Connecticut Next Generation Science Standards Assessment

2020–2021

Volume 4
Evidence of Reliability and Validity

CSDE

CONNECTICUT STATE
DEPARTMENT OF EDUCATION

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

1. INTRODUCTION AND OVERVIEW OF RELIABILITY AND VALIDITY EVIDENCE

The state of Connecticut implemented the Connecticut Next Generation Science Standards (NGSS) Assessment for operational use starting in the 2018–2019 school year. The Connecticut NGSS Assessment replaced the Connecticut Mastery Test (CMT) Science (administered to students in grades 5 and 8) and the Connecticut Academic Performance Test (CAPT) Science (administered to students in grade 10). The Connecticut NGSS Assessment is administered online to grades 5, 8, and 11 using a linear-on-the-fly (LOFT) test 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. For the 2020–2021 school year, remote testing forms were constructed to allow for assessing science among students taking the test remotely. Table 1 shows the complete list of tests for the operational test administration in spring 2021.

Table 1. Spring 2021 Assessment Modes

Language/Format	Assessment Mode	Grade
English	Online	5, 8, & 11
Spanish	Online	5, 8, & 11
English/Data Entry Interface (DEI)	Paper	5, 8, & 11
English/braille	Paper	5, 8, & 11
English and Spanish/remote	Online	5, 8, & 11

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

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

- ***Reliability.*** The reliability estimates are presented by grade and demographic subgroup. This section also includes conditional standard errors of measurement (CSEM) and classification accuracy and consistency results by grade.
- ***Content validity.*** This section presents evidence showing that all students' tests were constructed to measure the NGSS with a sufficient number of items targeting each area of the test blueprint.
- ***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 was used to evaluate the fit of the IRT model and to compare it with alternative models, including models with a simpler internal structure (e.g., unidimensional models) and models with a more elaborate internal structure.

- ***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.
- ***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. Test fairness is further monitored statistically using differential item functioning (DIF) analysis in tandem with content reviews by specialists.

1.1 Reliability

The term *reliability* refers to consistency in test scores. Reliability 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, he or she should receive consistent results. The reliability coefficient refers to the ratio of true score variance to observed score variance:

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

Another way to view reliability is to consider its relationship with the standard errors of measurement (SEM)—the smaller the standard error, the higher the precision of the test scores. For example, classical test theory 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 classical test theory SEM, which assumes a homoscedastic error, is derived from the classical notion expressed above as $\sigma_X\sqrt{1-\rho_{XX'}}$, where σ_X is the standard deviation of the scaled score, and $\rho_{XX'}$ is a reliability coefficient. Based on the definition of reliability, this formula can be derived as follows:

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

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

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

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

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

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

This shows that the SEM in the classical test theory 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 score at various score points. See Section 3, Reliability, for the derivation of heterogeneous measurement errors in IRT and a discussion of 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 and other modes of assessment” (p. 6). 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 (see 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 (see 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 his or her responses to items, this could affect item responses and inferences regarding abilities on the measured construct (see 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 test takers" (AERA, APA, & NCME, 2014, p. 15). This evidence is collected by surveying examinees 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*: 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 (see Section 3, Reliability; and Section 5, Evidence of Internal-External Structure, 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 examinees (see 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 multi-trait–multi-method matrix can be used. Additionally, test-criterion relationships indicate how accurately test scores predict criterion performance. The degree of accuracy depends mainly 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 5.2, Convergent and Discriminant Validity.

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 upon 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. As described in Volume 1, Annual Technical Report, and in this volume, test use should align with the intended purpose of the test.

Supporting a validity argument requires multiple sources of validity evidence. This enables one to evaluate whether sufficient evidence has been presented 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, subsequently, evidence that the scores can be used to support these inferences.

2. PURPOSE OF THE CONNECTICUT NEXT GENERATION SCIENCE STANDARDS ASSESSMENT

The primary purpose of Connecticut's Summative Assessment System is to yield accurate information on students' achievement of Connecticut's education standards. The Connecticut NGSS Assessment measures the science knowledge and skills of Connecticut students in grades 5, 8, and 11. The Connecticut State Department of Education (CSDE) provides an overview of the science assessment at https://portal.ct.gov/SDE/Student-Assessment/NGSS-Science/NGSS-Science. Information about the NGSS is available at www.nextgenscience.org.

The Connecticut NGSS Assessment supports instruction and student learning by measuring growth in student achievement. Assessments can be used as indicators to determine whether students in Connecticut are ready with the knowledge and skills that are essential for college education and careers.

Connecticut'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 tests are constructed to measure student proficiency in accordance with best practice as described in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014). Item development adheres to the principles of universal design in order to ensure that all students have access to the test content. Volume 2, Test Development, describes in more detail the Connecticut NGSS Assessment standards and test blueprints. Additional evidence of content validity can also be found in Section 4, Evidence of Content Validity. The Connecticut NGSS Assessment test scores are useful indicators for understanding individual students' academic achievement of the Connecticut content standards and for evaluating whether students' performance is improving over time. Additionally, both individual and aggregated scores can be used for measuring reliability of the test. A discussion of test-score reliability can be found in Section 3, Reliability.

The Connecticut NGSS Assessment is a criterion-referenced test that is designed to measure student performance on the NGSS in Connecticut Schools. As a comparison, norm-referenced tests are designed to compare or rank all students with one another. The Connecticut NGSS Assessment 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 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 must examine the reliability coefficients for and the validity of these test scores.

