Rhode Island and Vermont Multi-State Science Assessment

2020–2021

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 Multi-State Science Assessment (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 first year of operational test administration in spring 2021.

Table 1. Spring 2021 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 on the basis of 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 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 of the technical report, 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 to statistically monitor analyze test fairness further.

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, he or she 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 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 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 test takers 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

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 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 and 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 was introduced. If any aspect of the technology impedes or creates an advantage for a student in his or her responses to items, this could affect item responses and inferences regarding that student's abilities on the measured construct (refer to Volume 2, 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). 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 test takers have engaged in relevant performance strategies to answer the items correctly supports the validity of the test scores.

The third source of validity evidence is based on *internal structure*: the degree to which the relationships among test items and test components relate to the construct on which the proposed test scores are interpreted. 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 relationship 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. A multi-trait-multi-method matrix can be used to analyze both convergent and discriminant evidence.

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 relationship 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 aspect 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 Multi-State Science Assessment (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://ri.portal.cambiumast.com/-/media/project/client-portals/rhode-island/pdf/ri-ngsa-tam-2020-2021_final.pdf 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 of this technical report, 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 criterion-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 compare or rank 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 2021, 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,

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

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

The marginal reliability of science for the overall sample is reported by grade in Table 2 for both Rhode Island and Vermont, in Table 3 for Rhode Island, and in Table 4 for Vermont. The overall reliability ranged from 0.85 to 0.89, 0.85 to 0.89, and 0.86 to 0.88 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 Volume 1, Annual Technical Report, Section 5, Item Calibration). 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 1PL bifactor model (refer to Volume 1, Annual Technical Report, Section 6.1, Maximum Likelihood Function). 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	14,505	0.89
8	14,052	0.89
11	12,797	0.85

Table 3. Marginal Reliability Coefficients, Rhode Island

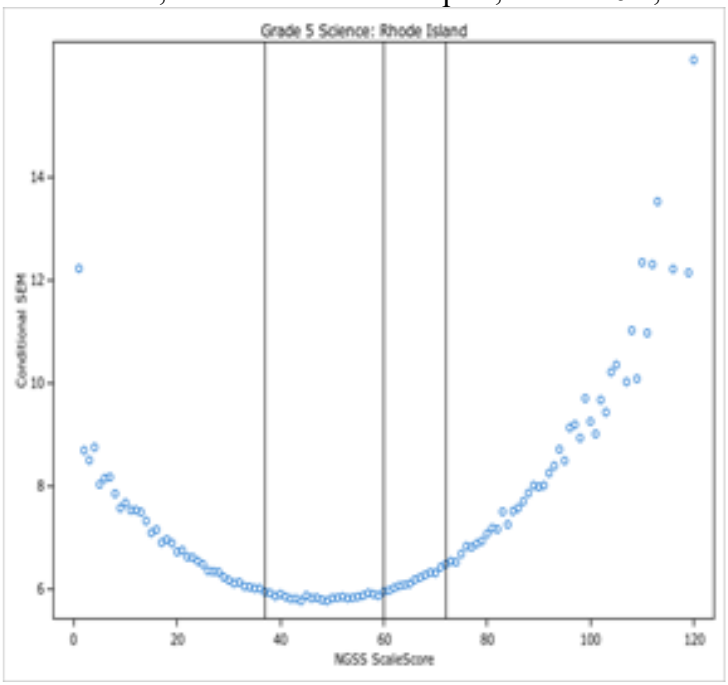
Grade	Sample Size	Reliability
5	9,231	0.89
8	8,715	0.89
11	8,173	0.85

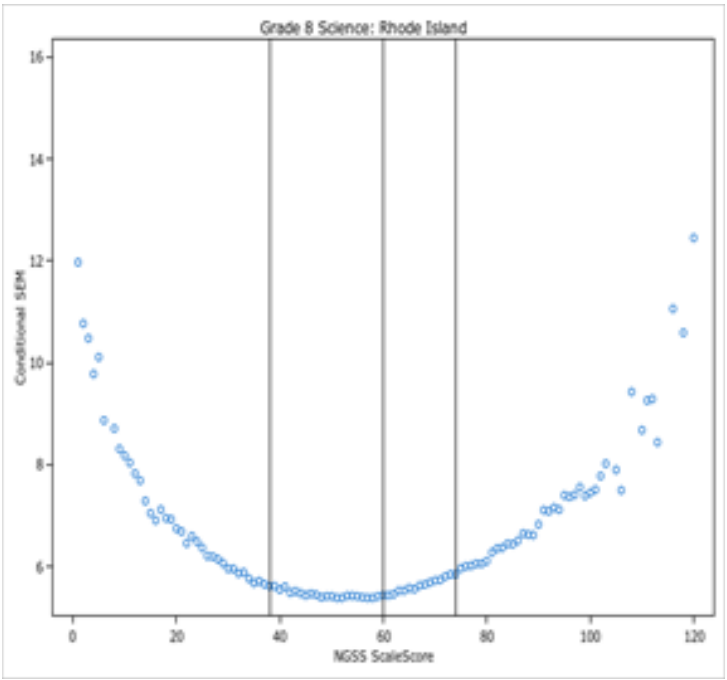
Table 4. Marginal Reliability Coefficients, Vermont

Grade	Sample Size	Reliability
5	5,274	0.88
8	5,337	0.88
11	4,624	0.86

3.1 STANDARD ERROR OF MEASUREMENT

The computation method for conditional standard errors of measurement (CSEM) has been described in Volume 1, Annual Technical Report, Section 6.4, Standard Errors of Estimate. Figure 1 through





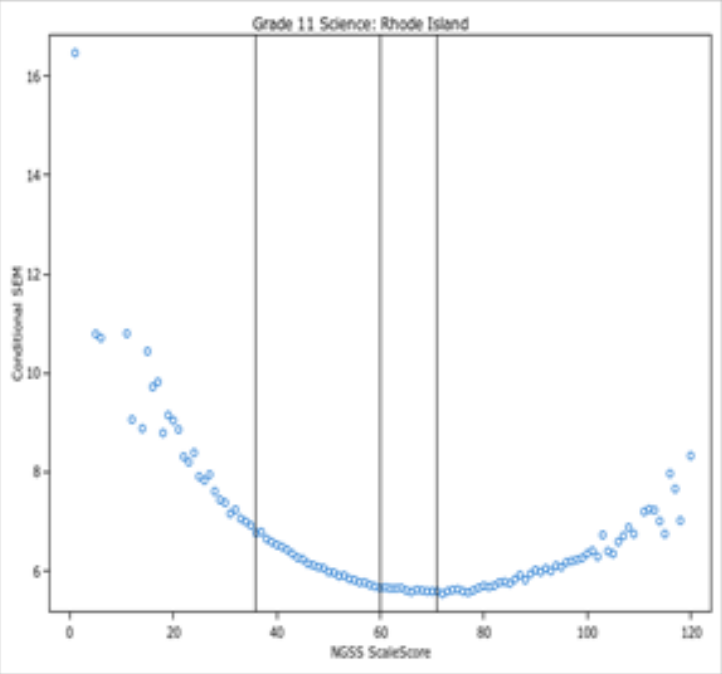
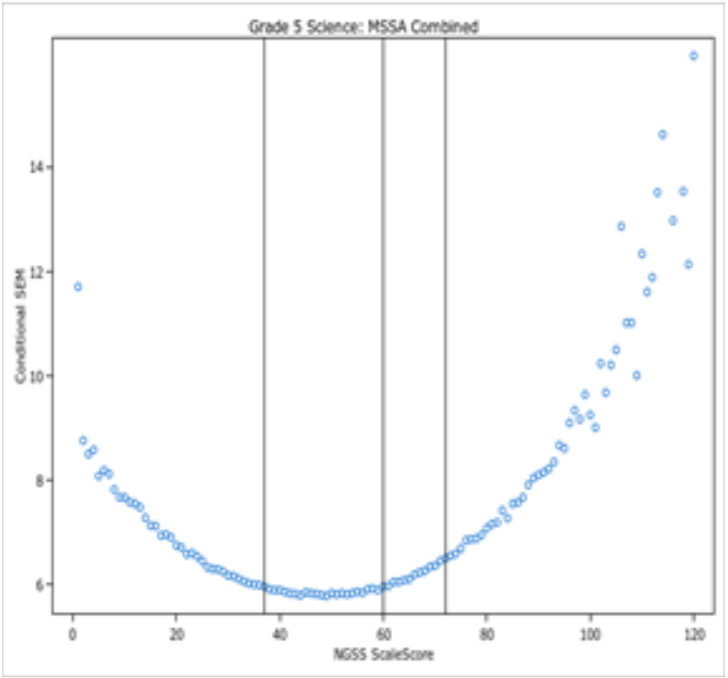
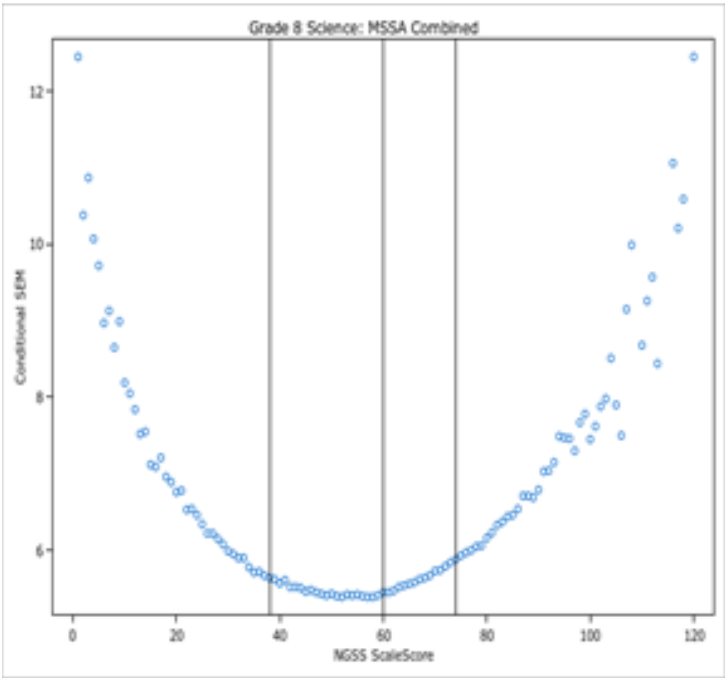

Grade 11 Science: Rhode Island
Conditional SEM
16
14
12
10
8
6
0
20
40
60
80
100
120
NGSS ScaleScore

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





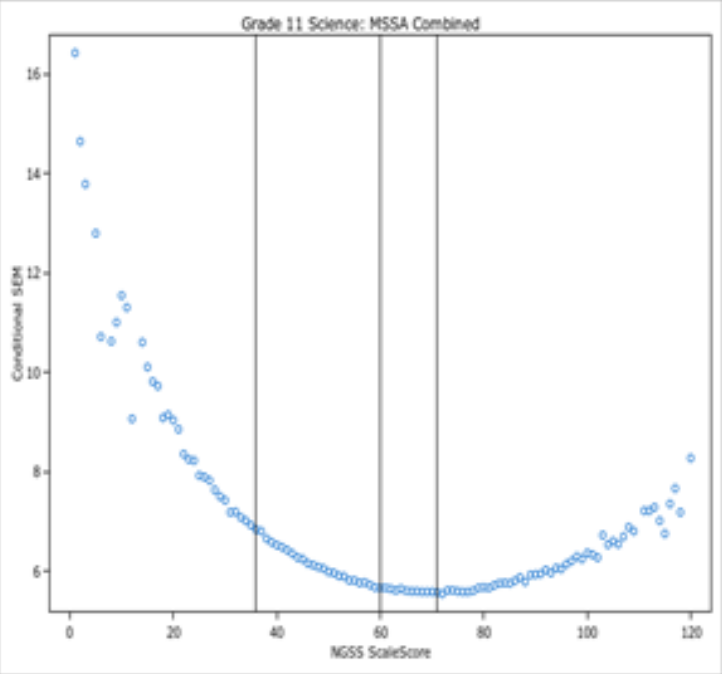
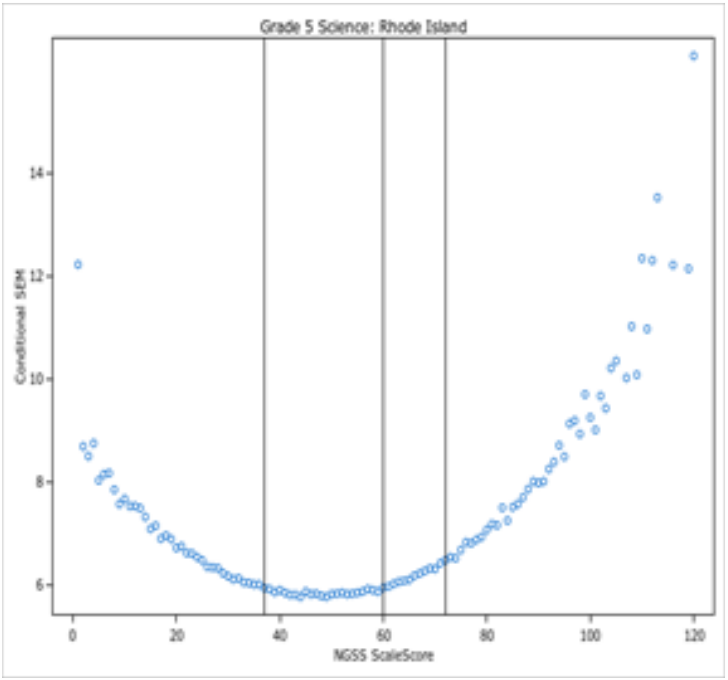
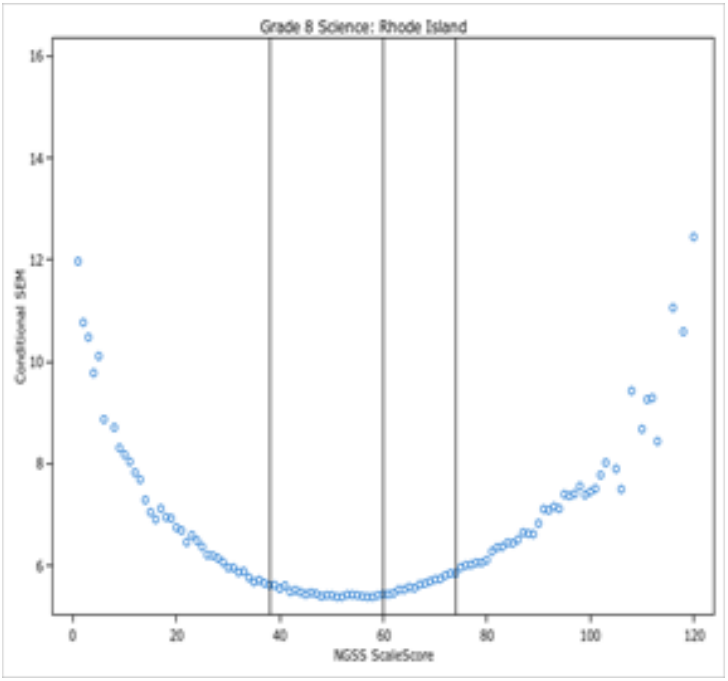

Grade 11 Science: MSSA Combined
Conditional SEM
16
14
12
10
8
6
0
20
40
60
80
100
120
NGSS ScaleScore

Figure 2. Conditional Standard Errors of Measurement, Rhode Island





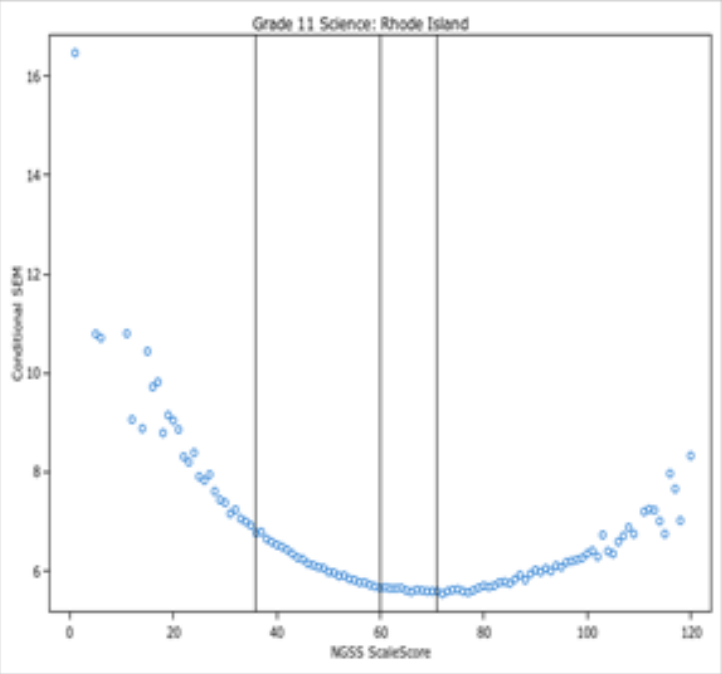
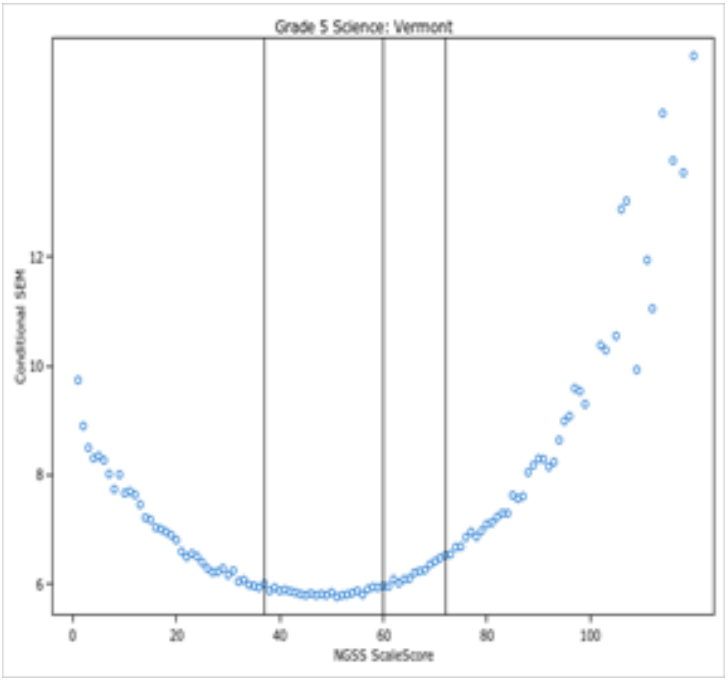
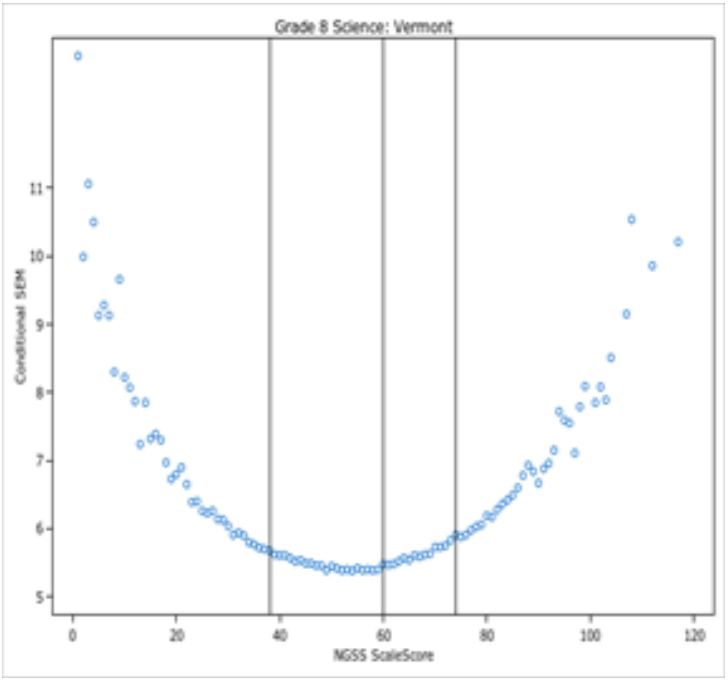
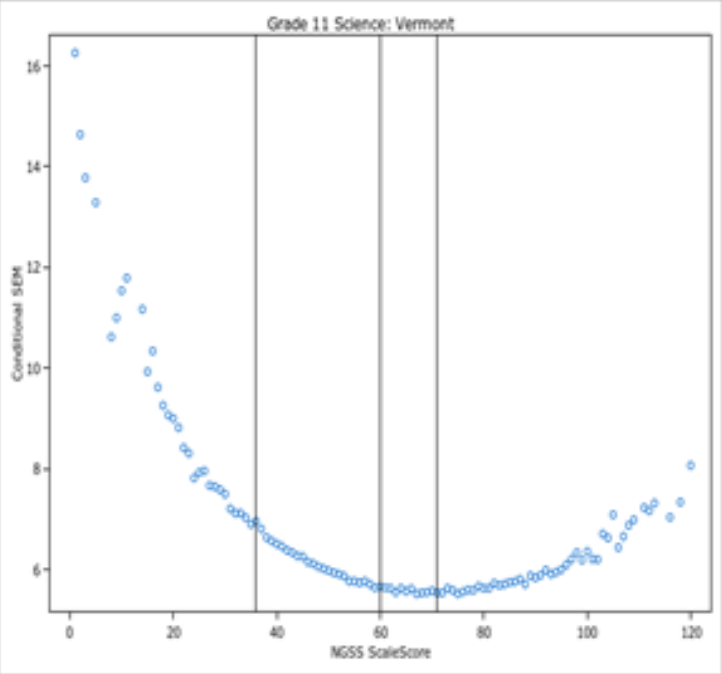

Grade 11 Science: Rhode Island
Conditional SEM
16
14
12
10
8
6
0
20
40
60
80
100
120
NGSS ScaleScore

Figure 3. Conditional Standard Errors of Measurement, Vermont






Grade 11 Science: Vermont
Conditional SEM
16
14
12
10
8
6
0
20
40
60
80
100
120
NGSS ScaleScore

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). Classification accuracy 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. Classification consistency 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, classification accuracy and consistency 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 a standard error of measurement (SEM) of $se(\hat{\theta}_j)$, and the estimated ability is distributed as $\hat{\theta}_j \sim N\left(\theta_j, se^2(\hat{\theta}_j)\right)$, assuming a normal distribution, where θ_j is the unknown true 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 classification accuracy (*CA*) at level l is estimated by

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

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

The classification accuracy for the pth 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 classification accuracy for the individual cut scores. In Rhode Island, the overall classification accuracy of the test ranged from 76.95% to 78.65%. In Vermont, the overall classification accuracy of the test ranged from 76.70% to 78.16%. The individual cut score accuracy rates were high across all grades and states, with the minimum value being 89.73% for grade 5 in Cut Score 2 for Vermont. This denotes that more than 89% of the time, CAI can accurately differentiate students between adjacent achievement levels in the spring 2021 Multi-State Science Assessment (MSSA). The classification accuracy for demographic subgroups is presented in Appendix C, Classification Accuracy and Consistency Index by Subgroups.

Table 5. Classification Accuracy Index, Rhode Island

Grade	Overall Accuracy (%)	Cut Accuracy (%)		
		Cut 1	*Cut 2*	*Cut 3*
5	78.01	91.67	91.64	94.51
8	78.65	91.43	91.70	95.49
11	76.95	92.81	89.86	94.07

Table 6. Classification Accuracy Index, Vermont

Grade	Overall Accuracy (%)	Cut Accuracy (%)		
		Cut 1	*Cut 2*	*Cut 3*
5	77.03	92.83	90.71	93.26
8	78.16	92.78	90.67	94.66
11	76.70	93.74	89.73	93.03

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 jth 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 classification consistency for the cuts. In Rhode Island, the overall classification consistency of the test ranged from 68.42% to 70.43%. In Vermont, the overall classification consistency of the test ranged from 68.21% to 69.73%. The individual cut score consistency rates were high across all grades and states with the minimum value being 85.55% for grade 5 in Cut Score 2 for Vermont. In all achievement levels, classification accuracy was slightly higher than classification consistency. Classification consistency rates can be lower than classification accuracy; the consistency is based on two tests with measurement errors, but the accuracy is based on one test with a measurement error and the true score. The classification consistency for demographic subgroups is presented in Appendix C, Classification Accuracy and Consistency Index by Subgroups.

Table 7. Classification Consistency Index, Rhode Island

Grade	Overall Consistency (%)	Cut Consistency (%)		
		Cut 1	*Cut 2*	*Cut 3*
5	69.97	88.31	88.36	92.28
8	70.43	88.04	88.36	93.64
11	68.42	89.98	85.72	91.58

Table 8. Classification Consistency Index, Vermont

Grade	Overall Consistency (%)	Cut Consistency (%)		
		Cut 1	*Cut 2*	*Cut 3*
5	68.69	89.88	86.98	90.64
8	69.73	89.81	87.02	92.46
11	68.21	91.25	85.55	90.20

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	6.46	-	-
	2	5.85	37	5.96
	3	6.15	60	5.96
	4	7.27	72	6.50
8	1	6.16	-	-
	2	5.46	38	5.64
	3	5.60	60	5.45
	4	6.34	74	5.88
11	1	7.62	-	-
	2	6.11	36	6.83
	3	5.62	60	5.66
	4	5.75	71	5.57

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	6.48	-	-
	2	5.85	37	5.94
	3	6.14	60	5.95
	4	7.28	72	6.48
8	1	6.13	-	-
	2	5.46	38	5.62
	3	5.60	60	5.44
	4	6.37	74	5.85
11	1	7.56	-	-
	2	6.12	36	6.77
	3	5.63	60	5.66
	4	5.77	71	5.59

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	6.41	-	-
	2	5.86	37	6.01
	3	6.17	60	5.97
	4	7.26	72	6.53
8	1	6.21	-	-
	2	5.47	38	5.68
	3	5.59	60	5.48
	4	6.31	74	5.91
11	1	7.74	-	-
	2	6.10	36	6.96
	3	5.59	60	5.66
	4	5.72	71	5.54

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 (ESS)	21	16
	Life Sciences (LS)	25	21
	Physics Sciences (PS)	24	29
8	Earth and Space Sciences (ESS)	15	20
	Life Sciences (LS)	29	23
	Physics Sciences (PS)	18	18
11	Earth and Space Sciences (ESS)	24	25
	Life Sciences (LS)	31	26
	Physics Sciences (PS)	25	23

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 was comprised of two components. The first component addressed the alignment of the 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).

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 0.67 for the combined states, Rhode Island, and Vermont, respectively. The disattenuated correlations ranged from 0.89 to 0.94, 0.92 to 0.95, and 0.86 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 around 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*	0.93	0.92
	LS	0.69	0.73*	0.94
	PS	0.66	0.67	0.69*
8	ESS	0.71*	0.93	0.93
	LS	0.67	0.73*	0.93
	PS	0.67	0.68	0.73*
11	ESS	0.68*	0.89	0.91
	LS	0.62	0.72*	0.90
	PS	0.58	0.59	0.60*

*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*	0.94	0.93
	LS	0.70	0.74*	0.95
	PS	0.67	0.68	0.70*
8	ESS	0.71*	0.95	0.94
	LS	0.68	0.73*	0.93
	PS	0.68	0.68	0.73*
11	ESS	0.66*	0.92	0.92
	LS	0.63	0.72*	0.92
	PS	0.57	0.60	0.59*

*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	0.75*	0.91	0.90
	LS	0.67	0.73*	0.92
	PS	0.64	0.65	0.68*
8	ESS	0.72*	0.91	0.91

Grade	Reporting Category	Earth and Space Sciences (ESS)	Life Sciences (LS)	Physical Sciences (PS)
	LS	0.66	0.73*	0.92
	PS	0.65	0.66	0.72*
11	ESS	0.70*	0.86	0.89
	LS	0.61	0.72*	0.87
	PS	0.59	0.58	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.

5.2 CONVERGENT AND DISCRIMINANT VALIDITY

According to Standard 1.16 of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014), it is necessary to provide evidence of convergent and discriminant validity evidence. 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		
			Cat1	*Cat2*	*Cat3*	*Cat1*	*Cat2*	*Cat3*	*Cat4*	*Cat1*	*Cat2*	*Cat3*
Science	5,184	Earth and Space Sciences (Cat1)	0.75*	0.91	0.90	0.84	0.76	0.82	0.86	0.80	0.92	0.86
		Life Sciences (Cat2)	0.67	0.73*	0.92	0.87	0.77	0.85	0.87	0.79	0.91	0.86
		Physical Sciences (Cat3)	0.64	0.65	0.68*	0.85	0.76	0.84	0.86	0.83	0.94	0.89
ELA		Reading (Cat1)	0.64	0.64	0.61	0.75*	0.89	0.92	0.95	0.80	0.92	0.85
		Writing (Cat2)	0.56	0.56	0.53	0.65	0.73*	0.84	0.89	0.78	0.91	0.84
		Listening (Cat3)	0.57	0.59	0.56	0.64	0.58	0.65*	0.94	0.79	0.91	0.85
		Research (Cat4)	0.65	0.65	0.62	0.72	0.66	0.66	0.76*	0.82	0.94	0.88
Mathematics		Concepts Procedures (Cat1)	0.65	0.63	0.64	0.64	0.62	0.59	0.67	0.87*	1.00	0.96
		Problem Solving, Modeling, and Data Analysis (Cat2)	0.63	0.61	0.61	0.63	0.60	0.58	0.65	0.76	0.61*	1.00
		Communicating Reasoning (Cat3)	0.61	0.60	0.60	0.60	0.59	0.56	0.63	0.74	0.69	0.67*

*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		
			Cat1	*Cat2*	*Cat3*	*Cat1*	*Cat2*	*Cat3*	*Cat4*	*Cat1*	*Cat2*	*Cat3*
Science	5,091	Earth and Space Sciences (Cat1)	0.72*	0.91	0.91	0.84	0.76	0.84	0.80	0.83	0.90	0.85
		Life Sciences (Cat2)	0.66	0.73*	0.91	0.85	0.75	0.85	0.81	0.84	0.92	0.86
		Physical Sciences (Cat3)	0.65	0.66	0.72*	0.82	0.76	0.84	0.80	0.84	0.92	0.84
ELA		Reading (Cat1)	0.62	0.63	0.61	0.76*	0.88	0.93	0.90	0.80	0.90	0.85
		Writing (Cat2)	0.55	0.55	0.55	0.66	0.74*	0.85	0.87	0.79	0.88	0.84
		Listening (Cat3)	0.54	0.55	0.55	0.63	0.56	0.59*	0.89	0.81	0.91	0.86
		Research (Cat4)	0.57	0.58	0.57	0.67	0.63	0.58	0.71*	0.79	0.89	0.83
Mathematics		Concepts Procedures (Cat1)	0.65	0.66	0.66	0.65	0.63	0.58	0.61	0.85*	1.00	0.97
		Problem Solving, Modeling, and Data Analysis (Cat2)	0.60	0.62	0.61	0.62	0.59	0.55	0.59	0.76	0.62*	1.00
		Communicating Reasoning (Cat3)	0.55	0.56	0.54	0.56	0.55	0.51	0.53	0.68	0.65	0.58*

*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 0.78 to 0.80 for Vermont.

