# Rhode Island and Vermont Multi-State Science Assessment

## 2021–2022

# Volume 4: Evidence of Reliability and Validity


# TABLE OF CONTENTS

# LIST OF TABLES

## LIST OF FIGURES



## LIST OF APPENDICES

# 1. INTRODUCTION AND OVERVIEW OF RELIABILITY AND VALIDITY EVIDENCE

The Rhode Island Department of Education (RIDE) works in partnership with the Vermont Agency of Education (VT AOE) to develop their science assessment, named the Multi-State Science Assessment (MSSA). The first operational year for MSSA was 2018–2019.

Unless explicitly stated otherwise, the term *Multi-State Science Assessment (MSSA)* will be used throughout this volume of the technical report to refer to both the Rhode Island Next Generation Science Assessment (RI NGSA) and the Vermont Science Assessment (VTSA), which together comprise the MSSA.

The MSSA is administered online to students in grades 5, 8, and 11 using a linear-on-the-fly test (LOFT) design. Accommodated versions are available for each grade, including braille and large print Data Entry Interface (DEI) forms. Spanish language versions of the tests are also available. Table 1 shows the complete list of tests for the operational test administration in spring 2022.

*Table 1. Spring 2022 Assessment Modes*

| Language/Format | Assessment Mode | Grade |
|---|---|---|
| English/LOFT | Online | 5, 8, and 11 |
| Spanish/LOFT | Online | 5, 8, and 11 |
| English/DEI-fixed | Paper | 5, 8, and 11 |
| English/braille-fixed | Online and Paper | 5, 8, and 11 |

Given the intended uses of these tests, both reliability and validity evidence are necessary to support appropriate inferences of student academic achievement from the MSSA scores. The analyses to support reliability and validity evidence reported in this volume were conducted based on test results for students whose scores were reported, including those students who took the online English language version and the accommodated versions of the MSSA.

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

1. ***Reliability.*** The reliability estimates are presented by grade and demographic subgroup. This section also includes the conditional standard error of measurement (CSEM) and 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 Next Generation Science Standards (NGSS) with a sufficient number of items targeting each area of the test blueprint.

3. ***Internal Structure Validity.*** Evidence is provided 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. As explained in detail in Volume 1,

Annual Technical Report, the IRT model is a multidimensional model with an overall dimension representing proficiency in science and nuisance dimensions that consider within-item local dependencies among scoring assertions. In this volume, evidence is provided with respect to the presence of item cluster effects. Additionally, confirmatory factor analysis (CFA) was used to evaluate the fit of the IRT model and to compare it to alternative models, including models with a simpler internal structure (e.g., unidimensional models) and models with more elaborate internal structures.

4. ***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.

5. ***Test Fairness.*** Fairness is an explicit concern during item development. Items are developed following the principles of universal design. Universal design removes barriers to provide access for the widest range of students possible. Specialists use differential item functioning (DIF) analysis in tandem with content reviews by specialists.

## 1.1 RELIABILITY

*Reliability* refers to consistency in test scores. Reliability can be defined as the degree to which an individual's deviation score remains 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, they should receive consistent results. The reliability coefficient refers to the ratio of true score variance to observed score variance:

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2},$$

Another way to view reliability is to consider its relationship with the standard errors 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:

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

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 $\sigma_X$ 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, $\sigma_X$, 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 the SEM in the CTT is assumed to be a homoscedastic error, irrespective of the standard deviation of a group.

In contrast, the SEMs in the 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 score at various score points. Refer to Section **Error! Reference source not found.**, **Error! Reference source not found.**, for the derivation of heterogeneous measurement errors in IRT, and how these errors are aggregated over the score distribution to obtain a single, marginal, IRT-based reliability coefficient.

## 1.2 VALIDITY

The term *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 a need for evidence and theory that support the 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 validity evidence is the relationship between the test content and the intended test construct (refer to Section **Error! Reference source not found.**, 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 Volume 2, Test Development, for details on the item development process and Section **Error! Reference source not found.**, Independent Alignment Study, for the results of an independent alignment study).

Technology-enhanced items should be examined to ensure that no construct-irrelevant variance is introduced. If any aspect of the technology impedes or creates an advantage for a student in their responses to items, this could affect item responses and inferences regarding that student's 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 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. Possible analyses to examine internal structure are dimensionality assessment, goodness-of-model-fit to data, and reliability analysis (refer to Section **Error! Reference source not found.**, **Error! Reference source not found.**, and Section **Error! Reference source not found.**, **Error! Reference source not found.**, for details). In addition, it is important to assess the degree to which 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 Volume 1, Annual Technical Report).

The fourth source of validity evidence 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 designed 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 mainly depends on 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 restriction may need to be considered in order to determine whether the conclusions of a test can be assumed for the larger population. Convergent and discriminant validity evidence are discussed in Section **Error! Reference source not found.**, **Error! Reference source not found.**.

The fifth source of validity evidence is the suggestion that the intended and unintended consequences of test use should be included in the test-validation process. Determining test validity 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 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 would interfere with the test's validity. As described in Volume 1, Annual Technical Report, and throughout this volume, 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.

## 2. Purpose of the Multi-State Science Assessment

The primary purpose of Rhode Island and Vermont's MSSA is to yield accurate information on students' achievement of Rhode Island's and Vermont's education standards. The MSSA measures the science knowledge and skills of Rhode Island and Vermont students in grades 5, 8, and 11.

The Rhode Island Department of Education (RIDE) and the Vermont Agency of Education (VT AOE) provide an overview of their science assessments at https://www.ride.ri.gov/InstructionAssessment/Assessment/NGSAAssessment.aspx and https://vt.portal.cambiumast.com/-/media/project/client-portals/vermont/pdf/2018/vtsa-tam-2020-2021_final.pdf. Information about the Next Generation Science Standards (NGSS) is available at: www.nextgenscience.org.

The MSSA supports instruction and student learning by measuring growth in student achievement. Assessments can be used as indicators to determine whether students in Rhode Island and Vermont possess the knowledge and skills that are essential for college education and career readiness.

The MSSA also provides evidence for the requirements of state and federal accountability systems. Test scores can be used 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 tests are constructed to measure student proficiency as described in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014). The MSSA was developed in compliance with the principles of universal design to ensure that all students have access to the test content. Volume 2, Test Development, describes the MSSA standards and test blueprints in more detail. Additional evidence of content validity can also be found in Section 0, Evidence of Content Validity. The MSSA test scores are useful indicators for understanding individual students' academic achievement of the MSSA content standards and evaluating whether students' performances are progressing over time. Additionally, both individual and aggregated scores can be used to measure test reliability. The reliability of the test scores can be found in Section 3, Reliability.

The MSSA is a standard-referenced test designed to measure students' performance on the NGSS in Rhode Island and Vermont schools. As a comparison, norm-referenced tests are designed to

rank or compare all students with one another. The Rhode Island and MSSA content standards and test blueprints are discussed in Volume 2, Test Development.

The scale score and relative strengths and weaknesses at the discipline level are 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 which teachers can use to tailor their instruction. To support their practical use across the state, we must 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 the science assessments for two reasons. First, in spring 2022, the science test was administered under a linear-on-the-fly test (LOFT) design. Potentially, each student received 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, because 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 $\sigma^2$ 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 for both Rhode Island and Vermont, in Table 3 for Rhode Island, and in Table 4 for Vermont. The overall reliability ranged from 0.86 to 0.89, 0.85 to 0.90, and 0.86 to 0.89 for the combined states, Rhode Island, and Vermont, respectively. Due to the new structure of the test, the Cambium Assessment, Inc. (CAI) also 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). CAI staff found that if the local dependencies among assertions pertaining to the same item are ignored, the marginal reliability would be inflated. Local dependencies can be ignored 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 6.1 of Volume 1, Annual Technical Report). Therefore, by ignoring the local dependencies, which are substantial for many item clusters, the reliability coefficient overestimates the true reliability of the test. 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 reported reliability coefficients may therefore overestimate the true reliability to some degree for these tests. The reliability coefficients are also reported for demographics subgroups in Appendix A, Student Demographics and Reliability Coefficients.