3. Reliability

Classical test-theory-based reliability indices are not appropriate for science assessments for two reasons. First, in spring 2021, the science test was administered under a linear-on-the-fly (LOFT) test design. Potentially, each student received a unique set of items, whereas classical test-theory-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 tests 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 Errors 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 0.88 to 0.89. Due to the new structure of the test, CAI has also explored the relationships between reliability and other important factors, such as the effect of nuisance dimensions (see Volume 1, Section 5, Annual Technical Report). It was found that if the local dependencies among assertions pertaining to the same item are ignored, the marginal reliability typically increases to 0.90 or above. 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 (see Volume 1, Section 6.1, Annual Technical Report). Obviously, by ignoring the local dependencies, which are substantial for many item clusters, the reliability coefficient is overestimating the true reliability of the test. Note, however, that local dependencies are also present to some degree in traditional assessments that make use of 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 reported reliability coefficients may be overestimating to some degree the true reliability of these tests. The reliability coefficients are also reported for demographics subgroups in Appendix A, Student Demographics and Reliability Coefficients.

Table 2. Marginal Reliability Coefficients

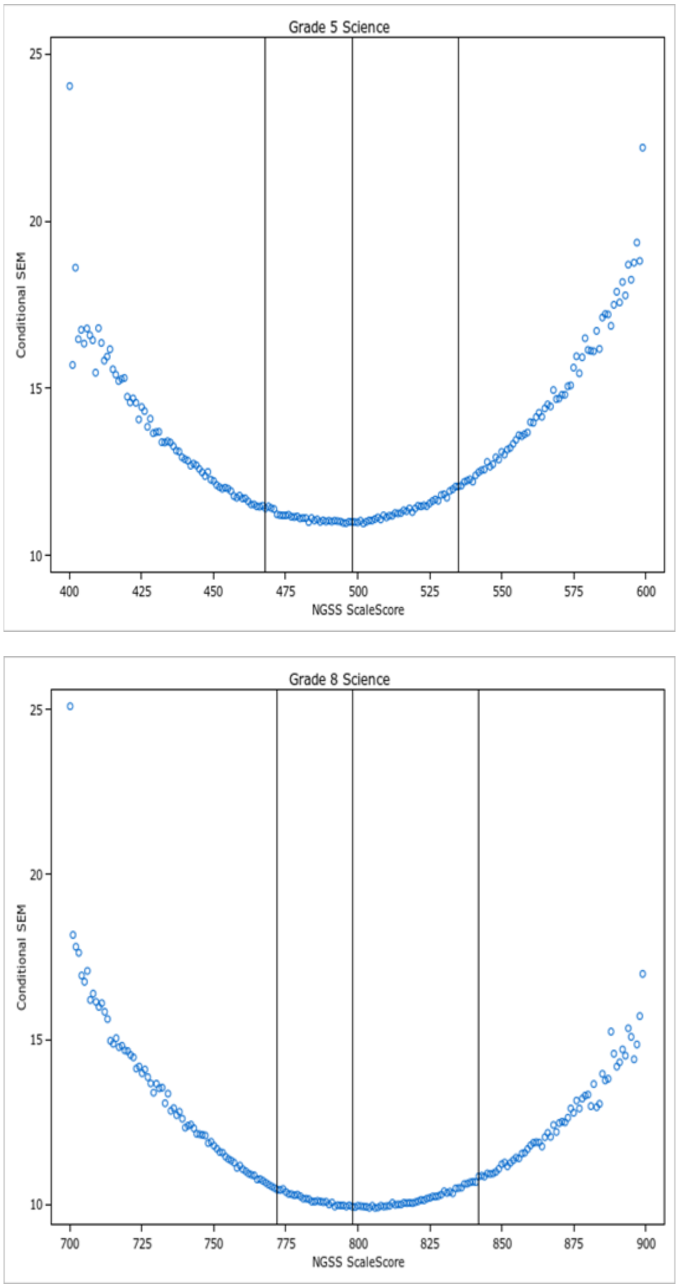
Grade	Sample Size	Reliability
5	34,938	0.88

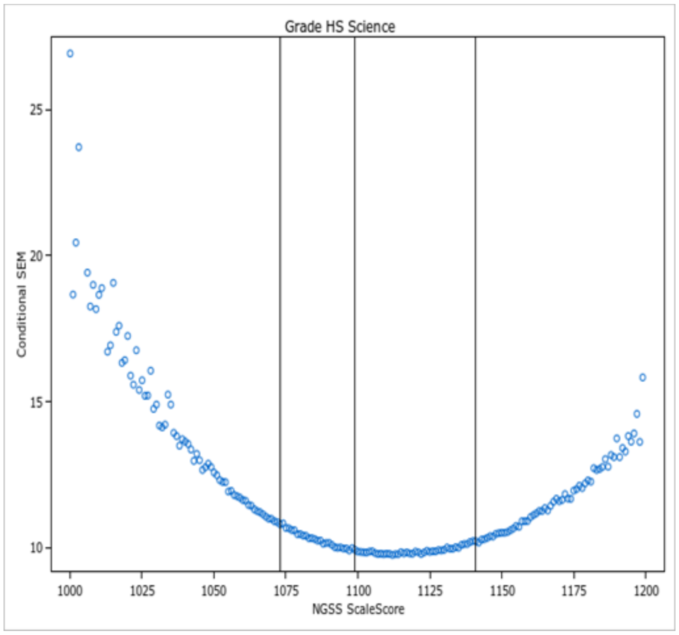
Grade	Sample Size	Reliability
8	36,391	0.89
11	29,789	0.88

3.1 Standard Error of Measurement

The computation method of conditional standard errors of measurement (CSEMs) has been 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 (the middle vertical line) 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





3.2 Reliability of Performance Classification

When student performance is reported in terms of performance 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 of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014).