Table 18. Correlations Across Spring 2021 ELA, Mathematics, and Science Scores, Vermont

Grade	N	ELA and Mathematics	ELA and Science	Mathematics and Science
5	5,184	0.79	0.80	0.78
8	5,091	0.78	0.78	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, Model Description of Volume 1, Annual Technical Report, presents a detailed description of the IRT model. The internal (latent) structure of the model is presented in Figure 4. 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 (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 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 $\hat{\sigma}^2_{\theta_{pooled}} = 0.69$).

Figure to **Error! Reference source not found.**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 $\eta_g = \frac{\sigma^2_g}{\sigma^2_{\theta_{RI}}+\sigma^2_g}$, $\eta_g = \frac{\sigma^2_g}{\sigma^2_{\theta_{VT}}+\sigma^2_g}$, and $\eta_g = \frac{\sigma^2_g}{\sigma^2_{\theta_{pooled}}+\sigma^2_g}$), 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 into account local dependencies among the assertions of an item cluster.

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

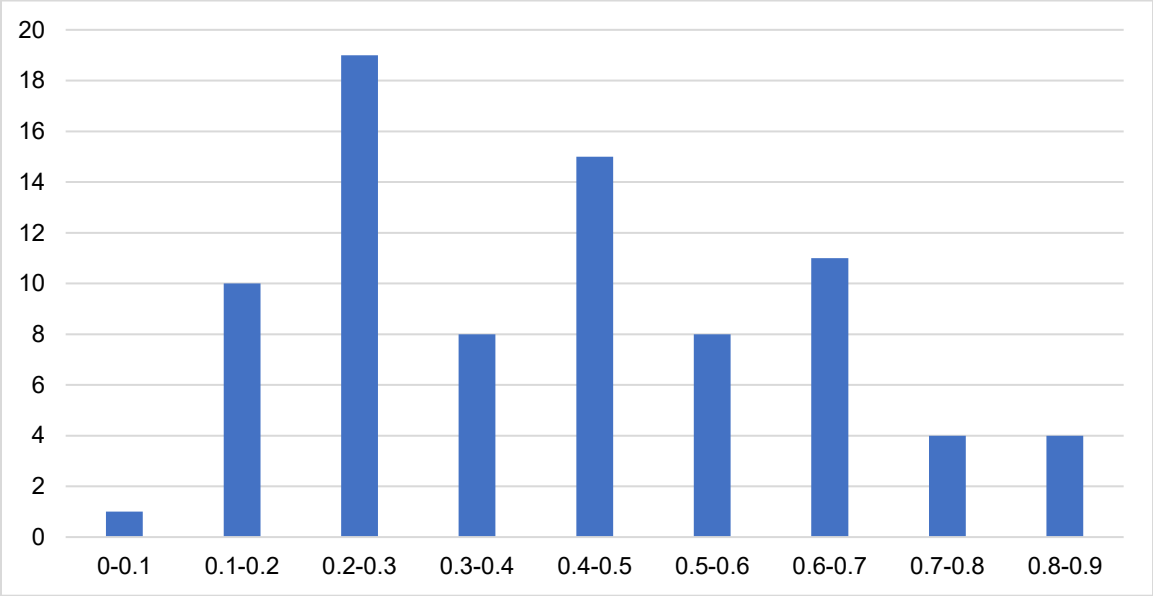


Figure 5. Cluster Variance Proportion for Operational Items in Middle School

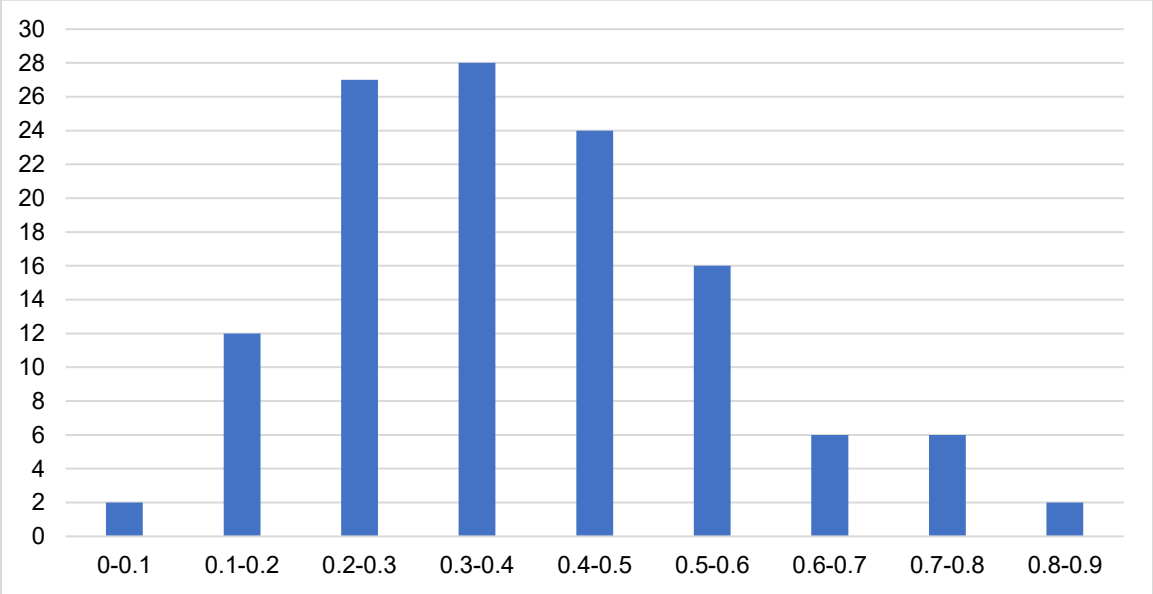
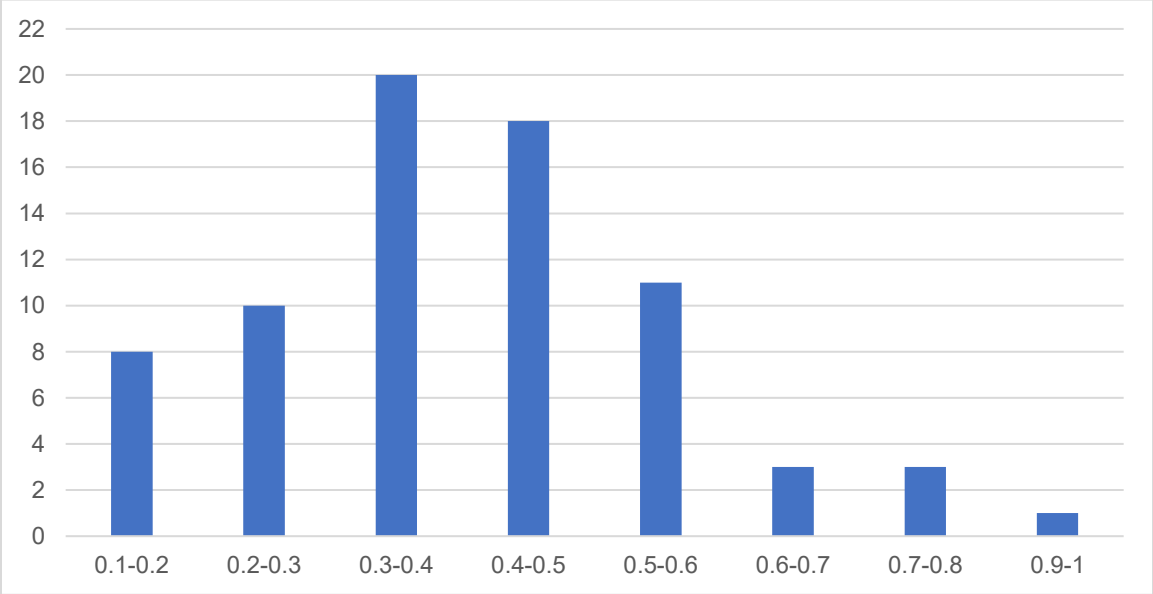


Figure 6. Cluster Variance Proportion for Operational Items in High School



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 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 Multi-State Science Assessment (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 follows:

- 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 7 through Figure 510 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"

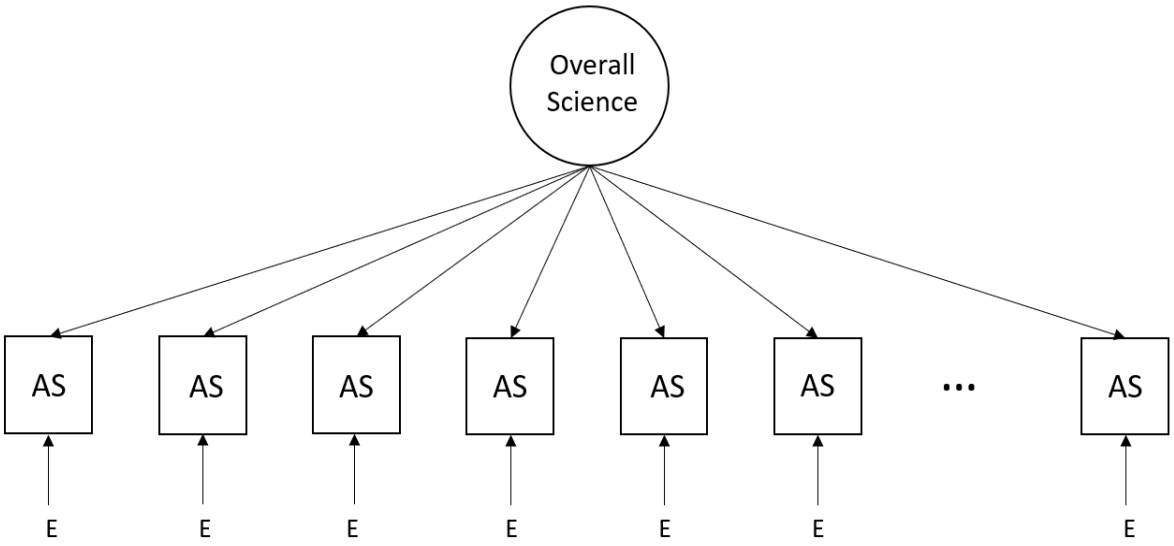


Figure 8. Second-Order Structural Model (Assertions-Disciplines-Overall): "Model 2"

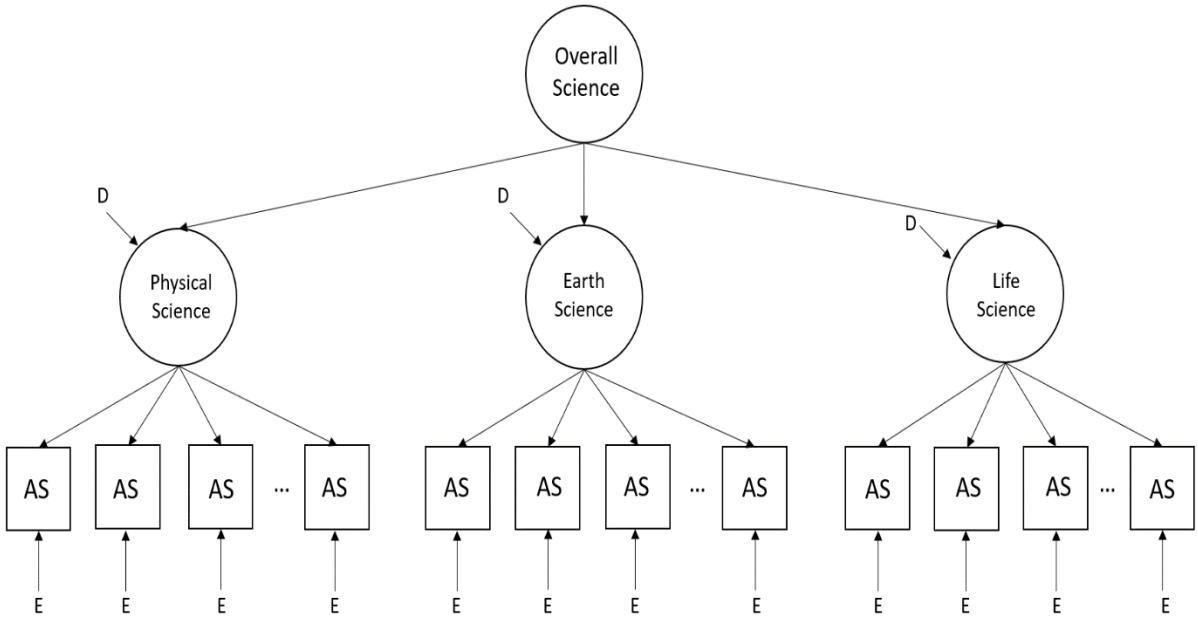


Figure 4. Second-Order Structural Model (Assertions-Clusters-Overall): "Model 3"

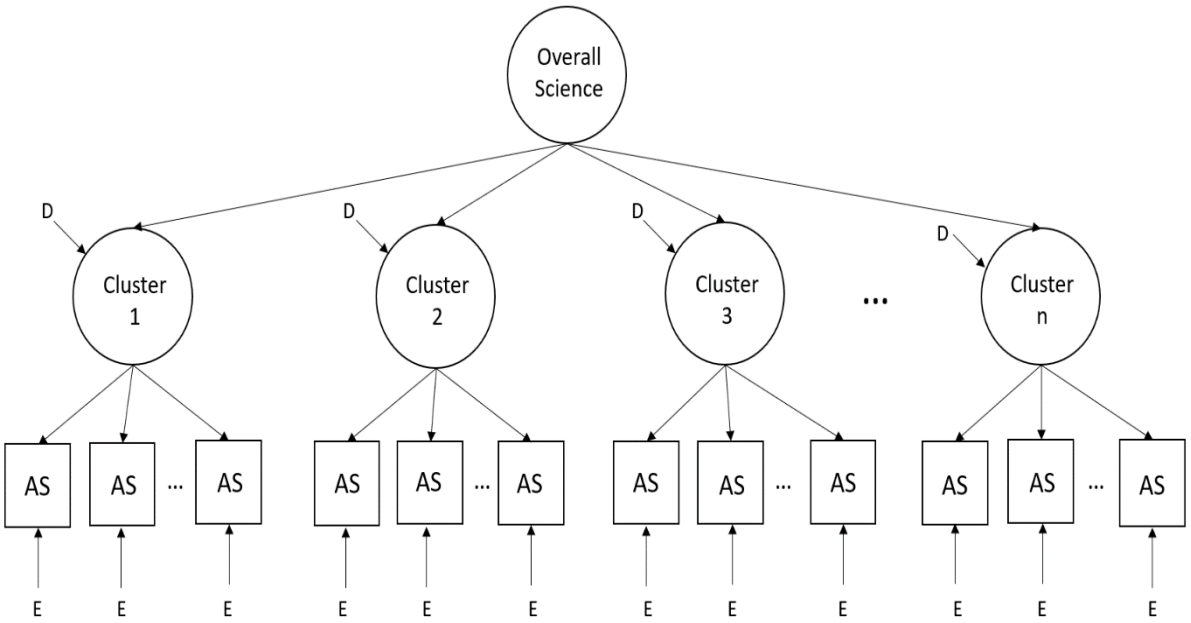
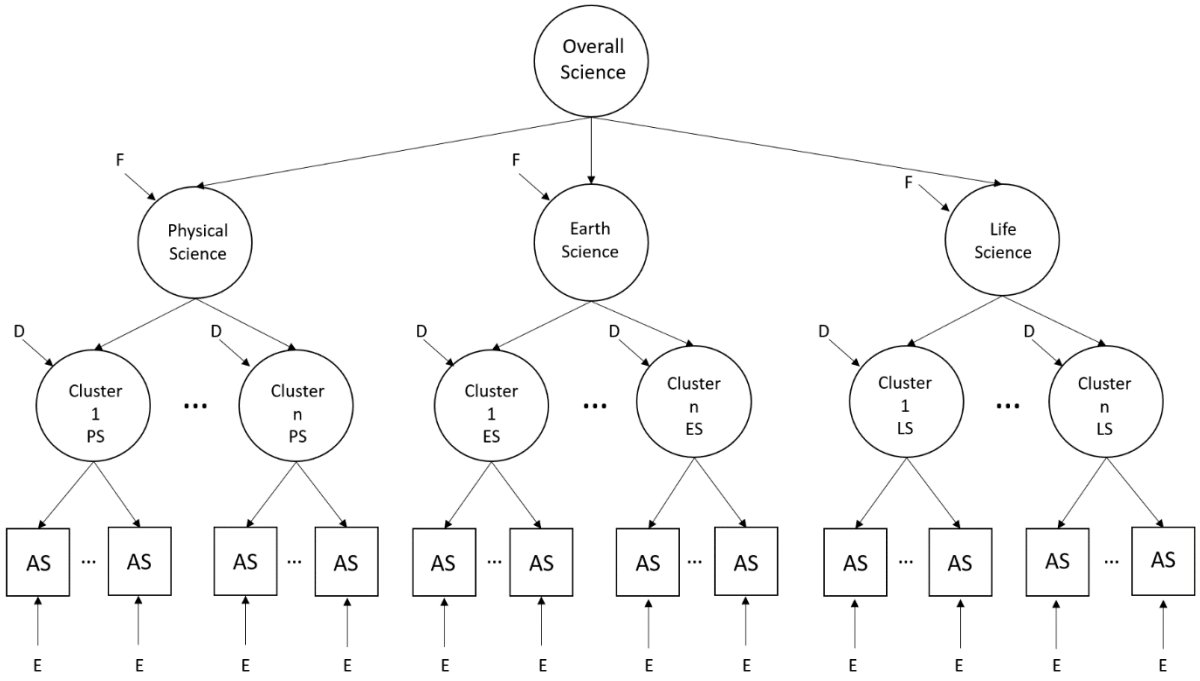


Figure 50. Third-Order Structural Model (Assertions-Clusters-Disciplines-Overall): "Model 4"



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. Hence, including the disciplines into the model (when clusters are taken into account) 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 taking into account 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

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 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 Volume 1, Annual Technical Report, Section 4.4, Differential Item Functioning Analysis.

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 (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, 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.

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.

Appendix A

Student Demographics and Reliability Coefficients

Student Demographics and Reliability Coefficients

Table A-1. Combined Marginal Reliability Coefficients by Demographic Subgroups

Group	Grade 5	Grade 8	Grade 11
All Students	0.89	0.89	0.85
Female	0.88	0.87	0.83
Male	0.89	0.89	0.87
African American	0.85	0.84	0.74
American Indian/Native Alaskan	0.87	0.84	0.72
Asian	0.88	0.90	0.86
Hispanic	0.86	0.85	0.78
Multi-Racial	0.89	0.90	0.86
Pacific Islander	0.88	0.87	0.81
White	0.88	0.88	0.86
Limited English Proficiency	0.76	0.70	0.51
Special Education	0.83	0.80	0.68
Economically Disadvantaged	0.86	0.85	0.78

Table A-2. Rhode Island Marginal Reliability Coefficients by Demographic Subgroups

Group	Grade 5	Grade 8	Grade 11
All Students	0.89	0.89	0.85
Female	0.88	0.88	0.83
Male	0.90	0.89	0.86
African American	0.85	0.84	0.74
American Indian/Native Alaskan	0.85	0.82	0.68
Asian	0.88	0.90	0.86
Hispanic	0.86	0.84	0.76
Multi-Racial	0.89	0.90	0.84
Pacific Islander	0.87	0.84	0.80
White	0.88	0.88	0.85
Limited English Proficiency	0.73	0.67	0.49
Special Education	0.81	0.78	0.67
Economically Disadvantaged	0.85	0.84	0.76

Table A-3. Vermont Marginal Reliability Coefficients by Demographic Subgroups

Group	Grade 5	Grade 8	Grade 11
All Students	0.88	0.88	0.86
Female	0.87	0.87	0.84
Male	0.89	0.89	0.88
African American	0.84	0.87	0.77
American Indian/Native Alaskan	0.90	0.89	0.77
Asian	0.87	0.89	0.86
Hispanic	0.88	0.90	0.87
Multi-Racial	0.89	0.89	0.89
Pacific Islander	0.91	0.88	0.85
White	0.88	0.88	0.86
Limited English Proficiency	0.85	0.84	0.72
Special Education	0.85	0.82	0.70
Economically Disadvantaged	0.87	0.86	0.82

Table A-4. Combined Scale Score Summary by Reporting Category, Grade 5

Reporting Category	Mean	SD	Min	Max	Reliability	SEM
Physical Sciences	50.30	20.26	1.16	119.9	0.69	11.05
Earth and Space Sciences	51.13	23.55	1.16	119.9	0.75	11.44
Life Sciences	50.61	22.69	1.16	119.9	0.73	11.44

Table A-5. Rhode Island Scale Score Summary by Reporting Category, Grade 5

Reporting Category	Mean	SD	Min	Max	Reliability	SEM
Physical Sciences	49.64	20.48	1.16	119.9	0.70	11.07
Earth and Space Sciences	49.85	23.24	1.16	119.9	0.75	11.38
Life Sciences	49.61	22.77	1.16	119.9	0.74	11.43

Table A-6. Vermont Scale Score Summary by Reporting Category, Grade 5

Reporting Category	Mean	SD	Min	Max	Reliability	SEM
Physical Sciences	51.44	19.84	1.16	119.9	0.68	11.02
Earth and Space Sciences	53.35	23.92	1.16	119.9	0.75	11.53
Life Sciences	52.36	22.46	1.16	119.9	0.73	11.46

Table A-7. Combined Scale Score Summary by Reporting Category, Grade 8

Reporting Category	Mean	SD	Min	Max	Reliability	SEM
Physical Sciences	52.50	18.79	1.07	119.91	0.73	9.70
Earth and Space Sciences	51.67	20.17	1.07	119.91	0.71	10.60
Life Sciences	51.43	20.64	1.07	119.91	0.73	10.48

Table A-8. Rhode Island Scale Score Summary by Reporting Category, Grade 8

Reporting Category	Mean	SD	Min	Max	Reliability	SEM
Physical Sciences	51.65	18.90	1.07	119.91	0.73	9.70
Earth and Space Sciences	50.79	19.99	1.07	119.91	0.71	10.60
Life Sciences	50.53	20.66	1.07	119.91	0.73	10.47

Table A-9. Vermont Scale Score Summary by Reporting Category, Grade 8

Reporting Category	Mean	SD	Min	Max	Reliability	SEM
Physical Sciences	53.90	18.51	1.07	119.91	0.72	9.7
Earth and Space Sciences	53.11	20.39	1.07	119.91	0.72	10.6
Life Sciences	52.92	20.51	1.07	119.91	0.73	10.5

Table A-10. Combined Scale Score Summary by Reporting Category, Grade 11

Reporting Category	Mean	SD	Min	Max	Reliability	SEM
Physical Sciences	55.86	16.16	1.16	119.97	0.60	10.04
Earth and Space Sciences	55.94	20.80	1.16	119.97	0.68	11.61
Life Sciences	55.23	20.96	1.16	119.97	0.72	10.93

Table A-11. Rhode Island Scale Score Summary by Reporting Category, Grade 11

Reporting Category	Mean	SD	Min	Max	Reliability	SEM
Physical Sciences	55.32	15.88	1.16	119.97	0.59	10.07
Earth and Space Sciences	54.80	20.19	1.16	119.97	0.66	11.60
Life Sciences	54.68	21.08	1.16	119.97	0.72	11.03

Table A-12. Vermont Scale Score Summary by Reporting Category, Grade 11

Reporting Category	Mean	SD	Min	Max	Reliability	SEM
Physical Sciences	56.82	16.61	1.16	119.97	0.63	9.98
Earth and Space Sciences	57.95	21.68	1.16	119.97	0.70	11.62
Life Sciences	56.20	20.73	1.16	119.97	0.72	10.76

Appendix B

Conditional Standard Error of Measurement

Conditional Standard Error of Measurement

Table B-1. CSEM at Each Scale Score Combined, Science Grade 5

Combined Science Grade 5

Scale Score	Achievement Level	CSEM
1	1	11.71
2	1	8.76
3	1	8.50
4	1	8.58
5	1	8.08
6	1	8.19
7	1	8.12
8	1	7.82
9	1	7.67
10	1	7.67
11	1	7.58
12	1	7.55
13	1	7.48
14	1	7.28
15	1	7.13
16	1	7.12
17	1	6.94
18	1	6.96
19	1	6.90
20	1	6.75
21	1	6.71
22	1	6.58
23	1	6.60
24	1	6.54
25	1	6.45
26	1	6.33
27	1	6.30
28	1	6.29
29	1	6.25
30	1	6.17
31	1	6.16
32	1	6.11

Combined Science Grade 5

Scale Score	Achievement Level	CSEM
33	1	6.06
34	1	6.02
35	1	6.00
36	1	5.99
37	2	5.96
38	2	5.91
39	2	5.89
40	2	5.89
41	2	5.86
42	2	5.83
43	2	5.82
44	2	5.79
45	2	5.85
46	2	5.83
47	2	5.82
48	2	5.80
49	2	5.78
50	2	5.83
51	2	5.81
52	2	5.83
53	2	5.81
54	2	5.83
55	2	5.86
56	2	5.84
57	2	5.91
58	2	5.92
59	2	5.89
60	3	5.96
61	3	5.97
62	3	6.05
63	3	6.05
64	3	6.08
65	3	6.10
66	3	6.19
67	3	6.23
68	3	6.26

Combined Science Grade 5

Scale Score	Achievement Level	CSEM
69	3	6.34
70	3	6.36
71	3	6.45
72	4	6.50
73	4	6.55
74	4	6.59
75	4	6.69
76	4	6.85
77	4	6.87
78	4	6.88
79	4	6.95
80	4	7.08
81	4	7.16
82	4	7.19
83	4	7.42
84	4	7.27
85	4	7.55
86	4	7.57
87	4	7.67
88	4	7.91
89	4	8.04
90	4	8.10
91	4	8.15
92	4	8.22
93	4	8.35
94	4	8.67
95	4	8.61
96	4	9.10
97	4	9.34
98	4	9.17
99	4	9.64
100	4	9.25
101	4	9.01
102	4	10.24
103	4	9.68
104	4	10.21

Combined Science Grade 5

Scale Score	Achievement Level	CSEM
105	4	10.50
106	4	12.87
107	4	11.02
108	4	11.02
109	4	10.01
110	4	12.34
111	4	11.61
112	4	11.89
113	4	13.52
114	4	14.63
116	4	12.98
118	4	13.54
119	4	12.14
120	4	16.14

Table B-2. CSEM at Each Scale Score Rhode Island, Science Grade 5

Rhode Island Science Grade 5

Scale Score	Achievement Level	CSEM
1	1	12.22
2	1	8.69
3	1	8.50
4	1	8.75
5	1	8.03
6	1	8.15
7	1	8.17
8	1	7.85
9	1	7.57
10	1	7.67
11	1	7.53
12	1	7.53
13	1	7.49
14	1	7.32
15	1	7.09
16	1	7.15
17	1	6.90
18	1	6.96
19	1	6.89
20	1	6.72
21	1	6.75
22	1	6.62
23	1	6.61
24	1	6.54
25	1	6.48
26	1	6.35
27	1	6.34
28	1	6.33
29	1	6.23
30	1	6.17
31	1	6.11
32	1	6.13
33	1	6.05
34	1	6.04

Rhode Island Science Grade 5

Scale Score	Achievement Level	CSEM
35	1	6.01
36	1	6.01
37	2	5.94
38	2	5.92
39	2	5.86
40	2	5.90
41	2	5.85
42	2	5.81
43	2	5.81
44	2	5.77
45	2	5.87
46	2	5.82
47	2	5.83
48	2	5.79
49	2	5.77
50	2	5.82
51	2	5.83
52	2	5.85
53	2	5.82
54	2	5.83
55	2	5.85
56	2	5.87
57	2	5.92
58	2	5.90
59	2	5.87
60	3	5.95
61	3	5.97
62	3	6.02
63	3	6.06
64	3	6.08
65	3	6.10
66	3	6.18
67	3	6.22
68	3	6.27
69	3	6.32
70	3	6.31

Rhode Island Science Grade 5

Scale Score	Achievement Level	CSEM
71	3	6.42
72	4	6.48
73	4	6.54
74	4	6.52
75	4	6.68
76	4	6.83
77	4	6.81
78	4	6.88
79	4	6.93
80	4	7.07
81	4	7.18
82	4	7.16
83	4	7.50
84	4	7.25
85	4	7.51
86	4	7.57
87	4	7.70
88	4	7.86
89	4	8.01
90	4	7.98
91	4	8.01
92	4	8.25
93	4	8.39
94	4	8.71
95	4	8.49
96	4	9.13
97	4	9.19
98	4	8.93
99	4	9.70
100	4	9.25
101	4	9.01
102	4	9.67
103	4	9.43
104	4	10.21
105	4	10.35
107	4	10.02

Rhode Island Science Grade 5

Scale Score	Achievement Level	CSEM
108	4	11.02
109	4	10.08
110	4	12.34
111	4	10.97
112	4	12.30
113	4	13.52
116	4	12.21
119	4	12.14
120	4	16.27

Table B-3. CSEM at Each Scale Score Vermont Science Grade 5

Vermont Science Grade 5

Scale Score	Achievement Level	CSEM
1	1	9.74
2	1	8.90
3	1	8.50
4	1	8.31
5	1	8.35
6	1	8.27
7	1	8.02
8	1	7.74
9	1	8.01
10	1	7.67
11	1	7.70
12	1	7.64
13	1	7.46
14	1	7.22
15	1	7.18
16	1	7.04
17	1	7.01
18	1	6.96
19	1	6.90
20	1	6.81
21	1	6.60
22	1	6.50
23	1	6.56
24	1	6.51
25	1	6.40
26	1	6.29
27	1	6.22
28	1	6.23
29	1	6.29
30	1	6.17
31	1	6.25
32	1	6.05
33	1	6.08
34	1	5.99

Vermont Science Grade 5

Scale Score	Achievement Level	CSEM
35	1	5.97
36	1	5.94
37	2	6.01
38	2	5.88
39	2	5.94
40	2	5.88
41	2	5.90
42	2	5.87
43	2	5.85
44	2	5.82
45	2	5.80
46	2	5.83
47	2	5.80
48	2	5.82
49	2	5.80
50	2	5.85
51	2	5.77
52	2	5.80
53	2	5.81
54	2	5.84
55	2	5.88
56	2	5.81
57	2	5.91
58	2	5.95
59	2	5.94
60	3	5.97
61	3	5.96
62	3	6.09
63	3	6.02
64	3	6.09
65	3	6.10
66	3	6.21
67	3	6.24
68	3	6.26
69	3	6.36
70	3	6.42

Vermont Science Grade 5

Scale Score	Achievement Level	CSEM
71	3	6.48
72	4	6.53
73	4	6.55
74	4	6.68
75	4	6.69
76	4	6.86
77	4	6.96
78	4	6.87
79	4	6.98
80	4	7.10
81	4	7.13
82	4	7.23
83	4	7.30
84	4	7.30
85	4	7.63
86	4	7.57
87	4	7.61
88	4	8.05
89	4	8.18
90	4	8.30
91	4	8.29
92	4	8.15
93	4	8.24
94	4	8.64
95	4	8.99
96	4	9.08
97	4	9.59
98	4	9.54
99	4	9.30
102	4	10.38
103	4	10.29
105	4	10.55
106	4	12.87
107	4	13.02
109	4	9.93
111	4	11.94