*Table 2. Combined Marginal Reliability Coefficients*

| Grade | Sample Size | Reliability |
|---|---|---|
| **5** | 15,520 | 0.89 |
| **8** | 16,046 | 0.89 |
| **11** | 14,268 | 0.86 |

*Table 3. Marginal Reliability Coefficients, Rhode Island*

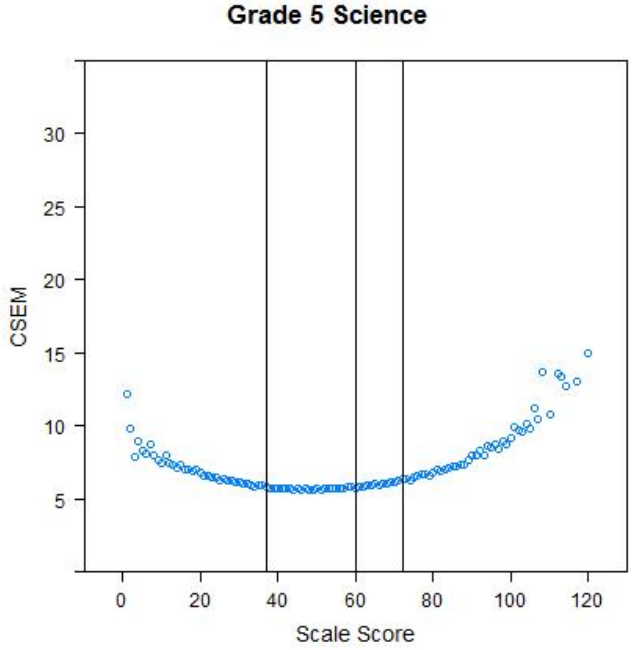
| Grade | Sample Size | Reliability |
|---|---|---|
| **5** | 9,957 | 0.89 |
| **8** | 10,300 | 0.90 |
| **11** | 9,096 | 0.85 |

*Table 4. Marginal Reliability Coefficients, Vermont*

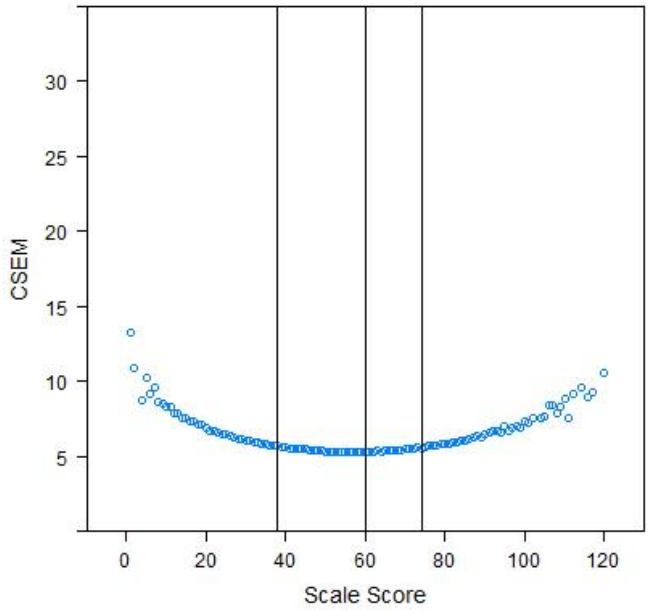
| Grade | Sample Size | Reliability |
|---|---|---|
| **5** | 5,563 | 0.89 |
| **8** | 5,746 | 0.89 |
| **11** | 5,172 | 0.86 |

## 3.1 STANDARD ERROR OF MEASUREMENT

The computation method for CSEMs has been described in Section 6.4 of Volume 1, Annual Technical Report. Figure 1 through



**Grade 8 Science**




**Grade 11 Science**

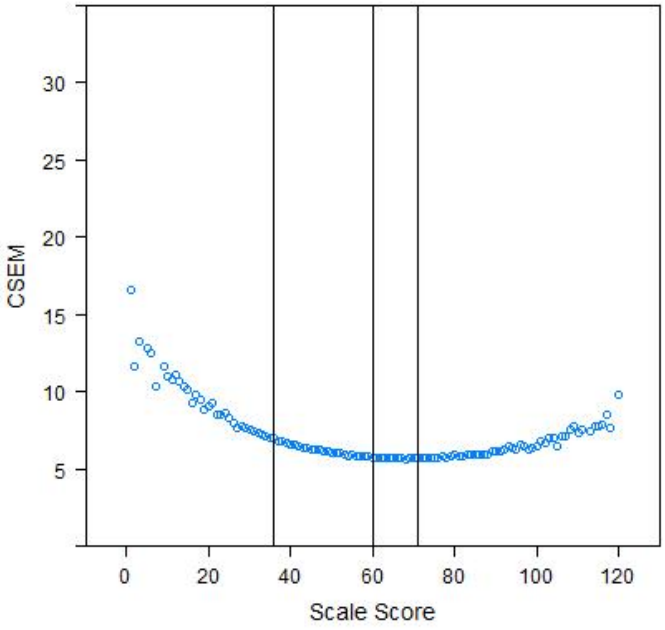
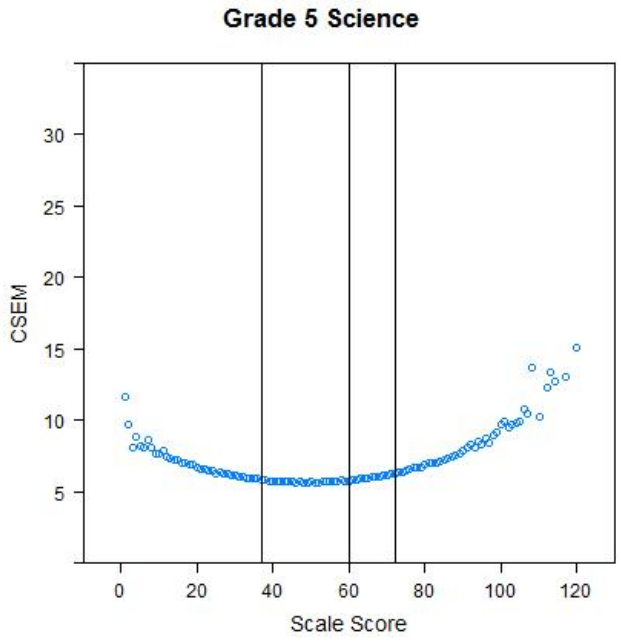
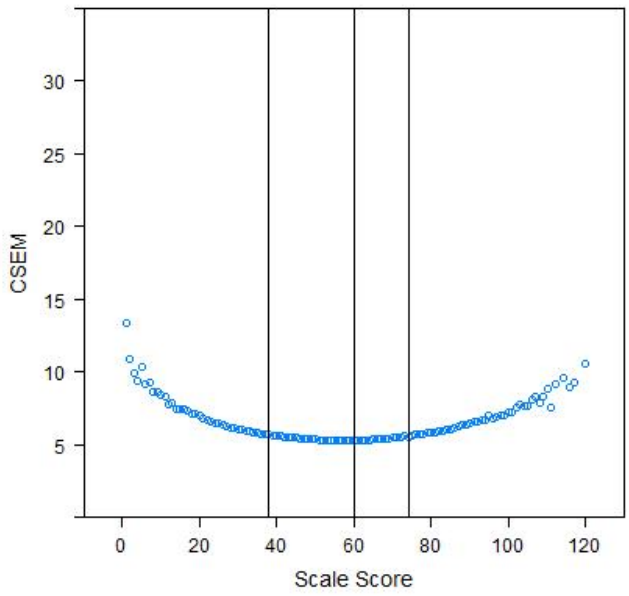


Figure 3 present the average CSEM for each scale score. The standard errors near the proficiency cut score (the middle vertical line) were low for all grades, which is a desirable test property. The CSEM at each scale score is reported in Appendix B, Conditional Standard Error of Measurement.

*Figure 1. Conditional Standard Errors of Measurement, Combined*



**Grade 8 Science**




**Grade 11 Science**

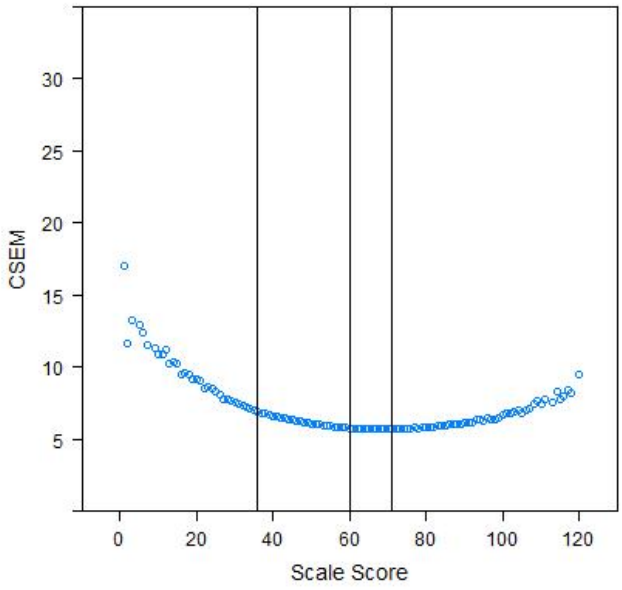
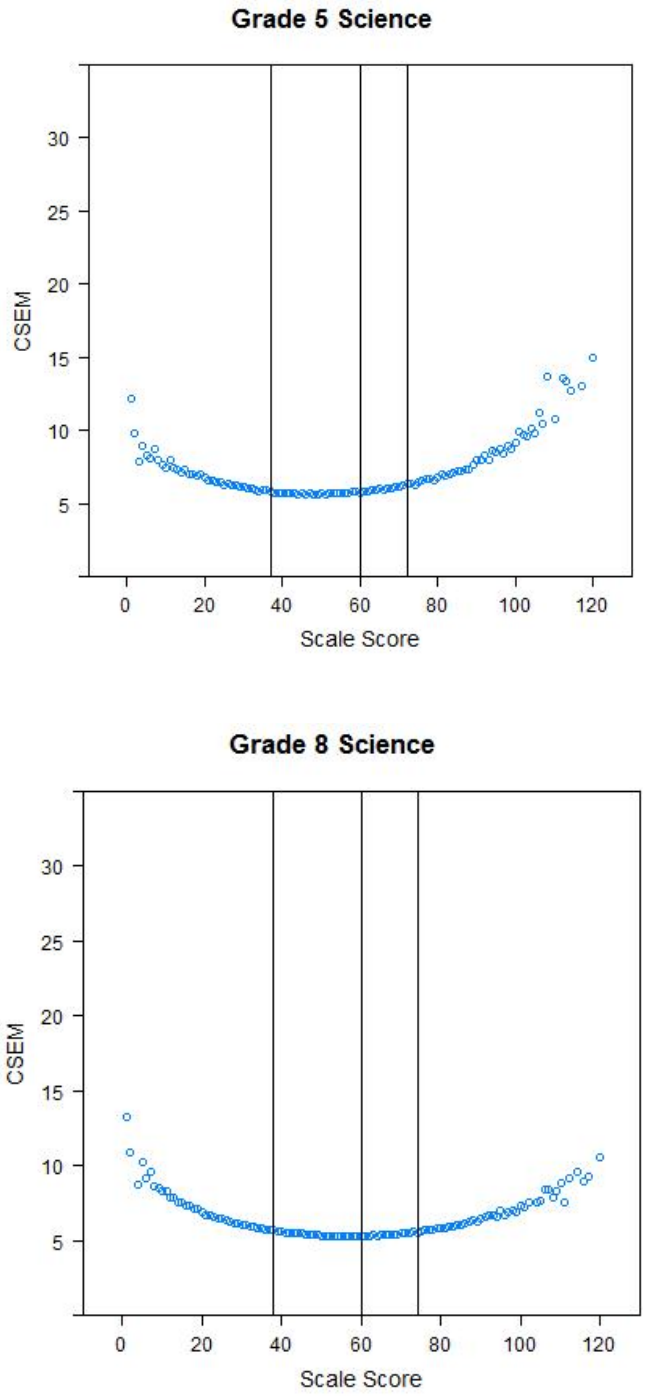