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

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

where c_{Ll} and c_{Ul} denote the score corresponding to the lower and upper limits of performance 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

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

where pl_j is the jth student's observed performance level. The classification accuracy (CA) at level l is estimated as

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

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

The CA for the pth cut 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 cuts. The overall CA of the test ranges from 76.82% to 77.89%. The individual cut accuracy rates are high across all grades and forms, with the minimum value being 90.22% for grade 11. It denotes that more than 90% of the time we can accurately differentiate students between adjacent performance levels in the spring 2021 Connecticut NGSS Assessment. The CA for demographic subgroups is presented in Appendix C, CA and Consistency Index by Subgroups.

Table 3. Classification Accuracy Index

Grade	**Overall Accuracy (%)**	**Cut Accuracy (%)**		
		Cut 1	*Cut 2*	*Cut 3*
5	76.82	92.23	90.75	93.80
8	77.89	91.27	90.65	95.93
11	76.97	91.27	90.22	95.44

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 performance level l in test one, and p_{jk} is the probability of the true score at performance level k in test two for the jth student. The classification consistency index for the cuts (CCC) and overall classification consistency (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 4 provides the classification consistency for the cuts. The overall CC of the test ranges from 67.81% to 69.36%. The individual cut consistency rates are high across all grades and forms, with the minimum value being 86.22% for grade 11. In all performance levels, CA is 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 accuracy and consistency rates for each performance level are higher for the levels with a smaller standard error. The CC for demographic subgroups is presented in Appendix C, Classification Accuracy and Consistency Index by Subgroups.

Table 4. Classification Consistency Index

Grade	Overall Consistency (%)	Cut Consistency (%)		
		Cut 1	*Cut 2*	*Cut 3*
5	67.81	89.07	86.97	91.31
8	69.36	87.84	86.80	94.23
11	68.26	87.96	86.22	93.53

3.3 PRECISION AT CUT SCORES

Table 5 presents the mean CSEM at each performance level by grade. The table also includes performance level cut scores and associated CSEM. The CSEM at each scale score is reported in Appendix B, Conditional Standard Error of Measurement.

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

Grade	Performance Level	Mean CSEM	Cut Score (Scale Score)	CSEM at Cut Score
5	1	12.47	-	-
	2	11.13	468	11.41
	3	11.33	498	11.01
	4	13.38	535	12.07
8	1	11.67	-	-
	2	10.15	772	10.45
	3	10.11	798	9.94
	4	11.64	842	10.85
11	1	11.87	-	-
	2	10.29	1,073	10.79
	3	9.87	1,099	9.92
	4	11.02	1,141	10.22

4. Evidence of Content Validity

This section demonstrates that the knowledge and skills assessed by the Connecticut NGSS Assessment are representative of the content standards of the larger knowledge domain. We describe the content standards for the Connecticut NGSS Assessment and discuss the test development process and mapping Connecticut NGSS Assessment 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 Connecticut NGSS Assessment was aligned to the NGSS, adopted by Connecticut in 2015. The standards are available for review at the following URL: https://portal.ct.gov/SDE/Science/Science-Standards-and-Resources. 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, of the 2020–2021 Connecticut NGSS Technical Report.

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

Table 6. Number of Items for Each Discipline

Grade	Reporting Category	Item Cluster	Stand-Alone Item
5	Earth and Space Sciences (ESS)	28	33
	Life Sciences (LS)	32	36
	Physics Sciences (PS)	34	40
8	Earth and Space Sciences (ESS)	39	36
	Life Sciences (LS)	51	42
	Physics Sciences (PS)	41	36
11	Earth and Space Sciences (ESS)	11	19
	Life Sciences (LS)	36	51
	Physics Sciences (PS)	15	27

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 non-profit Wisconsin Center for Education Products and Services (WCEPS) conducted an alignment study in July 2019. The study comprised two components. The first component addressed the alignment of the Memorandum of Understanding (MOU) item bank, shared by all states that are part of the MOU. In a second component, 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 (see Volume 2).

The results of the alignment study are presented in Appendix F, Alignment Study Executive Summary.

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 the test dimensionality. The Connecticut NGSS Assessment is modeled with the Rasch testlet model (Wang & Wilson, 2005). The item response theory (IRT) model is a high-dimensional model that incorporates a nuisance dimension for each item cluster (and stand-alone items with four or more assertions) in addition to an overall dimension representing overall proficiency. This approach is innovative and quite different from the traditional approach of ignoring local dependencies. Validity evidence for 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 with alternative models, including those with a simpler internal structure (i.e., unidimensional models without cluster effects) and models with a more elaborate internal structure.

Another pathway is to explore 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 section.

5.1 Correlations Among Discipline Scores

Table 7 presents the observed and disattenuated correlation matrix of the discipline scores. The observed correlations range from 0.69 to 0.72, and disattenuated correlations range from 0.98 to 0.99.

In some instances, the observed correlations were lower than one might expect. As previously noted, however, the correlations were subject to a large amount of measurement error at the discipline level due to the limited number of items from which the scores were derived. Consequently, 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	Reporting Category	Earth and Space Sciences (ESS)	Life Sciences (LS)	Physical Sciences (PS)
5	ESS	0.73*	0.99	0.98
	LS	0.71	0.71*	0.99
	PS	0.69	0.69	0.68*
8	ESS	0.72*	0.98	0.98
	LS	0.71	0.73*	0.98
	PS	0.70	0.70	0.70*
11	ESS	0.71*	0.99	0.99
	LS	0.72	0.74*	0.98
	PS	0.69	0.69	0.68*

Note. *Diagonal value represents marginal reliability for each discipline. Observed correlations are below the diagonal, and disattenuated are above. Disattenuated correlations larger than 1 were truncated to 1.