Vermont Science Grade 5

Scale Score	Achievement Level	CSEM
112	4	11.05
114	4	14.63
116	4	13.76
118	4	13.54
120	4	15.68

Table B-4. CSEM at Each Scale Score Combined, Science Grade 8

Combined Science Grade 8

Scale Score	Achievement Level	CSEM
1	1	12.45
2	1	10.38
3	1	10.87
4	1	10.07
5	1	9.72
6	1	8.97
7	1	9.13
8	1	8.65
9	1	8.99
10	1	8.19
11	1	8.05
12	1	7.84
13	1	7.52
14	1	7.55
15	1	7.12
16	1	7.09
17	1	7.21
18	1	6.96
19	1	6.89
20	1	6.76
21	1	6.78
22	1	6.53
23	1	6.54
24	1	6.46
25	1	6.34
26	1	6.22
27	1	6.22
28	1	6.15
29	1	6.08
30	1	5.99
31	1	5.95
32	1	5.90
33	1	5.90
34	1	5.78

Combined Science Grade 8

Scale Score	Achievement Level	CSEM
35	1	5.71
36	1	5.72
37	1	5.67
38	2	5.64
39	2	5.62
40	2	5.57
41	2	5.61
42	2	5.52
43	2	5.52
44	2	5.51
45	2	5.46
46	2	5.48
47	2	5.45
48	2	5.43
49	2	5.41
50	2	5.43
51	2	5.40
52	2	5.39
53	2	5.42
54	2	5.41
55	2	5.42
56	2	5.40
57	2	5.39
58	2	5.39
59	2	5.41
60	3	5.45
61	3	5.45
62	3	5.47
63	3	5.52
64	3	5.54
65	3	5.56
66	3	5.58
67	3	5.62
68	3	5.64
69	3	5.67
70	3	5.73

Combined Science Grade 8

Scale Score	Achievement Level	CSEM
71	3	5.74
72	3	5.79
73	3	5.84
74	4	5.88
75	4	5.93
76	4	5.97
77	4	6.00
78	4	6.05
79	4	6.06
80	4	6.16
81	4	6.23
82	4	6.33
83	4	6.37
84	4	6.44
85	4	6.46
86	4	6.54
87	4	6.71
88	4	6.71
89	4	6.69
90	4	6.79
91	4	7.03
92	4	7.04
93	4	7.15
94	4	7.49
95	4	7.47
96	4	7.46
97	4	7.30
98	4	7.67
99	4	7.78
100	4	7.45
101	4	7.62
102	4	7.88
103	4	7.98
104	4	8.51
105	4	7.90
106	4	7.50

Combined Science Grade 8

Scale Score	Achievement Level	CSEM
107	4	9.15
108	4	9.99
110	4	8.68
111	4	9.26
112	4	9.57
113	4	8.44
116	4	11.06
117	4	10.21
118	4	10.59
120	4	12.45

Table B-5. CSEM at Each Scale Score Rhode Island, Science Grade 8

Rhode Island Science Grade 8

Scale Score	Achievement Level	CSEM
1	1	11.97
2	1	10.77
3	1	10.48
4	1	9.78
5	1	10.11
6	1	8.87
8	1	8.71
9	1	8.31
10	1	8.18
11	1	8.04
12	1	7.83
13	1	7.69
14	1	7.29
15	1	7.05
16	1	6.91
17	1	7.12
18	1	6.95
19	1	6.93
20	1	6.75
21	1	6.69
22	1	6.46
23	1	6.60
24	1	6.49
25	1	6.37
26	1	6.21
27	1	6.20
28	1	6.15
29	1	6.07
30	1	5.96
31	1	5.96
32	1	5.87
33	1	5.89
34	1	5.77
35	1	5.68

Rhode Island Science Grade 8

Scale Score	Achievement Level	CSEM
36	1	5.72
37	1	5.66
38	2	5.62
39	2	5.62
40	2	5.55
41	2	5.61
42	2	5.49
43	2	5.52
44	2	5.48
45	2	5.44
46	2	5.47
47	2	5.45
48	2	5.40
49	2	5.42
50	2	5.42
51	2	5.39
52	2	5.39
53	2	5.43
54	2	5.43
55	2	5.42
56	2	5.40
57	2	5.39
58	2	5.39
59	2	5.42
60	3	5.44
61	3	5.44
62	3	5.46
63	3	5.53
64	3	5.53
65	3	5.58
66	3	5.56
67	3	5.63
68	3	5.65
69	3	5.69
70	3	5.73
71	3	5.74

Rhode Island Science Grade 8

Scale Score	Achievement Level	CSEM
72	3	5.81
73	3	5.85
74	4	5.85
75	4	5.97
76	4	6.01
77	4	6.02
78	4	6.06
79	4	6.06
80	4	6.11
81	4	6.28
82	4	6.36
83	4	6.37
84	4	6.45
85	4	6.44
86	4	6.51
87	4	6.65
88	4	6.63
89	4	6.62
90	4	6.83
91	4	7.11
92	4	7.09
93	4	7.16
94	4	7.12
95	4	7.40
96	4	7.37
97	4	7.41
98	4	7.56
99	4	7.39
100	4	7.45
101	4	7.51
102	4	7.78
103	4	8.02
105	4	7.90
106	4	7.50
108	4	9.43
110	4	8.68

Rhode Island Science Grade 8

Scale Score	Achievement Level	CSEM
111	4	9.26
112	4	9.29
113	4	8.44
116	4	11.06
118	4	10.59
120	4	12.45

Table B-6. CSEM at Each Scale Score Vermont, Science Grade 8

Vermont Science Grade 8

Scale Score	Achievement Level	CSEM
1	1	12.94
2	1	9.99
3	1	11.06
4	1	10.50
5	1	9.13
6	1	9.28
7	1	9.13
8	1	8.30
9	1	9.66
10	1	8.22
11	1	8.07
12	1	7.87
13	1	7.24
14	1	7.85
15	1	7.32
16	1	7.39
17	1	7.30
18	1	6.97
19	1	6.73
20	1	6.80
21	1	6.90
22	1	6.65
23	1	6.39
24	1	6.40
25	1	6.26
26	1	6.23
27	1	6.26
28	1	6.14
29	1	6.13
30	1	6.04
31	1	5.91
32	1	5.94
33	1	5.90
34	1	5.80

Vermont Science Grade 8

Scale Score	Achievement Level	CSEM
35	1	5.77
36	1	5.72
37	1	5.70
38	2	5.68
39	2	5.62
40	2	5.61
41	2	5.61
42	2	5.57
43	2	5.52
44	2	5.54
45	2	5.49
46	2	5.49
47	2	5.46
48	2	5.46
49	2	5.39
50	2	5.45
51	2	5.42
52	2	5.39
53	2	5.40
54	2	5.38
55	2	5.42
56	2	5.39
57	2	5.40
58	2	5.39
59	2	5.40
60	3	5.48
61	3	5.47
62	3	5.48
63	3	5.52
64	3	5.57
65	3	5.54
66	3	5.61
67	3	5.59
68	3	5.62
69	3	5.63
70	3	5.73

Vermont Science Grade 8

Scale Score	Achievement Level	CSEM
71	3	5.73
72	3	5.75
73	3	5.83
74	4	5.91
75	4	5.88
76	4	5.91
77	4	5.98
78	4	6.03
79	4	6.06
80	4	6.19
81	4	6.17
82	4	6.28
83	4	6.36
84	4	6.42
85	4	6.49
86	4	6.60
87	4	6.78
88	4	6.93
89	4	6.84
90	4	6.67
91	4	6.88
92	4	6.96
93	4	7.15
94	4	7.72
95	4	7.59
96	4	7.55
97	4	7.11
98	4	7.79
99	4	8.09
101	4	7.85
102	4	8.08
103	4	7.89
104	4	8.51
107	4	9.15
108	4	10.54
112	4	9.86

Vermont Science Grade 8

Scale Score	Achievement Level	CSEM
117	4	10.21

Table B-7. CSEM at Each Scale Score Combined, Science Grade 11

Combined Science Grade 11

Scale Score	Achievement Level	CSEM
1	1	16.42
2	1	14.64
3	1	13.78
5	1	12.79
6	1	10.71
8	1	10.62
9	1	11.00
10	1	11.54
11	1	11.30
12	1	9.06
14	1	10.60
15	1	10.10
16	1	9.81
17	1	9.72
18	1	9.08
19	1	9.14
20	1	9.03
21	1	8.85
22	1	8.35
23	1	8.24
24	1	8.22
25	1	7.92
26	1	7.89
27	1	7.82
28	1	7.63
29	1	7.50
30	1	7.42
31	1	7.18
32	1	7.19
33	1	7.08
34	1	7.01
35	1	6.92
36	2	6.83
37	2	6.80

Combined Science Grade 11

Scale Score	Achievement Level	CSEM
38	2	6.65
39	2	6.58
40	2	6.52
41	2	6.48
42	2	6.42
43	2	6.35
44	2	6.27
45	2	6.24
46	2	6.15
47	2	6.12
48	2	6.08
49	2	6.05
50	2	5.98
51	2	5.96
52	2	5.90
53	2	5.89
54	2	5.81
55	2	5.81
56	2	5.76
57	2	5.77
58	2	5.72
59	2	5.67
60	3	5.66
61	3	5.66
62	3	5.64
63	3	5.61
64	3	5.64
65	3	5.60
66	3	5.59
67	3	5.59
68	3	5.58
69	3	5.58
70	3	5.58
71	4	5.57
72	4	5.54
73	4	5.61

Combined Science Grade 11

Scale Score	Achievement Level	CSEM
74	4	5.61
75	4	5.59
76	4	5.58
77	4	5.58
78	4	5.60
79	4	5.66
80	4	5.67
81	4	5.66
82	4	5.71
83	4	5.75
84	4	5.76
85	4	5.75
86	4	5.80
87	4	5.87
88	4	5.78
89	4	5.92
90	4	5.93
91	4	5.94
92	4	6.02
93	4	5.96
94	4	6.06
95	4	6.04
96	4	6.14
97	4	6.20
98	4	6.29
99	4	6.24
100	4	6.36
101	4	6.33
102	4	6.27
103	4	6.72
104	4	6.53
105	4	6.60
106	4	6.54
107	4	6.69
108	4	6.88
109	4	6.80

Combined Science Grade 11

Scale Score	Achievement Level	CSEM
111	4	7.21
112	4	7.21
113	4	7.28
114	4	7.01
115	4	6.75
116	4	7.35
117	4	7.66
118	4	7.18
120	4	8.27

Table B-8. CSEM at Each Scale Score Rhode Island, Science Grade 11

Combined Science Grade 11

Scale Score	Achievement Level	CSEM
1	1	16.47
5	1	10.79
6	1	10.71
11	1	10.80
12	1	9.06
14	1	8.88
15	1	10.44
16	1	9.72
17	1	9.82
18	1	8.79
19	1	9.15
20	1	9.04
21	1	8.86
22	1	8.31
23	1	8.20
24	1	8.39
25	1	7.91
26	1	7.83
27	1	7.95
28	1	7.61
29	1	7.44
30	1	7.38
31	1	7.16
32	1	7.24
33	1	7.06
34	1	7.00
35	1	6.92
36	2	6.77
37	2	6.79
38	2	6.65
39	2	6.59
40	2	6.53
41	2	6.49
42	2	6.43

Combined Science Grade 11

Scale Score	Achievement Level	CSEM
43	2	6.35
44	2	6.27
45	2	6.24
46	2	6.15
47	2	6.12
48	2	6.08
49	2	6.06
50	2	5.98
51	2	5.97
52	2	5.90
53	2	5.91
54	2	5.84
55	2	5.82
56	2	5.77
57	2	5.77
58	2	5.72
59	2	5.69
60	3	5.66
61	3	5.67
62	3	5.65
63	3	5.65
64	3	5.66
65	3	5.61
66	3	5.58
67	3	5.62
68	3	5.61
69	3	5.59
70	3	5.59
71	4	5.59
72	4	5.54
73	4	5.59
74	4	5.62
75	4	5.63
76	4	5.59
77	4	5.57
78	4	5.61

Combined Science Grade 11

Scale Score	Achievement Level	CSEM
79	4	5.66
80	4	5.70
81	4	5.68
82	4	5.70
83	4	5.77
84	4	5.78
85	4	5.75
86	4	5.83
87	4	5.92
88	4	5.82
89	4	5.94
90	4	6.02
91	4	5.98
92	4	6.05
93	4	6.00
94	4	6.11
95	4	6.08
96	4	6.18
97	4	6.20
98	4	6.23
99	4	6.26
100	4	6.35
101	4	6.41
102	4	6.29
103	4	6.73
104	4	6.40
105	4	6.35
106	4	6.59
107	4	6.71
108	4	6.88
109	4	6.75
111	4	7.20
112	4	7.25
113	4	7.23
114	4	7.01
115	4	6.75

Combined Science Grade 11

Scale Score	Achievement Level	CSEM
116	4	7.97
117	4	7.66
118	4	7.02
120	4	8.33

Table B-9. CSEM at Each Scale Score Vermont, Science Grade 11

Vermont Science Grade 11

Scale Score	Achievement Level	CSEM
1	1	16.26
2	1	14.64
3	1	13.78
5	1	13.29
8	1	10.62
9	1	11.00
10	1	11.54
11	1	11.79
14	1	11.17
15	1	9.93
16	1	10.34
17	1	9.62
18	1	9.26
19	1	9.07
20	1	9.00
21	1	8.82
22	1	8.42
23	1	8.31
24	1	7.82
25	1	7.93
26	1	7.96
27	1	7.67
28	1	7.65
29	1	7.59
30	1	7.50
31	1	7.21
32	1	7.12
33	1	7.12
34	1	7.03
35	1	6.90
36	2	6.96
37	2	6.81
38	2	6.64
39	2	6.57

Vermont Science Grade 11

Scale Score	Achievement Level	CSEM
40	2	6.51
41	2	6.46
42	2	6.38
43	2	6.34
44	2	6.27
45	2	6.26
46	2	6.15
47	2	6.12
48	2	6.06
49	2	6.02
50	2	5.98
51	2	5.94
52	2	5.91
53	2	5.87
54	2	5.77
55	2	5.77
56	2	5.74
57	2	5.77
58	2	5.71
59	2	5.64
60	3	5.66
61	3	5.64
62	3	5.63
63	3	5.55
64	3	5.63
65	3	5.57
66	3	5.62
67	3	5.52
68	3	5.54
69	3	5.55
70	3	5.58
71	4	5.54
72	4	5.54
73	4	5.63
74	4	5.59
75	4	5.52

Vermont Science Grade 11

Scale Score	Achievement Level	CSEM
76	4	5.56
77	4	5.60
78	4	5.59
79	4	5.67
80	4	5.63
81	4	5.63
82	4	5.73
83	4	5.69
84	4	5.71
85	4	5.75
86	4	5.76
87	4	5.81
88	4	5.71
89	4	5.88
90	4	5.84
91	4	5.89
92	4	5.99
93	4	5.91
94	4	5.95
95	4	6.00
96	4	6.10
97	4	6.21
98	4	6.34
99	4	6.19
100	4	6.36
101	4	6.21
102	4	6.20
103	4	6.71
104	4	6.63
105	4	7.09
106	4	6.44
107	4	6.66
108	4	6.88
109	4	6.99
111	4	7.23
112	4	7.17

Vermont Science Grade 11

Scale Score	Achievement Level	CSEM
113	4	7.32
116	4	7.04
118	4	7.34
120	4	8.07

Appendix C

Classification Accuracy and Consistency Index by Subgroups

Classification Accuracy and Consistency Index by Subgroups

Table C-1. Classification Accuracy by Demographic Subgroup, Rhode Island

Group	N	Overall (%)	By Cut (%)			By Level (%)			
			Cut 1	*Cut 2*	*Cut 3*	*L1*	*L2*	*L3*	*L4*
Grade 5									
All Students	9140	78.01	91.67	91.64	94.51	86.37	79.85	59.67	80.1
Female	4473	77.63	91.38	91.32	94.76	85.64	79.81	59.73	80.45
Male	4653	78.37	91.98	91.94	94.26	87.03	79.91	59.61	79.83
African American	743	78.73	92.11	92.18	94.24	88.35	80.04	58.66	79.85
American Indian/Alaskan Native	65	83.51	87.1	96.96	99.42	87.43	80	57.71	91.61
Asian	313	77.45	94.34	90.51	92.37	87.78	80.69	58.57	80.75
Hispanic	2527	80.68	88.69	94.27	97.61	87.38	79.17	59.84	77.04
Multi-Racial	457	80.6	88.03	94.88	97.6	87	78.5	59.62	75.38
Pacific Islander	18	76.17	93.57	89.79	92.59	84.36	80.34	59.76	80.58
White	5017	78.83	89.45	94.01	95.25	90.5	77.44	60.16	62.82
Limited English Proficiency (LEP)	1111	84.37	86.37	98.2	99.79	88.11	78.25	58.12	72.39
Non-LEP	8029	77.13	92.4	90.74	93.78	85.64	80.02	59.69	80.11
Special Education (SPED)	1369	84.67	88.32	97.33	98.99	89.98	77.53	57.93	77.52
Non-SPED	7771	76.83	92.26	90.64	93.73	84.38	80.13	59.74	80.15
Economically Disadvantaged	4104	80.45	88.57	94.22	97.57	87.1	79.22	59.44	75.74
Non-Economically Disadvantaged	5036	76.02	94.2	89.54	92.03	84.55	80.38	59.75	80.78
Grade 8									
All Students	8566	78.65	91.43	91.7	95.49	84.55	80.45	67.47	80.61
Female	4117	78.53	91.66	91.26	95.58	83.3	80.98	67.72	79.7
Male	4437	78.73	91.21	92.09	95.4	85.42	79.88	67.24	81.39
African American	772	80.71	87.47	94.61	98.62	86.22	79.18	66.35	72.06

Group	N	Overall (%)	By Cut (%)			By Level (%)			
			Cut 1	*Cut 2*	*Cut 3*	*L1*	*L2*	*L3*	*L4*
American Indian/Alaskan Native	62	83.13	88.1	95.98	99.04	88.67	79.74	67.69	-
Asian	234	80.38	93.61	91.97	94.76	83.44	81.67	68.9	88.26
Hispanic	2273	80.02	87.57	94.1	98.32	84.69	79.61	67.29	77.75
Multi-Racial	409	79.33	91.35	92.3	95.64	84.65	80.47	67.57	84.41
Pacific Islander	15	80.44	90.02	92.64	97.78	80.05	83.99	74.62	-
White	4801	77.47	93.84	89.97	93.62	83.42	81.03	67.48	80.31
Limited English Proficiency (LEP)	808	83.33	84.85	98.61	99.86	87.01	76.16	65.61	57.84
Non-LEP	7758	78.17	92.11	90.98	95.04	83.62	80.74	67.48	80.64
Special Education (SPED)	1182	81.85	85.19	97.4	99.25	85.97	78.12	65.96	79.58
Non-SPED	7384	78.14	92.43	90.79	94.89	83.86	80.76	67.52	80.63
Economically Disadvantaged	3423	80.51	87.83	94.44	98.23	85.29	80.09	66.96	76.18
Non-Economically Disadvantaged	5143	77.41	93.82	89.87	93.67	83.07	80.71	67.61	81.13
Grade 11									
All Students	7996	76.95	92.81	89.86	94.07	73.17	81.25	59.93	84.4
Female	3926	76.7	93.59	89.04	93.86	72.85	81.6	59.82	83.08
Male	4057	77.19	92.08	90.65	94.27	73.4	80.88	60.07	85.48
African American	672	77.78	87.7	92.55	97.4	74.03	81.29	60.45	79.3
American Indian/Alaskan Native	36	78.3	88.25	92.8	97.03	64.48	81.66	60.92	66.06
Asian	270	76.28	96.21	88.37	91.43	75.61	80.92	59.96	83.37
Hispanic	1833	77.46	88.01	92.28	97.01	73.68	80.83	60.07	80.35
Multi-Racial	270	77.83	92.41	90.75	94.5	74.15	82.78	61.37	85.66
Pacific Islander	18	74.92	91.26	89.19	94.18	NA	78.43	58.06	77.7
White	4897	76.63	95.19	88.6	92.61	72.07	81.4	59.81	84.96
Limited English Proficiency (LEP)	494	77.41	80.2	97.53	99.63	74.68	78.77	53.48	77.74
Non-LEP	7502	76.92	93.65	89.36	93.7	72.76	81.47	59.96	84.43

Group	N	Overall (%)	By Cut (%)			By Level (%)			
			Cut 1	*Cut 2*	*Cut 3*	*L1*	*L2*	*L3*	*L4*
Special Education (SPED)	921	78.98	84.02	95.88	99.02	73.8	81.43	58.27	85.7
Non-SPED	7075	76.69	93.96	89.08	93.42	72.86	81.22	59.97	84.38
Economically Disadvantaged	2703	77.33	88.48	91.93	96.77	73.29	81.21	60.13	80.09
Non-Economically Disadvantaged	5293	76.75	95.03	88.81	92.69	73.01	81.28	59.88	84.92

Table C-2. Classification Accuracy by Demographic Subgroup, Vermont

Group	N	Overall (%)	By Cut (%)			By Level (%)			
			Cut 1	*Cut 2*	*Cut 3*	*L1*	*L2*	*L3*	*L4*
Grade 5									
All Students	5214	77.03	92.83	90.71	93.26	86.43	80.08	59.61	78.36
Female	2435	76.88	92.49	90.59	93.58	84.79	80.39	59.27	77.93
Male	2602	77.14	93.01	90.79	93.12	87.44	79.71	59.9	78.56
African American	135	83.27	89.92	94.75	98.57	88.02	83.04	59.04	88.31
American Indian/Alaskan Native	18	80.64	91.62	92.93	95.73	83.71	78.48	-	81.68
Asian	124	75.46	91.52	89.16	94.53	84.66	77.58	57.65	76.06
Hispanic	122	78.38	93.24	91.84	93.08	84.01	81.68	62.42	80.87
Multi-Racial	168	78.93	93.46	91.18	94.11	86.22	82.03	59.61	85.09
Pacific Islander	6	77.15	88.13	94.52	94.5	99.36	64.98	67.11	-
White	4641	76.77	92.93	90.57	93.04	86.44	79.97	59.58	78.06
Limited English Proficiency (LEP)	123	79.26	89.48	93.1	96.51	84.33	81.13	60.94	68.49
Non-LEP	5091	76.98	92.91	90.65	93.18	86.52	80.05	59.6	78.46
Special Education (SPED)	981	83.05	88.93	95.95	98.1	89.14	79.42	59.23	77.88
Non-SPED	4233	75.64	93.74	89.49	92.14	84.2	80.2	59.64	78.38
Economically Disadvantaged	1324	79.66	90.81	92.57	96.14	86.97	80.8	59.58	76.47
Non-Economically Disadvantaged	3890	76.14	93.52	90.08	92.28	86.06	79.82	59.62	78.59
Grade 8									
All Students	5216	78.16	92.78	90.67	94.66	83.85	80.98	67.61	79.21
Female	2433	77.65	93.01	89.72	94.88	82.78	81.14	67.34	78.41
Male	2584	78.6	92.48	91.58	94.5	84.52	80.84	67.91	79.71
African American	148	81.78	89.39	95.01	97.35	86.62	81.47	66.75	73.61
American Indian/Alaskan Native	12	80.35	87.14	94.54	98.67	84.38	79.73	64.1	90.55

Group	N	Overall (%)	By Cut (%)			By Level (%)			
			Cut 1	*Cut 2*	*Cut 3*	*L1*	*L2*	*L3*	*L4*
Asian	108	75.34	92.04	90.31	92.94	78.7	76.97	67.83	81.58
Hispanic	130	78.17	92.78	91.34	94.01	85.57	78.83	67.67	82.34
Multi-Racial	160	80.07	92.93	90.86	96.24	87.18	81.43	68.29	82.88
Pacific Islander	36	80.69	95.58	88.15	96.91	87.34	84.83	62.41	74.95
White	4622	78.02	92.9	90.53	94.55	83.52	81.06	67.63	79.01
Limited English Proficiency (LEP)	74	80.26	84.66	96.08	99.52	83.55	78.47	65.1	99.99
Non-LEP	5142	78.13	92.9	90.6	94.59	83.86	81.02	67.62	79.17
Special Education (SPED)	897	82.37	86.55	96.84	98.96	86.08	79.81	67.23	79.51
Non-SPED	4319	77.29	94.08	89.39	93.77	81.81	81.19	67.63	79.2
Economically Disadvantaged	1144	79.38	90.27	91.98	97.1	84.73	80.47	67.4	78.43
Non-Economically Disadvantaged	4072	77.82	93.49	90.31	93.98	83.42	81.14	67.65	79.29
Grade 11									
All Students	4509	76.7	93.74	89.73	93.03	74.43	81.36	60	83.89
Female	2028	76.46	94.64	88.54	93.06	72.48	81.87	60.29	82.9
Male	2288	77	93.04	90.81	92.95	75.83	80.91	59.69	84.58
African American	95	79.57	88.46	93.82	97.19	73.09	83.29	63.88	76.37
American Indian/Alaskan Native	12	73.66	92.63	83.9	96.78	70.14	74.59	65.41	86.25
Asian	112	76.48	96.07	89.81	90.41	80.62	82	60.75	84.3
Hispanic	119	75.95	94.08	88.82	92.85	82.92	80.86	59.96	83.79
Multi-Racial	107	76.58	92.68	90.44	93.29	67.57	80.41	61.54	88.28
Pacific Islander	5	80.6	98.68	85.12	96.8	NA	94.18	57.7	99.27
White	4059	76.66	93.81	89.66	92.99	74.29	81.33	59.88	83.79
Limited English Proficiency (LEP)	31	80.34	92.33	88.7	99.18	76.57	85.07	56.89	-
Non-LEP	4478	76.67	93.75	89.73	92.99	74.4	81.32	60.02	83.89
Special Education (SPED)	546	78.3	85.12	94.18	98.91	77.29	79.76	58.06	85.93

Group	N	Overall (%)	By Cut (%)			By Level (%)			
			Cut 1	*Cut 2*	*Cut 3*	*L1*	*L2*	*L3*	*L4*
Non-SPED	3963	76.47	94.92	89.11	92.22	73.28	81.71	60.07	83.85
Economically Disadvantaged	721	76.49	91.5	89.77	95.03	73.89	81.36	59.74	81.56
Non-Economically Disadvantaged	3788	76.74	94.16	89.72	92.65	74.59	81.36	60.05	84.1

Table C-3. Classification Consistency by Demographic Subgroup, Rhode Island

Group	N	Overall (%)	By Cut (%)			By Level (%)			
			Cut 1	*Cut 2*	*Cut 3*	*L1*	*L2*	*L3*	*L4*
Grade 5									
All Students	9140	69.97	88.31	88.36	92.28	79.13	72.21	48.82	70.59
Female	4473	69.45	87.94	87.91	92.56	77.4	72.66	49.09	69.15
Male	4653	70.46	88.7	88.77	92	80.66	71.76	48.57	71.77
African American	743	73.31	83.35	92.78	96.69	81.92	70.48	46.89	63.9
American Indian/Alaskan Native	65	76.66	81.74	95.78	98.97	81.65	73.17	38.39	88.34
Asian	313	69.11	91.87	86.7	89.37	77.08	74.22	45.99	74.44
Hispanic	2527	73.28	84.21	91.95	96.57	81.68	71.93	47.13	63.13
Multi-Racial	457	71.13	88.73	89.16	92.24	81.04	72.82	48.33	71.48
Pacific Islander	18	71.03	84.95	91.81	93.46	82.1	69.38	52.71	47.1
White	5017	67.66	90.94	85.82	89.57	74.7	72.42	49.54	71.44
Limited English Proficiency (LEP)	1111	77.97	80.84	97.37	99.66	84.96	67.77	37.91	41.2
Non-LEP	8029	68.86	89.34	87.11	91.26	76.9	72.69	49	70.67
Economically Disadvantaged	1369	78.64	83.63	96.22	98.56	87.02	67.24	44.55	62.08
Non-Economically Disadvantaged	7771	68.44	89.13	86.98	91.18	75.12	72.82	48.99	70.78
Grade 8									
All Students	8566	70.43	88.04	88.36	93.64	75.8	73.6	57.29	69.63
Female	4117	70.18	88.33	87.71	93.75	73.48	74.72	56.77	69.07
Male	4437	70.62	87.77	88.94	93.51	77.46	72.43	57.79	70.06
African American	772	73.5	82.89	92.44	97.99	79.9	72.76	53.18	57.3
American Indian/Alaskan Native	62	75.87	82.63	94.47	98.61	83.63	69.12	62.95	27.39
Asian	234	71.95	90.65	88.39	92.48	70.47	75.75	57.23	80.85
Hispanic	2273	72.11	82.71	91.63	97.55	77.53	72.58	55.16	61.28

Group	N	Overall (%)	By Cut (%)			By Level (%)			
			Cut 1	*Cut 2*	*Cut 3*	*L1*	*L2*	*L3*	*L4*
Multi-Racial	409	71.18	87.87	89.17	93.77	77.06	72.6	57.64	74.93
Pacific Islander	15	71.87	86.6	88.7	96.09	74.23	74.88	68.5	12.01
White	4801	68.93	91.36	86	91.06	71.62	74.27	57.92	69.75
Limited English Proficiency (LEP)	808	76.61	78.77	98.01	99.78	83.78	64.14	47.96	32.33
Non-LEP	7758	69.79	89.01	87.35	93	73.14	74.29	57.39	69.68
Economically Disadvantaged	1182	74.81	79.42	96.36	98.92	81.18	68.93	54.69	61.55
Non-Economically Disadvantaged	7384	69.73	89.42	87.08	92.79	73.43	74.24	57.39	69.78
Grade 11									
All Students	7996	68.42	89.98	85.72	91.58	52.96	76.91	47.7	75.66
Female	3926	68.03	90.95	84.56	91.3	50.09	77.44	47.7	72.84
Male	4057	68.79	89.05	86.83	91.84	55.05	76.36	47.7	78.02
African American	672	69.02	82.84	89.29	96.2	55.89	78.21	45.56	59.49
American Indian/Alaskan Native	36	69.69	83.9	89.36	95.97	46.22	80.95	32.95	57.61
Asian	270	67.82	94.64	83.53	88.2	49.57	75.29	48.12	77.42
Hispanic	1833	68.91	83.54	88.93	95.71	55.63	78.06	44.87	65.54
Multi-Racial	270	69.22	89.34	86.74	91.96	56.43	77.51	50.54	74.13
Pacific Islander	18	66.67	87.74	85.88	91.94	29.8	77.14	46.48	73.24
White	4897	68.14	93.2	84.07	89.53	49.03	76.1	48.33	77.13
Limited English Proficiency (LEP)	494	69.12	73.19	96.33	99.45	58.6	76.21	29.13	70.44
Non-LEP	7502	68.37	91.08	85.02	91.06	51.66	76.96	47.87	75.68
Economically Disadvantaged	921	70.79	78.23	93.72	98.53	58.23	78.63	36.09	74.78
Non-Economically Disadvantaged	7075	68.11	91.5	84.68	90.68	50.9	76.62	48.1	75.68