*Figure 2. Conditional Standard Errors of Measurement, Rhode Island*



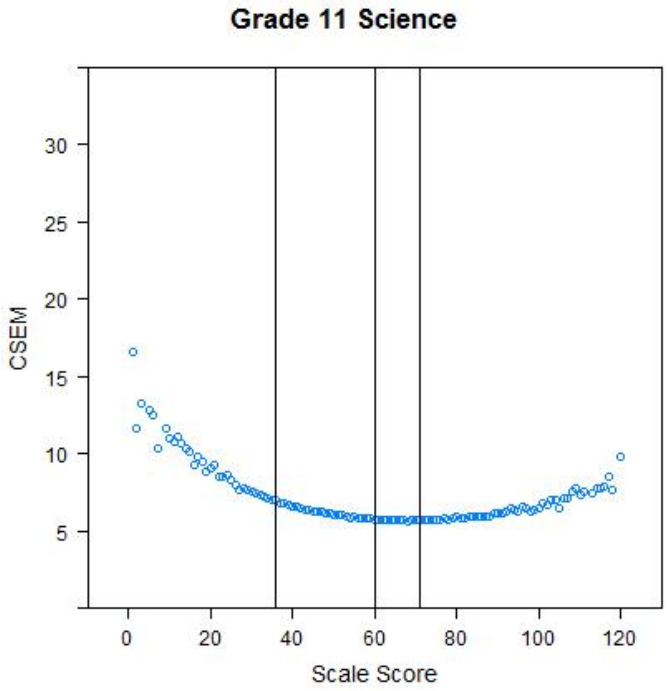
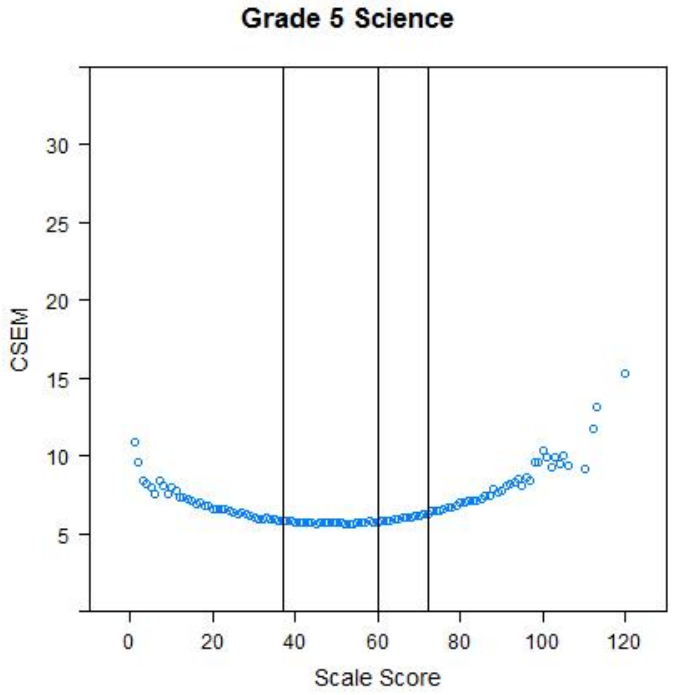
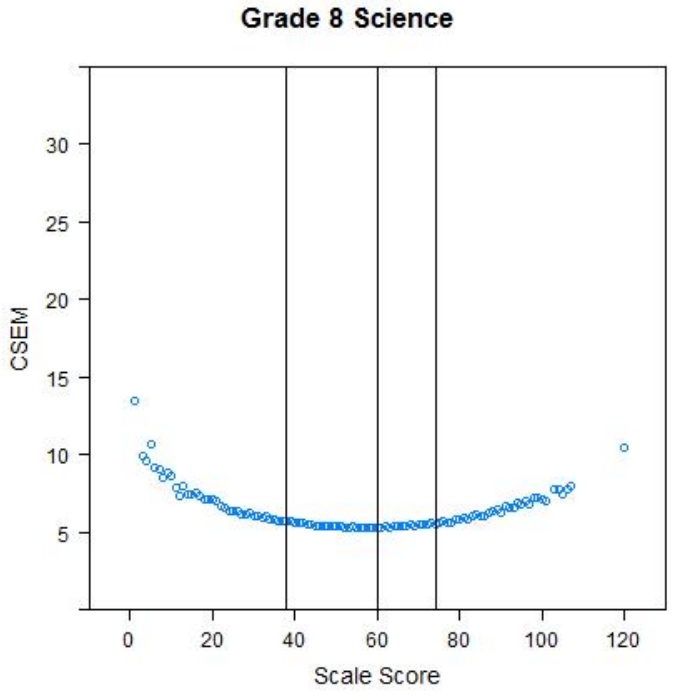
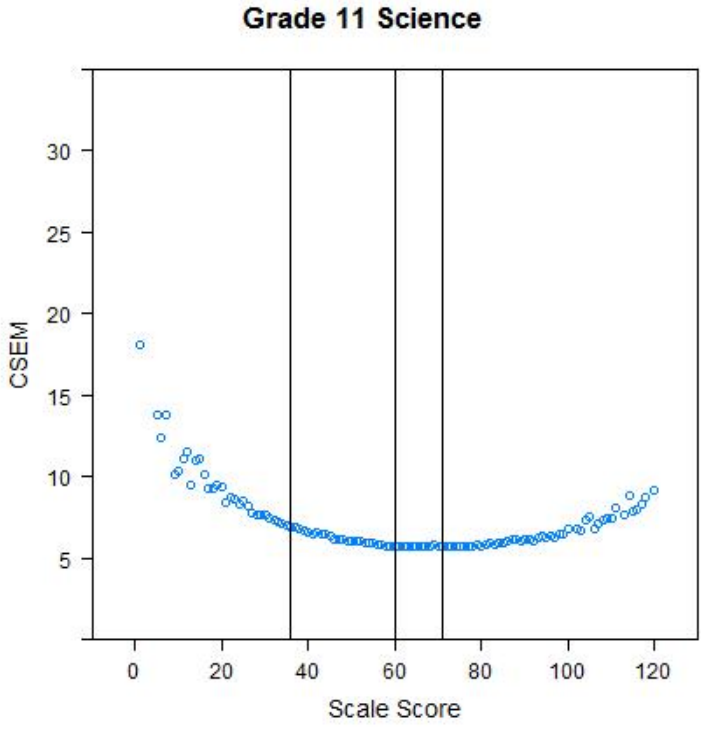

Grade 11 Science
CSEM
30
25
20
15
10
5
0
20
40
60
80
100
120
Scale Score

*Figure 3. Conditional Standard Errors of Measurement, Vermont*




Grade 11 Science
CSEM
30
25
20
15
10
5
0
20
40
60
80
100
120
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 performance 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 form taken and the classifications that would be made based on the students' true scores if, hypothetically, 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, the true ability is unknown, and students are not administered an alternate, equivalent form. Therefore, CA and CC are estimated based on 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 an 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 $\theta_j$ 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 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 to score 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 $j$th student's observed achievement level. The CA 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 accuracy for the $p$th cut score (CAC) 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 as seen below:

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

Table 5 and Table 6 provide the CA for the individual cut scores. In Rhode Island, the overall CA of the test ranged from 76.24% to 79.47%. In Vermont, the overall classification accuracy of the test ranged from 76.19% to 78.97%. The individual cut score accuracy rates were high across all grades and states, with the minimum value being 90.41% for grade 11 in Cut Score 2 for Rhode Island. This denotes that more than 90% of the time, CAI can accurately differentiate students between adjacent achievement levels in the spring 2022 MSSA. The CA for demographic subgroups is presented in Appendix C, Classification Accuracy and Consistency Indices by Subgroups.