5.2 Convergent and Discriminant Validity

According to Standard 1.16 of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014), evidence of convergent and discriminant validity must be provided. 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 Connecticut NGSS Assessment, 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., correlation of science disciplines within science) will correlate more positively than subscores 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 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 between reporting categories from science, ELA, and mathematics are presented in Table 8 and Table 9. On the diagonal, the reliability coefficient of the reporting category is shown. Correlations across subjects are presented only for grades 5 and 8 since ELA and mathematics assessments are administered only in grades 3–8.

Table 8. Correlations Across Subjects, Grade 5

Subject	Number of Students	Reporting Category	Science			English Language Arts (ELA)			Mathematics		
			Cat1	*Cat2*	*Cat3*	*Cat1*	*Cat2*	*Cat3*	*Cat1*	*Cat2*	*Cat3*
Science	34,467	Earth and Space Sciences (ESS) (Cat1)	0.73*	0.99	0.98	0.92	0.93	0.92	0.88	1.00	0.95
		Life Sciences (LS) (Cat2)	0.71	0.71*	0.99	0.94	0.96	0.93	0.86	1.00	0.93
		Physical Sciences (PS) (Cat3)	0.69	0.69	0.68*	0.91	0.93	0.90	0.87	1.00	0.93
ELA		Reading (Cat1)	0.69	0.69	0.65	0.76*	1.00	0.98	0.85	1.00	0.92
		Listening (Cat2)	0.61	0.62	0.59	0.67	0.59*	0.99	0.86	1.00	0.92
		Writing and Research (Cat3)	0.72	0.72	0.69	0.79	0.70	0.85*	0.86	1.00	0.93
Mathematics		Concepts and Procedures (Cat1)	0.71	0.68	0.67	0.69	0.62	0.75	0.89*	1.00	1.00
		Problem Solving, Modeling, and Data Analysis (Cat2)	0.67	0.65	0.64	0.67	0.60	0.72	0.78	0.58*	1.00
		Communicating and Reasoning (Cat3)	0.67	0.64	0.63	0.66	0.59	0.71	0.78	0.72	0.68*

Note. *Diagonal value represents the reliability coefficient of the reporting category. Observed correlations are below the diagonal, and disattenuated are above. Disattenuated correlations larger than 1 were truncated to 1.

Table 9. Correlations Across Subjects, Grade 8

Subject	Number of Students	Reporting Category	Science			English Language Arts (ELA)			Mathematics		
			Cat1	*Cat2*	*Cat3*	*Cat1*	*Cat2*	*Cat3*	*Cat1*	*Cat2*	*Cat3*
Science	34,968	Earth and Space Sciences (ESS) (Cat1)	0.72*	0.98	0.98	0.87	0.89	0.87	0.87	0.99	0.91
		Life Sciences (LS) (Cat2)	0.70	0.72*	0.98	0.88	0.89	0.87	0.87	0.98	0.91
		Physical Sciences (PS) (Cat3)	0.69	0.69	0.70*	0.86	0.89	0.87	0.88	0.99	0.91
ELA		Reading (Cat1)	0.65	0.66	0.64	0.77*	1.00	0.99	0.86	0.97	0.91
		Listening (Cat2)	0.59	0.59	0.58	0.69	0.61*	1.00	0.88	1.00	0.92
		Writing and Research (Cat3)	0.66	0.67	0.65	0.78	0.70	0.80*	0.88	0.98	0.92
Mathematics		Concepts and Procedures (Cat1)	0.69	0.69	0.69	0.71	0.64	0.73	0.87*	1.00	1.00
		Problem Solving, Modeling, and Data Analysis (Cat2)	0.63	0.63	0.62	0.65	0.59	0.67	0.76	0.57*	1.00
		Communicating and Reasoning (Cat3)	0.61	0.61	0.60	0.63	0.56	0.65	0.76	0.68	0.62*

Note. *Diagonal value represents the reliability coefficient of the reporting category. Observed correlations are below the diagonal, and disattenuated are above. 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. Correlations are presented in Table 10 and are relatively high, between 0.78 and 0.83.

Table 10. Correlations Across Spring 2021 ELA, Mathematics, and Science Scores

Grade	*N*	English Language Arts (ELA) & Mathematics	ELA & Science	Mathematics & Science
5	34,467	0.82	0.83	0.81
8	34,968	0.80	0.78	0.80

5.3 CLUSTER EFFECTS

The Connecticut NGSS Assessment is modeled with 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 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. The psychometric approach for the assessment is innovative and quite different from the traditional approach of ignoring local dependencies. The validity evidence for 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 participate in the MOU item sharing agreement and the states that rely on the science 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. As a comparison, the estimated variance parameter of the overall dimension for Connecticut elementary school in 2019 was $\hat{\sigma}^2_{\theta CT} = 0.78$.

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 estimated variance parameter of the overall dimension for Connecticut middle school in 2019 was $\hat{\sigma}^2_{\theta CT} = 0.78$.

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 estimated variance parameter of the overall dimension for Connecticut high school in 2019 was $\hat{\sigma}^2_{\theta CT} = 0.83$.

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 $\eta_g = \frac{\hat{\sigma}_g^2}{\hat{\sigma}_{\theta_{CT}}^2+\hat{\sigma}_g^2}$), where $\hat{\sigma}_{\theta_{CT}}^2$ is the variance estimate of the overall proficiency of Connecticut students. 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 is observed. 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