Table C-4. Classification Consistency by Demographic Subgroup, Vermont

Group	N	Overall (%)	By Cut (%)			By Level (%)			
			Cut 1	*Cut 2*	*Cut 3*	*L1*	*L2*	*L3*	*L4*
Grade 5									
All Students	5214	68.69	89.88	86.98	90.64	78.13	72.59	48.59	69.26
Female	2435	68.41	89.37	86.85	91.05	75.7	73.56	47.55	68.74
Male	2602	68.88	90.17	87.04	90.44	79.87	71.71	49.33	69.26
African American	135	76.22	85.42	92.56	97.82	83.8	75.16	46.27	57.58
American Indian/Alaskan Native	18	72.73	87.73	90.42	94.09	77.36	75.22	31.31	83.9
Asian	124	67.19	88.4	85.33	92.44	75.96	71.23	45.39	67.91
Hispanic	122	69.31	90.19	88.06	89.85	76.3	73.8	48.96	72.67
Multi-Racial	168	71.04	90.67	87.46	91.76	80.74	74.07	49.13	73.43
Pacific Islander	6	69.21	85.72	90.85	90.82	84.06	57.18	67.15	16.54
White	4641	68.39	90.03	86.79	90.35	77.79	72.47	48.67	69.05
Limited English Proficiency (LEP)	123	71.11	84.97	90.25	95.12	79.86	72.07	47.52	58.61
Non-LEP	5091	68.63	90	86.9	90.53	78.06	72.61	48.6	69.36
Economically Disadvantaged	981	76.3	84.36	94.18	97.33	85.03	70.8	45.85	66.72
Non-Economically Disadvantaged	4233	66.92	91.16	85.31	89.09	73.03	72.95	48.79	69.4
Grade 8									
All Students	5216	69.73	89.81	87.02	92.46	74.44	74.25	56.97	69.34
Female	2433	69.12	90.16	85.76	92.75	71.22	74.82	56.98	67.67
Male	2584	70.26	89.38	88.19	92.27	76.18	73.93	57.11	70.09
African American	148	74.46	84.93	92.85	96.47	80.83	75.57	52.06	61.42
American Indian/Alaskan Native	12	72.34	80.97	93.61	97.55	79.85	61.72	63.32	84.85
Asian	108	66.62	89.38	86.34	90.43	68.77	70.16	57.97	69.77
Hispanic	130	69.65	89.55	88.21	91.43	79.32	69.76	58.68	70.32

Group	N	Overall (%)	By Cut (%)			By Level (%)			
			Cut 1	*Cut 2*	*Cut 3*	*L1*	*L2*	*L3*	*L4*
Multi-Racial	160	71.64	89.78	86.89	94.6	76.21	75.57	55.74	74.87
Pacific Islander	36	73.28	93.26	83.7	96.08	86.6	75.65	47.89	77.32
White	4622	69.56	89.98	86.83	92.3	73.73	74.36	57.05	69.15
Limited English Proficiency (LEP)	74	71.8	78.14	94.34	99.21	76.78	69.23	51.19	78.46
Non-LEP	5142	69.7	89.98	86.91	92.37	74.36	74.32	56.99	69.32
Economically Disadvantaged	897	75.08	80.98	95.47	98.5	81.36	71.1	52.2	66.28
Non-Economically Disadvantaged	4319	68.62	91.64	85.26	91.21	69	74.83	57.19	69.44
Grade 11									
All Students	4509	68.21	91.25	85.55	90.2	53.79	76.19	48.51	76.32
Female	2028	67.78	92.45	83.9	90.13	51.63	76.67	48.59	73.82
Male	2288	68.73	90.32	87.04	90.25	55.35	75.87	48.2	78.39
African American	95	70.55	83.42	90.5	95.89	50.93	80.35	49	59.15
American Indian/Alaskan Native	12	63.06	89.3	78.22	94.25	50.11	71.18	47.38	68.98
Asian	112	68.02	94.54	85.25	86.74	59	74.69	50.91	76.3
Hispanic	119	66.95	91.25	84.28	90.01	57.72	73.54	50.78	75.29
Multi-Racial	107	68.01	90	86.35	90.42	46.16	75.5	51.7	77.22
Pacific Islander	5	74.63	97.46	81.23	94.26	3.99	79.44	56.13	87.47
White	4059	68.21	91.37	85.49	90.15	53.9	76.2	48.25	76.41
Limited English Proficiency (LEP)	31	72.06	88.69	84.23	98.46	61.07	80.99	41.3	5.71
Non-LEP	4478	68.18	91.26	85.56	90.15	53.71	76.15	48.54	76.34
Economically Disadvantaged	546	70.04	79.72	91.61	98.3	56.62	78.24	38.87	74.5
Non-Economically Disadvantaged	3963	67.96	92.83	84.72	89.09	52.59	75.79	48.95	76.35

Appendix D

Science Clusters Cognitive Lab Report

Science Cluster Cognitive Interviews

Fran Stancavage

Susan Cole

March 2018

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

Cambium Assessment, Inc (CAI) and a group of states are developing methods to measure student learning of Next Generation Science Standards (NGSS) and other standards derived from the K–12 science framework. Educators involved in the development of the framework and the standards encourage measuring learning using integrated tasks that require a student's sustained concentration on a realistic science or engineering task. This set of cognitive interviews was undertaken early in the development process to test and refine our approach to developing item clusters to measure NGSS and related performance expectations (PEs).

The approach taken for each cluster was to identify a *phenomenon* to be explained, modeled, described, or analyzed (as appropriate for the performance expectation) and have a sequence of interrelated, often interdependent items (some containing multiple interactions) that build to support the completion of a task.

This set of cognitive interviews was designed to provide data on newly developed item clusters aligned with the NGSS. We evaluated 12 clusters, four designed for elementary school, four designed for middle school, and four designed for high school. Each cluster contained one to five items, many with separately scored sub-items. Per the request of the item development team, the labs focused on the following questions:

- How long did students take to respond to each cluster?
- How well did students score on each item and on each cluster overall?
- What aspects of the items were confusing to students?
- What reasoning skills did students display as they worked their way through each item?

A limitation of the cognitive lab analysis was that many of the students had limited exposure to content covered in the clusters, particularly the clusters on German Pyramid Candle (elementary school), Morning Fog (middle school), Texas Weather (middle school), Saving the Tuna (high school), and Tomcods (high school). To partially offset this lack of formal instruction, students were provided with a one- or two-page hard-copy lesson on the relevant science content for each cluster. Some of the later cognitive interviews were conducted in schools in which the teachers had received substantial training in teaching the new standards.

The remainder of this report includes an overview of methods, a description of the study sample, a discussion of the findings for each of the 12 clusters, and a final section on the students' overall perceptions of the science clusters.

2. METHODS

2.1 STUDY DESIGN

Between January and May 2017, cognitive interviews were conducted with 18 elementary school students, 12 middle school students, and 15 high school students. The interviews lasted one and one-half hours, and each student was presented with all four clusters for their grade level. The order of the clusters was rotated so that the risk of student fatigue or missing responses was distributed across the clusters.

Students were encouraged to think out loud while they were responding to the items (concurrent think-aloud), and interviewers were instructed to use follow-up probes to clarify and expand on what each student said (or what each student was observed to do). To preclude the possibility that students' responses to later items would be influenced by probing on earlier items, probes were only administered after students had completed all the items in a cluster.

At the start of the interview, the interviewer trained the student on the concurrent think-aloud technique. The interviewer first modeled the technique and then had the student practice on one or, if necessary, two items. Lower grade multiple-choice mathematics items were used for the modeling and practice.

After the think-aloud training, students were provided with a hard-copy lesson on the relevant science content, as described previously. The item development team developed the lessons, and the interviewer collected the hard copy before the student stared the cluster.

At the end of the cognitive interview, each student was asked three general questions: (1) whether the student had studied any of the cluster topics in school, (2) whether the student had taken tests that look similar and/or used similar tools, and (3) how hard the student thought this test was.

2.2 TRAINING AND PILOT TESTING

Five interviewers (and one backup interviewer) were trained for the project. Since all the interviewers were experienced in the cognitive interview technique, the training primarily focused on reviewing the content of the clusters and familiarizing the interviewers with the test platform and the specifics of the interview protocols. Project leads provided a separate two-hour training for the protocol at each grade level.

Additionally, at each grade level, an experienced team member conducted a pilot interview to fine tune the protocol and, especially, to determine the number of clusters that could be covered in one interview and hence the number of students that would be required to adequately test the clusters. The pilot administrations confirmed that, at each grade level, all four clusters could be covered in a single one and one-half hour interview. Thus, for each cluster, we ultimately had data on 12 to 18 students.

2.3 STUDY SAMPLE

Students were primarily drawn from the San Francisco Bay area. Utah also contributed students for the elementary school sample, and Connecticut contributed students for the high school sample.

The Utah students were particularly valuable to the study because they were in schools where teachers were receiving Next Generation Science Standards (NGSS) training from an NGSS author.

To recruit students in the San Francisco Bay area, the project manager and a designated scheduler at the Cambium Assessment, Inc (CAI) worked with a recruitment firm. This firm used a household-based approach to recruitment and employed an CAI-developed recruitment screener. Having recognized that exposure to inquiry-based science would be limited, we targeted higher achieving students with the expectation that they would be the most likely to have received this instruction and have benefited from it. We tried to recruit students whose parents reported the students' grades as being mostly As and/or Bs in science. We balanced the sample on gender and ethnicity (white/non-white).

In Utah and Connecticut, the CAI program manager worked directly with designated school districts to recruit students near Salt Lake City and Hartford, respectively. The cognitive interviews were conducted at the CAI offices in San Mateo, California, and on-site at the schools in Utah and Connecticut. The characteristics of the sample are summarized in Table 1 and shown by student in the Appendix.

Table 1. Characteristics of Sample, by Grade Level

Characteristic	Elementary School (*n* = 18)	Middle School (*n* = 12)	High School (*n* = 15)
Location			
California	12	12	12
Connecticut	N/A	N/A	3
Utah	6	N/A	N/A
Grade Level			
Grade 5	15	N/A	N/A
Grade 6	3[1]	N/A	N/A
Grade 8	N/A	7	N/A
Grade 9	N/A	5	N/A
Grade 10	N/A	N/A	1[2]
Grade 11	N/A	N/A	13
Grade 12	N/A	N/A	1[2]
Gender			
Male	13	6	5
Female	5	6	10
Parent or Teacher Reported Ethnicity			
African American	1	2	1
Asian	2	3	1
Hispanic	1	1	5
White	13	6	6

Characteristic	Elementary School (*n* = 18)	Middle School (*n* = 12)	High School (*n* = 15)
Other	1	0	1
Prefer not to answer	0	0	1
Parent-Reported Achievement in Science[3]			
Mostly As	7	11	7
Mostly Bs	5	1	5

[1] Utah students
[2] Connecticut students
[3] Data for California subjects only

3. FINDINGS

We begin this section with a summary of findings that includes key take-aways from the cognitive interviews and basic performance statistics for each of the 12 clusters.

The summary is followed by a detailed discussion of cognitive interview findings for each of the 12 clusters. Each cluster-level discussion starts with a summary of student performance, a list of task demands, and an image of the cluster stimulus. These are followed by an item-by-item discussion that, for each item, displays the item text, summarizes score patterns, and addresses students' comprehension and reasoning.

The discussion of findings ends with a summary of students' general perceptions of the science clusters, as expressed at the end of the cognitive interviews.

3.1 SUMMARY OF FINDINGS

3.1.1 Key Take-Aways

Feasibility of Cluster Approach

Results from the cognitive interviews suggest that it is feasible to incorporate item clusters into standardized science tests. On average, the clusters took 12 minutes to complete, and students reported being familiar with the format conventions and tools used in the clusters and appeared to easily navigate the clusters' interactive features and response formats.

- When questioned at the end of the cognitive interviews, nearly all students at each grade level reported that they had taken online tests that used similar page layouts, multimedia, and tools (e.g., page layouts with stimulus on the left and items on the right; embedded video; scroll bars; Back, Next, and Zoom in/Zoom out buttons; drop-down menus; and connect line and Add Arrow tools).

- Further, interviewers noted that students at all grade levels appeared comfortable navigating the clusters and, generally speaking, understood how to interact with the simulations and the response formats. When students experienced confusion, it was due to idiosyncratic problems with specific simulations or test items.

Relationship to Content Knowledge

Across grade levels, most students who participated in the cognitive interviews found the greatest challenge to be their lack of relevant content knowledge or experience applying science and engineering practices. This is not unexpected given that the clusters were built to measure NGSS constructs, and most of the students in the sample had not been exposed to NGSS-based instruction.

- Utah students, who were specifically included in the elementary school sample because they came from schools in which teachers were receiving NGSS training from an NGSS author, did better on all clusters. Details are given in the next subsection, where we summarize student performance by cluster.

Many students commented on their lack of relevant content knowledge during the think-alouds, and, when questioned at the end of the interview, students reported that they lacked prior

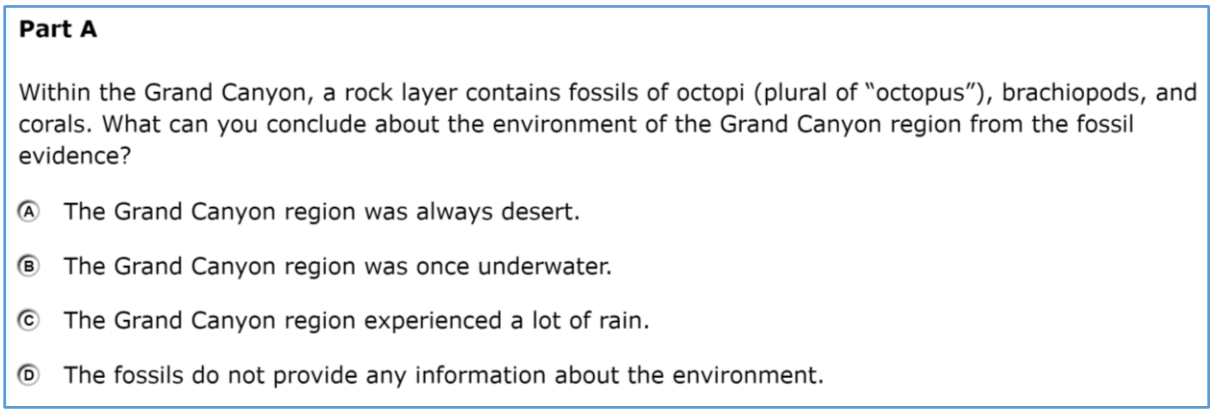
instruction in most of the topics covered by the clusters. If they had studied those topics, they said that it was at less depth than required to be successful. For example, one high school student said, in reference to the Blood Sugar Regulation cluster, that she had reviewed molecule concentrations but never discussed how they are impacted by meals, "not that in-depth, more gone over these and what they do for the body."

- By contrast, one of the Utah students said he had studied all four elementary school topics. "At the beginning of the year we studied the heat one and how we can help make a motor turn something on, like a light bulb. I thought of that. Maybe it was just backwards, the light was helping the fan to spin. The light was turning or making it spin by the energy it was producing. I remember last year, in 4th grade, we studied the Grand Canyon and the animals, and we did a little bit this year, and the animals that were living in the walls like trilobite and some others like starfish. We saw this video of this hole that was in Arizona, and there were tons of fossils in it. I think we studied a little bit on the terrarium one . . . We studied a little bit about [the desert plants]. About how each plant could survive."

Measuring Intended Constructs

In general, students who received credit on a given item (and some who did not) displayed a reasoning process that aligned with the skills that the item was intended to measure.

- This held true even for standard multiple-choice or multi-select items. For example, thinking aloud as he responded to this question in the Redwall Limestone cluster,



Part A

Within the Grand Canyon, a rock layer contains fossils of octopi (plural of "octopus"), brachiopods, and corals. What can you conclude about the environment of the Grand Canyon region from the fossil evidence?

Ⓐ The Grand Canyon region was always desert.

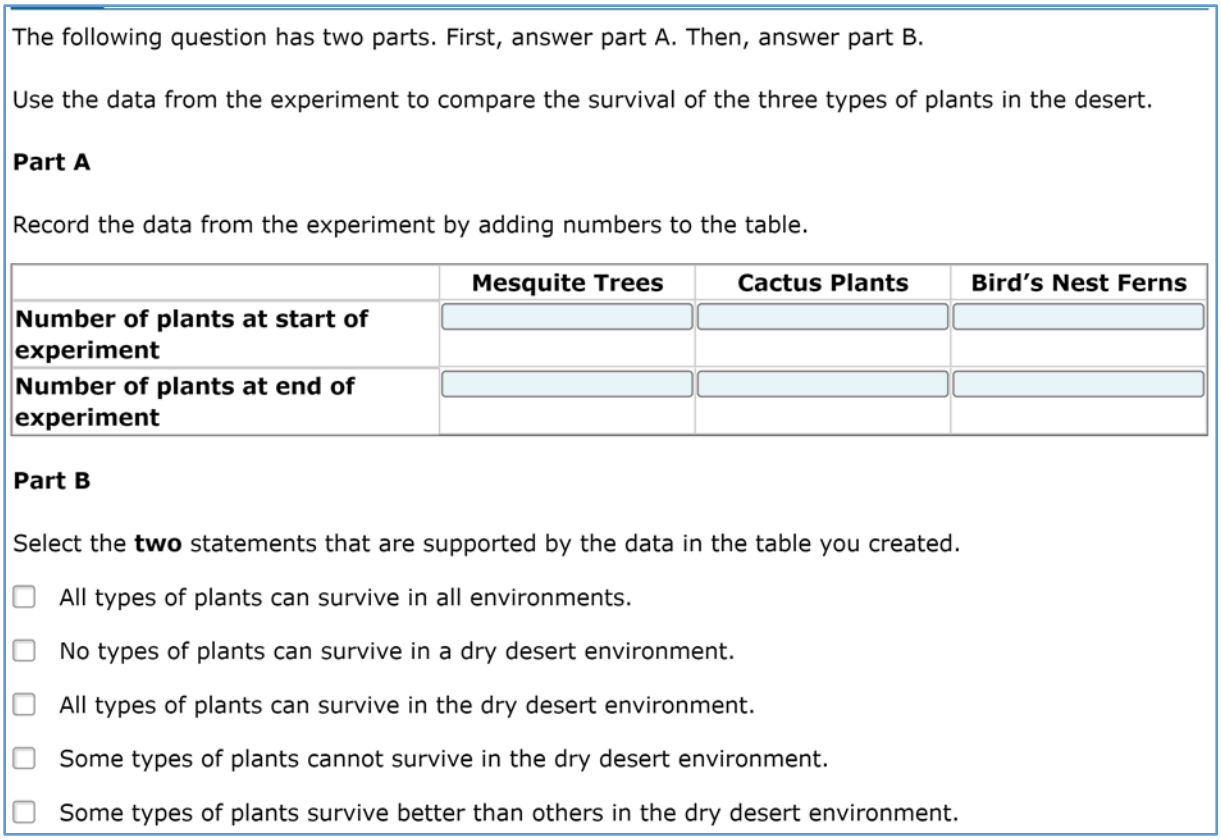
Ⓑ The Grand Canyon region was once underwater.

Ⓒ The Grand Canyon region experienced a lot of rain.

Ⓓ The fossils do not provide any information about the environment.

  one elementary school student first read option A, *[t]he Grand Canyon region was always desert*, out loud. Then he said he wanted to check the next option and read *[t]he Grand Canyon region was once underwater*. The student said that option B could be the answer, "but the first option [A] is not because it said in the question [the fossils] were sea animals." The student then read option C, *[t]he Grand Canyon region experienced a lot of rain*, and option D, *[t]he fossils do not provide any information about the environment*. He said that the answer couldn't be option D because "[the question] doesn't have anything to do with the animals that are living today." He said it probably wasn't option C because "even if it rained, [but] it wasn't an ocean, then the coral couldn't live there." The student concluded that the correct answer had to be B.

- In another example, an elementary school student explained her response to Part B of this two-part item from the Desert Plants cluster



The following question has two parts. First, answer part A. Then, answer part B.

Use the data from the experiment to compare the survival of the three types of plants in the desert.

Part A

Record the data from the experiment by adding numbers to the table.

	Mesquite Trees	Cactus Plants	Bird's Nest Ferns
Number of plants at start of experiment			
Number of plants at end of experiment			

Part B

Select the **two** statements that are supported by the data in the table you created.

- ☐ All types of plants can survive in all environments.
- ☐ No types of plants can survive in a dry desert environment.
- ☐ All types of plants can survive in the dry desert environment.
- ☐ Some types of plants cannot survive in the dry desert environment.
- ☐ Some types of plants survive better than others in the dry desert environment.

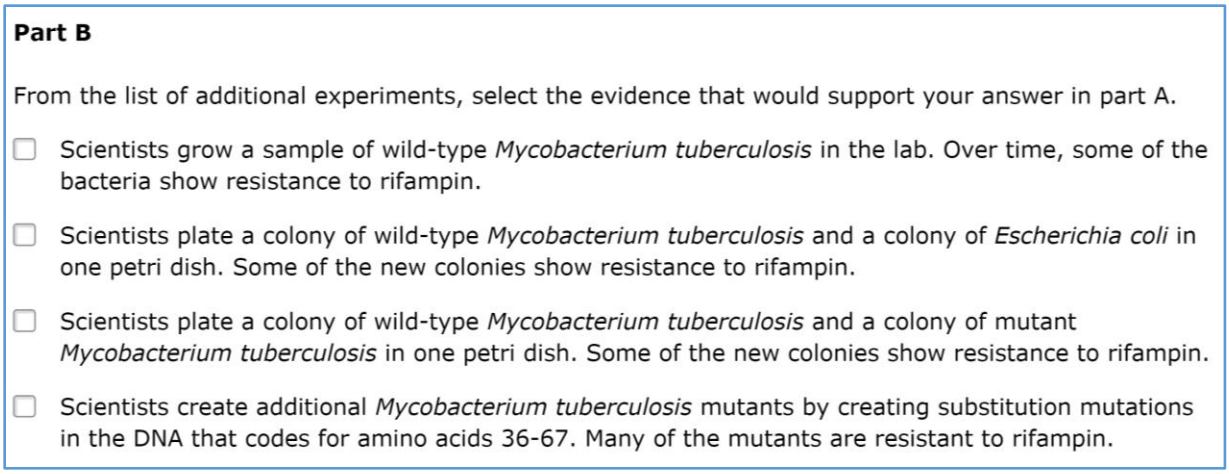
by saying that she chose the second-to-last option (*[s]ome types of plants cannot survive in the dry desert environment*) because "at the start of the experiment, there was a total of 5 bird's nest ferns, and then they all died, and also because one of the mesquite trees – they died – but I mean, most of them still remained." And she chose the last option (*[s]ome types of plants survive better than others in the dry desert environment*) because "out of all 3 of the plants, the cactus all lived instead of dying." She shared that she did not choose the first option (*[a]ll types of plants can survive in all environments*) because "As you can see, some of them died – like the bird's nest ferns and the mesquite trees." She shared that she did not choose the second option (*[n]o types of plants can survive in a dry desert environment*) "because the cactus – they still lived." She shared that she did not choose the third option (*[a]ll types of plants can survive in the dry desert environment*) "because the bird's nest ferns died."

There were exceptions where students gained or lost credit for non-construct relevant reasons, but these were related to specific item flaws that could be fixed before the items were used operationally.

General Recommendations for Improvements

While the validity of the general approach was supported by the cognitive lab findings, there were flaws in specific types of items that can and should be remediated before using the items operationally:

- Students needed more cueing on multi-select items such as the following:



Earning a score point for this item required correctly selecting both the first and the last options, but most students stopped after choosing one response. This type of error could be minimized by adding "mark all that apply" to the item stem.

- Students interactions with simulations should be checked to make sure that the simulations are functioning as intended. For example, a flaw in the simulation for the Texas Weather cluster allowed some students—who knew the proper tools for measuring each phenomenon (e.g., wind speed)—to lose credit for correctly matching tools with phenomena. This occurred because, when these students ran the simulation, they simply manipulated the tools and overlooked the drop-down menu for choosing the phenomenon they intended to measure. The simulation ran as intended under these conditions, so there was nothing to cue the students that they were inadvertently losing points.

- Scoring rubrics should be reviewed to make sure that they are constructed in a consistent manner and conform to the task demands they are intended to measure. In the cognitive interviews, some rubrics awarded a point for meeting a single, straightforward criterion, while others required that the student do several things correctly. For example, in item 1 in the Galilean Moons cluster, students got 1 score point for each of the moons for which they correctly measured the maximum distance from Jupiter. On the other hand, in item 1 of the Redwall Limestone cluster, students had to correctly identify six different animals as being found, or not found, in Arizona to earn any credit.

We recommend that the second type of rubric (requiring students to do several things correctly) be limited to cases in which integration across knowledge is the construct of interest.

3.1.2 Cluster Score Distributions and Average Time to Complete, by Grade Level

Elementary School Clusters

As shown in Table 2, average time to complete the elementary school clusters ranged from six minutes for the Redwall Limestone cluster to 12 minutes for the Desert Plants cluster.

Table 2. Maximum Score and Average Time to Complete: Elementary School Clusters

Cluster Name	Maximum Score	Average Time to Complete
Desert Plants	9	12
German Pyramid Candle	4	9
Redwall Limestone	4	6
Terrarium Matter Cycle	9	11

Table 3 and Table 4 show the score distributions for elementary school clusters with maximum scores of four and nine, respectively.

The Redwall Limestone cluster was easy for all students, with 12 students (71%) earning three or 4 score points. Utah students did even better, with half earning the maximum score of four points and two others earning 3 points.

The Desert Plants cluster was also relatively easy, with 15 students (83%) earning at least four of the nine points possible. All six Utah students earned scores in this range. Further, two Utah students were the only ones who earned the maximum score of eight, and four of the five students who earned at least seven points were from Utah.

The Terrarium Matter Cycle cluster was harder for all students, with only four students (22%) earning at least four of the nine points possible. Half of the Utah students earned scores in this range. No student earned the full nine points on this cluster, but the highest scoring student was a Utah student who earned seven points.

The German Pyramid Candle was the hardest cluster, with only one student (from Utah) earning the maximum score of four points (and none earning 3 points). Further, seven students (41%) earned no credit, but only one Utah student was included in this group.

Table 3. Number of Students Attaining Cluster Total Scores in Specified Range: Elementary School Clusters with Maximum Score = 4

Cluster Name	Score 4–3	Score 2–1	Score 0
German Pyramid Candle	1	9	7
Redwall Limestone	12	4	1

Note. For both clusters, $n = 17$.

Table 4. Number of Students Attaining Cluster Total Scores in Specified Range: Elementary School Clusters with Maximum Score = 9

Cluster Name	Score 9–7	Score 6–4	Score 3–1	Score 0
Desert Plants	5	10	2	1
Terrarium Matter Cycle	1	3	13	1

Note. For both clusters, $n = 18$.

Middle School Clusters

As shown in Table 5, the average time to complete the middle school clusters ranged from 10 minutes for the Galilean Moons cluster to 14 minutes for the Texas Weather cluster.

Table 5. Maximum Score and Average Time to Complete: Middle School Clusters

Cluster Name	Maximum Score	Average Time to Complete
Galilean Moons	9	10
Hippos	10	10
Morning Fog	9	12
Texas Weather	11	14

Table 6 through Table 8 show the score distributions for middle school clusters with maximum scores of nine, 10, or, 11, respectively.

Students performed best on the Galilean Moons cluster with five students (42%) earning at least seven points and an additional four students (33%) earning between six and four points.

The Hippos cluster was also fairly easy, with seven students (58%) earning four or more points.

The Morning Fog and Texas Weather clusters (maximum scores nine and 11, respectively) were both challenging for students. Only five students (43%) earned scores greater than three on Morning Fog, and only four students (33%) earned scores greater than three on the Texas Weather cluster.

Table 6. Number of Students Attaining Cluster Total Sores in Specified Range: Middle School Clusters with Maximum Score = 9

Cluster Name	Score 9–7	Score 6–4	Score 3–1	Score 0
Galilean Moons	5	4	3	0
Morning Fog	2	3	7	0

Note. For both clusters, $n = 12$.

Table 7. Number of Students Attaining Cluster Total Scores in Specified Range: Middle School Clusters with Maximum Score = 10

Cluster Name	Score 10–7	Score 6–4	Score 3–1	Score 0
Hippos	2	5	3	0

Note. $n = 10$.

Table 8. Number of Students Attaining Cluster Total Scores in The Specified Range: Middle School Clusters with Maximum Score = 11

Cluster Name	Score 11–7	Score 6–4	Score 3–1	Score 0
Texas Weather	0	4	8	0

Note. $n = 12$.

High School Clusters

As shown in Table 9, the average time to complete the high school clusters ranged from 10 minutes for the Tuberculosis cluster to 19 minutes for the Blood Sugar Regulation cluster.

Table 9. Maximum Score and Average Time to Complete: High School Clusters

Cluster Name	Maximum Score	Average Time to Complete
Blood Sugar Regulation	7	19
Saving the Tuna	7	14
Tomcods	8	17
Tuberculosis	5	10

Table 10 through Table 12 show the score distributions for high school clusters with maximum scores of five, seven, or eight, respectively.

Students found all the high school clusters challenging but performed the worst on the Tomcods cluster. Only one student (7%) earned a score greater than three on this eight-point cluster, and four students (31%) earned no credit. Similarly, there were four students in both the Tuberculosis and Saving the Tuna clusters who earned no credit. No one earned more than 5 points on the seven-point Blood Sugar Regulation cluster, but scores for most students (9 out of 12) were solidly in the mid-range of 5 to 3 points.

Table 10. Number of Students Attaining Cluster Total Scores in Specified Range: High School Clusters with Maximum Score = 5

Cluster Name	Score 5–4	Score 3–1	Score 0
Tuberculosis	1	9	4

Note. $n = 14$.

Table 11. Number of Students Attaining Cluster Total Scores in Specified Range: High School Clusters with Maximum Score = 7

Cluster Name	Score 7–6	Score 5–3	Score 2–1	Score 0
Blood Sugar Regulation	0	9	3	1
Saving the Tuna	1	2	5	4

Note. Blood Pressure Regulation $n = 13$; Saving the Tuna $n = 12$.

Table 12. Number of Students Attaining Cluster Total Scores in Specified Range: High School Clusters with Maximum Score = 8

Cluster Name	Score 8–6	Score 5–4	Score 3–1	Score 0
Tomcods	0	1	9	4

Note. $n = 14$.

3.2 DETAILED DISCUSSION BY CLUSTER: ELEMENTARY SCHOOL

3.2.1 Cluster 1: Desert Plants

Performance Summary

The median time to complete the Desert Plants cluster was 11.5 minutes. Table 13 and Table 14 indicate the number of students attaining cluster total scores and item scores within the specified ranges, respectively.

Table 13. Number of Students Attaining Cluster Scores in Specified Range: Desert Plants

Score 9–7	Score 6–4	Score 3–1	Score 0
5	10	2	1

Note. Maximum score = 9; *n* = 18.