*Table 5. Classification Accuracy Index, Rhode Island*

| Grade | Overall Accuracy (%) | Cut Accuracy (%) | | |
|---|---|---|---|---|
| | | *Level 2 Cut* | *Level 3 Cut* | *Level 4 Cut* |
| **5** | 77.97 | 92.10 | 91.63 | 94.07 |
| **8** | 79.47 | 90.80 | 92.72 | 95.93 |
| **11** | 76.24 | 90.41 | 91.11 | 94.52 |

*Table 6. Classification Accuracy Index, Vermont*

| Grade | Overall Accuracy (%) | Cut Accuracy (%) | | |
|---|---|---|---|---|
| | | *Level 2 Cut* | *Level 3 Cut* | *Level 4 Cut* |
| **5** | 77.45 | 92.39 | 91.22 | 93.66 |
| **8** | 78.97 | 91.60 | 91.80 | 95.54 |
| **11** | 76.19 | 92.33 | 90.53 | 93.10 |

### 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 test one, and $p_{jk}$ is the probability of the true score at achievement level $k$ in test two for the $j$th student. The classification consistency index for the cut scores (CCC) and overall CC were estimated in a way similar to CAC and CA.

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

and

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

Table 7 and

Table 8 provide the CC for the cuts. In Rhode Island, the overall CC of the test ranged from 67.64% to 71.52%. In Vermont, the overall CC of the test ranged from 67.63% to 70.76%. The individual cut score consistency rates were high across all grades and states with the minimum value being 86.65% for grade 11 in Cut Score 3 for Vermont. In all achievement levels, CA was slightly higher than CC. CC rates can be lower than CA; 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. The CC for demographic subgroups is presented in Appendix C, Classification Accuracy and Consistency Indices by Subgroups.

*Table 7. Classification Consistency Index, Rhode Island*

| Grade | Overall Consistency (%) | Cut Consistency (%) | | |
|---|---|---|---|---|
| | | *Level 2 Cut* | *Level 3 Cut* | *Level 4 Cut* |
| **5** | 69.95 | 88.95 | 88.30 | 91.71 |
| **8** | 71.52 | 87.15 | 89.81 | 94.26 |
| **11** | 67.64 | 86.82 | 87.47 | 92.27 |

*Table 8. Classification Consistency Index, Vermont*

| Grade | Overall Consistency (%) | Cut Consistency (%) | | |
|---|---|---|---|---|
| | | *Level 2 Cut* | *Level 3 Cut* | *Level 4 Cut* |
| **5** | 69.15 | 89.37 | 87.63 | 91.08 |

| Grade | Overall Consistency (%) | Cut Consistency (%) | | |
|---|---|---|---|---|
| | | *Level 2 Cut* | *Level 3 Cut* | *Level 4 Cut* |
| **8** | 70.76 | 88.27 | 88.45 | 93.70 |
| **11** | 67.63 | 89.34 | 86.65 | 90.37 |

## 3.3 PRECISION AT CUT SCORES

Table 9 through Table 11 present the mean CSEM at each achievement level by grade. The table also includes achievement level cut scores and associated CSEM. The CSEM at each scale score is reported in Appendix B, Conditional Standard Error of Measurement.

*Table 9. Achievement Levels and Associated Conditional Standard Error of Measurement, Combined*

| Grade | Achievement Level | Mean CSEM | Cut Score (Scale Score) | CSEM at Cut Score |
|---|---|---|---|---|
| **5** | 1 | 6.39 | - | - |
| | 2 | 5.74 | 37 | 5.84 |
| | 3 | 6.00 | 60 | 5.79 |
| | 4 | 7.04 | 72 | 6.34 |
| **8** | 1 | 6.30 | - | - |
| | 2 | 5.45 | 38 | 5.72 |
| | 3 | 5.42 | 60 | 5.37 |
| | 4 | 6.06 | 74 | 5.58 |
| **11** | 1 | 7.86 | - | - |
| | 2 | 6.28 | 36 | 6.98 |
| | 3 | 5.76 | 60 | 5.80 |
| | 4 | 5.96 | 71 | 5.72 |

*Table 10. Achievement Levels and Associated Conditional Standard Error of Measurement, Rhode Island*

| Grade | Achievement Level | Mean CSEM | Cut Score (Scale Score) | CSEM at Cut Score |
|---|---|---|---|---|
| **5** | 1 | 6.40 | - | - |
| | 2 | 5.74 | 37 | 5.85 |
| | 3 | 6.01 | 60 | 5.80 |

| Grade | Achievement Level | Mean CSEM | Cut Score (Scale Score) | CSEM at Cut Score |
|---|---|---|---|---|
| | 4 | 7.03 | 72 | 6.37 |
| **8** | 1 | 6.30 | - | - |
| | 2 | 5.45 | 38 | 5.70 |
| | 3 | 5.42 | 60 | 5.37 |
| | 4 | 6.08 | 74 | 5.58 |
| **11** | 1 | 7.87 | - | - |
| | 2 | 6.29 | 36 | 7.01 |
| | 3 | 5.75 | 60 | 5.79 |
| | 4 | 5.95 | 71 | 5.71 |

*Table 11. Achievement Levels and Associated Conditional Standard Error of Measurement, Vermont*

| Grade | Achievement Level | Mean CSEM | Cut Score (Scale Score) | CSEM at Cut Score |
|---|---|---|---|---|
| **5** | 1 | 6.37 | - | - |
| | 2 | 5.75 | 37 | 5.83 |
| | 3 | 6.00 | 60 | 5.76 |
| | 4 | 7.06 | 72 | 6.29 |
| **8** | 1 | 6.29 | - | - |
| | 2 | 5.46 | 38 | 5.74 |
| | 3 | 5.43 | 60 | 5.36 |
| | 4 | 6.01 | 74 | 5.58 |
| **11** | 1 | 7.83 | - | - |
| | 2 | 6.27 | 36 | 6.92 |
| | 3 | 5.77 | 60 | 5.80 |
| | 4 | 5.97 | 71 | 5.73 |

# 4. EVIDENCE OF CONTENT VALIDITY

This section demonstrates how the knowledge and skills assessed by the Multi-State Science Assessment (MSSA) are representative of the content standards of the larger knowledge domain. This section also describes the content standards for the MSSA and discusses the test development process and the mapping of MSSA 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 MSSA was aligned to the Next Generation Science Standards (NGSS), adopted by the Rhode Island Department of Education (RIDE) and the Vermont Agency of Education (VT AOE) in 2013. The NGSS are available for review at the following URLs: https://www.ride.ri.gov/instructionassessment/science.aspx#44942047-next-generation-science-standards for Rhode Island and https://education.vermont.gov/student-learning/content-areas/science for Vermont. Blueprints were developed to ensure that the test and items were aligned to the prioritized standards they were intended to measure. A complete description of the blueprint and test development process can be found in Volume 2, Test Development.

Table 12 presents the disciplines by grade, and the number of operational items administered measuring each discipline.

*Table 12. Number of Items for Each Discipline*

| Grade | Reporting Category | Item Clusters | Stand-Alone Items |
|---|---|---|---|
| 5 | Earth and Space Sciences | 20 | 30 |
| | Life Sciences | 32 | 32 |
| | Physical Sciences | 28 | 45 |
| 8 | Earth and Space Sciences | 22 | 28 |
| | Life Sciences | 16 | 38 |
| | Physical Sciences | 24 | 37 |
| 11 | Earth and Space Sciences | 17 | 25 |
| | Life Sciences | 33 | 33 |
| | Physical Sciences | 23 | 29 |

## 4.2 INDEPENDENT ALIGNMENT STUDY

While it is critically important to develop and strictly enforce an item development process that works to ensure alignment of test items to content standards, it is also important to independently verify the alignment of test items to content standards. The WebbAlign team of the not-for-profit Wisconsin Center for Education Products and Services (WCEPS) conducted an alignment study

in July 2019. The study was comprised of two components. The first component addressed the alignment of the Memorandum of Understanding (MOU) item bank, shared by all states that were part of the MOU. In the second component, an alignment was investigated for each state participating in the study, in the context of their state-specific blueprint and item bank, which is a particular state-vetted subset of items from the shared MOU item bank (refer to Volume 2, Test Development). The executive summary of the study is presented at Appendix F, Alignment Study Executive Summary.

# 5. EVIDENCE OF INTERNAL-EXTERNAL STRUCTURE

This section explores the internal structure of the assessment is explored using the scores provided at the discipline level. The relationship between discipline scores is just one indicator of test dimensionality. The Multi-State Science Assessment (MSSA) 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 will focus on the presence of cluster effects and how substantial they are. Additionally, confirmatory factor analysis is used to evaluate the fit of the IRT model and to compare the model to alternative models, including those with simpler internal structures (i.e., unidimensional models without cluster effects) and models with a more elaborate internal structure.

Another pathway to consider is exploring 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 Section 5.1, Correlations Among Discipline Scores.

## 5.1 CORRELATIONS AMONG DISCIPLINE SCORES

Table 13 through Table 15 present the observed and disattenuated correlation matrix of the discipline scores. The observed correlations ranged from 0.58 to 0.69, 0.57 to 0.70, and 0.58 to 0.68 for the combined states, Rhode Island, and Vermont, respectively. The disattenuated correlations ranged from 0.90 to 0.93, 0.91 to 0.94, and 0.87 to 0.92 for the combined states, Rhode Island, and Vermont, respectively.

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 became very high. The disattenuated correlations were roughly 0.9 or higher, supporting the use of a psychometric model that does not include a separate dimension for each of the three disciplines.