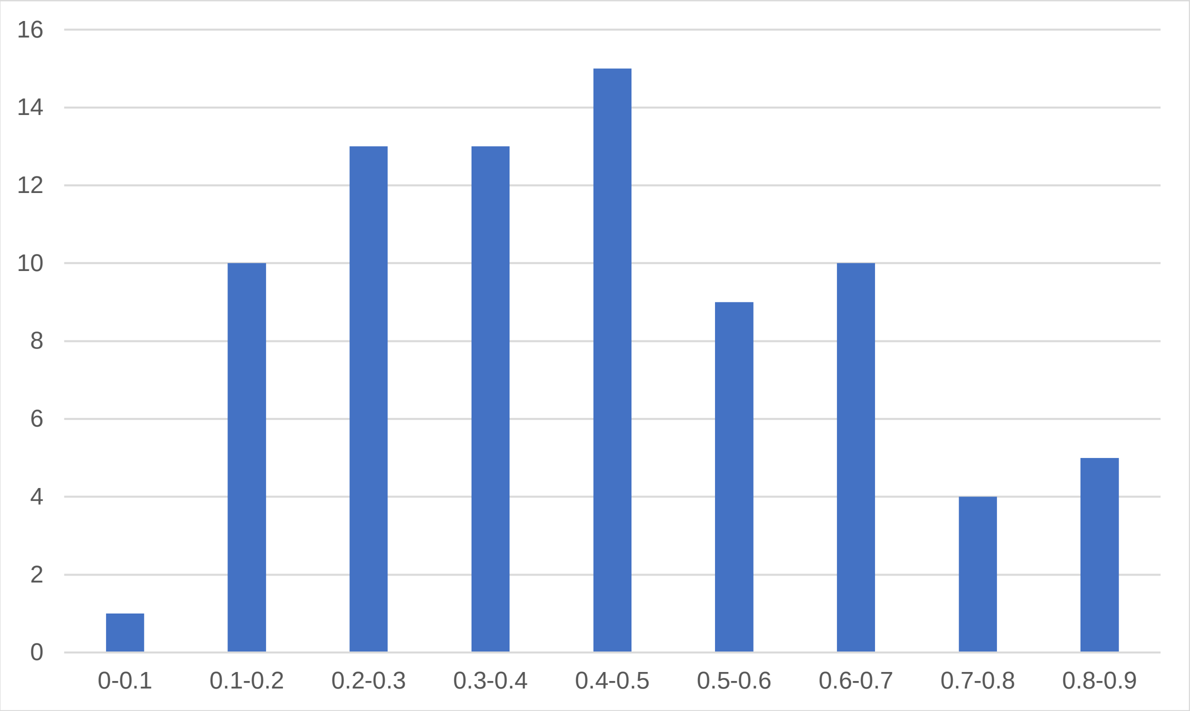


Figure 3. Cluster Variance Proportion for Operational Items in Middle School

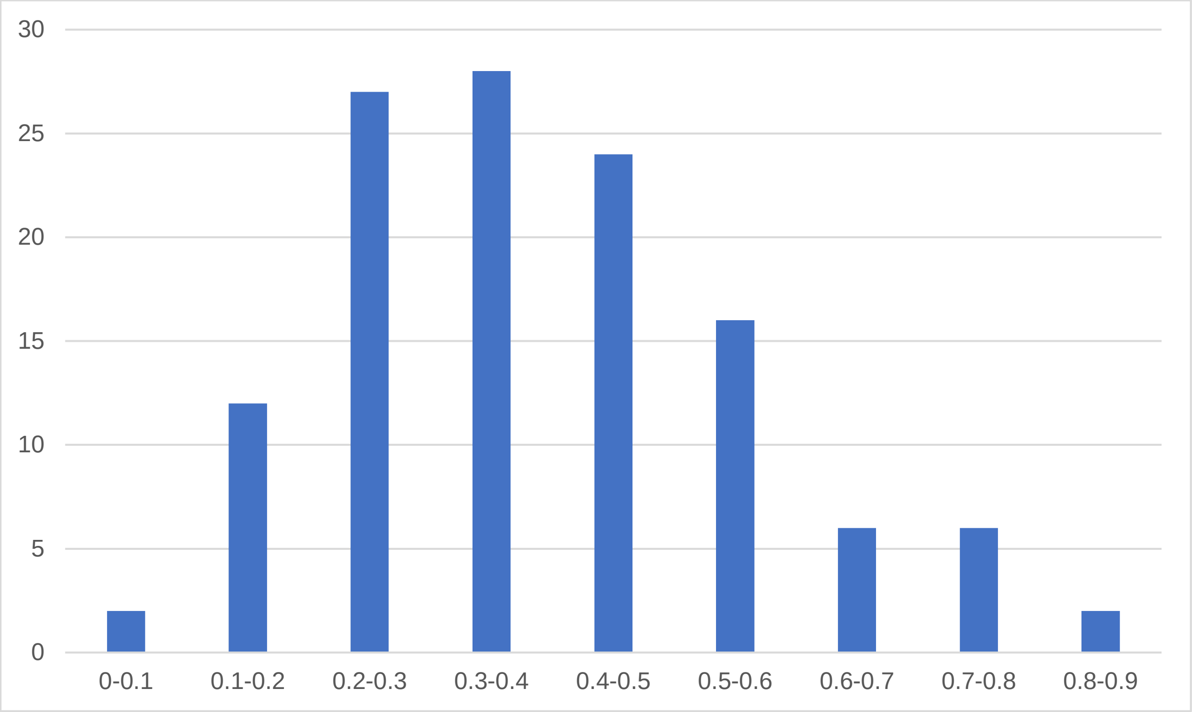
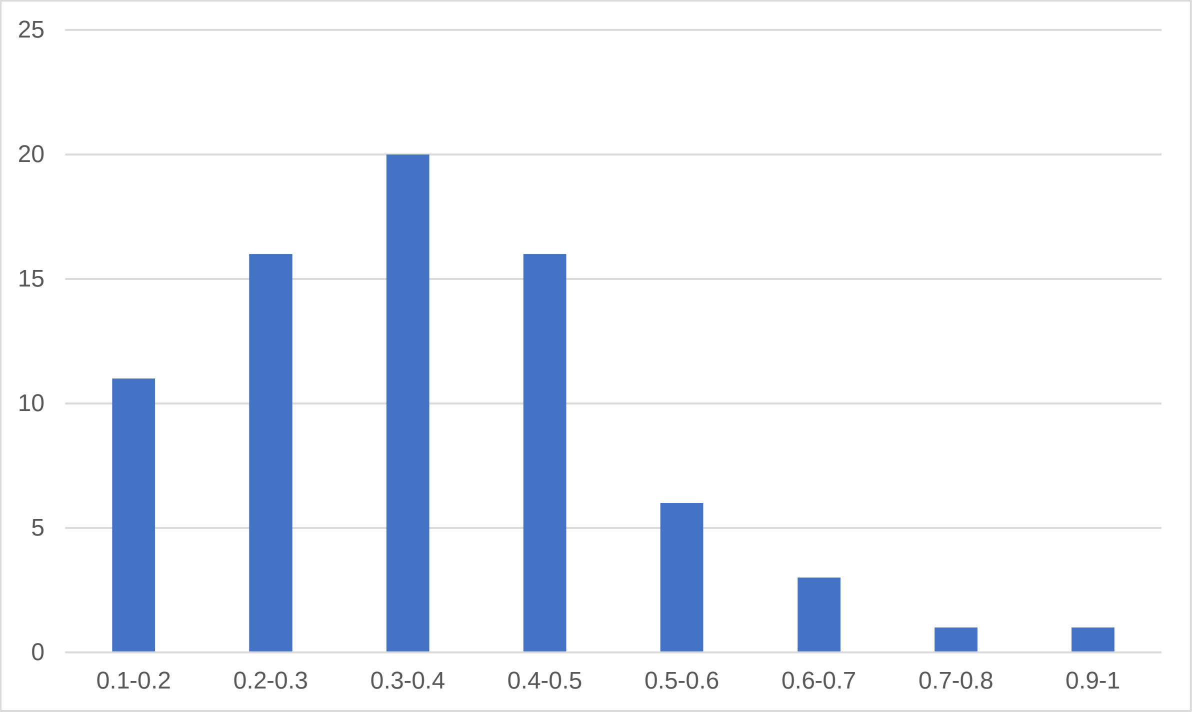