Table 14. Number of Students Attaining Item Scores in Specified Range, by Item: Desert Plants

	Maximum Item Score	Score 1	Score 0
Item 1 (Part A)	1	12	6
Item 1 (Part B)	1	13	5
Item 2 (Part B)	1	3	15

	Maximum Item Score	Score 3	Score 2–1	Score 0
Item 2 (Part A)	3	2	13	3
Item 3	3	14	3	1

Note. *n* = 18.

Students did relatively well on this cluster, but Item 2 was much more challenging than Items 1 or 3.

Task Demands

The following are task demands of the Desert Plants cluster:

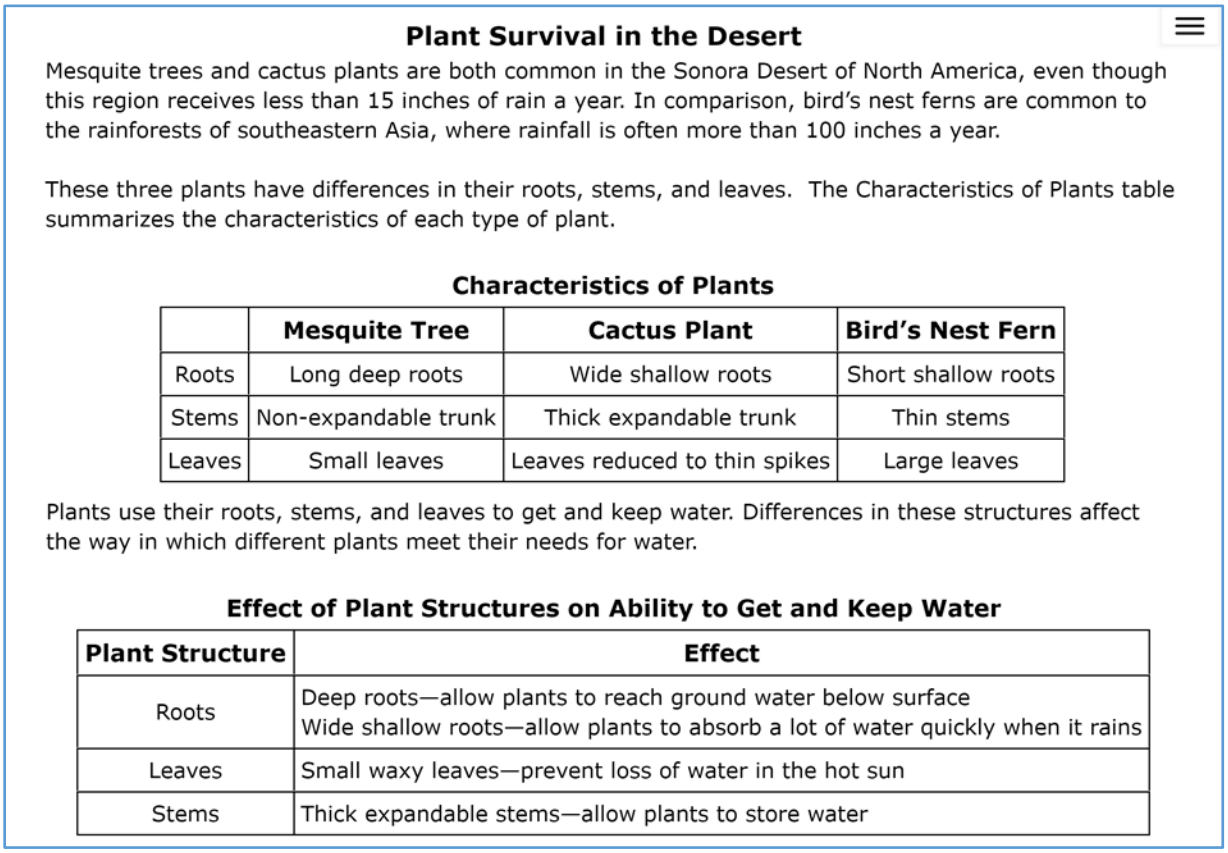
- Organize or summarize data to highlight trends and patterns and/or determine relationships between the traits of an organism and survival in its environment.
- Understand and generate simple bar graphs or tables that document patterns, trends, or relationships between traits of an organism and its survival in a particular environment.

- Identify patterns or evidence in the data that support inferences about characteristics of an organism and those of its environment.
- Based on the provided data, identify or describe a claim regarding the relationship between the characteristics of an organism and survival in a particular environment.
- Evaluate the evidence to sort relevant from irrelevant information regarding survival of an organism in a particular environment.

Stimulus

The stimulus for the Desert Plants cluster is shown in Figure 1.

Figure 1. Stimulus: Desert Plants



Plant Survival in the Desert

Mesquite trees and cactus plants are both common in the Sonora Desert of North America, even though this region receives less than 15 inches of rain a year. In comparison, bird's nest ferns are common to the rainforests of southeastern Asia, where rainfall is often more than 100 inches a year.

These three plants have differences in their roots, stems, and leaves. The Characteristics of Plants table summarizes the characteristics of each type of plant.

Characteristics of Plants

	Mesquite Tree	Cactus Plant	Bird's Nest Fern
Roots	Long deep roots	Wide shallow roots	Short shallow roots
Stems	Non-expandable trunk	Thick expandable trunk	Thin stems
Leaves	Small leaves	Leaves reduced to thin spikes	Large leaves

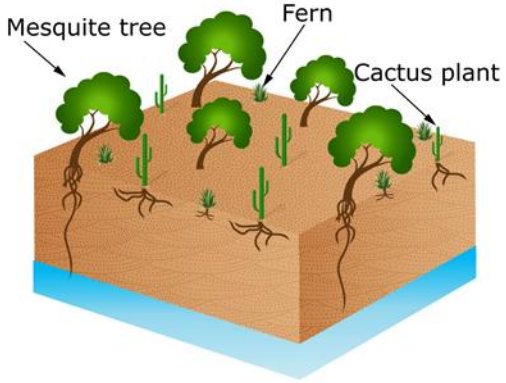
Plants use their roots, stems, and leaves to get and keep water. Differences in these structures affect the way in which different plants meet their needs for water.

Effect of Plant Structures on Ability to Get and Keep Water

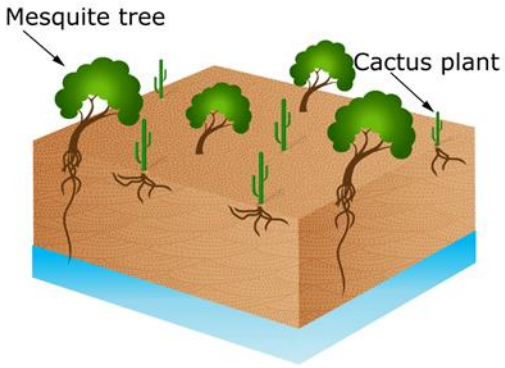
Plant Structure	Effect
Roots	Deep roots—allow plants to reach ground water below surface Wide shallow roots—allow plants to absorb a lot of water quickly when it rains
Leaves	Small waxy leaves—prevent loss of water in the hot sun
Stems	Thick expandable stems—allow plants to store water

To test how different characteristics affect a plant's ability to survive with less than 15 inches of rain a year, scientists planted Mesquite trees, cactus plants, and bird's nest ferns in a desert environment. A year later, they recorded how many of each type of plant survived.

Start of Experiment



End of Experiment



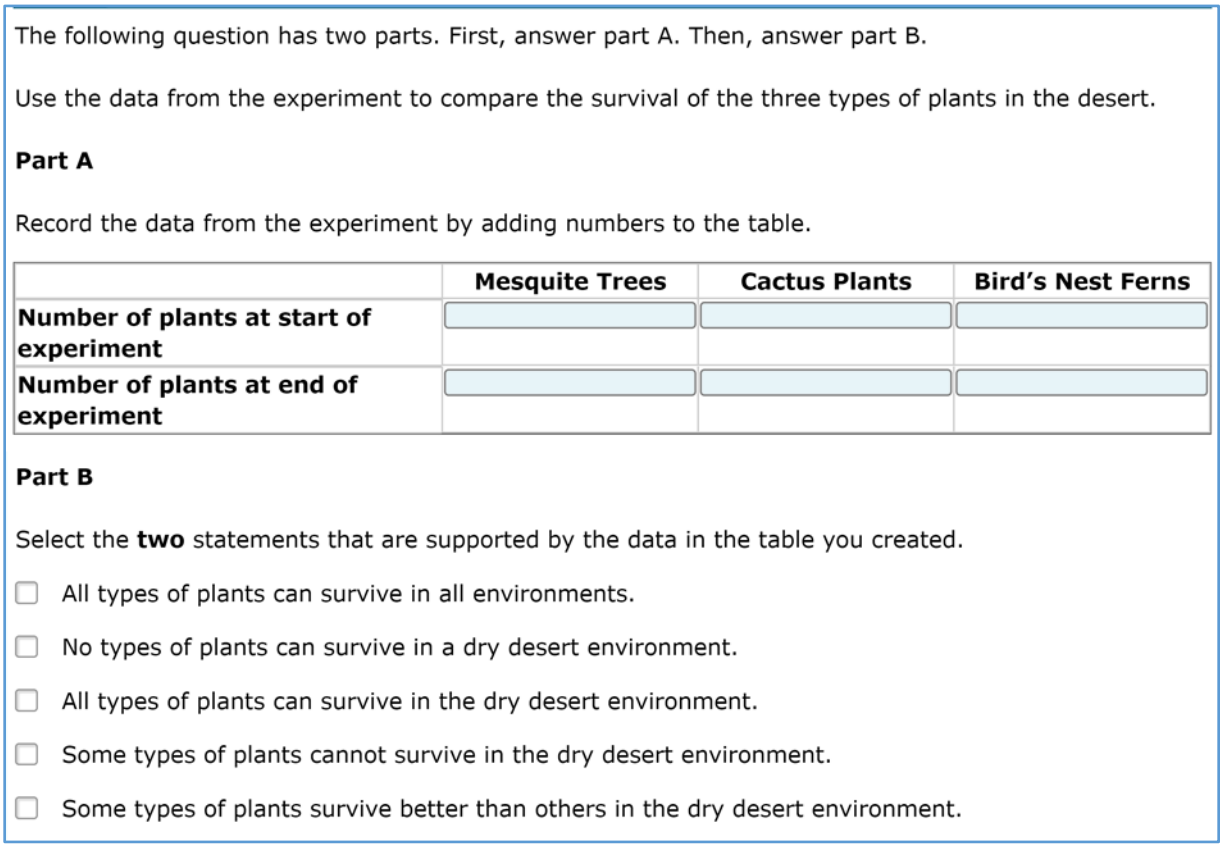
In the questions that follow you will construct an argument for why some plants survive better in the desert than others.

Details by Item

Item 1

Item 1 of the Desert Plants cluster is shown in Figure 2.

Figure 2. Item 1: Desert Plants



The following question has two parts. First, answer part A. Then, answer part B.

Use the data from the experiment to compare the survival of the three types of plants in the desert.

Part A

Record the data from the experiment by adding numbers to the table.

	Mesquite Trees	Cactus Plants	Bird's Nest Ferns
Number of plants at start of experiment			
Number of plants at end of experiment			

Part B

Select the **two** statements that are supported by the data in the table you created.

- ☐ All types of plants can survive in all environments.
- ☐ No types of plants can survive in a dry desert environment.
- ☐ All types of plants can survive in the dry desert environment.
- ☐ Some types of plants cannot survive in the dry desert environment.
- ☐ Some types of plants survive better than others in the dry desert environment.

Item 1 (Part A)

SCORES

Half of the California students (six) and all of the Utah students (six) earned credit (1 score point) on Part A.

COMPREHENSION

Those students who received credit for this item did not appear to be confused by any features of the item.

However, the students who did not receive credit seemed to have a general lack of comprehension of what was being asked. For example,

- one student wrote incoherent sentences instead of numbers;
- a second student decided to start at 27 "as a random number to start with"; and

- a third student said, "For mesquite trees, I got the start of experiment 1, do you see you start with 1, and at the end I saw how much they had altogether, and I got 3, so I was guessing that's how much it was." For the cactus plants, the student said, "I thought the same thing—they started off with 1 then ended with 3." For the bird's nest ferns, he said, "I was thinking the same thing because I was looking at the characteristics of plants—you start with 1 then you end with 3."

REASONING

The 12 students who earned credit all made sensible use of the experiment data.

> For example, one student said she counted the trees, plants, and ferns in the *Start of the Experiment* exhibit and began entering the numbers in the first row of the table. She explained, "I put 5 mesquite trees, because when I counted, there was 5 [at the beginning of the experiment]. When I counted the cactus, there was 5. And then the same for bird's nest ferns." She counted the trees, plants, and ferns in the *End of the Experiment* exhibit and began entering the numbers in the second row of the table. The student noted that there were four mesquite trees, explaining that this was "[b]ecause one of them had died during the experiment. And then for the cactus plants, the number stayed the same, at 5, because they normally live there, like, a lot, and they really don't need a lot of water to survive. And then the bird ferns all died during the experiment, so then that is a total of 0."

Item 1 (Part B)

SCORES

Thirteen students, including five of the six Utah students, earned credit (1 point) on Part B, which required them to identify two statements that are supported by the table in Part A. (One of these students did not receive credit for Part A but understood the general concept.)

COMPREHENSION

No features of this item appeared to confuse students.

REASONING

Most students used credible reasoning from evidence to reach a solution.

> For example, one student chose the second-to-last option (*[s]ome types of plants cannot survive in the dry desert environment*) because "at the start of the experiment, there was a total of five bird's nest ferns and then they all died, and also because one of the mesquite trees – they died – but I mean, most of them still remained." And she chose the last option (*[s]ome types of plants survive better than others in the dry desert environment*) because "out of all three of the plants, the cactus all lived instead of dying." She shared that she did not chose the first option (*[a]ll types of plants can survive in all environments*) because "As you can see, some of them died – like the bird's nest ferns and the mesquite trees." She shared that she did not choose the second option (*[n]o types of plants can survive in a dry desert environment*) "because the cactus – they still lived." She shared that she did not choose the third option (*[a]ll types of plants can survive in the dry desert environment*) "because the bird's nest ferns died."

Item 2

Item 2 of the Desert Plants cluster is shown in Figure 3.

Figure 3. Item 2: Desert Plants

The following question has two parts. First, answer part A. Then, answer part B.

Determine which traits of the three types of plants affect their survival in the desert.

Part A

The three tables show traits of each type of plant from the experiment. Select the boxes to identify whether each trait helps or does not help each plant survive in the desert.

Mesquite Tree Traits

	Helps Survival	Does Not Help Survival
Long deep roots	☐	☐
Non-expandable trunk	☐	☐
Small leaves	☐	☐

Cactus Plant Traits

	Helps Survival	Does Not Help Survival
Wide shallow roots	☐	☐
Thick stem	☐	☐
Thin spikes as leaves	☐	☐

Bird's Nest Fern Traits

	Helps Survival	Does Not Help Survival
Short shallow roots	☐	☐
Thin stem	☐	☐
Large leaves	☐	☐

Part B

Type a number into each box to identify the number of traits that help or do not help the plants survive, based on the tables in part A.

	Helps Survival	Does Not Help Survival
Mesquite Trees		
Cactus Plants		
Bird's Nest Ferns		

Item 2 (Part A)

SCORES

Points were awarded based on the number of plants for which the student correctly identified the traits that help the plant survive. Two students earned 3 score points (full credit) on Part A, six students earned 2 score points, and seven students earned 1 score point.

COMPREHENSION

No features of this item appeared to confuse students.

REASONING

Nine of the students used the *Characteristics of Plants* and *Effects of Plant Structures on Ability to Get and Keep Water* tables, and at least three of these students also referred to the exhibits showing plants that were alive at the beginning and end of the experiment. However, they did not necessarily interpret all the data correctly. For example, the following student referenced the information in the stimulus tables frequently and appropriately but misinterpreted some of the data. She did not appear to use the exhibits on the start and end of the experiment to check her understanding of which traits help or hinder survival.

- For the mesquite tree she said, "the mesquite tree has long deep roots and also has small leaves," and checked *Helps Survival* for roots and leaves. She continued, "The [mesquite] plant—I don't think that the non-expandable trunk will help. It says that thick expandable stems allow plants to store water, except the tree doesn't have one, so it can't store a lot of water, so I don't think that will help it survive." She checked *Does Not Help Survival* for the non-expandable trunk.

- For the cactus plant she said, "The cactus plant traits, it says it has wide shallow roots that allow the plant to absorb lots of water when it rains. So that would help it survive." She checked *Helps Survival* for roots. She continued, "The thick trunk also will, but thick stem would do that." She checked *Helps Survival* for trunk. She continued, "Then thin spikes as leaves—that probably wouldn't help them a lot." She checked *Does Not Help Survival* for leaves.

- For the bird's nest fern she said, "So for the bird's nest fern traits, it has shallow roots, and shallow roots allow it to absorb a lot of water when it rains, so that would probably help survive." She checked *Helps Survival* for roots. She continued, "A thin stem—that would probably not help it survive since the thin stem would not be able to hold a lot of water to help it survive." She checked *Does Not Help Survival* for the stem. She continued, "Then large leaves—that would probably be good. And small waxy leaves have lots of water in the hot sun. Yep." She checked *Helps Survival* for leaves.

Seven students made little or no use of the data in the stimulus and based their reasoning for Part A on prior knowledge or conjecture.

Item 2 (Part B)

SCORES

On Part B, most students quickly filled out the table on the number of traits that help or do not help each plant survive based on their responses in Part A.

However, only three students completed all six cells correctly, as required to earn credit (1 score point) on Part B.

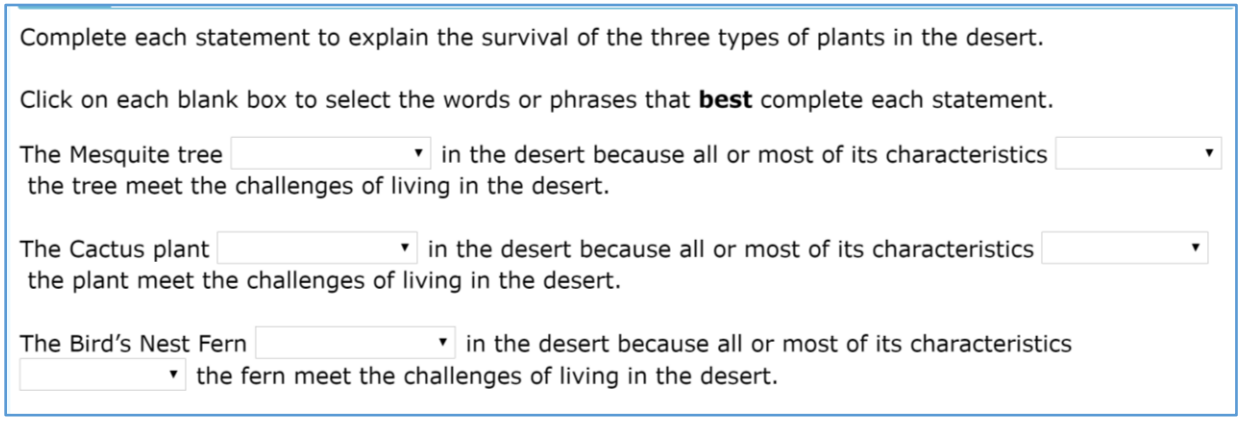
COMPREHENSION

On Part B, three students wrote the types of traits in the response fields (e.g., long deep roots) rather than the number of traits as indicated in the instructions. One student also wrote some extraneous text. One other student wrote text that was mostly incoherent.

Item 3

Item 3 of the Desert Plants cluster is shown in Figure 4.

Figure 4. Item 3: Desert Plants



SCORES

Students earned 1 point for each statement they completed correctly. Fourteen students completed all three statements correctly and earned full credit. This included all six of the Utah students.

Sixteen students earned a score point for the statement on the mesquite tree. Sixteen students earned a score point for the statement on the cactus plant, and 15 students earned a score point for the statement on the bird's nest fern.

COMPREHENSION

All students navigated through this item with ease.

REASONING

Most students used their answers to previous questions in the cluster to select responses from the drop-down menus. At least five students used information from the stimulus, and three students used prior knowledge.

The following is an example of a student who reasoned appropriately from the evidence in the stimulus to respond to Item 3:

> The student selected *survived well* for mesquite tree, explaining that this was "because all or most of its characteristics helped the tree meet the challenges of living in the desert; because the characteristics, such as having the long deep roots and the small leaves can help it survive in the desert." She selected *survived best* for cactus plant, "because all or most of its characteristics helped it meet the challenges of living in the desert; because, of all of the plants, it stayed alive, and the characteristics such as having wide shallow roots and thick stems helped it live." The student selected *did not survive* for bird's nest fern, noting that "only one of its traits helped, and the rest—the two other ones—did not help it." Then she selected the answers for the second part of each item, choosing *helped* for mesquite tree, *helped* for cactus plant, and *did not help* for bird's nest fern.

3.2.2 Cluster 2: German Pyramid Candle

Performance Summary

The median time to complete the German Pyramid Candle cluster was nine minutes. Table 15 and Table 16 indicate the number of students attaining cluster total scores and item scores within the specified ranges, respectively.

Table 15. Number of Students Attaining Cluster Total Scores in Specified Range: German Pyramid Candle

Score 4–3	Score 2–1	Score 0
1	9	7

Note. Maximum score = 4. *n* = 17; one student ran out of time before attempting this cluster.

Table 16. Number of Students Attaining Item Scores in Specified Range, by Item: German Pyramid Candle

	Maximum Item Score	Score 2	Score 1	Score 0
Item 1	2	3	5	9

	Maximum Item Score	Score 1	Score 0
Item 2	1	2	15
Item 3	1	5	12

Note. n = 17; one student ran out of time before attempting this cluster.

This was the most difficult of the elementary school clusters; only one student (from Utah) earned full credit (4 points).

Task Demands

The following are task demands of the German Pyramid Candle cluster:

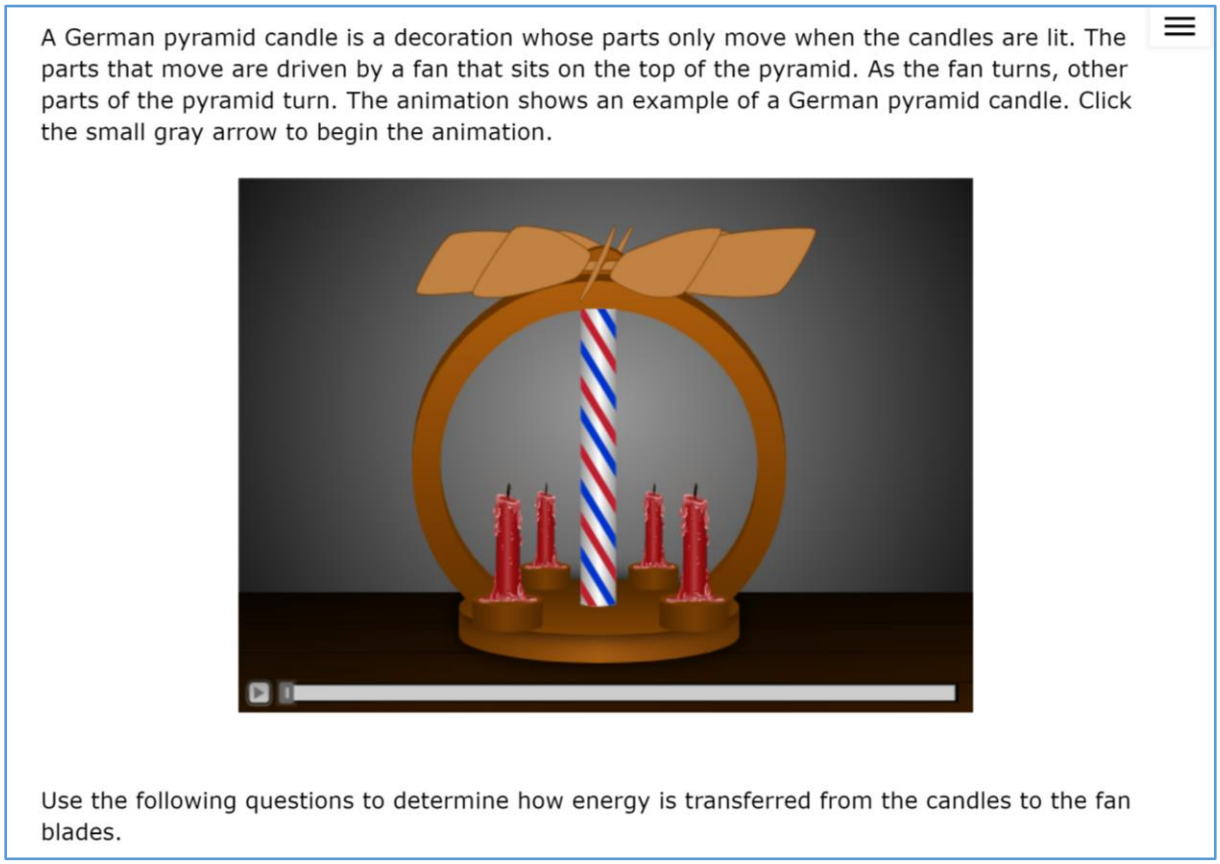
- Identify from a list, including distractors, the materials/tools needed for an investigation of how energy is transferred from place to place through heat, sound, light, or electric currents.
- Identify the outcome data that should be collected in an investigation of how energy is transferred from one place to another through heat, sound, light, or electric currents.
- Make and/or record observations about the transfer of energy from one place to another via heat, sound, light, or electric currents.
- Interpret and/or communicate the data from an investigation.

- Select, describe, or illustrate a prediction made by applying the findings from an investigation.

Stimulus

The stimulus for the German Pyramid Candle cluster is shown in Figure 5.

Figure 5. Stimulus: German Pyramid Candle

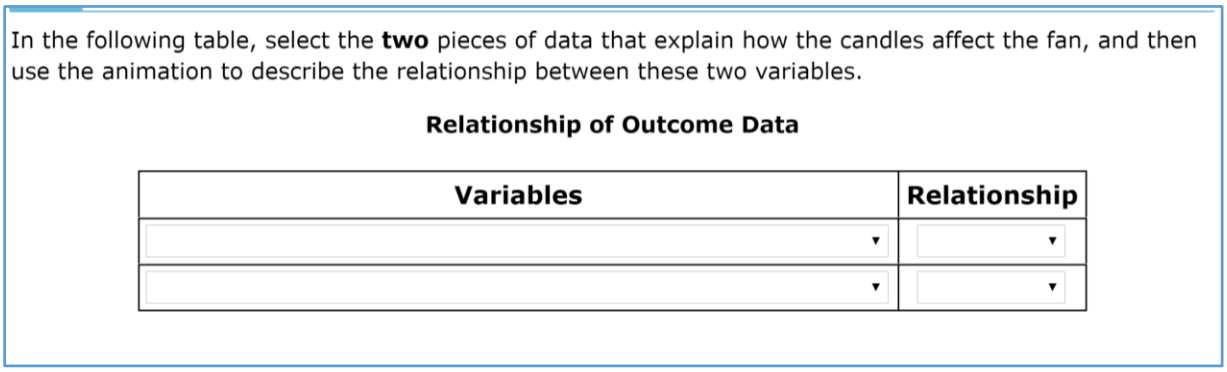


Details by Item

Item 1

Item 1 of the German Pyramid Candle cluster is shown in Figure 6.

Figure 6. Item 1: German Pyramid Candle


In the following table, select the **two** pieces of data that explain how the candles affect the fan, and then use the animation to describe the relationship between these two variables.

Relationship of Outcome Data

Variables	Relationship

SCORES

Two (Utah) students earned full credit (2 score points) on this item, which required students to identify two variables that explain the influence of the candles on the fan and then describe the relationship between these variables.

Seven other students earned partial credit for selecting the two correct variables but not correctly specifying the relationships—five were Utah students.

Additional students selected at least one of the correct variables.

A total of 13 students correctly selected the temperature of the air between the blades and the candles as one of the variables, and eight students correctly selected the rotation speed of the blade.

COMPREHENSION

Students clearly did not understand how to describe the relationship between the two variables as only four students entered any responses to this part of the question. It is not clear how much of the confusion was because the students did not understand how energy was transferred and how much of the confusion was due to not understanding what the question was asking.

Five students were hesitant about the entire item, and two students tried to guess at the relationships between the two variables because they did not really understand what "the relationship" meant.

REASONING

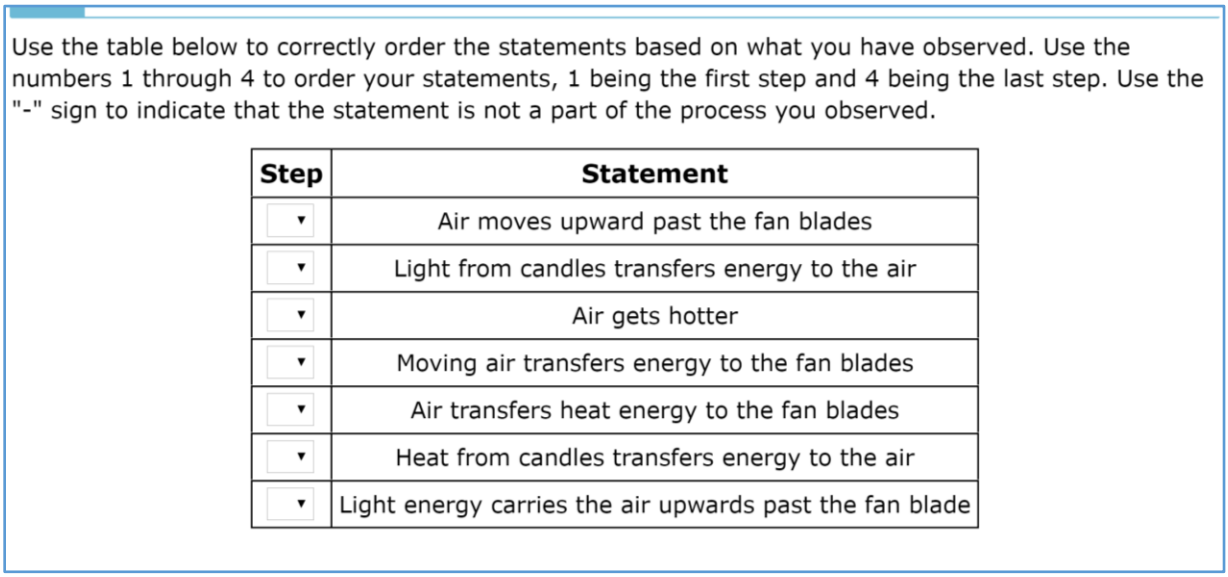
Most students tried to reason their way to a solution but lacked the content knowledge to do so without error. The following shows the reasoning process for one student who exemplifies this:

> The student said, "The first variable is probably going to be *brightness* because if they're more brighter, it probably means that it's hotter. And for relationship, I'm going to do *increase* because I think it turns because something is taking in the heat energy and it's using the heat energy from the candles to rotate the fan, and that's why the brightness of the candles would probably increase the speed of the rotation of the fans. And so for variable two, I'm going to do the *temperature of the air between the blades and the candles* – I chose that because if the air is colder or cooler, it's probably not going to rotate that much because it takes in the heat energy that the candles create and it rotates them . . . And if it's like hot or warm, it's probably going to rotate faster . . . if I'm correct. And for the relationship, I'm going to do decrease because if it's slower or cooler, it's probably going to be less . . . or not as fast as if it was warmer."

