasing grades. For grade 8, all fit indices for Models 3 and 4 indicated good model fit for all three forms. For grade 7, all fit indices for Models 3 and 4 indicated good fit for two out of the six forms, and the degree of misfit for the other four forms was small. For grade 6, all three forms had fit indices

[1] For very few assertions per form and models, some error variances for the assertions were slightly below 0. For grade 6, 1–2 assertions per form and model had error variance below 0, with the lowest error variance being -0.027. For grade 7, Forms 1, 2, 5, and 6 each had one negative error variance for one assertion in Models 3 and 4, with the lowest error variance being -0.099. Form 4 had 1–2 assertions with negative error variance in each model, and the lowest error variance was -0.102. For grade 8, there were no assertions with negative error variances for any of the forms or models.

above the threshold values for at least one of the absolute fit indices for Models 3 and 4. The amount of misfit was small for the RMSEA but more substantial for the SRMR for two out of the three forms.

Table 14. Fit Measures per Model and Form, Grade 6

Model	Form	CFI	TLI	RMSEA	SRMR
Model 1 Assertions-Overall Science (one-factor model)	**1**	0.995	0.995	**0.106**	**0.163**
	2	0.997	0.997	**0.093**	**0.148**
	3	0.995	0.995	**0.109**	**0.161**
Model 2 Assertions-Disciplines-Overall Science (second-order model)	**1**	0.996	0.996	**0.089**	**0.144**
	2	0.998	0.998	**0.078**	**0.128**
	3	0.997	0.997	**0.087**	**0.135**
Model 3 Assertions-Clusters-Overall Science (second-order model)	**1**	0.998	0.998	**0.065**	**0.107**
	2	0.999	0.999	0.056	**0.095**
	3	0.998	0.998	**0.067**	**0.104**
Model 4 Assertions-Clusters-Disciplines-Overall Science (third-order model)	**1**	0.998	0.998	**0.065**	**0.107**
	2	0.999	0.999	0.056	**0.095**
	3	0.998	0.998	**0.067**	**0.104**

Note. Numbers in bold do not meet the criteria for goodness-of-fit.

Table 15. Fit Measures per Model and Form, Grade 7

Model	Form	CFI	TLI	RMSEA	SRMR
Model 1 Assertions-Overall Science (one-factor model)	**1**	**0.892**	**0.889**	0.060	0.074
	2	**0.938**	**0.936**	**0.083**	**0.109**
	3	**0.940**	**0.939**	0.052	0.065
	4	**0.937**	**0.936**	**0.068**	**0.114**
	5	**0.939**	**0.937**	**0.093**	**0.119**
	6	**0.898**	**0.895**	0.056	0.071
Model 2 Assertions-Disciplines-Overall Science (second-order model)	**1**	**0.908**	**0.906**	0.055	0.073
	2	0.962	0.961	**0.065**	**0.088**
	3	0.950	**0.949**	0.048	0.063
	4	0.955	0.954	0.058	**0.094**
	5	0.959	0.957	**0.077**	**0.103**
	6	**0.906**	**0.903**	0.054	0.070
Model 3 Assertions-Clusters-Overall Science (second-order model)	**1**	**0.938**	**0.937**	0.046	0.072
	2	0.974	0.973	0.054	**0.082**
	3	0.967	0.966	0.039	0.055

Model	Form	CFI	TLI	RMSEA	SRMR
	4	0.977	0.976	0.041	0.072
	5	0.975	0.974	0.060	**0.089**
	6	**0.932**	**0.930**	0.046	0.072
Model 4 Assertions-Clusters-Disciplines-Overall Science (third-order model)	**1**	**0.939**	**0.937**	0.045	0.072
	2	0.974	0.973	0.054	**0.082**
	3	0.967	0.966	0.039	0.055
	4	0.977	0.976	0.041	0.072
	5	0.975	0.974	0.060	**0.089**
	6	**0.932**	**0.930**	0.046	0.072

Note. Numbers in bold do not meet the criteria for goodness-of-fit.

Table 16. Fit Measures per Model and Form, Grade 8

Model	Form	CFI	TLI	RMSEA	SRMR
Model 1 Assertions-Overall Science (one-factor model)	**1**	**0.929**	**0.927**	0.043	0.060
	2	0.959	0.958	0.042	0.056
	3	**0.943**	**0.941**	0.052	0.074
Model 2 Assertions-Disciplines-Overall Science (second-order model)	**1**	**0.934**	**0.932**	0.041	0.060
	2	0.963	0.963	0.040	0.056
	3	0.950	**0.949**	0.049	0.072
Model 3 Assertions-Clusters-Overall Science (second-order model)	**1**	0.953	0.952	0.034	0.057
	2	0.974	0.973	0.034	0.054
	3	0.970	0.969	0.038	0.064
Model 4 Assertions-Clusters-Disciplines-Overall Science (third-order model)	**1**	0.953	0.952	0.034	0.057
	2	0.974	0.974	0.033	0.053
	3	0.970	0.969	0.038	0.064

Note. Numbers in bold do not meet the criteria for goodness-of-fit.