*Table 13. Correlations Among Disciplines, Combined*

| Grade | Reporting Category | Earth and Space Sciences (ESS) | Life Sciences (LS) | Physical Sciences (PS) |
|---|---|---|---|---|
| **5** | ESS | 0.75* | 0.92 | 0.93 |
| | LS | 0.69 | 0.74* | 0.92 |
| | PS | 0.68 | 0.67 | 0.71* |
| **8** | ESS | 0.72* | 0.93 | 0.93 |
| | LS | 0.69 | 0.77* | 0.92 |
| | PS | 0.67 | 0.69 | 0.73* |
| **11** | ESS | 0.65* | 0.91 | 0.90 |
| | LS | 0.62 | 0.71* | 0.91 |
| | PS | 0.58 | 0.61 | 0.64* |

*The diagonal values are marginal reliabilities for each discipline, below the diagonal are the observed correlations, and above the diagonal are the disattenuated correlations.

*Table 14. Correlations Among Disciplines, Rhode Island*

| Grade | Reporting Category | Earth and Space Sciences (ESS) | Life Sciences (LS) | Physical Sciences (PS) |
|---|---|---|---|---|
| **5** | ESS | 0.75* | 0.93 | 0.93 |
| | LS | 0.69 | 0.74* | 0.93 |
| | PS | 0.68 | 0.67 | 0.71* |
| **8** | ESS | 0.72* | 0.94 | 0.94 |
| | LS | 0.70 | 0.77* | 0.93 |
| | PS | 0.68 | 0.70 | 0.73* |
| **11** | ESS | 0.63* | 0.91 | 0.92 |
| | LS | 0.60 | 0.70* | 0.92 |
| | PS | 0.57 | 0.61 | 0.63* |

*The diagonal values are marginal reliabilities for each discipline, below the diagonal are the observed correlations, and above the diagonal are the disattenuated correlations.

*Table 15. Correlations Among Disciplines, Vermont*

| Grade | Reporting Category | Earth and Space Sciences (ESS) | Life Sciences (LS) | Physical Sciences (PS) |
|---|---|---|---|---|
| **5** | ESS | 0.76* | 0.91 | 0.92 |
| | LS | 0.68 | 0.73* | 0.91 |
| | PS | 0.67 | 0.66 | 0.71* |
| **8** | ESS | 0.72* | 0.91 | 0.90 |

| Grade | Reporting Category | Earth and Space Sciences (ESS) | Life Sciences (LS) | Physical Sciences (PS) |
|---|---|---|---|---|
| | LS | 0.67 | 0.77* | 0.90 |
| | PS | 0.64 | 0.67 | 0.72* |
| 11 | ESS | 0.67* | 0.90 | 0.87 |
| | LS | 0.63 | 0.72* | 0.88 |
| | PS | 0.58 | 0.61 | 0.66* |

*The diagonal values are marginal reliabilities for each discipline, below the diagonal are the observed correlations, and above the diagonal are the disattenuated correlations.

## 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 evidence of convergent and discriminant validity. Part of providing validity evidence is 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 MSSA, which could easily permit 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 would be smaller in magnitude due to the larger 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. The pattern was generally consistent with the *a priori* expectation that subscores within a test correlate higher than correlations between tests measuring a different construct. The correlations among the reporting category scores, both observed (below the shaded cells that form a diagonal) and corrected for attenuation (above the shaded cells that form a diagonal) are presented in Table 16 and Table 17. The shaded cells contain the reliability coefficient of the reporting category. Correlations across subjects are presented for grades 5 and 8 only because English language arts (ELA) and mathematics assessments were administered to grades 3–8 only. Only Vermont's correlations are presented, as there was no data available for Rhode Island.

*Table 16. Correlations Across Subjects, Grade 5 Vermont*

| Subject | Number of Students | Reporting Category | Science | | | English Language Arts (ELA) | | | | Mathematics | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | *ESS* | *LS* | *PS* | *R* | *W* | *L* | *R* | *CP* | *PS* | *CR* |
| **Science** | 5,502 | Earth and Space Sciences (ESS) | 0.75* | 0.91 | 0.92 | 0.85 | 0.78 | 0.82 | 0.85 | 0.83 | 0.90 | 0.86 |
| | | Life Sciences (LS) | 0.67 | 0.73* | 0.91 | 0.86 | 0.77 | 0.84 | 0.85 | 0.79 | 0.88 | 0.84 |
| | | Physical Sciences (PS) | 0.67 | 0.66 | 0.71* | 0.85 | 0.77 | 0.84 | 0.84 | 0.82 | 0.90 | 0.86 |
| **ELA** | | Reading (R) | 0.64 | 0.64 | 0.62 | 0.75* | 0.85 | 0.91 | 0.94 | 0.78 | 0.87 | 0.83 |
| | | Writing (W) | 0.58 | 0.56 | 0.55 | 0.63 | 0.72* | 0.84 | 0.86 | 0.79 | 0.87 | 0.81 |
| | | Listening (L) | 0.58 | 0.58 | 0.58 | 0.65 | 0.58 | 0.66* | 0.91 | 0.79 | 0.87 | 0.85 |
| | | Research (R) | 0.64 | 0.63 | 0.61 | 0.70 | 0.64 | 0.65 | 0.75* | 0.80 | 0.88 | 0.85 |
| **Mathematics** | | Concepts Procedures (CP) | 0.67 | 0.63 | 0.65 | 0.64 | 0.63 | 0.61 | 0.65 | 0.88* | 0.99 | 0.97 |
| | | Problem Solving, Modeling, and Data Analysis (PS) | 0.64 | 0.62 | 0.62 | 0.62 | 0.61 | 0.58 | 0.62 | 0.76 | 0.67* | 1.00 |
| | | Communicating Reasoning (CR) | 0.63 | 0.60 | 0.60 | 0.60 | 0.58 | 0.58 | 0.62 | 0.76 | 0.71 | 0.70* |

*Diagonal values represent the reliability coefficient of the reporting category. Observed correlations are below the diagonal and disattenuated are above. The disattenuated correlations larger than 1 were truncated to 1.

*Table 17. Correlations Across Subjects, Grade 8 Vermont*

| Subject | Number of Students | Reporting Category | Science | | | English Language Arts (ELA) | | | | Mathematics | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | *ESS* | *LS* | *PS* | *R* | *W* | *L* | *R* | *CP* | *PS* | *CR* |
| **Science** | 5,606 | Earth and Space Sciences (ESS) | 0.72* | 0.91 | 0.90 | 0.82 | 0.75 | 0.80 | 0.81 | 0.83 | 0.90 | 0.85 |
| | | Life Sciences (LS) | 0.67 | 0.77* | 0.90 | 0.83 | 0.74 | 0.82 | 0.80 | 0.80 | 0.88 | 0.84 |
| | | Physical Sciences (PS) | 0.64 | 0.67 | 0.72* | 0.81 | 0.76 | 0.80 | 0.79 | 0.83 | 0.91 | 0.85 |
| **ELA** | | Reading (R) | 0.60 | 0.62 | 0.59 | 0.74* | 0.89 | 0.91 | 0.90 | 0.80 | 0.90 | 0.82 |
| | | Writing (W) | 0.54 | 0.55 | 0.55 | 0.65 | 0.73* | 0.84 | 0.88 | 0.81 | 0.88 | 0.82 |
| | | Listening (L) | 0.53 | 0.56 | 0.53 | 0.62 | 0.56 | 0.61* | 0.88 | 0.79 | 0.90 | 0.83 |
| | | Research (R) | 0.58 | 0.59 | 0.56 | 0.65 | 0.63 | 0.58 | 0.71* | 0.79 | 0.88 | 0.82 |
| **Mathematics** | | Concepts Procedures (CP) | 0.65 | 0.65 | 0.65 | 0.64 | 0.64 | 0.58 | 0.62 | 0.86* | 1.00 | 0.97 |
| | | Problem Solving, Modeling, and Data Analysis (PS) | 0.62 | 0.63 | 0.63 | 0.63 | 0.61 | 0.57 | 0.60 | 0.78 | 0.66* | 1.00 |
| | | Communicating Reasoning (CR) | 0.57 | 0.58 | 0.57 | 0.56 | 0.56 | 0.51 | 0.55 | 0.72 | 0.67 | 0.63* |

*Diagonal values represent the reliability coefficient of the reporting category. Observed correlations are below the diagonal and disattenuated 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: ELA, mathematics, and science. The correlations presented in Table 18 were relatively high, from 0.77 to 0.80 for Vermont.

*Table 18. Correlations Across Spring 2022 ELA, Mathematics, and Science Scores, Vermont*

| Grade | N | ELA and Mathematics | ELA and Science | Mathematics and Science |
|---|---|---|---|---|
| **5** | 5,502 | 0.79 | 0.80 | 0.78 |
| **8** | 5,606 | 0.78 | 0.77 | 0.78 |

## 5.3 CLUSTER EFFECTS

The MSSA is modeled with the Rasch testlet model (Wang & Wilson, 2005). The IRT model is a high-dimensional model, incorporating a nuisance dimension for each item cluster, in addition to a dimension representing overall proficiency. Section 5.1 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 9. 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, and Turhan (2018) confirmed that both the item difficulty parameters and the cluster variances are recovered well for the Rasch testlet model under a variety of conditions. Cluster effects with a range of magnitudes were recovered well. The results obtained by Rijmen et al. (2018) confirmed earlier findings reported in the literature (e.g., Bradlow, Wainer, & Wang, 1999) under conditions that were chosen to closely resemble the assessment. 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.

We examined the distribution of cluster variances obtained from the 2019 IRT calibrations for the entire bank used across all states that participated in the Memorandum of Understanding (MOU) item-sharing agreement and the states that relied on the science Independent College and Career Readiness (ICCR) item pool.