Item 2

Item 2 of the German Pyramid Candle cluster is shown in Figure 7.

Figure 7. Item 2: German Pyramid Candle



SCORES

All but one student observed the whole animation, but only two (Utah) students earned credit (1 score point) on this item by correctly ordering the steps based on what they observed in the animation.

COMPREHENSION

One student did not seem to understand that he was to order the steps, and it was not clear how he selected the numbers for his responses.

REASONING

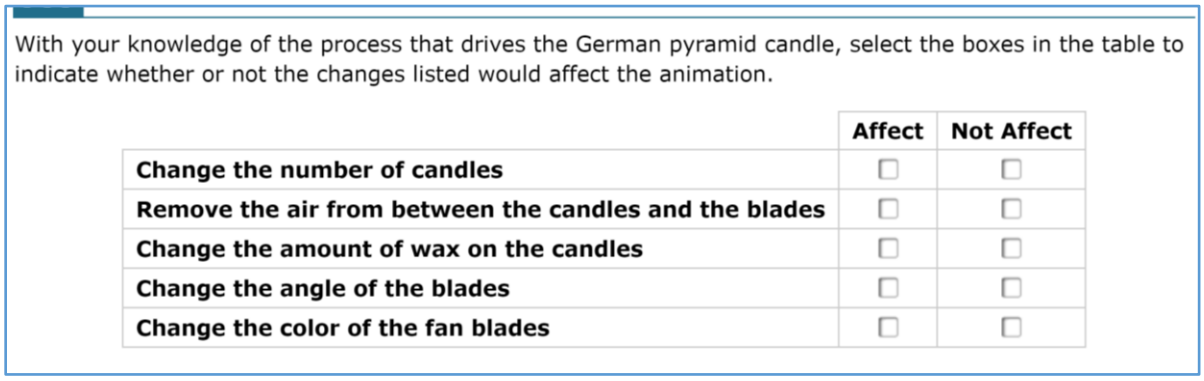
Students had the same issues with lack of content knowledge as they did with Item 1.

> For example, one student correctly chose *[h]eat from candles transfers energy to the air* for step 1 (noting that "the energy carries the air upward past the fan"), but faltered after that. She chose *[a]ir transfers heat energy to the blades* for step 2, noting that it "was going to the fan blades." For step 3, the student initially chose *[a]ir moves upward past the fan blades* but changed it to *[l]ight energy carries the air upwards past the fan blade*. When prompted later to explain why she changed her answer, she explained, "Because it made more sense if hot air moved upward past the fan blades, but it was just air, so I was thinking light energy carries the air upward past the fan blades because first the energy goes to the fan blades and then the light energy from the candles goes past the fans." For step 4, she thought for a moment and said, "I think this (*air gets hotter*), and chose it," explaining "because it goes around more."

Item 3

Item 3 of the German Pyramid Candle cluster is shown in Figure 8.

Figure 8. Item 3: German Pyramid Candle


With your knowledge of the process that drives the German pyramid candle, select the boxes in the table to indicate whether or not the changes listed would affect the animation.

	Affect	Not Affect
Change the number of candles	☐	☐
Remove the air from between the candles and the blades	☐	☐
Change the amount of wax on the candles	☐	☐
Change the angle of the blades	☐	☐
Change the color of the fan blades	☐	☐

Scores

Five students earned credit (1 score point) for this item.

Nine other students correctly classified four of the five changes, but earned no credit, based on the scoring rubric.

Comprehension

No features of this item appeared to confuse students.

Reasoning

As with the other items in this cluster, students needed prior content knowledge to reason their way to a correct solution. For example, one student, who had most of the requisite knowledge, said,

> "For the first one, the *change in number of candles*, I think that, with more heat and light, I think it will affect it a little bit more by making the blades spin faster. *Removing the air from between the candle and blades*, I think that will affect it because the GPC probably takes in the air from what's underneath it. For the third one, the *change in the amount of wax on the candles*, I think that will not affect it because the wax just increases the duration of the candle, which wouldn't affect it. *Change the angle of the blades*, I don't think that would affect it because if you just turn the blades over to at least an angle where it looks like it's even, I don't think that will affect it either. *Change the color of the fan blades*, I don't think changing the color of the fan blades would affect it because it's just color, and it's for decoration most of the time."

3.2.3 Cluster 3: Redwall Limestone

Performance Summary

The median time to complete the Redwall Limestone cluster was six minutes. Table 17 and Table 18 indicate the number of students attaining cluster total scores and item scores within the specified ranges, respectively.

Table 17. Number of Students Attaining Cluster Total Scores in Specified Range: Redwall Limestone

Score 4–3	Score 2–1	Score 0
12	4	1

Note. Maximum score = 4; *n* = 17; one student ran out of time before attempting this cluster.

Table 18. Number of Students Attaining Item Score in Specified Range, by Item: Redwall Limestone

	Score 1	Score 0
Item 1	13	4
Item 2	13	4
Item 3 (Part A)	14	3
Item 3 (Part B)	7	10

Note. Maximum score for each item = 1; *n* = 17; one student ran out of time before attempting this cluster.

Task Demands

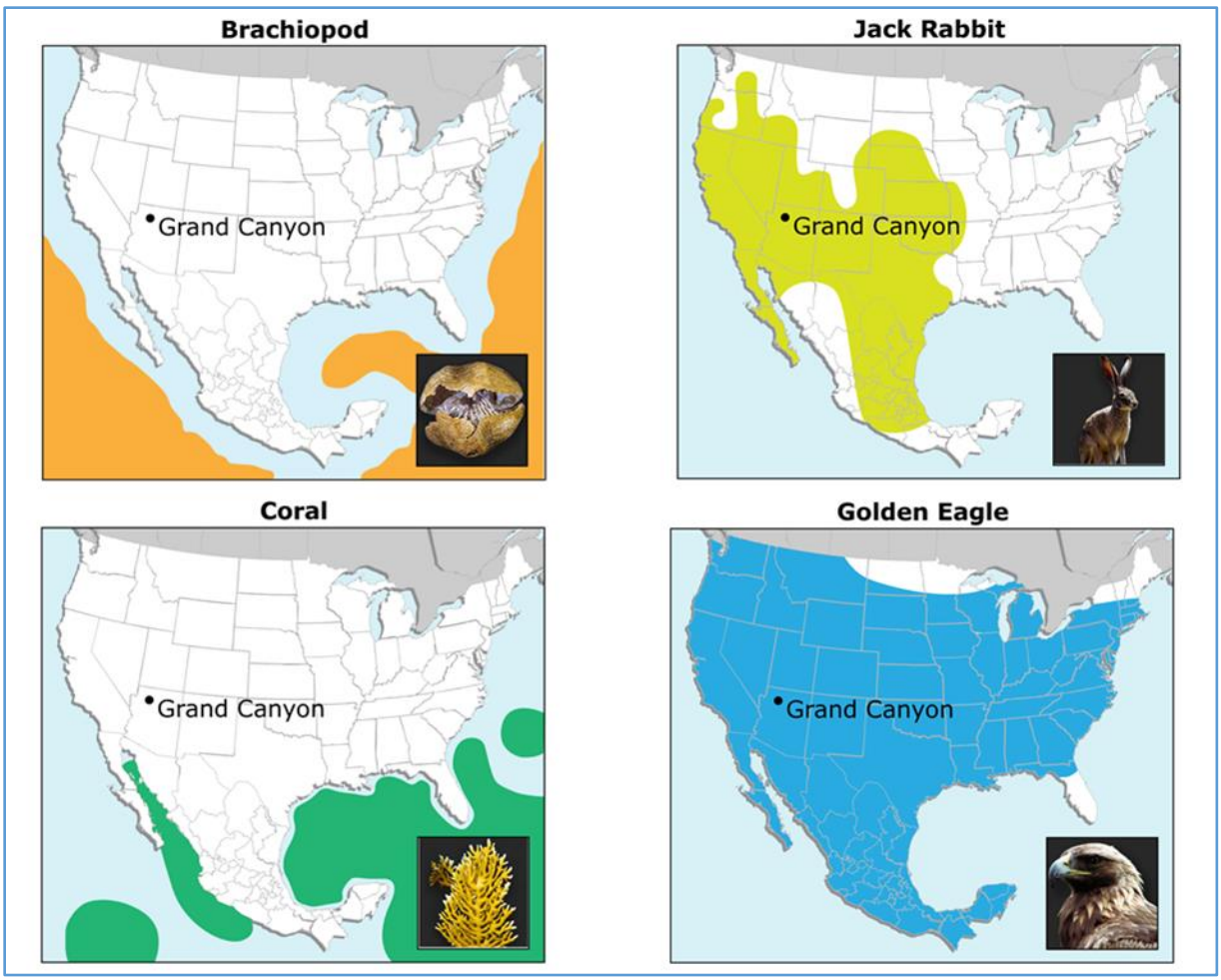
The following are task demands of the Redwall Limestone cluster:

- Organize or summarize data to highlight trends, patterns, or correlations between plant and animal fossils and the environments in which they lived.
- Generate graphs or tables that document patterns, trends, or correlations in the fossil record.
- Identify evidence in the data that support inferences about plant and animal fossils and the environments in which they lived.

Stimulus

The stimulus for the Redwall Limestone cluster is shown in Figure 9.

Figure 9. Stimulus: Redwall Limestone



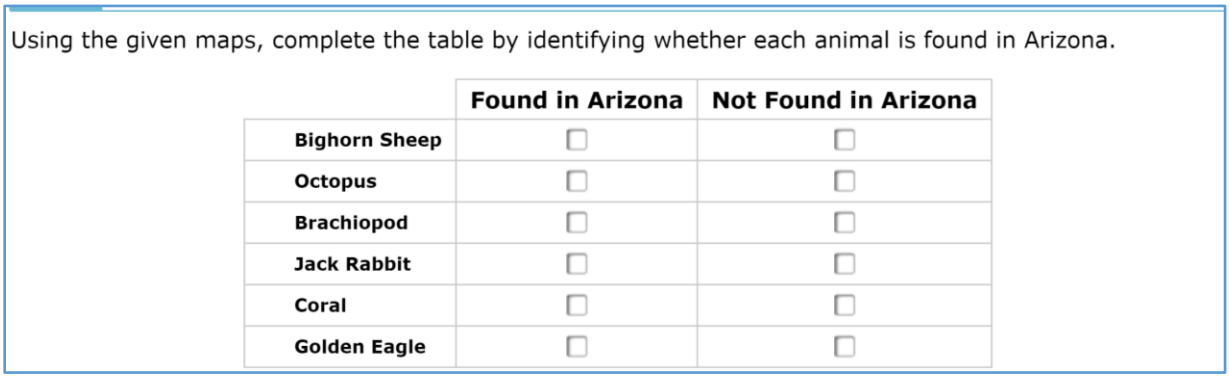
Despite some incorrect responses, nearly all the students seemed comfortable navigating through the maps to decide where the animals are found and filling out the tables in Items 1 and 2. One student did not make any use of the maps.

Details by Item

Item 1

Item 1 of the Redwall Limestone cluster is shown in Figure 10.

Figure 10. Item 1: Redwall Limestone



SCORES

Thirteen students earned credit (1 score point) on this item.

COMPREHENSION

No features of this item appeared to confuse students.

REASONING

Ten of the 13 students who earned credit showed evidence in the think-aloud of using the maps to reason their way to a solution, as intended.

For example, one student

- selected *Found in Arizona* for bighorn sheep "because the map that it gives you shows you that it's located in Arizona."
- selected *Not Found in Arizona* for octopus, explaining that "It's found in oceans – not really in the state."
- selected *Not Found in Arizona* for brachiopod, noting, with a laugh, "Because it's in the oceans, not the state – like the octopus . . . octopi."
- selected *Found in Arizona* for jack rabbit "because the map that it gives you shows it's located in Arizona."
- selected *Not Found in Arizona* for coral because "the map that it gives you has those green things that shows you that it's not located in Arizona."
- selected *Found in Arizona* for the golden eagle, noting that "the blue is all over the United States, so yeah, it's in Arizona."

Among the four students who did not earn credit for this item, each mis-located two of the six animals. The think-alouds showed that three of these students formed their answers based on background knowledge and some educated guessing rather than using the maps.

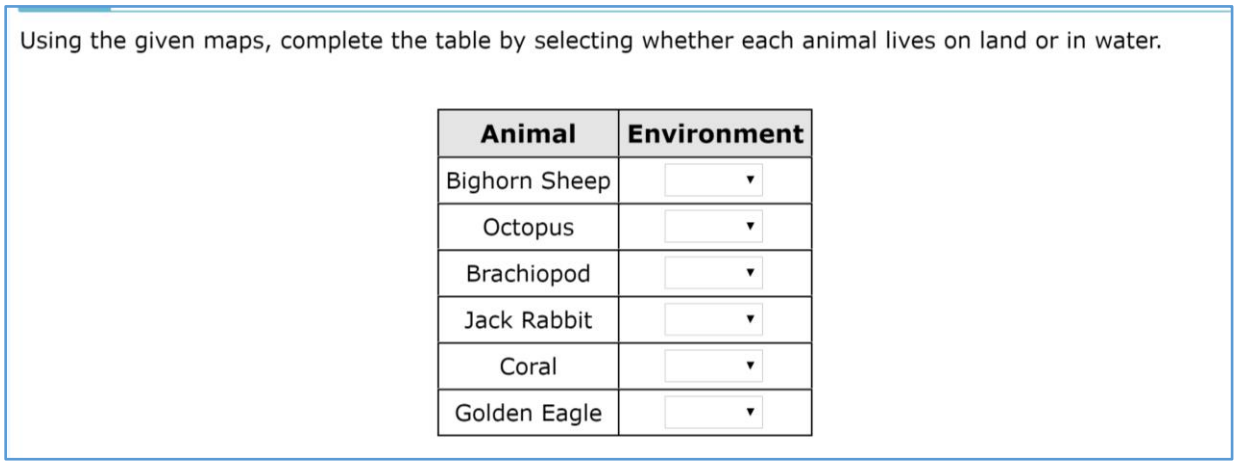
For example, one student

- selected *Not Found in Arizona* for bighorn sheep because "When I went to Arizona, I've never seen a bighorn sheep over there, so I really think it is not in there."
- selected *Found in Arizona* for jack rabbit, explaining that "it's in there because I've seen one when I went to Arizona."
- selected *Not Found in Arizona* for coral. This choice appeared to be at random, marked after the student said, "I've never heard of that animal too because in school we don't really learn about coral and so yeah I've never heard of it and I don't know if they're ever in Arizona, so . . ."
- selected *found in Arizona* for golden eagle because "I think it's in Arizona because our school mascot is the golden eagle and they always say golden eagles are from Arizona."

Item 2

Item 2 of the Redwall Limestone cluster is shown in Figure 11.

Figure 11. Item 2: Redwall Limestone



SCORES

Thirteen students earned credit (1 score point) on this item.

COMPREHENSION

No features of this item appeared to confuse students. All students worked through the item fairly quickly, and three of the students commented that it was easy.

REASONING

Among the 13 students who earned credit, most did not appear to make much use of the maps in formulating their responses, apparently because they felt that they could easily respond based on background knowledge about the animals.

> For example, one student shared that she knows bighorn sheep live on land and that octopi are living in the water. But then she noted that she wasn't sure about coral, adding, "Sometimes you see coral on the beach or somewhere else, and so I don't know if it's land or water. But maybe it was washed up on the beach, so I was thinking water."

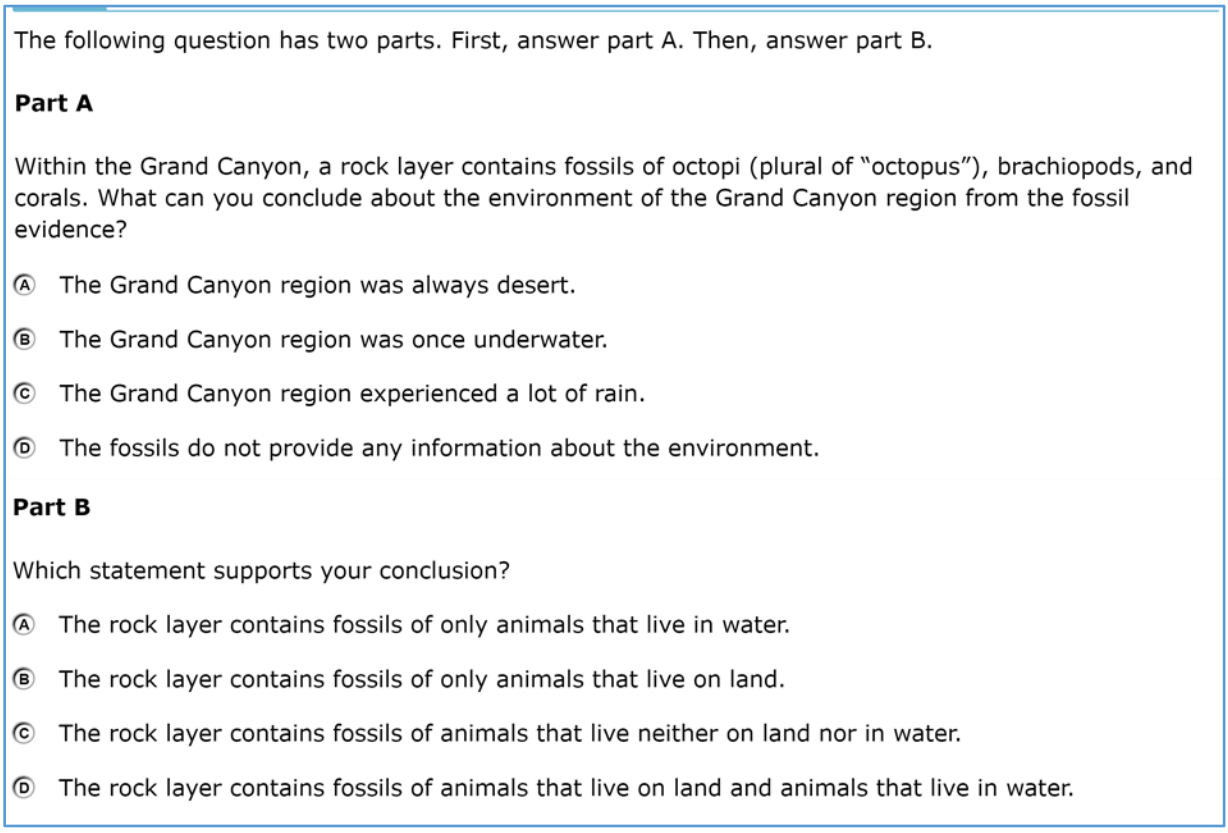
Students who did not earn credit for this item mis-located either the brachiopod or the coral; one student also mis-located the golden eagle. These students also relied on background knowledge for their responses. For example, one student explained his choices as follows:

- The bighorn sheep "is on land because I don't think he'll make it in the water."
- The octopus "has to live in the water to survive."
- The brachiopod "has to live in the water because it looks like a jellyfish and jellyfishes have to live in the water, so I thought maybe that does too, and I looked at the picture and thought it has to live in the water."
- "I looked at [the jack rabbit], and that's a land animal, and regular rabbits live on land, and that's why I picked that one."
- "[The coral] has to be on land because it kind of looks like a tree and trees have to be on land."
- "Birds and eagles are on land, so I picked that eagle to be on land, so I just knew it from my knowledge."

Item 3

Item 3 of the Redwall Limestone cluster is shown in Figure 12.

Figure 12. Item 3: Redwall Limestone



The following question has two parts. First, answer part A. Then, answer part B.

Part A

Within the Grand Canyon, a rock layer contains fossils of octopi (plural of "octopus"), brachiopods, and corals. What can you conclude about the environment of the Grand Canyon region from the fossil evidence?

Ⓐ The Grand Canyon region was always desert.

Ⓑ The Grand Canyon region was once underwater.

Ⓒ The Grand Canyon region experienced a lot of rain.

Ⓓ The fossils do not provide any information about the environment.

Part B

Which statement supports your conclusion?

Ⓐ The rock layer contains fossils of only animals that live in water.

Ⓑ The rock layer contains fossils of only animals that live on land.

Ⓒ The rock layer contains fossils of animals that live neither on land nor in water.

Ⓓ The rock layer contains fossils of animals that live on land and animals that live in water.

Item 3 (Part A)

Scores

Fourteen students earned credit (1 score point) on this sub-item.

There was no common theme to the wrong answers—there were three possible wrong answers, and each of the three students who failed to earn credit chose a different one.

Comprehension

Among the three students who did not earn full credit for the sub-item, one student appeared not to understand what the question was asking. She said she was confused on how to respond because "I thought it was going to ask me 'does it usually rain there?' and it doesn't usually rain there because it's in Arizona."

Reasoning

The 14 students who earned credit for this sub-item (1 score point) all appeared to evaluate the possible response option against credible criteria as they reasoned their way to a solution.

For example, one student first read option A, *[t]he Grand Canyon region was always desert*, out loud. Then he said he wanted to check the next option and read *[t]he Grand Canyon region was once underwater*. The student said that option B could be the answer, "but the first option [A] is not because it said in the question [the fossils] were sea animals." The student then read option C, *[t]he Grand Canyon region experienced a lot of rain*, and option D, *[t]he fossils do not provide any information about the environment*. He said that it can't be option D because "[the question] doesn't have anything to do with the animals that are living today." He said it probably wasn't option C because "even if it rained, [but] it wasn't an ocean, then the coral couldn't live there." The student concluded that the correct answer had to be B.

Item 3 (Part B)

SCORES

Seven students earned credit (1 score point) on this sub-item.

COMPREHENSION

Among the 10 students who did not earn credit on this sub-item, most appeared to be confused as to what the question was asking. Rather than associating the question with Part A, these students appeared to be trying to answer a separate question about the types of animal fossils that might be found in the canyon walls. Further, they did not seem to know where to look for information that would help them answer the question; they tended to reference the list of *current-day* animals mentioned in the stimulus, and to do so irrespective of whether these animals were found in Arizona. Consequently, nine of these 10 students selected option D, *[t]he rock layer contains fossils of animals that live on land and animals that live in water*, using reasoning such as the following:

One student said, "obviously C, *the rock layer contains fossils of animals that live neither on land nor in water*, is wrong, it's not only water because they have jack rabbits, the goat-ram thing, and the eagle so that's not true." For option B, *the rock layer contains fossils of only animals that live on land*," he said: "that's not true, there are octopus, coral and brachiopod." He read out loud response option C a second time, *the rock layer contains fossils of animals that live neither on land nor in water*, and said "the bird does live on land and it flies a lot, but it's still on land, so it has to be D, *the rock layer contains fossils of animals that live on land and animals that live in water*."

Some students also seemed to have problems with the structure of the answer choices (A, or B, or neither A nor B, or both A and B).

For example, one student said, "What I found confusing was this one since I was looking at D and it said, 'live in water' at the end, just like A, so I was looking at it, and I figured out that it said lived on land AND on water. It kind of confused me just looking at the end that both of them said 'live in water.'"

REASONING

The seven students who earned credit for this sub-item all appeared to use credible criteria in reasoning their way to a solution.

For example, one student read out loud the stem and option A, *[t]he rock layer contains fossils of only animals that live in water*. He said that it could be that one, but he wanted to read the other options. He read out loud option B, *[t]he rock layer contains fossils of only animals that live on land*. The student said, "no, it wouldn't be that one because the answer [to Part A] doesn't have anything to do with that." He read option C, *[t]he rock layer contains fossils of animals that live neither on land nor in water*, and said it couldn't be the right answer, because the question says that [the rock layer] has sea animals. He read option D*, [t]he rock layer contains fossils of animals that live on land and animals that live in water*. The student said that "the question never said anything about that part" and chose A.

3.2.4 Cluster 4: Terrarium Matter Cycle

Performance Summary

The median time to complete the Terrarium Matter Cycle cluster was 11 minutes. Table 19 and Table 20 indicate the number of students attaining cluster total scores and item scores within the specified ranges, respectively.

Table 19. Number of Students Attaining Cluster Total Scores in Specified Range: Terrarium Matter Cycle

Score 9–7	Score 6–4	Score 3–1	Score 0
1	3	13	1

Note. Maximum score = 9; *n* = 18.

Table 20. Number of Students Attaining Item Scores in Specified Range, by Item: Terrarium Matter Cycle

	Maximum Item Score	Score 1	Score 0
Item 1 (Part A)	1	3	15
Item 1 (Part B)	1	6	12
Item 2 (Part A)	1	8	7
Item 2 (Part C)	1	1	17
Item 2 (Part D)	1	1	17
Item 3	1	7	11

	Maximum Item Score	Score 3	Score 2–1	Score 0
Item 2 (Part B)	3	3	10	5

Note. n = 18

Earning credits on this cluster was challenging for the students. Two of the Utah students earned the most credit (seven and six credits respectively), likely reflecting their greater exposure to NGSS-based instruction.

Task Demands

The following are task demands of the Terrarium Matter Cycle cluster:

- Select or identify from a collection of potential model components, including distractors, the parts of a model needed to describe the movement of matter among plants, animals, decomposers, and the environment.

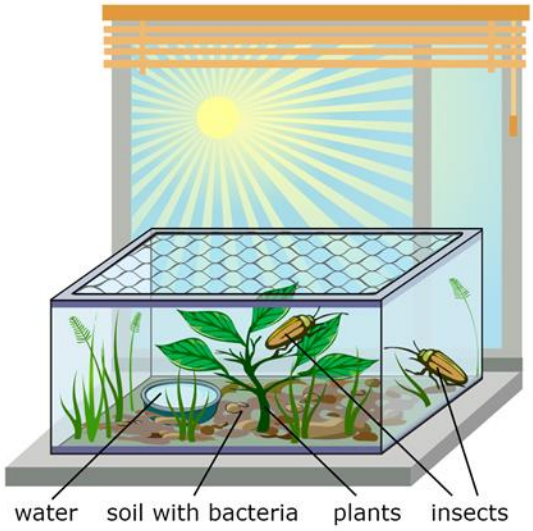
- Manipulate the components of a model to demonstrate properties, processes, and/or events that result in the movement of matter among plants, animals, decomposers, and the environment including the relationships of organisms and/or the cycle(s) of matter and/or energy.
- Articulate, describe, illustrate, select, or identify the relationships among components of a model that describe the movement of matter among plants, animals, decomposers, and the environment.
- Make predictions about the effects of changes in model components including the substitution, elimination, or addition of matter and/or an organism and the result.

Stimulus

The stimulus for the Terrarium Matter Cycle cluster is shown in Figure 13.

Figure 13. Stimulus: Terrarium Matter Cycle

A science class sets up four terrariums on a sunny windowsill. Each terrarium contains water and insects. Each one also contains a combination of gravel, soil with bacteria, and/or plants according to the Terrarium Setups table.



Terrarium Setups

	Terrarium 1	Terrarium 2	Terrarium 3	Terrarium 4
Soil			X	X
Gravel	X	X		
Plants		X		X

The students observe the terrariums every 5 days for 15 total days and record observations of the insects and plants. Their data are shown in the Terrarium Observations diagrams.

Terrarium 1 Observations

Day	Insects
1	Alive
5	Not alive
10	Not alive
15	Not alive

Terrarium 2 Observations

Day	Insects	Plants
1	Alive	Alive
5	Alive	Alive
10	Alive	Not alive
15	Not alive	Not alive

Terrarium 3 Observations

Day	Insects
1	Alive
5	Not alive
10	Not alive
15	Not alive

Terrarium 4 Observations

Day	Insects	Plants
1	Alive	Alive
5	Alive	Alive
10	Alive	Alive
15	Alive	Alive

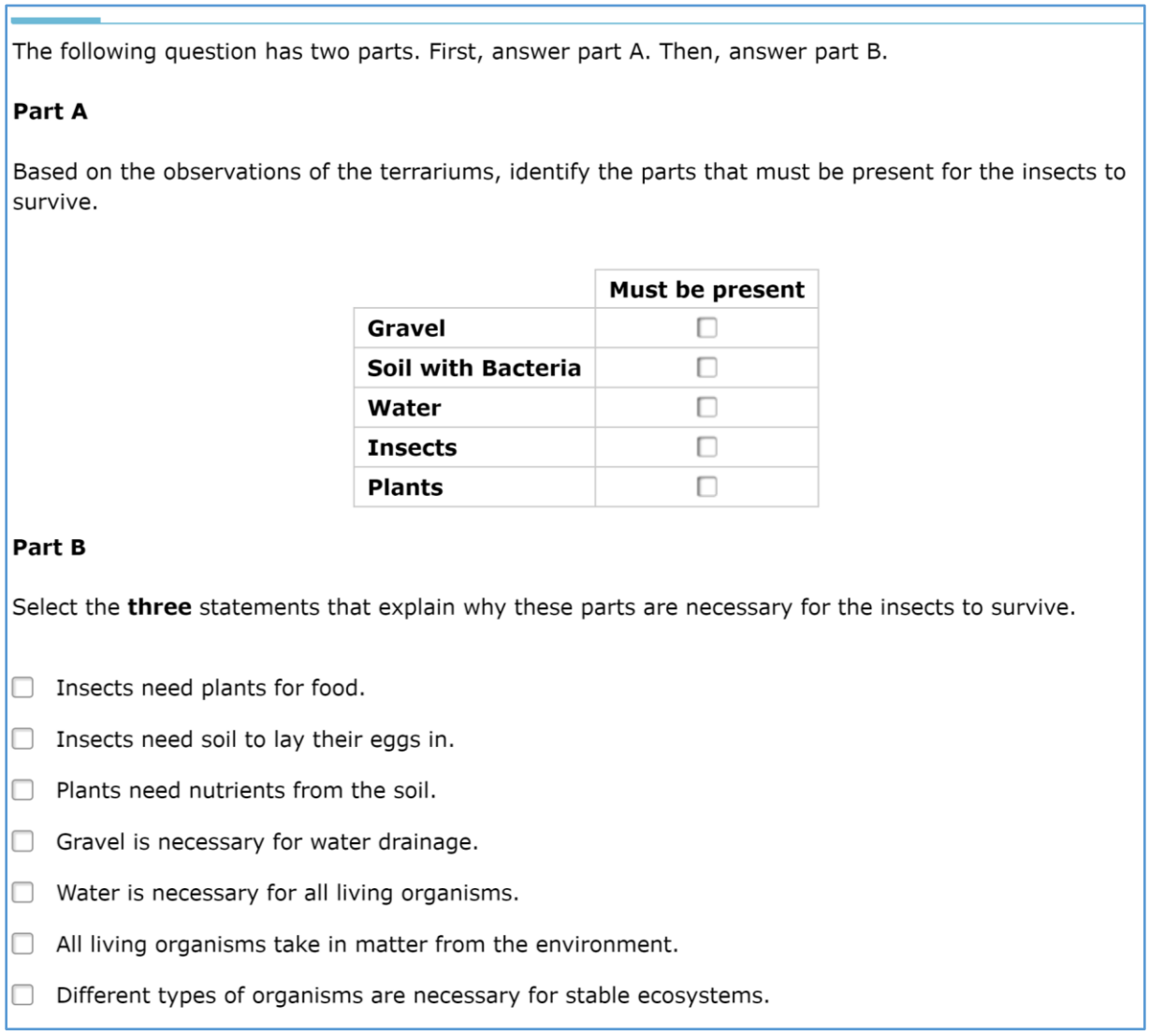
In the following questions, you will develop a model to show why the insects only survive under certain environmental conditions.