For Models 3 and 4, grade 6 showed some degree of misfit across all three forms according to the measures of absolute model fit, especially for the SRMR. Further examination indicated that the lack of fit could be attributed to a single item that was common to all three grade 6 forms in this factor-analysis study. After removing that item, only two forms had two or more clusters per discipline. The fit for both forms improved drastically in Models 3 and 4, with all fit measures, except the SRMR for one form, meeting the criteria for model fit. The SRMR value that exceeded the threshold value did so barely, with a value of 0.083. Table 17 shows the fit measures for grade 6 after removal of the item causing misfit. Note that, unlike Models 3 and 4, Models 1 and 2 still did not meet the criteria of model fit after removing the item.

Table 17. Fit Measures per Model and Form—Grade 6—One Cluster Removed[2]

Model	Form	CFI	TLI	RMSEA	SRMR
Model 1 Assertions-Overall Science (one-factor model)	**1**	0.977	0.976	**0.094**	**0.130**
	2	0.974	0.973	**0.082**	**0.118**
Model 2 Assertions-Disciplines-Overall Science (second-order model)	**1**	0.986	0.986	**0.072**	**0.106**
	2	0.985	0.984	**0.062**	**0.094**
Model 3 Assertions-Clusters-Overall Science (second-order model)	**1**	0.992	0.991	0.057	**0.083**
	2	0.991	0.991	0.048	0.072
Model 4 Assertions-Clusters-Disciplines-Overall Science (third-order model)	**1**	0.992	0.991	0.057	**0.083**
	2	0.991	0.991	0.048	0.072

Note. Numbers in bold do not meet the criteria for goodness-of-fit.

Table 18 shows the estimated correlations among disciplines for Model 4 (third-order model). The correlations are all very high, ranging between 0.913 and 1. The high correlations between the disciplines in Model 4 indicate that, after considering the cluster effects, the disciplines did not add much to the model. This may explain why Model 4 did not show an improvement in fit compared to Model 3. Overall, the findings support the IRT model used for calibration.

Table 18. Model Implied Correlations per Form for the Disciplines in Model 4

Grade	Form	Discipline	Earth and Space Sciences (ESS)	Life Sciences (LS)
6	**1**	Physical Sciences (PS)	0.999	0.941
		Earth and Space Sciences (ESS)	–	0.940
	2	Physical Sciences (PS)	1.000	0.964
		Earth and Space Sciences (ESS)	–	0.964
	3	Physical Sciences (PS)	0.975	0.923
		Earth and Space Sciences (ESS)	–	0.947
7	**1**	Physical Sciences (PS)	0.983	0.947
		Earth and Space Sciences (ESS)	–	0.937
	2	Physical Sciences (PS)	0.978	0.972
		Earth and Space Sciences (ESS)	–	0.951
	3	Physical Sciences (PS)	0.955	0.936
		Earth and Space Sciences (ESS)	–	0.966
	4	Physical Sciences (PS)	0.938	0.913
		Earth and Space Sciences (ESS)	–	0.973

[2] One assertion per model in form 1 and one assertion on three of the models in form 2 had error variances below 0, with the lowest error variance being −0.027.

Grade	Form	Discipline	Earth and Space Sciences (ESS)	Life Sciences (LS)
	5	Physical Sciences (PS)	0.931	0.944
		Earth and Space Sciences (ESS)	–	0.965
	6	Physical Sciences (PS)	0.941	0.928
		Earth and Space Sciences (ESS)	–	0.967
8	1	Physical Sciences (PS)	0.971	0.971
		Earth and Space Sciences (ESS)	–	0.970
	2	Physical Sciences (PS)	0.956	0.958
		Earth and Space Sciences (ESS)	–	0.935
	3	Physical Sciences (PS)	0.966	0.978
		Earth and Space Sciences (ESS)	–	0.988

5.4.2 Conclusion

The models with no cluster effects provided the highest degrees of misfit across forms and grades (Models 1 and 2), indicating that the cluster effects need to be taken into account as additional latent variables. On the other hand, once the cluster effects are accounted for, a single science dimension is sufficient (Model 3): including additional dimensions for the science disciplines (Life Science, Physical Science, Earth and Space Sciences) did not improve model fit and the correlations among those three dimensions are very high (Model 4). Model 3, with a single overall dimension for Science and additional latent variables to account for the effect of item clusters, provided the best balance between model fit and parsimony.