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 0.57 and a mean value of 0.92. The median value was slightly smaller than the estimated variance parameters of the overall dimension ($\hat{\sigma}^2_{\theta_{RI}} = 0.84$, $\hat{\sigma}^2_{\theta_{VT}} = 0.75$, and $\hat{\sigma}^2_{\theta_{pooled}} = 0.81$).

For middle school, the estimated value of the cluster variances of all operational, scored items ranged from 0 to 4.63, with a median value of 0.46 and a mean value of 0.68. The median value

was slightly smaller than the estimated variance parameters of the overall dimension ( $\hat{\sigma}^2_{\theta_{RI}} = 0.79$, $\hat{\sigma}^2_{\theta_{VT}} = 0.77$, and $\hat{\sigma}^2_{\theta_{pooled}} = 0.78$).

For high school, the estimated value of the cluster variances of all operational, scored items ranged from 0.11 to 7.75, with a median value of 0.45 and a mean value of 0.65. The median value was slightly smaller than the estimated variance parameters of the overall dimension ( $\hat{\sigma}^2_{\theta_{RI}} = 0.67$, $\hat{\sigma}^2_{\theta_{VT}} = 0.71$, and $\hat{\sigma}^2_{\theta_{pooled}} = 0.69$).

**Error! Reference source not found.** through Figure 6 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 $\eta_g = \frac{\sigma_g^2}{\sigma^2_{\theta_{RI}} + \sigma_g^2}$, $\eta_g = \frac{\sigma_g^2}{\sigma^2_{\theta_{VT}} + \sigma_g^2}$, and $\eta_g = \frac{\sigma_g^2}{\sigma^2_{\theta_{pooled}} + \sigma_g^2}$), where $\sigma^2_{\theta_{RI}}$ and $\sigma^2_{\theta_{VT}}$ are the variance estimates of the overall proficiency of students in Rhode Island and Vermont, respectively, and $\hat{\sigma}^2_{\theta_{pooled}}$ is the pulled variance estimate of both states. The variance proportion shows the relative magnitude of the variance of a cluster compared to the variance of the overall dimension. For instance, if the variance proportion of a cluster is larger than 0.5, then the cluster variance is larger than the overall variance; otherwise, the cluster variance is smaller than the overall variance. For all three grade bands, a wide range of cluster variances was observed. These results indicated 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.

*Figure 4. Cluster Variance Proportion for Operational Items in Elementary School*

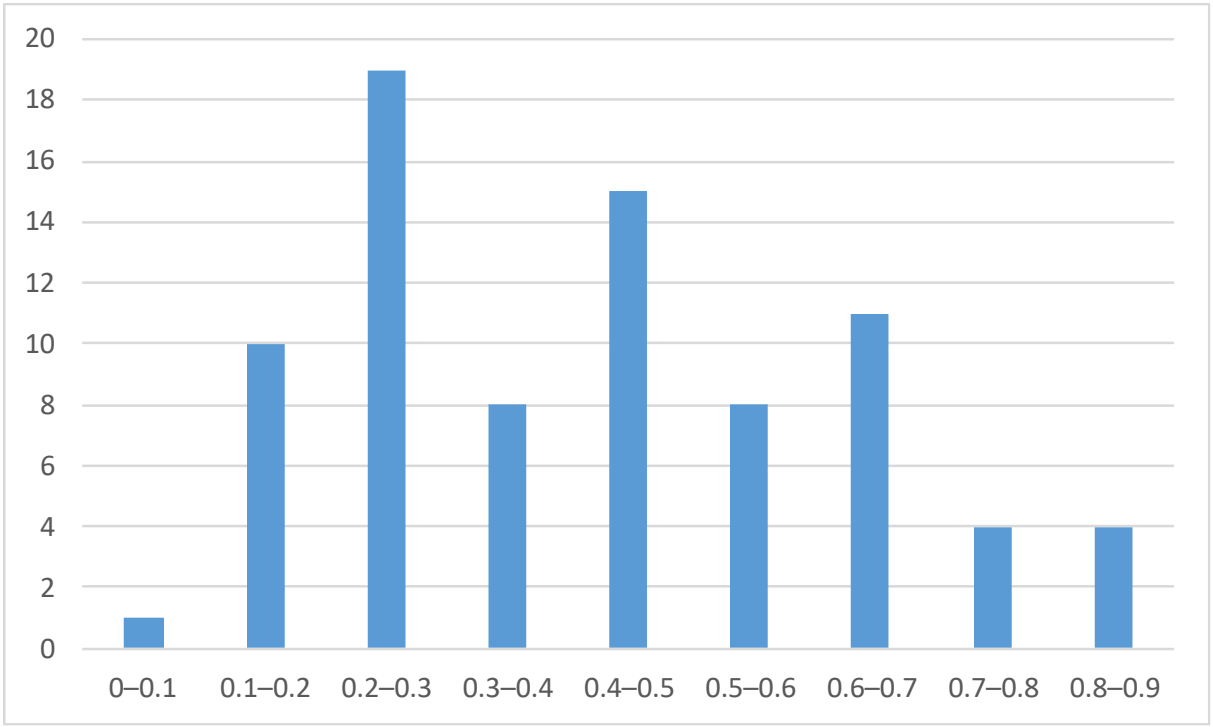


*Figure 5. Cluster Variance Proportion for Operational Items in Middle School*

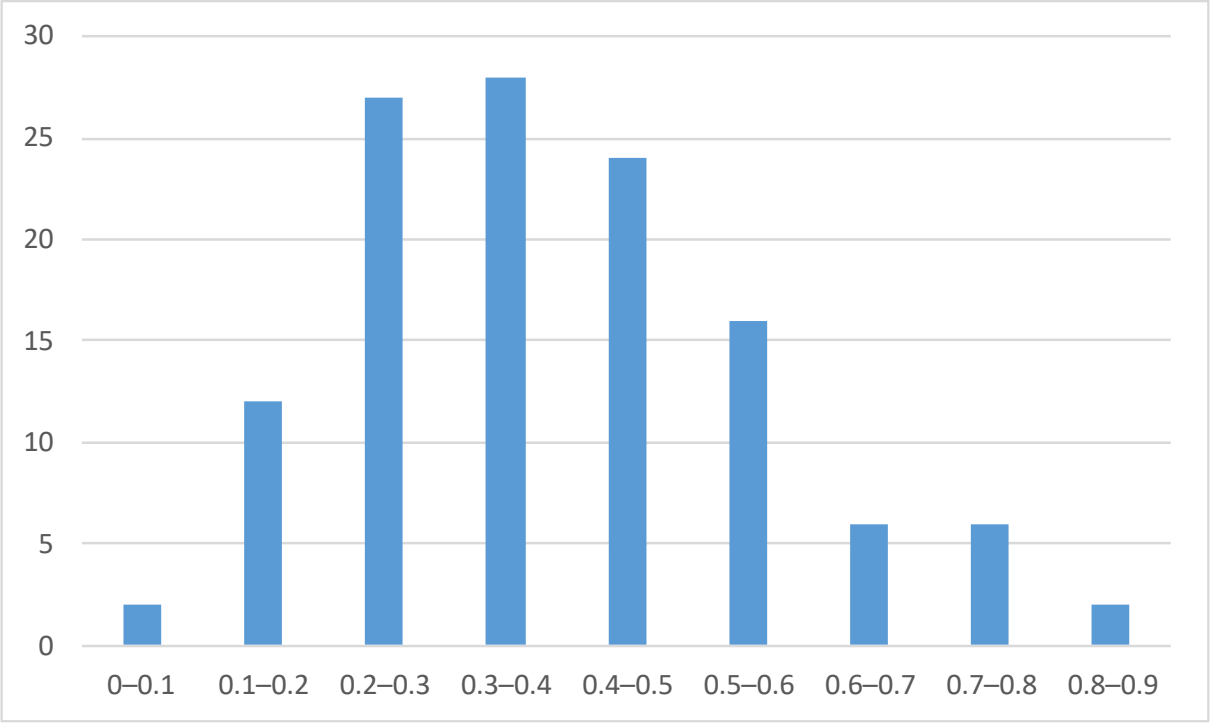
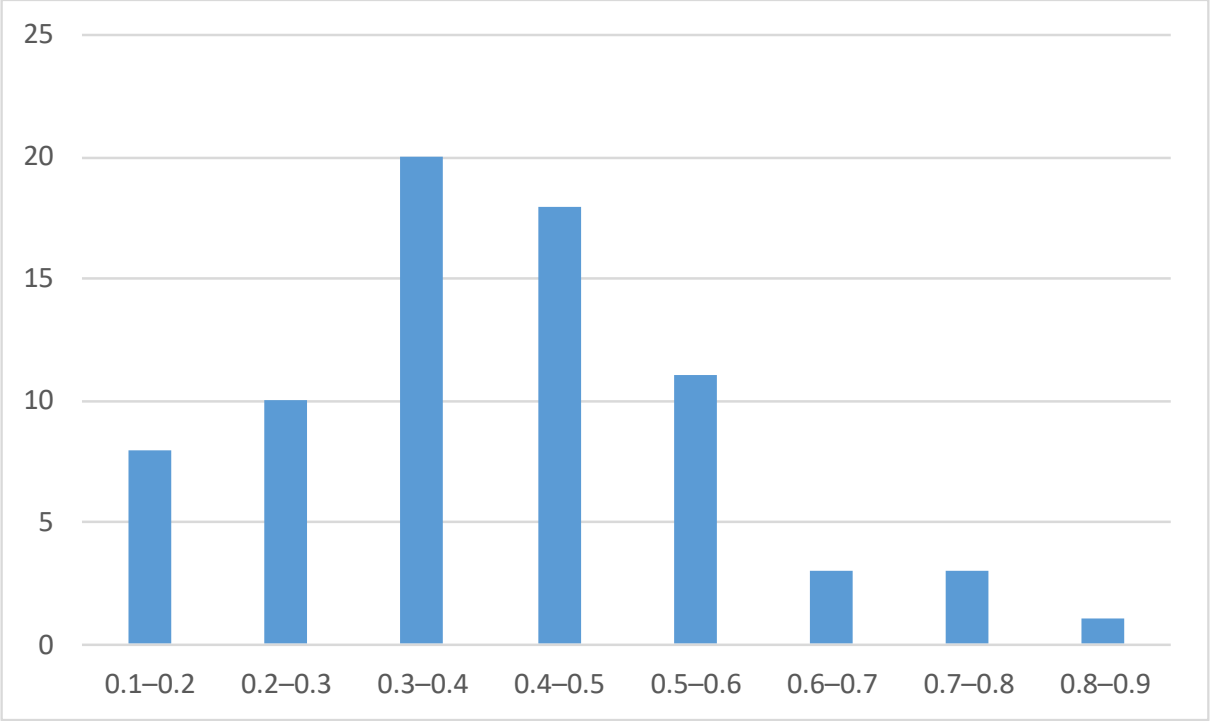