Details by Item

Item 1

Item 1 of the Terrarium Matter Cycle cluster is shown in Figure 14.

Figure 14. Item 1: Terrarium Matter Cycle



The following question has two parts. First, answer part A. Then, answer part B.

Part A

Based on the observations of the terrariums, identify the parts that must be present for the insects to survive.

	Must be present
Gravel	☐
Soil with Bacteria	☐
Water	☐
Insects	☐
Plants	☐

Part B

Select the **three** statements that explain why these parts are necessary for the insects to survive.

- ☐ Insects need plants for food.
- ☐ Insects need soil to lay their eggs in.
- ☐ Plants need nutrients from the soil.
- ☐ Gravel is necessary for water drainage.
- ☐ Water is necessary for all living organisms.
- ☐ All living organisms take in matter from the environment.
- ☐ Different types of organisms are necessary for stable ecosystems.

Item 1 (Part A)

SCORES

Three students earned credit (1 score point) on this sub-item, which required them to correctly identify all four of the elements that must be present for the insects to survive. Ten other students correctly identified three of the four parts.

COMPREHENSION

Several students had trouble with the concept that the organism itself (i.e., insects) was one of the things that had to be present for that organism to survive. Six students gave a response that correctly identified soil with bacteria, water, and light as essential, but left out insects. Some others chose insects, but interpreted it as other insects, or were not sure.

> For example, when the interviewer asked after the think-aloud, "You weren't sure whether to click insects or not here. Could you tell me a little about that?" One student said, "Yeah. Would it be the insects themselves? Or would it be different insects? Like you'd put two cockroaches in there with a ladybug. Or you'd put two ladybugs with a spider. I don't know. If insects have to be there to survive, then yes, but if it is different insects and they'd be harmless, then I'd say no, they don't need to be there. So maybe more description there."

REASONING

The three students who received credit for the sub-item displayed the type of reasoning from evidence that was expected, although their reasoning was not necessarily correct in every detail.

> For example, one student said, "I know a class sets up four terrariums by a sunny windowsill, so light can get in to help the plants. I know plants have a photosynthesis process, and they need the sun to make food. There are also insects so they can eat, and water so they can drink, and soil so they can have a stable root because I know that plants don't need soil to grow. In terrarium 3 and 4 there is soil, and in terrarium 1 and 2 there is gravel, and in 2 and 4 there are plants. A student observes the terrarium every 5 days for 15 days and records observation. Three times he observes them to collect observation—like the two living things in there, like the insects and the plants, and the data is shown on the diagram. I can see that the day 1 the insects are alive because in terrarium 1 there is only gravel, but no plants, so they don't have anything to eat, so they can only survive about a day. Day 1, the insects are alive because—they are alive for three checks because they have gravel and plants . . . . The plants dying would probably be because maybe gravel is not strong to hold their roots. If the plants die, so do the insects. In terrarium 3, the insects are alive, and they all die on the next days because they don't have any plants to eat. And then terrarium 4 has plants and soil, so it has plenty for the insects to eat, and it is a good support for the plants, so if they both stay alive, they can feed off each other."

Many students who did not receive credit made only limited use of the experimental data provided in the stimulus and relied entirely or primarily on background knowledge.

> For example, for *Gravel*, one student said, "I don't think it should be present because, if you just need gravel, you would have nothing to do with the soil in there." For *Soil with Bacteria* the student said, "It must be present because a lot of plants and flowers, they need soil—and they also have bacteria in it or something." For *Water*, the student said, "It definitely needs to be present because with just sun and soil, it won't let it grow because every plant needs water, soil, and sun." For *Insects*, the student said, "Yeah, because bees like going on sunflowers, so yeah it could be present." For *Plants*, the student said, "Not so much cause if you're going to grow one it's already present . . . ." When asked if this was from the student's prior knowledge, she agreed.

Item 1 (Part B)

SCORES

Six students earned credit (1 score point) on this sub-item, which required students to correctly identify all three of the statements that explained why the elements in Part A are necessary for the insects to survive. Ten other students correctly identified two of the three statements.

COMPREHENSION

No features of this item appeared to confuse students.

REASONING

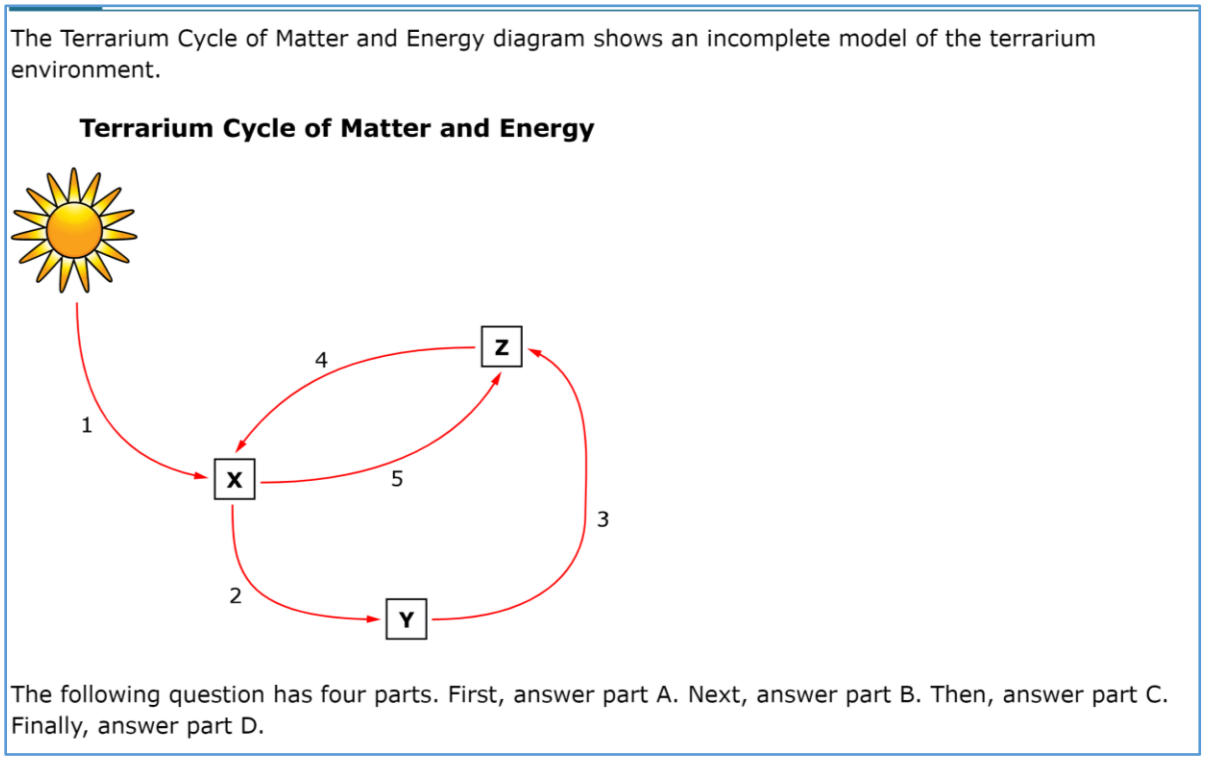
Students reasoned from background knowledge, but not necessarily content area knowledge gained in school.

> For example, one student selected option 1, and when asked how she knew, the student said, "if insects don't have food or water they'll die, and I know that just from background knowledge." The student selected option 3 because, "plants need nutrients from the soil, or they will die too… I just used my background knowledge." Student selected option 4 (*[g]ravel is necessary for water drainage*) and when asked how she knew, she said, "Just from learning it in school, I've just heard it before."

Item 2

Item 2 of the Terrarium Matter Cycle cluster is shown in Figure 15.

Figure 15. Item 2: Terrarium Matter Cycle



Part A

Select the boxes to identify X, Y, and Z.

	X	Y	Z
Gravel	☐	☐	☐
Soil with Bacteria	☐	☐	☐
Water	☐	☐	☐
Insects	☐	☐	☐
Plants	☐	☐	☐

Part B

Select the boxes to identify X, Y, and Z as a producer, consumer, or decomposer.

X: ▾
Y: ▾
Z: ▾

Part C

Select the **two** numbers that represent arrows in the model to show when matter or energy is moved from the environment to organisms.

☐ 1
☐ 2
☐ 3
☐ 4
☐ 5

Part D

Carbon dioxide and water are missing from this model. If added, where would the arrow be pointing?

Ⓐ from X toward Y
Ⓑ from Y toward Z
Ⓒ from the environment toward X
Ⓓ from the environment toward Z

Students generally did not understand the *Terrarium Cycle of Matter and Energy* diagram in Item 2. One student did not answer any of the parts in Item 2.

Item 2 (Part A)

SCORES

Only three students earned full credit (3 score points) on Part A, which required selecting correct labels for X, Y, and Z. Ten other students earned 1 score point. Two of the three students who earned full credit were from Utah.

COMPREHENSION

Six students said Part A was confusing. They appeared not to understand the conventions of the diagram and possibly also did not understand the concept of matter and energy cycle.

> For example, one student said, "I don't get this question . . . I think it's missing something—the soil, the water, and insects that give it nutrients or something." The student attempted to click the diagram, thinking it might be interactive. She then moved on to Part A, read it aloud, and said, "I think for number 1 it's sun, then X is going to be *water*, and then this is going to be *insects*, and then this is going to be *plants*." After checking X for *Water*, the student also checked X for *Insects* and X for *Plants*. She then realized that she had overwritten her response to X twice and went back to check X for *Water*, Y for *Insects*, and Z for *Plants*.

Only one of the Utah students thought this sub-item was confusing; the remaining five Utah students did not express confusion or appear to guess at the interpretation of the diagram.

Item 2 (Part B)

SCORES

Eight students earned credit (1 score point) in Part B by correctly identifying X, Y, and Z as a producer, consumer, or decomposer. Seven other students identified one of the components correctly.

COMPREHENSION

Only one student expressed confusion on Part B, and this appeared to relate more to confusion over the producer, consumer, and decomposer roles than to the wording of the item. The student said:

> "What was confusing on this was B, because I forgot which one was that, so I was looking, and I thought about what was a producer, and I remembered that [it] was something that helps it grow. And X was the soil and bacteria, so X would have been the producer. The consumer got me confused because I didn't remember learning about the consumer. So, I was thinking it probably was the plants since I knew the decomposer was the one who would help the things decompose into the ground, and that was probably the insects. So, I knew that Y was the consumer."

REASONING

The reasoning of students who received credit for Part B indicated that they did know the facts of the natter and energy cycle, whether or not they understood the letters in the response choices as referencing the diagram.

> For example, one student said, "X is a *producer*, Y is a *consumer*, and Z has to be *decomposer* . . . X is producer because sunlight goes to the plants, and then the plants produce food for themselves and others, Y is consumer because the consumer eats the producer, and Z is decomposer, because after the consumer dies, the decomposer decomposes it and turns it into soil."

Item 2 (Part C)

SCORES

Only one (Utah) student earned credit (1 score point) on Part C, which required that students select both the arrows in the model that showed where matter or energy is moved from the environment to organisms. Nine other students correctly selected the arrow from the sun to X, but not the arrow from Z to X.

COMPREHENSION

The vocabulary used in this sub-item, particularly "environment," "organism," and "matter," was unfamiliar to several of the students.

> For example, one student did not understand the term "matter." The student said he was confused by "questions that had things to do with 'matter' because I know what matter is, but we started learning in science class, and I haven't fully gotten the sense of matter yet."

Confusion may also have arisen from the way in which the term "environment" is used, namely, to refer to the inanimate environment only.

REASONING

Most students tried to reason their way to a solution, but their content knowledge was too limited to allow them to identify both correct arrows. For example:

> One student said, "I'm going to say one of my answers is '1' because of light energy maybe is being moved from the environment, from the sun – I'm pretty sure that's part of the environment, and I'm pretty sure a plant is an organism. And for my second number I'm trying to think about what I can say . . . because the plant has matter, I'm pretty sure, or everything has matter. And a plant is an organism, and it says matter or energy, and the matter is being given or moved from the plant to the insect."
>
> Another student said, "I chose 2 and 3 since those are the necessary parts since the soil went in a circle to the soil. From the soil to the plants and from the plant to the insect. Since I thought that was the most important part. If it was 4 and 2, it would just be the same thing, but I thought 2 and 3 would be better and make more sense since the insect would be going to the soil and then the soil would make the plants and that wouldn't really make sense." The interviewer asks the student, "What do you think the question is asking?" The student

said, “It is showing that energy is moved from the environment to the organisms and I chose those since the matter in the sun is giving the soil energy to make the plants grow and that would keep going around. The plants would be decomposed or eaten by the bugs.”

Item 2 (Part D)

SCORES

Only three students earned credit (1 score point) on Part D, which asked where the arrow would be pointed if carbon dioxide and water were added to the model. Interestingly, eight students incorrectly indicated that the arrow would point from X toward Y.

COMPREHENSION

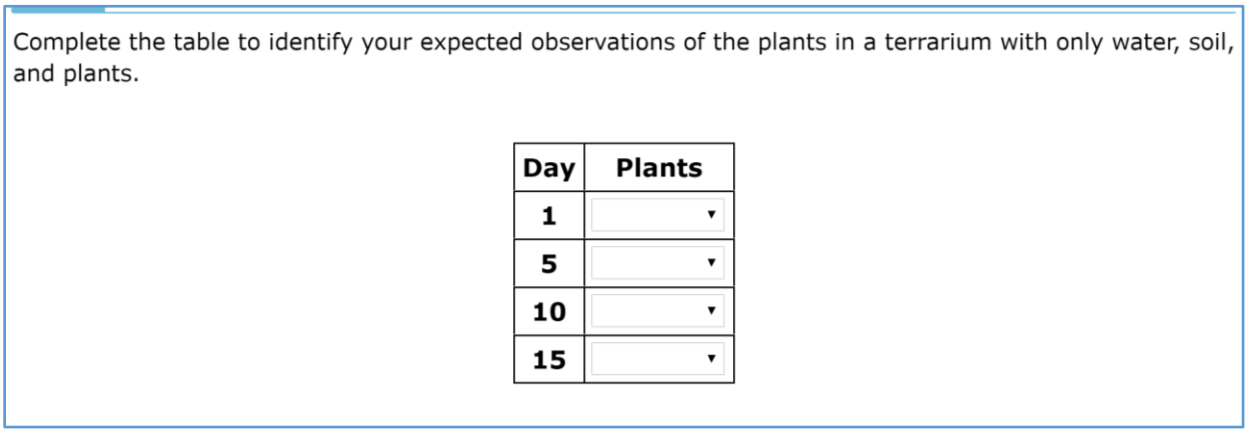
Several students simply lacked the content knowledge to answer this question.

> For example, one student said, "because I had to find from X toward Y – I had to know that the insects carried the carbon dioxide to the plants, but then also carry it to the soil."

Item 3

Item 3 of the Terrarium Matter Cycle cluster is shown in Figure 16.

Figure 16. Item 3: Terrarium Matter Cycle



SCORES

Seven students earned credit (1 score point) on this item.

COMPREHENSION

No issues with comprehension of the item were noted.

REASONING

Some students applied the information provided in the experiment to help them answer this question, although not all students were able to interpret the information from the experiment correctly.

> An example of using the experimental information correctly was a student who said, "This question is asking me to see how the plants, what I would observe if the plants were in a terrarium with water, soil, and plants. Plants would be plants, and soil would be soil, and water would be something to keep the plants alive. So, day 1 they would probably be alive.

> After 5 days, as long as plants are supplied by water and sun, they'd be alive. On day 10, they'd probably still be alive because of the ecosystem in the terrarium. On day 15, they could really be either, but I think that this question wants you to say, if they have everything they need, they'd be alive." After completing the cluster, when the interviewer asked the student if he used any information from the left side of the screen, the student said, "I used a lot of information from the left side of the screen because in terrarium 4 they stayed alive for 15 whole days, and just having soil, plants and water was not on that chart, but I bet they had it. I thought, since they stayed alive on that one, they'd stay alive in this one."
>
> Another student used the data from the terrarium experiment but without seeming to comprehend how to interpret the data. He said, "What I found confusing was on [day] 5 that [the terraria] were tied, and that 2 of them were alive and 2 of them were not alive. So that made it really confusing since I didn't know which one to choose."

At least 10 students, however, including some of those who earned credit, used only their prior content knowledge and/or personal experience to respond.

> For example, one student said, "Day 1: *alive*. I think I'll put *alive*. My plants have been alive for 2 weeks." She clicked *Alive* for days 1, 5, and 10. "*Alive*. I don't know if they're going to be alive so I'm going to try *Not Alive* (clicked *Not Alive* for day 15), I don't know. I've had tomatoes that lasted like months and months."

3.3 DETAILED DISCUSSION BY CLUSTER: MIDDLE SCHOOL

3.3.1 Cluster 1: Galilean Moons

Performance Summary

The median time to complete the Galilean Moons cluster was 10 minutes. Table 21 and Table 22 indicate the number of students attaining cluster total scores and item scores within the specified ranges, respectively.

Table 21. Number of Students Attaining Cluster Total Scores in Specified Range: Galilean Moons

Score 9–7	Score 6–4	Score 3–1	Score 0
5	4	3	0

Note. Maximum score = 9; *n* = 12.

Table 22. Number of Students Attaining Item Scores in Specified Range, by Item: Galilean Moons

	Maximum Item Score	Score 4–3	Score 2–1	Score 0
Item 1	4	7	1	4
Item 2	4	7	4	1

	Maximum Item Score	Score 1	Score 0
Item 3	1	3	9

Note. n = 12.

Task Demands

The following are task demands of the Galilean Moons cluster:

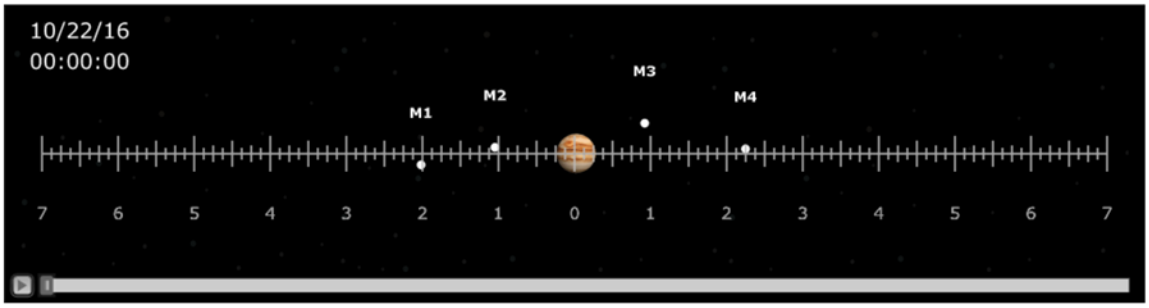
- Make simple calculations using given data to estimate the properties (e.g., mass, surface temperature, diameter) and locations of different solar system objects relative to a given reference point/object (Item 1).
- Calculate or estimate or identify properties of objects or relationships among objects in the solar system, based on data from one or more sources (Item 2).
- Given a partial model of objects in the solar system, identify objects or relationships that can be represented in the model or the reasons why they cannot be represented in the model (Item 3).

Stimulus

The stimulus for the Galilean Moons cluster is shown in Figure 17.

Figure 17. Stimulus: Galilean Moons

Four of Jupiter's closest moons can be seen orbiting the planet by using a low-powered telescope. A ruler on the lens of the telescope is used to take measurements. The animation shows the movements of the moons and Jupiter over the course of several days. Click on the small gray arrow at the bottom left of the picture to begin the animation.



The table shows data on each of the moons.

Data on Galilean Moons

	Diameter (km)	Mean Distance from Jupiter (km)	Orbital Period (days)
Callisto	4,800	2,000,000	16.7
Europa	3,318	700,000	3.5
Ganymede	5,262	1,000,000	7.2
Io	3,630	400,000	1.8

Details by Item

Item 1

Item 1 of the Galilean Moons cluster is shown in Figure 18.

Figure 18. Item 1: Galilean Moons

Use the measuring tool on the animation to determine each moon's maximum distance from Jupiter.

Complete the table by entering the measurements to the closest 0.25 mark.

	Maximum Distance from Jupiter in Animation
M1	
M2	
M3	
M4	

SCORES

This item was relatively easy for students; six students earned 4 score points (full credit), and one other student earned 3 score points. However, four students earned no credit (including one student who skipped over the item without attempting to answer it).

Eight of the 12 students seemed comfortable manipulating the simulation and re-watched, with appropriate pauses, to figure out each moon's distances from Jupiter. Some also re-watched the simulation while responding to Item 2.

One student neglected to watch the simulation at all.

COMPREHENSION

Although, the introduction to the stimulus states that "A ruler on the lens of the telescope is used to take measurements," five students did not understand the measuring tool, or the units used on the tool.

> One of these students used the mean distance from Jupiter in kilometers from the *Data on Galilean Moons* table for her responses to the item. The student said that the instructions suggested using a measuring tool, but she did not see a measuring tool.
>
> Another student said, "I thought the numbers [going across the lens on the animation] were extremely confusing. I think that if they're trying to take it to orbital days, then they have to make the length longer, but if it takes 16.7 days—well that's orbit. I don't know, it's just super confusing. They should say that the numbers represent the length of time or the number of days."

At least two students were confused by the instructions "to the closest 0.25 mark."

REASONING

The seven students who earned three or 4 score points all showed evidence in the think-aloud of using the animation in the manner intended to formulate their response.

> For example, one student said that she was going to follow one moon at a time "because I can't follow all of them at the same time." As she watched the animation a second time, she noted where each of the moons was, narrating aloud, "M2 is around the 1.5 mark. M4 is around the 2.5 mark." She then paused the video, studied the text of Item 1, and began entering the data. When she reached the response field for M3, she said, "I'll just leave it at 7, because it went a little past 7 but not too far."

Item 2

Item 2 of the Galilean Moons cluster is shown in Figure 19.

Figure 19. Item 2: Galilean Moons

SCORES

This item was also relatively easy for students; seven students received full credit (4 score points), and only one student received no credit.

COMPREHENSION

No features of this item appeared to confuse students.

REASONING

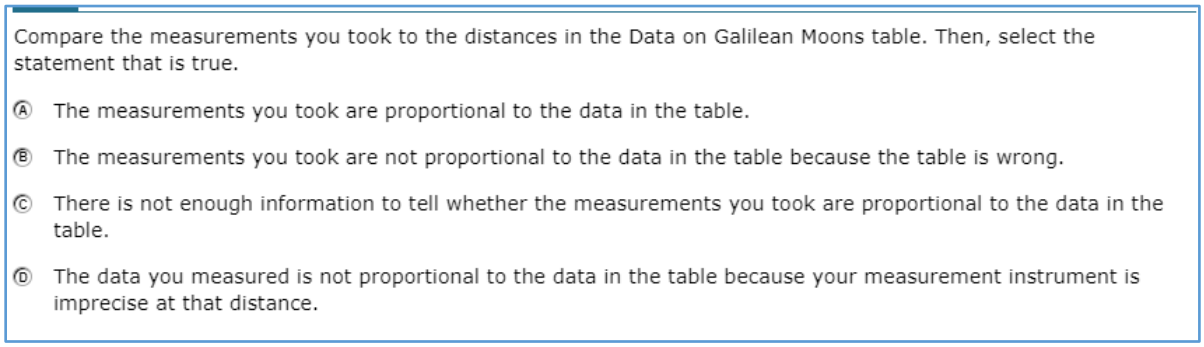
Nearly all the students reasoned their way to a solution using the stimulus materials as intended.

> For example, one student stated she was going to look for the mean distance from Jupiter [on the *Data on Galilean Moons* table] and use what she got from the previous question—the maximum distance for each moon. The student selected M3 for Callisto "because it is the farthest away and has the largest mean distance." She noted that Europa has the third "biggest" mean and, looking for the third largest maximum distance, deduced that M4 must be Europa. Seeing that Ganymede has the second largest mean distance, the student selected M1. The last moon left (Io) was identified by default as M2.

Item 3

Item 3 of the Galilean Moons cluster is shown in Figure 20.

Figure 20. Item 3: Galilean Moons



SCORES

This item was much more challenging than the other items in the cluster, and only three students selected the correct response that the data the student measured are not proportional to the data in the table due to the differences in measurement accuracy.

The nine students who did not earn credit for this item were fairly evenly distributed across the distractors (four students chose C, three chose A, and two chose B), suggesting that they really were at a loss to understand how to explain the differences between their measurements and the data in the table.

COMPREHENSION

Two students said that they did not know the meaning of "proportional," and, based on the item responses, it's likely that a number of others did not fully understand the concept of proportional.

Although not mentioned, students may also not have understood what it meant that "your measurement instrument is imprecise."

REASONING

Even students who selected the right answer, may not have done so with full comprehension.

> For example, one student read through all the answers, then started eliminating answers. First, she eliminated A and B, then decided the answer was D because the ruler measured the distance in the animation, but the table gave the distances in kilometers.

3.3.2 Cluster 3: Hippos

Performance Summary

The median time to complete the Hippos cluster was 10 minutes. Table 23 and Table 24 indicate the number of students attaining cluster total scores and item scores within the specified ranges, respectively.

Table 23. Number of Students Attaining Cluster Total Scores in Specified Range: Hippos

Score 10–7	Score 6–4	Score 3–1	Score 0
2	5	3	0

Note. Maximum score = 10; *n* = 10; two students ran out of time before completing this cluster.

Table 24. Number of Students Attaining Item Scores in the Specified Range, by Item: Hippos

	Maximum Item Score	Score 4–3	Score 2–1	Score 0
Item 1	4	1	9	0
Item 5	3	1	4	5

	Maximum Item Score	Score 1	Score 0
Item 2	1	5	5
Item 3	1	7	3
Item 4	1	3	7

Note. *n* = 10; two students ran out of time before completing this cluster.

Task Demands

The following are task demands of the Hippos cluster:

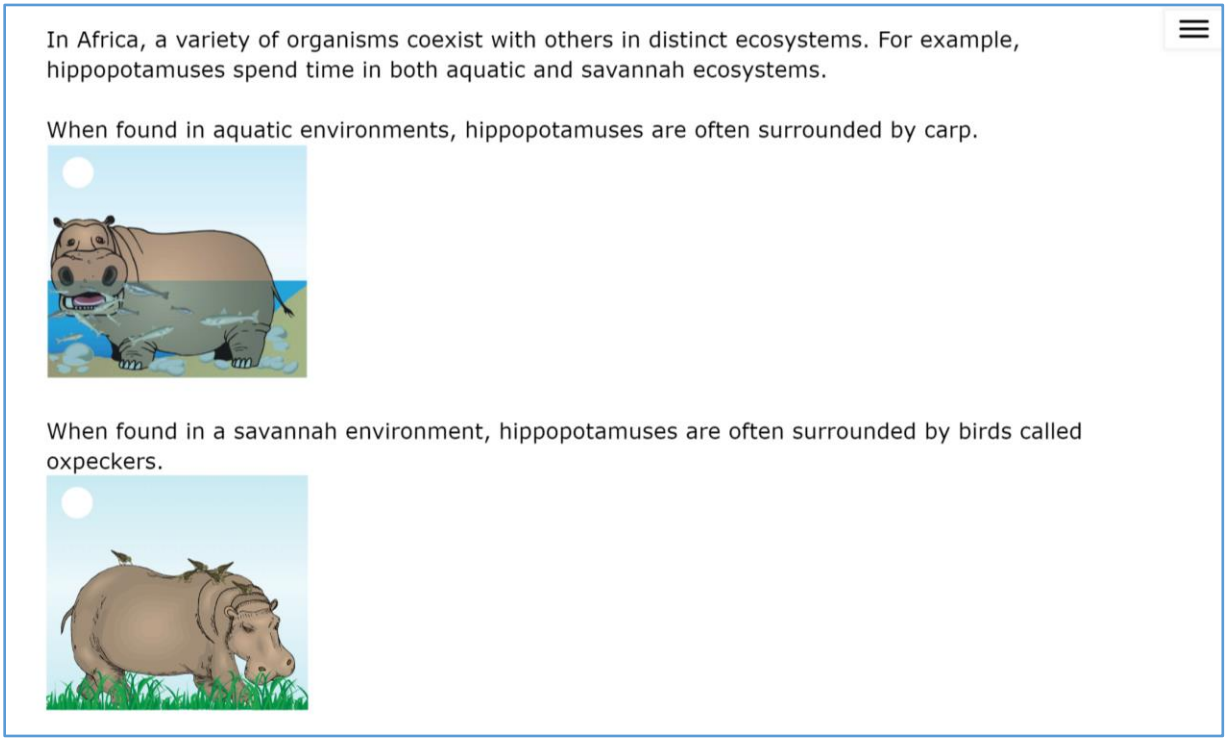
- Articulate, describe, illustrate, or select the relationships or interactions to be explained. This may entail sorting relevant from irrelevant information or features (Item 1).
- Express or complete a causal chain common or distinct across organisms or environments. This may include indicating directions of causality in an incomplete model such as a flow chart or diagram or completing cause and effect chains (Item 2).
- Express or complete a causal chain common or distinct across organisms or environments. This may include indicating directions of causality in an incomplete model such as a flow chart or diagram or completing cause and effect chains (Item 3).

- Articulate, describe, illustrate, or select the relationships or interactions to be explained. This may entail sorting relevant from irrelevant information or features (Item 4).
- Use an explanation to predict interactions among different organisms or in different environments (Item 5).

Stimulus

The stimulus for the Hippos cluster is shown in Figure 21.