Figure 4. Cluster Variance Proportion for Operational Items in High School



5.4 Confirmatory Factor Analysis

In Section 5.3, Cluster Effects, evidence is presented for the existence of substantial cluster effects. In this section, the internal structure of the IRT model used for calibrating the item parameters is further evaluated using confirmatory factor analysis. 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 confirmatory factor analysis 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. Every student is only responding to 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 the NGSS framework, except for Utah. As a result, the student responses of these other states are not readily amenable for the application of confirmatory factor analysis techniques.

The 2018 Utah operational field test for science 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 confirmatory factor analysis. The Utah science standards, even though the standards are grade-specific for middle school, were developed under a framework similar to the one developed for the NGSS, and a crosswalk is available between both sets of standards. Utah is part of the MOU, and many of the other states that take part 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 with respect to the internal structure of the Connecticut NGSS Assessment.

In 2018, Utah’s science assessments comprised a set of fixed-form tests per grade, and all items in these 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 the 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 11 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 of the grades.

Table 11. Numbers of Forms, Clusters per Discipline (Range Across Forms), Assertions per Form (Range Across Forms), and Students per Form (Range Across Forms)

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

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*		
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 performance. 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 performance.

CAI used confirmatory factor analysis 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, only one factor represented overall science performance. All assertions are indicators of this overall proficiency factor. The first model was a testlet model where 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 performance. This was a second-order model with science disciplines rather than 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 cluster, and clusters were indicators of the corresponding science discipline, with disciplines being indicators of the overall science performance.

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 differs from the calibration model where all the discrimination parameters of the assertions were set to 1.

Figure 5. One-Factor Structural Model (Assertions-Overall): “Model 1”

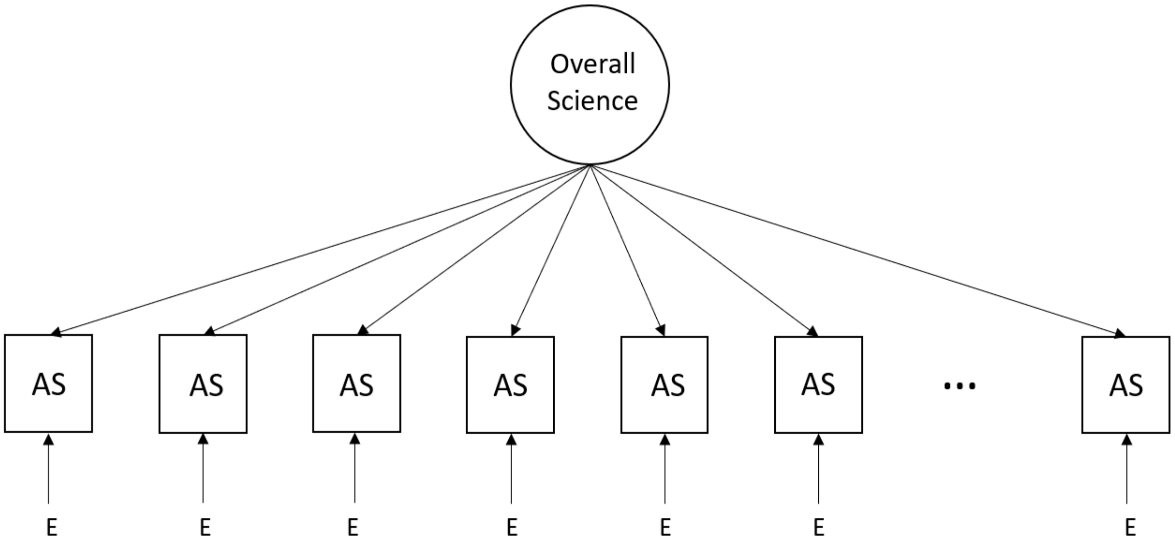


Figure 6. Second-Order Structural Model (Assertions-Disciplines-Overall): “Model 2”