*Figure 6. Cluster Variance Proportion for Operational Items in High School*



## 5.4 CONFIRMATORY FACTOR ANALYSIS

In Section 5.3, Cluster Effects, evidence is presented for the existence of substantial cluster effects. In this section, the internal structure of the IRT model used for calibrating the item parameters is further evaluated using confirmatory factor analysis (CFA). In addition, alternative models are considered, including models with a simpler internal structure (e.g., unidimensional models) and models with a more elaborate internal structure.

Estimation methods for the CFA of 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 test (LOFT) design results in sparse data matrices. Because every student responded to only a small number of items relative to the size of the item pool, data were 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 the Next Generation Science Standards (NGSS) framework, except for Utah. As a result, the student responses from other states were not readily amenable to the application of CFA techniques.

The 2018 Utah operational field test for science used one set of fixed-form tests for each grade. Therefore, the data for each fixed-form test were complete, and the fixed-form tests were amenable to CFA. Even though the standards are grade-specific for middle school, the Utah science standards 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 MOU, and many of the other states that participate in the MOU also use the middle school items developed for and owned by Utah. Taken together, analyzing the fixed science forms that were administered in Utah in 2018 can provide evidence regarding the internal structure of the MSSA.

In 2018, Utah's science assessments comprised a set of fixed-form tests per grade, and all items in those forms were clusters. The number of fixed-form tests varied by grade, but within each grade the total number of clusters was the same across forms. However, some items were rejected during rubric validation or data review and were removed from this analysis. All students with a "completed" status were included in the factor analysis. The percentage of students per grade that had a status other than "completed" was less than 0.85%. Table 19 summarizes the number of forms included in this analysis, the number of clusters per discipline (range across forms), the number of assertions (range across forms), and the number of students (range across forms) for each grade.

*Table 19. Range Across Forms for Number of Forms, Clusters per Discipline, Number of Assertions per Form, and Number of Students per Form*

| Grade | Number of Fixed Forms | Number of Clusters per Discipline in Each Form | | | Number of Assertions per Form | Number of Students per Form |
|---|---|---|---|---|---|---|
| | | *Physical Sciences* | *Earth and Space Sciences* | *Life Sciences* | | |
| **6** | 3 | 2 | 2–3 | 2–3 | 74–83 | 6,804–6,881 |
| **7** | 6 | 2 | 2 | 5 | 83–89 | 3,822–3,890 |
| **8** | 3 | 6–7 | 2 | 2 | 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 a cluster are indicators of the cluster, and each cluster is an indicator of overall science achievement. Because assertions are not pure indicators of a specific factor, each assertion has a corresponding error component. Similarly, clusters include an error component, indicating they are not pure indicators of the overall science achievement.

CAI used a CFA to evaluate the fit of the second-order model described above to student data from spring 2018. Three additional structural models were included in the analysis as well. In the first model, there is only one factor representing overall science achievement. All assertions are indicators of this overall proficiency factor. The first model is a testlet model where all cluster variances are zero. In the second model, assertions are indicators of the corresponding science discipline, and each discipline is an indicator of the overall science achievement. This is a second-order model with science disciplines rather than clusters as first-order factors. This model does not take the cluster effects into account. In the last, most general model, assertions are indicators of the corresponding cluster, and clusters are 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)

**Error! Reference source not found.**7 through Figure 1010 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 are different from the calibration model for which all the discrimination parameters of the assertions were set to 1.

*Figure 7. One-Factor Structural Model (Assertions-Overall): "Model 1"*

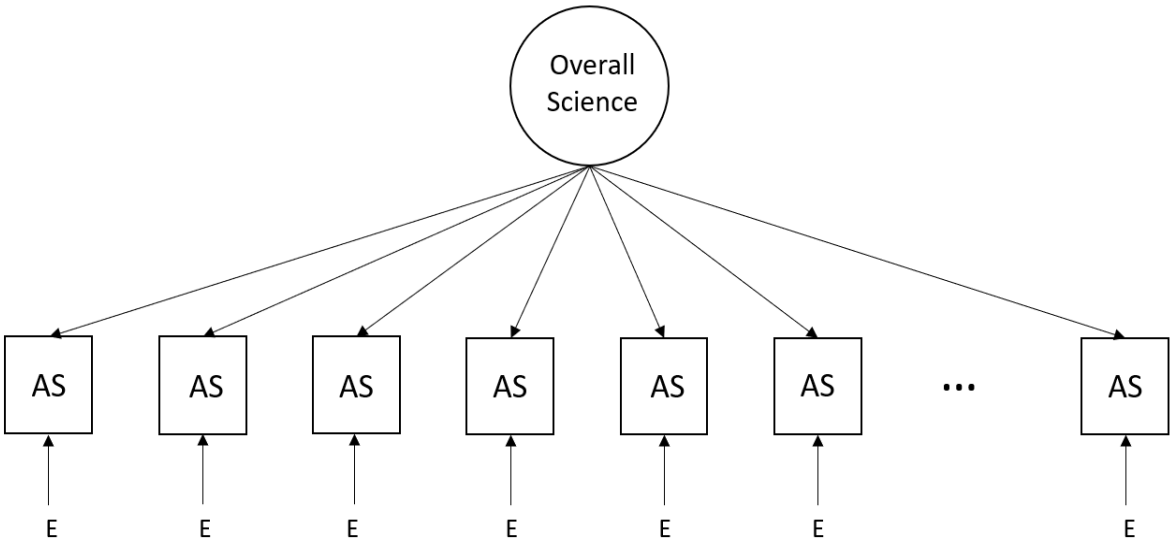



*Figure 8. Second-Order Structural Model (Assertions-Disciplines-Overall): "Model 2"*

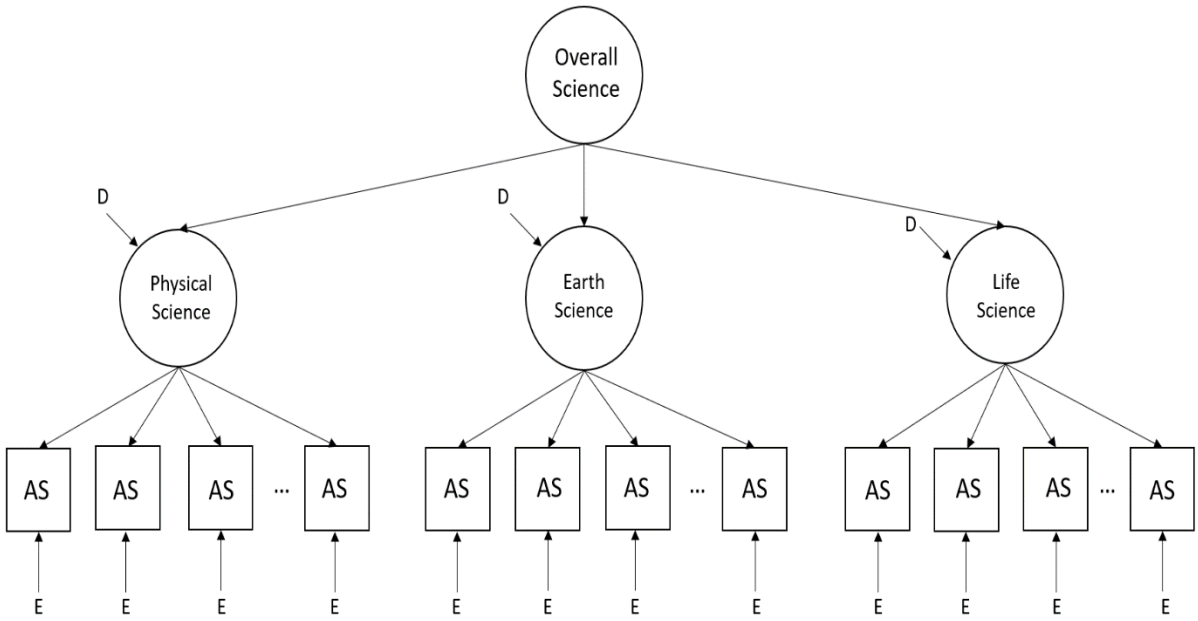


*Figure 9. Second-Order Structural Model (Assertions-Clusters-Overall): "Model 3"*