Overall, the findings support the use of the Rasch testlet model as the IRT calibration model and the reporting of an overall score directly computed from all the items a student took. Because there are enough items within each discipline in the test blueprint, discipline subscores can be reported at the individual level although they may not provide much unique information from the total score for most students. However, many stakeholders often desire information about student performance in addition to a single overall score. Note that it is not uncommon to provide subscores at the individual level even when the assessment is essentially unidimensional in a psychometric sense. For example, based on the dimensionality analyses for the Smarter Balanced Assessment, there is evidence suggesting “no consistent and pervasive multidimensionality was demonstrated” (Smarter Balanced Assessment Consortium, 2016, p.182) yet individual claim scores are routinely reported in addition to overall ELA and Mathematics scores.

6. FAIRNESS IN CONTENT

The principles of universal design (UD) provide standards for test designs that minimize the impact of construct-irrelevant factors when assessing student achievement. UD removes barriers to enable access for the widest possible range of students. CAI applies the following seven principles of UD during the process of test development (Thompson, Johnstone, & Thurlow, 2002):

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

Test development specialists receive extensive training in UD and apply its principles to the development of all test materials. In the review process, North Dakota educators and stakeholders verify adherence to the principles of UD.

6.1 COGNITIVE LABORATORY STUDIES

In 2017, when the development of item clusters began for the states participating in the Memorandum of Understanding (MOU), cognitive lab studies were carried out to evaluate and refine the process of developing item clusters aligned to three-dimensional science standards. Results of these studies confirmed the feasibility of the approach. Item clusters were completed within 12 minutes on average, and students reported familiarity with the format conventions and online tools used in the item clusters. The students appeared to easily navigate the item clusters' interactive features and response formats. In general, students who received credit on a given item displayed a reasoning process that aligned with the skills that the item was intended to measure.

A second set of cognitive lab studies was carried out in 2018 and 2019 to determine whether students using braille could understand the task demands of selected accommodated three-dimensional science-aligned item clusters and navigate the interactive features of these item clusters in a manner that allowed them to fully display their knowledge and skills relative to the constructs of interest. In general, both the students who relied entirely on braille or Job Access With Speech (JAWS) and those who had some vision and were able to read the screen with magnification were able to find the information they needed to respond to the questions, navigate the various response formats, and finish within a reasonable amount of time. The item clusters were clearly different from (and more complex than) other tests with which the students were familiar; however, the study recommended that students should be given adequate time to practice with at least one sample item cluster before taking the summative test. The study's findings also proposed tool-specific recommendations for accessibility for visually impaired students. The

reports for both cognitive laboratory studies are presented in Appendix D, Science Clusters Cognitive Lab Report, and Appendix E, Braille Cognitive Lab Report.

6.2 STATISTICAL FAIRNESS IN ITEM STATISTICS

Differential item functioning (DIF) analysis was conducted with other states that field-tested the items for the initial item bank. A thorough content review was performed in those states. The procedures surrounding this review of items for bias is further described in Section 4.4 of Volume 1, Annual Technical Report, along with the DIF analysis process for the NDSA for Science.

7. SUMMARY

This report is intended to provide a collection of reliability and validity evidence that supports 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 sub-group levels, showing that the reliability of all tests is in line with acceptable industry standards.
- ***Content validity.*** Evidence of content validity is provided to support the assertion that content coverage on each test is consistent with the test specifications of the blueprint across testing modes.
- ***Internal structural validity.*** Evidence of internal structural validity is provided to support the selection of a measurement model, the tenability of model assumptions, and the reporting of an overall score and subscores at the reporting-category level.
- ***Relationship of test scores to external variables.*** Evidence of convergent and discriminant validity is provided to support the relationship between the test and other measures that are intended to assess similar and different constructs.
- ***Test fairness.*** Items are developed according to the principles of universal design, which removes barriers in order to enable access for the widest possible range of students. Evidence of test fairness is provided using DIF analysis in tandem with content reviews by specialists.

8. REFERENCES

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

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*, 153–168.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). New York: The Guilford Press.

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

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

Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement*, *30*, 3–21.

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

National Center for Education Statistics. (2010). *Statistical methods for protecting personally identifiable information in aggregate reporting* (Statewide Longitudinal Data System Technical Brief, Brief 3). Retrieved from https://nces.ed.gov/pubs2011/2011603.pdf

Rijmen, F. (2009). *Three multidimensional models for testlet-based tests: Formal relations and an empirical comparison.* Educational Testing Service (ETS) Research Rep. No. RR–09–37, Princeton, NJ: ETS.

Rijmen, F., Jiang, T., & Turhan, A. (2018, April). *An item response theory model for new science assessments.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

Smarter Balanced Assessment Consortium. (2016). *2013-2014 Technical Report*. Retrieved from https://portal.smarterbalanced.org/library/en/2013-14-technical-report.pdf.

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). Minneapolis: University of Minnesota, National Center on Educational Outcomes. Retrieved from https://nceo.umn.edu/docs/onlinepubs/synth44.pdf.

Wang, W. C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement*, *29*(2), 126–149.

Yung, Y. F., Thissen, D., & McLeod, L. D. (1999). On the relationship between the higher-order factor model and the hierarchical factor model. *Psychometrika*, *64*, 113–128.