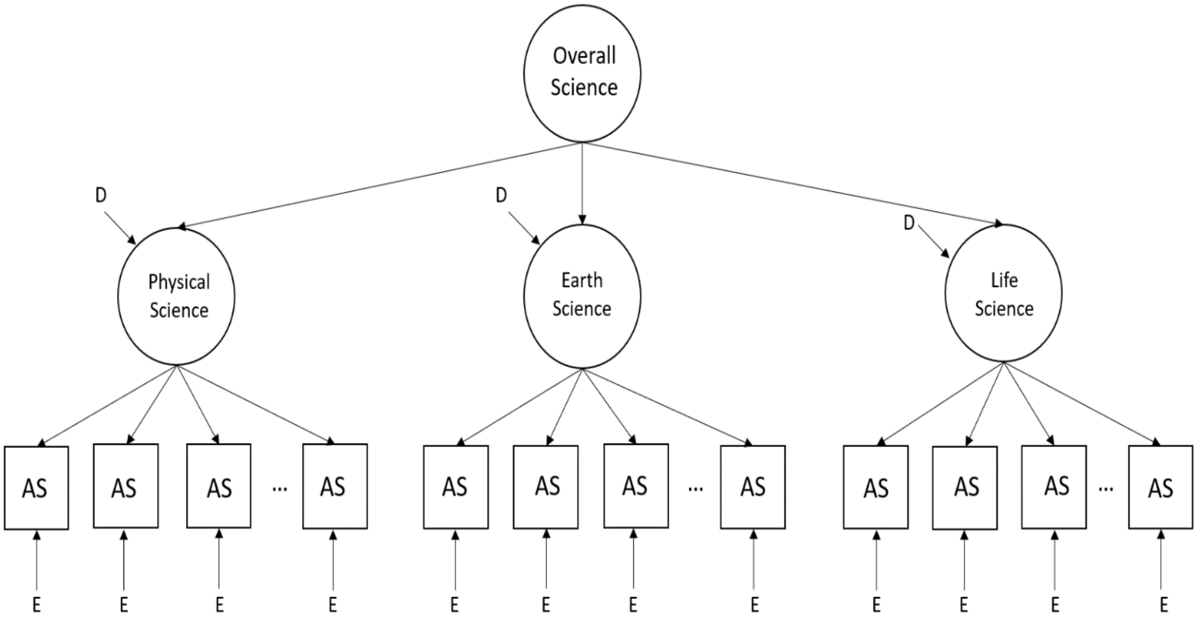


Figure 7. Second-Order Structural Model (Assertions-Clusters-Overall): "Model 3"

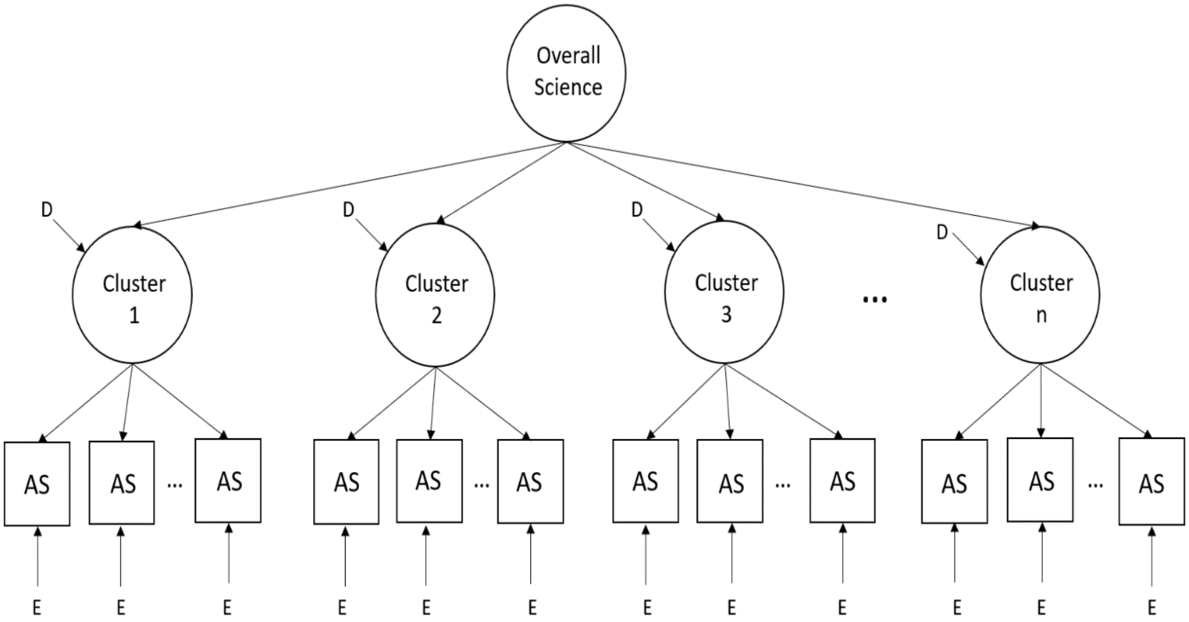
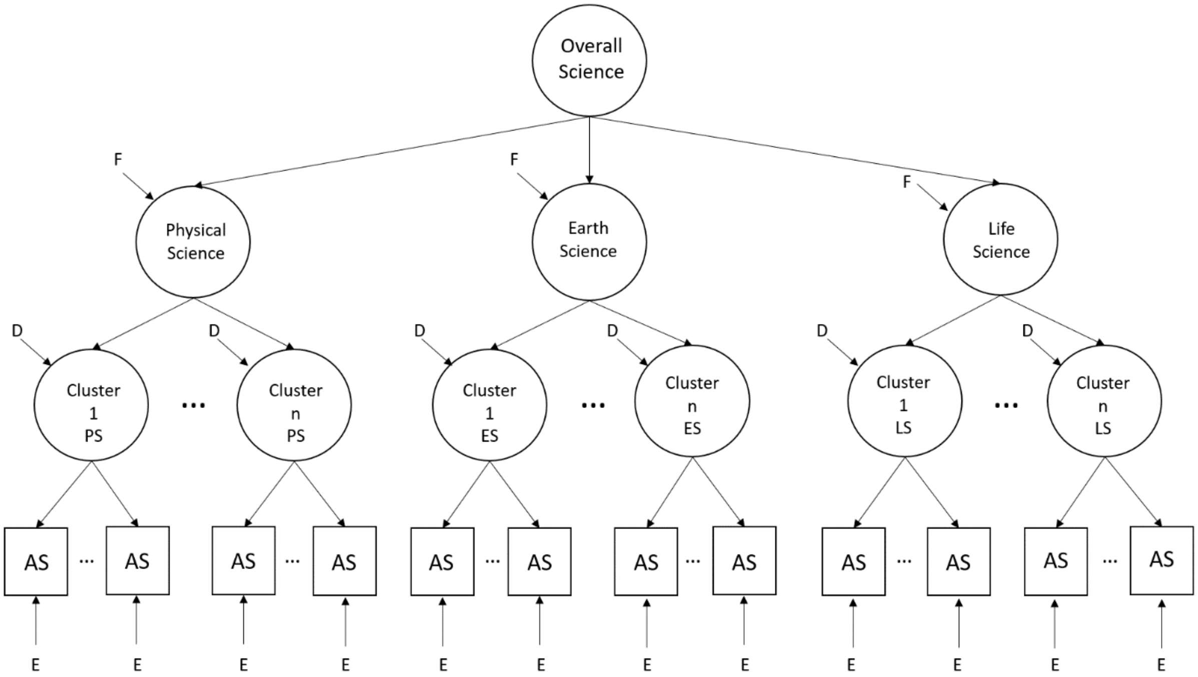