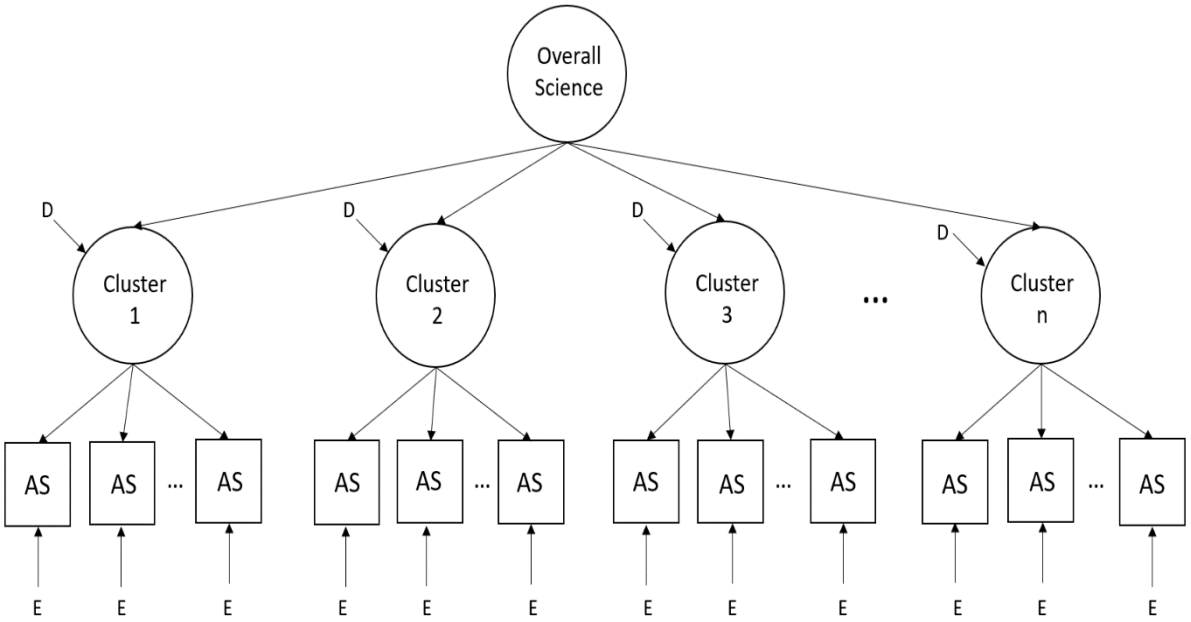
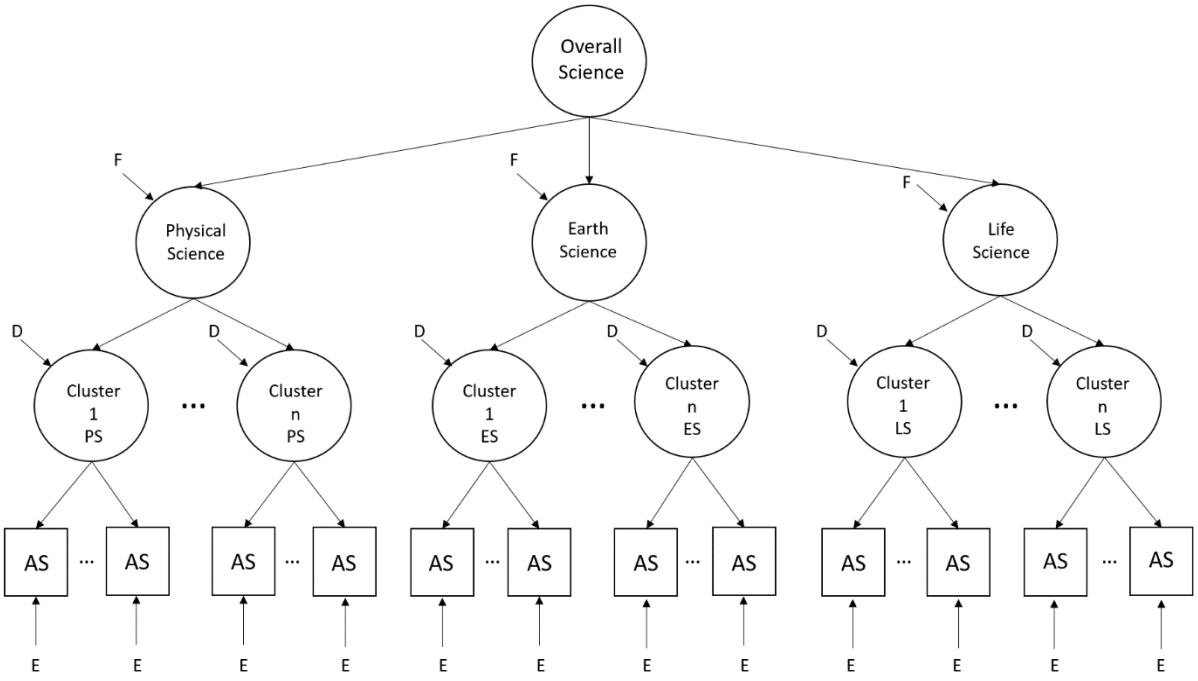



*Figure 10. Third-Order Structural Model (Assertions-Clusters-Disciplines-Overall): "Model 4"*



### 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). The CFI and TLI are relative fit indices, meaning they evaluate model fit by comparing the model of interest to a baseline model. The RMSEA and SRMR are indices of absolute fit. Table 20 provides a list of these measures along with the corresponding thresholds that indicate a good fit.

*Table 20. 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 21 through Table 23 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 20. The following conclusions can be drawn across all grades and models:

- Model 1 showed the most misfit across grades and forms.
- Across forms, Model 3 generally showed more improvement in model fit relative to Model 1 than Model 2 did (i.e., higher values for the CFI and TLI and lower values for the RMSEA and SRMR). This means that accounting for the clusters resulted in a higher improvement in model fit over a single factor model than 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 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 and 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 21. Fit Measures per Model and Form, Grade 6*

| Model | Form | CFI | TLI | RMSEA | SRMR |
|---|---|---|---|---|---|
| **Model 1** Assertions-Overall (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 (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 (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 (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 22. Fit Measures per Model and Form, Grade 7*

| Model | Form | CFI | TLI | RMSEA | SRMR |
|---|---|---|---|---|---|
| **Model 1** Assertions-Overall (one-factor model) | **1** | **0.892** | **0.889** | 0.06 | 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 (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 (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 (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 23. Fit Measures per Model and Form, Grade 8*

| Model | Form | CFI | TLI | RMSEA | SRMR |
|---|---|---|---|---|---|
| **Model 1** Assertions-Overall (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 (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 (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 (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 that were part of this factor analysis study. After removing this item, there were only two forms that 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 24**Error! Reference source not found.** shows the fit measures for grade 6 after removing the item that caused the 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 24. Fit Measures per Model and Form – 6th Grade – One Cluster Removed*[2]

| Model | Form | CFI | TLI | RMSEA | SRMR |
|---|---|---|---|---|---|
| **Model 1** Assertions-Overall (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 (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 (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 (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 25**Error! Reference source not found.** shows the estimated correlations among disciplines for Model 4 (third-order model). The correlations were all very high and ranged between 0.913 and 1. The high correlations between the disciplines in Model 4 indicated 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 25. 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 |

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

| Grade | Form | Discipline | Earth and Space Sciences (ESS) | Life Sciences (LS) |
|---|---|---|---|---|
| | | Earth and Space Sciences (ESS) | – | 0.973 |
| | 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 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):

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 have received extensive training on the principles of universal design and apply these principles in the development of all test materials. Rhode Island and Vermont educators and stakeholders verified adherence to the principles of universal design throughout the review process.

## 6.1 COGNITIVE LABORATORY STUDIES

In 2017, when the development of item clusters for the states that were part of the Memorandum of Understanding (MOU) started, cognitive lab studies were conducted to evaluate and refine the process of developing item clusters aligned to the Next Generation Science Standards (NGSS). The results of the cognitive lab studies confirmed the feasibility of the approach. Item clusters were completed within 12 minutes on average, and students reported being familiar with the format conventions and online tools used in the item clusters. They 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 were conducted in 2018 and 2019 to determine if students using braille could understand the task demands of selected accommodated three-dimensional science standards-aligned item clusters and navigate the interactive features of these 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 and/or the Job Access With Speech (JAWS) screen-reading software 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 clusters were different from (and more complex than) other tests with which the students were familiar; however, the study recommended that students be given adequate time to practice with at least one

sample cluster before taking the summative test. The study also resulted in tool-specific recommendations for accessibility for visually impaired students. The reports of both sets of cognitive lab 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

A 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 details surrounding the review of those items for bias along with the DIF analysis process for the MSSA are described further in Section 4.4, Differential Item Functioning Analysis, of Volume 1, Annual Technical Report.

# 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 was in line with acceptable industry standards.
- ***Content Validity.*** Evidence is provided to support the assertion that content coverage on each test was consistent with the 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 model assumptions, and the reporting of an overall score and subscores at the reporting-category levels.
- ***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 intended to assess similar constructs, as well as between the test and other measures intended to assess different constructs.
- ***Test Fairness.*** Items were developed following the principles of universal design, which removed barriers to provide access for the widest range of students possible. Evidence of test fairness is provided statistically 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, NY: 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*(1), 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 Measuremen*t, *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.