Figure 8. Third-Order Structural Model (Assertions-Clusters-Disciplines-Overall): "Model 4"



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). 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 12 provides a list of these measures along with the corresponding thresholds indicating a good fit.

Table 12. 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 13 through Table 15 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 12. Across all grades and models, the following conclusions can be drawn:

- Model 1 shows the most misfit across grades and forms.
- Across forms, Model 3 generally shows more improvement in model fit relative to Model 1 than Model 2 does (i.e., higher values for CFI and TLI and lower values for 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 does 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. Hence, including the disciplines in the model (when clusters are taken into account) did not improve model fit.
- Overall model fit for Models 3 and 4 decreases with decreasing grades. For grade 8, all fit indices for Models 3 and 4 indicate good model fit for all three forms. For grade 7, all fit indices for Models 3 and 4 indicate good fit for two out of the six forms, and the degree of misfit for the other four forms is small. For grade 6, all three forms have fit indices above the threshold values for at least one of the absolute fit indices for Models 3 and 4.

[1] For very few assertions per form and models, some error variances 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.

The amount of misfit is small for the RMSEA but more substantial for the SRMR for two out of the three forms.

Table 13. 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 14. 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**	0.060	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 (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
	4	0.977	0.976	0.041	0.072
	5	0.975	0.974	0.060	**0.089**

Model	**Form**	**CFI**	**TLI**	**RMSEA**	**SRMR**
	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 15. 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 16 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 16. Fit Measures per Model and Form, Grade 6, with 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 17 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 taking into account the cluster effects, the disciplines do 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 17. Model-Implied Correlations per Form for the Disciplines in Model 4

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

[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	Life Sciences
	5	Physical Sciences	0.931	0.944
		Earth and Space Sciences	–	0.965
	6	Physical Sciences	0.941	0.928
		Earth and Space Sciences	–	0.967
8	1	Physical Sciences	0.971	0.971
		Earth and Space Sciences	–	0.970
	2	Physical Sciences	0.956	0.958
		Earth and Space Sciences	–	0.935
	3	Physical Sciences	0.966	0.978
		Earth and Space Sciences	–	0.988

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. In the review process, adherence to the principles of universal design is verified by Connecticut educators and stakeholders.

6.1 Cognitive Laboratory Studies

In 2017, when the development of item clusters for the states that are part of the MOU began, cognitive lab studies were carried out to evaluate and refine the process of developing item clusters aligned to the NGSS. 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 carried out in 2018 and 2019 to determine if students using braille can understand the task demands of selected accommodated three-dimensional science standards–aligned item clusters and can navigate the interactive features of these clusters in a manner that allows 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 clearly different from (and more complex than) other tests with which the students were familiar, however; and 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

DIF analyses were 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 this review of items for bias is further described in Volume 1, Section 4.4, Annual Technical Report, along with the DIF analysis process for the Connecticut NGSS Assessment.

7. Summary

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

- ***Reliability.*** Various measures of reliability are provided at the aggregate and subgroup levels, showing that the reliability of all tests is in line with acceptable industry standards.
- ***Content validity.*** Evidence is provided to support the assertion that content coverage on each 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 are developed following the principles of universal design, which removes 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 (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (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.

Rijmen, F. (2009). *Three multidimensional models for testlet-based tests: Formal relations and an empirical comparison*. (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.

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 http://education.umn.edu/NCEO/OnlinePubs/Synthesis44.html.

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.