Figure 21. Stimulus: Hippos

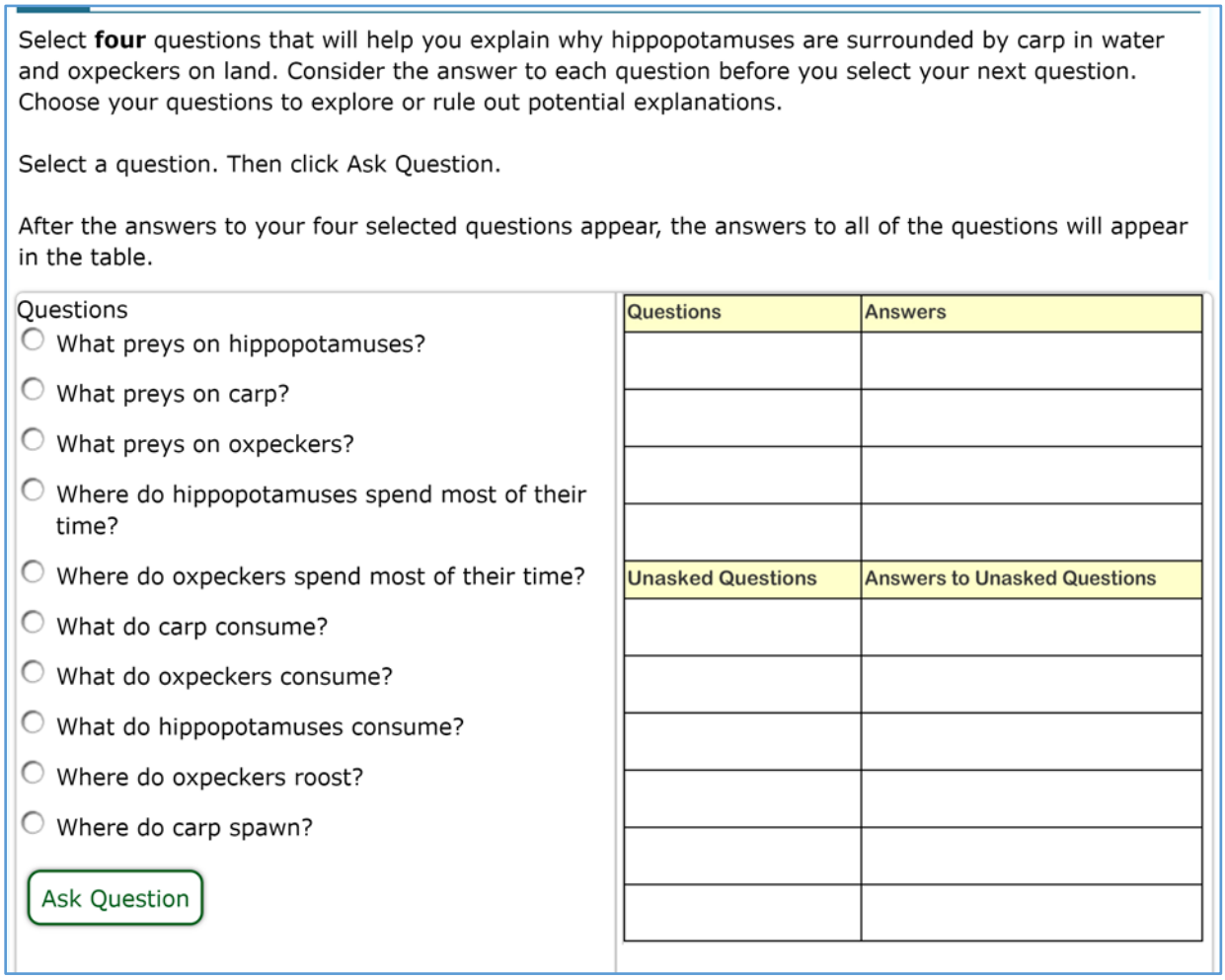


Details by Item

Item 1

Item 1 of the Hippos cluster is shown in Figure 22.

Figure 22. Item 1: Hippos



SCORES

Every student earned some credit on this item:

- One student earned 4 points (full credit).
- Three students earned 3 points.
- Six students earned 2 points.
- One student earned 1 point.

COMPREHENSION

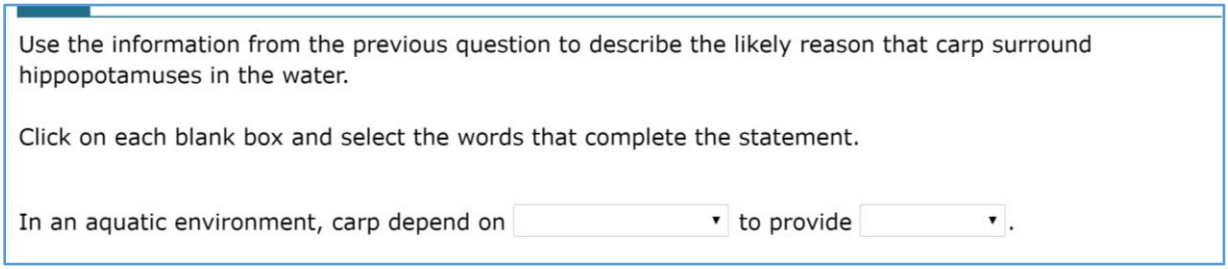
As evidenced from their reasoning in the think-alouds, students understood that they were to choose questions they thought would be helpful to explain the relationships between hippos and oxpeckers or carp, although, as can be seen from the score distribution, they did not necessarily know what those questions would be. Two students, however, commented on the fact that being asked to choose questions seemed like a waste of time in light of the fact that answers eventually were populated for all the questions.

Three students did not initially understand that they had to click "Ask Question" and could only ask one question at a time; one student initially thought that she had to type the text of the question rather than select from the list.

Item 2

Item 2 of the Hippos cluster is shown in Figure 23.

Figure 23. Item 2: Hippos



SCORES

Half of the students (five) received credit for this item.

COMPREHENSION

Students found this item easy to comprehend, and they had sufficient knowledge of transactional relationships among animals to understand the concept behind the item.

Score variance on this item (and the next) came from the "to provide" response; students found it obvious that the response for the first drop-down box should be Hippopotamuses.

REASONING

Most students reasoned appropriately from the information in Item 1 to determine their response.

> For example, one student said, "In an aquatic environment, carp depend on . . . so why would a carp depend on the hippopotamus? [Referring back to question 1:] So what preys on hippos? I don't need that. Where do they spend their time? I don't need that. Where do oxpeckers spend most of their time? On the bodies of host mammals. What do hippos consume? Grass and plants. Where do oxpeckers roost? On the bodies of host mammals. Oh, so I believe that in the aquatic environment, carp depend on hippos to provide . . . food . . . Because they eat fleas, dead skin, parasites, and mucous."

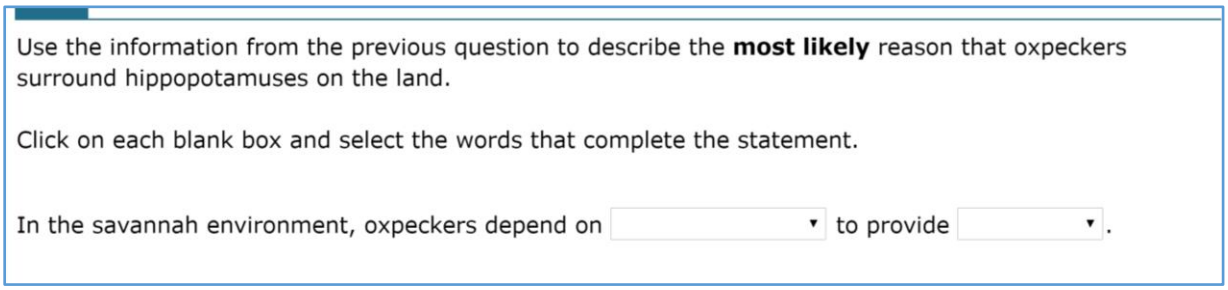
Those who did not respond correctly simply made wrong inferences from the data—some of which were wrong but plausible.

> For example, one student explained why he selected protection by saying, "hippopotamuses are a much bigger animal than the fish and could provide protection from the crocodile." The student noted that, in Item 1, one of the answers indicated that crocodiles, snakes and larger fish prey on carp.

Item 3

Item 3 of the Hippos cluster is shown in Figure 24.

Figure 24. Item 3: Hippos



Scores

Seven students received credit for this item.

Comprehension

This item is very similar to Item 2, and the same observations about comprehension apply.

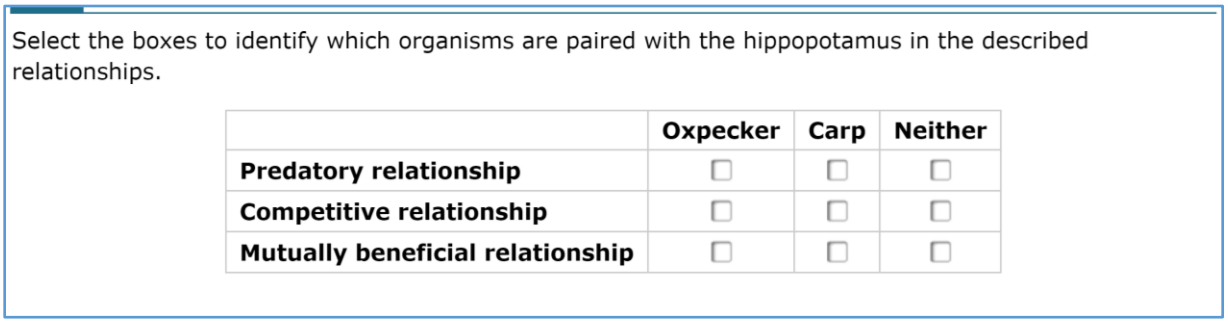
Reasoning

This item is very similar to Item 2, and the same observations about reasoning apply.

Item 4

Item 4 of the Hippos cluster is shown in Figure 25.

Figure 25. Item 4: Hippos



SCORES

Three students earned credit on this item, which required that all three answers about organisms in relationships with hippos be correct. The fewest students (two) correctly identified the answer for *Competitive relationship*.

COMPREHENSION

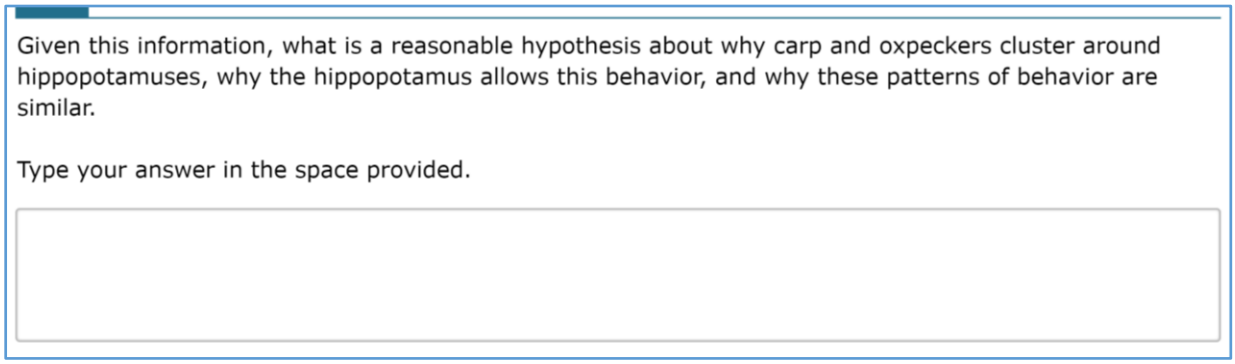
Although students generally understood the concept of transactional relationship among animals, some lacked prior knowledge of the terms used in the item.

> For example, one student said that "mutually beneficial" was the only relationship mentioned in the sample lesson. He did not know if the predatory and competitive relationships were "interchangeable or how it worked."

Item 5

Item 5 of the Hippos cluster is shown in Figure 26.

Figure 26. Item 5: Hippos



SCORES

One student earned full credit (3 score points) by providing correct hypotheses for each of the three questions posed in the item stem.

Four other students provided a correct hypothesis for at least one of the questions.

COMPREHENSION

There were no comprehension issues with this item.

REASONING

Some students failed to address the task of formulating hypotheses altogether. Others made appropriate use of the information gathered from the previous items in formulating their responses, but, given that their understanding of the previous items was not necessarily correct, these misunderstandings could carry over into this item.

3.3.3 Cluster 3: Morning Fog

Performance Summary

The median time to complete the Morning Fog cluster was 12 minutes. Table 25 and Table 26 indicate the number of students attaining cluster total scores and item scores within the specified ranges, respectively.

Table 25. Number of Students Attaining Cluster Total Scores in Specified Range: Morning Fog

Score 9–7	Score 6–4	Score 3–1	Score 0
2	3	7	0

Note. Maximum score = 9; n = 12.

Table 26. Number of Students Attaining Item Scores in Specified Range, by Item: Morning Fog

	Maximum Item Score	Score 7–6	Score 5–3	Score 2–1	Score 0
Item 1 (Parts A–C)	7	0	10	2	0

	Maximum Item Score	Score 2	Score 1	Score 0
Item 1 (Part D)	2	3	0	9

Note. n = 12.

Task Demands

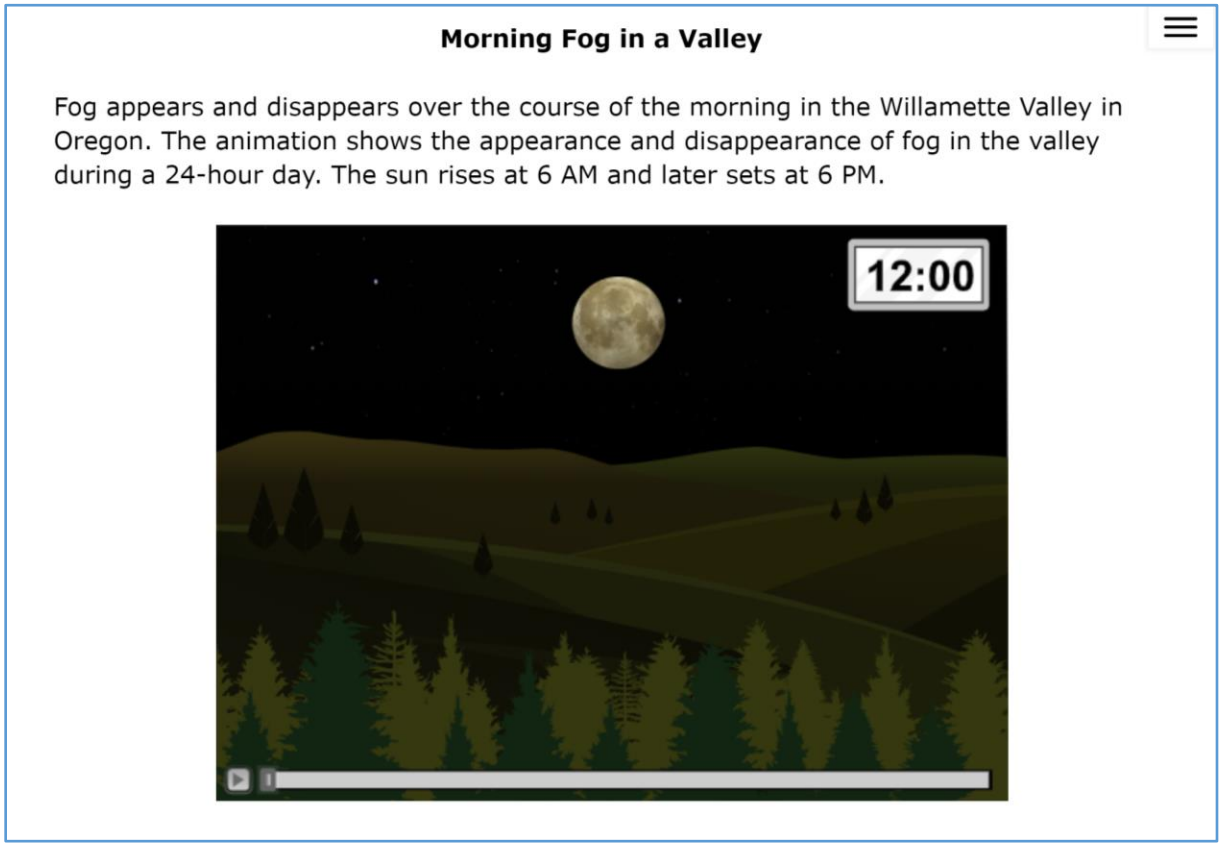
The following are task demands of the Morning Fog cluster:

- Select or identify from a collection of potential model components including distractors, the components needed to model the model of evaporation, condensation, transpiration, precipitation, or other behaviors of water molecules during the water cycle.
- Assemble or complete, from a collection of potential model components, an illustration or flow chart that represents the phenomenon. This does not include labeling an existing diagram.
- Given models or diagrams of the phenomenon, identify the parts of the model and how they change in each scenario OR identify the properties of the model that cause the change.

Stimulus

The stimulus for the Morning Fog cluster is shown in Figure 27.

Figure 27. Stimulus: Morning Fog

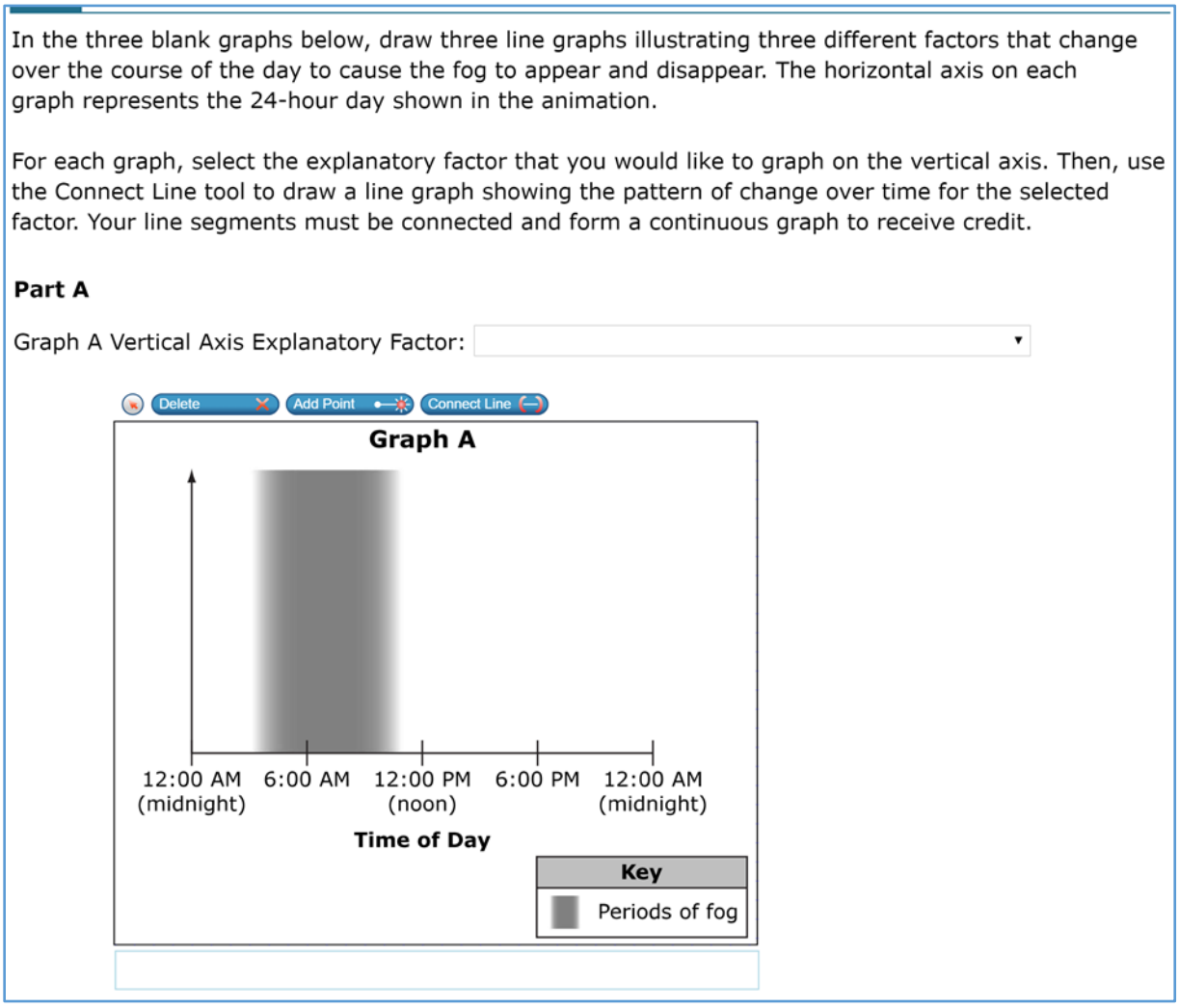


Details by Item

Item 1

Item 1 of the Morning Fog cluster is shown in Figure 28.

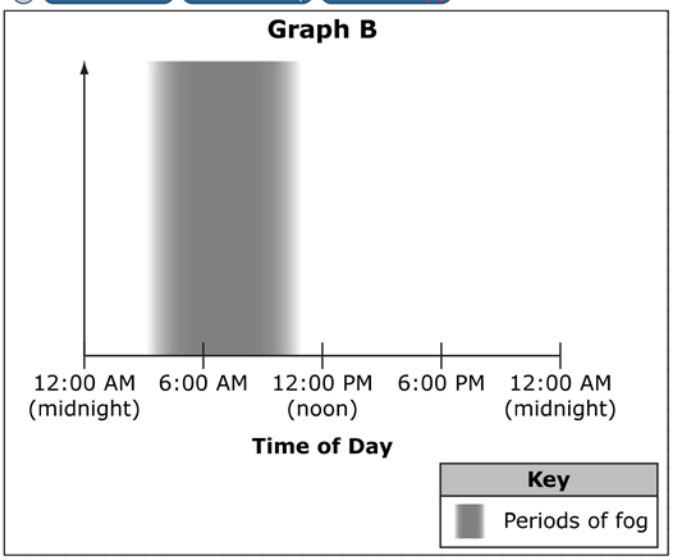
Figure 28. Item 1: Morning Fog



Part B

Graph B Vertical Axis Explanatory Factor:

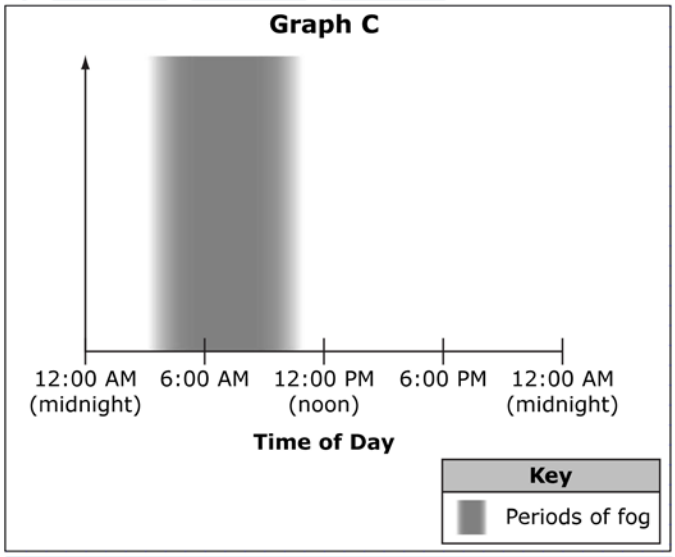
Delete Add Point Connect Line



Part C

Graph C Vertical Axis Explanatory Factor:

Delete Add Point Connect Line



Part D

The process described in ▾ causes the process described in ▾, which causes the process described in ▾.

Item 1 (Parts A–C)

SCORES

Parts A–C were scored as a unit.

Students could earn up to 6 points for correctly drawing three-line graphs showing how weather factors affecting fog formation changed over the course of the day; they could earn up to 3 points for correctly identifying the explanatory factor associated with each of the processes they chose to graph.

Half of the students (six) earned some credit for their graphs, but none earned full credit.

- Six earned points for graphing a decrease in the evening in one or more of the following: sunlight intensity, temperature, and/or proportion of water in the air
- Six earned points for graphing sunlight intensity, showing both an increase in the morning and a decrease in the evening.

No one earned points for graphing either the proportion of water in the air declining as the fog forms and increasing as the fog dissipates, or the temperature decreasing when the fog begins to form and rising when the fog dissipates.

Four students did not earn any credits for their graphs, and their graphs did not resemble the correct answers: they included horizontal lines, a single line that ascended, and dots with no connecting line.

All but two of the students earned at least two out of the three possible score points for the explanatory factors. The numbers of students earning points for correctly identifying each explanatory factor were as follows:

- Sunlight intensity (nine students)
- Air temperature (eight students)
- Proportion of water in the air in gas form (nine students)

COMPREHENSION

Eight students were confused about how to draw the line graphs, including four who did not understand that they had to define the value of the y-axis. The following are examples of think-alouds from students who were confused by the graphs:

- "I have no idea. I don't understand this graph. It's confusing. Since there's nothing on the left, the vertical. (referring to the y-axis). The three factors that can change, I have no idea what they mean by that. I feel like they're not giving enough information for me to understand. I'm so confused. The three different factors are what—the nighttime? What's the difference between the graphs? Wouldn't they all be the same? Oh, three different factors." (The student apparently didn't see the explanatory factor drop-down menu until this point.)

- The student re-read the part of the question that discusses "showing the pattern of change over time for the selected factor" and commented, "yeah, that really doesn't make sense, how they want me to connect the line. If I saw this on a test, I would just freak out because I wouldn't know how I was supposed to draw a line graph to represent this."

- "How do you represent how much fog? I'm guessing"—the student clicked to create some points—"I'm guessing it'd be something like that." The student clicked around some more and then connected the points. "I guess that's what I'm gonna say, because this really doesn't make sense how they want you to draw a graph. If anything, they should have increments and a chart of how high the fog rises or how much of whatever is in the air."

Six students were initially unclear about how to use the pull-down menu of explanatory factors, but mostly figured out how to use them.

Two students had a somewhat better understanding of Parts A–C after they read Part D and went back and changed some of their answers in Parts A–C.

> For example, after reading Part D, one student realized that each graph was meant to represent a different factor. When asked, the student said that he misunderstood the question and picked the same factor for all three graphs at first because he didn't know what was meant by the term "explanatory factor," and thought the question was just asking about the fog.

REASONING

Half of the students (six) re-watched the animation while drawing the line graphs.

An example of correct reasoning from the animation comes from the student who earned the most score points on parts A–C (7 points). She indicated that she chose Proportion of Water in the Air for her first graph because it was "the one that related to the fog the most." When asked to explain more about her graph, the student said she looked at the animation "to see the intensity of the fog and when it decreased" and that's why she made the graph increasing then decreasing. "First increasing from 3 to 6 [A.M.], then decreasing from 6 to 8."

Item 1 (Part D)

SCORES

Only three students earned the two possible core points by correctly responding that variations in sunlight intensity affect air temperature, which, in turn, affects the proportion of water in the air in gas form (water cycle).

COMPREHENSION

Since most students were confused by Parts A–C, they also had trouble understanding what they were being asking to do in Part D.

3.3.4 Cluster 4: Texas Weather

Performance Summary

The median time to complete the Texas Weather cluster was 14 minutes. Table 27 and Table 28 indicate the number of students attaining cluster total scores and items scores within the specified ranges, respectively.

Table 27. Number of Students Attaining Cluster Total Scores in Specified Range: Texas Weather

Score 11–7	Score 6–4	Score 3–1	Score 0
0	4	8	0

Note. Maximum score = 11; *n* = 12.

Table 28. Number of Students Attaining Item Scores in Specified Range, by Item: Texas Weather

	Maximum Item Score	Score 8–7	Score 6–4	Score 3–1	Score 0
Item 1 (Part A)	8	0	2	8	2

	Maximum Item Score	Score 1	Score 0
Item 1 (Part B)	1	1	11
Item 2	1	4	6
Item 3	1	6	3

Note. n= 12 for Item 1, Parts A and B; 11 for Item 2, and 10 for Item 3. One student did not scroll down to Items 2 and 3, and one student gave up and refused to attempt Item 3.

Task Demands

The following are task demands of the Texas Weather cluster:

- Describe, illustrate, or select tools, locations, and/or methods to use in investigations of phenomena related to interactions of air masses. This should show how or where measurements will be taken (Item 1).
- Identify, select, or describe the relevance of particular data or sources relevant to the process of weather forecasting (Item 1).
- Predict the effects of given changes in the air masses' interactions on subsequent weather (Item 2).
- Predict the effects of given changes in the air masses' interactions on subsequent weather (Item 3).

Stimulus

The stimulus for the Texas Weather cluster is shown in Figure 29.

Figure 29. Stimulus: Texas Weather

The weather in Austin turned cold and wet around 3:00 p.m. yesterday. Following is the hour-by-hour weather report for Austin.

	Noon	1:00 PM	2:00 PM	3:00 PM	4:00 PM	5:00 PM
Temperature	80° F	75° F	70° F	68° F	66° F	65° F
Chance of rain	0%	30%	50%	95%	100%	100%
Humidity	80%	85%	88%	92%	95%	96%
Wind	SE 9 MPH	SE 10 MPH	SE 9 MPH	NW 12 MPH	NW 13 MPH	NW 12 MPH
Pressure	32.0 inHG	30.3 inHG	29.9 inHG	29.0 inHG	28.7 inHG	28.5 inHG

As you work through the following questions, you will gather the information needed to explain the cause of this weather pattern.

Details by Item

Item 1

Item 1 of the Texas Weather cluster is shown in Figure 30.

Figure 30. Item 1: Texas Weather

Part A

The following question has two parts. First, answer part A. Then, answer part B.

Use the simulator to take measurements that will help you determine what caused Austin's afternoon weather.

You will be scored on your selections, so be sure to:

- specify what you are looking for,
- use the appropriate tools to look for them,
- keep taking measurements until you know what caused the weather, and
- stop taking measurements when you have all the information you need.

You may take a maximum of 8 measurements.

Checking for a(n) Air Mass

Location 1

Time of day 3pm

Tool 1 Thermometer

Tool 2 Barometer

Take Measurement

Measurement Number	Location	Checking For	Time of Day	Temperature	Wind Speed	Wind Direction	Pressure

Part B

From the measurements that you have taken, indicate up to two measurements (by "Measurement Number" from the result table in the simulation) that provide sufficient evidence for the claim in the first column. Be sure to select "None" if the measurements do not provide sufficient evidence of a claim.

	1	2	3	4	5	6	7	8	None
A low pressure air mass moved west towards Austin.	☐	☐	☐	☐	☐	☐	☐	☐	☐
A high pressure front moved south towards Austin.	☐	☐	☐	☐	☐	☐	☐	☐	☐
A cold front moved north towards Austin.	☐	☐	☐	☐	☐	☐	☐	☐	☐
Precipitation moved into Austin from the east.	☐	☐	☐	☐	☐	☐	☐	☐	☐

Item 1 (Part A)

SCORES

Part A was extremely difficult for students, and the randomness of earned points across students suggests that none of the students really understood what they were supposed to do with the simulator, either because they didn't have the requisite content knowledge or they were confused by the manner in which the simulator was presented.

Four of the points in the scoring rubric for Part A involve the parameters that the student chooses for trials on the simulator or matching the right tools with the right parameters, but many students failed to change the parameter on successive trials and simply focused on manipulating the tools. Four students used air mass (the default) for all of their measurements, and two students used primarily air mass. Consequently, score points based on choice of parameter or match between parameter and tools may not be meaningful. That said,

- nine students earned 1 score point for selecting air mass as the parameter on at least one trial;
- no students earned a score point for matching the correct tools with air mass;
- no students earned a score point for selecting movement as the parameter; and
- two students earned a score point for matching the correct tools with movement on at least one trial.

The four remaining points for Part A were awarded for measuring the correct factor at the proper locations and/or time and for doing so using the correct tools.

- Three students earned a point for at least one trial checking for movement measured at locations 3, 4, or 5.
- A different student earned a point for at least one trial checking for air mass measured at 1 p.m. at locations 3, 4, or 5.

The criterion statements in this section of the rubric were inconsistent. The criterion on which three students earned a point was the most permissive in that it specified a location, but not a time.

COMPREHENSION

Seven students did not initially understand what actions they were supposed to take to run trials on the simulator. Seven other students were unfamiliar with some of the measuring tools and did not know what they measured. Another student took only one measurement because he did not understand how to take more measurements.

The instructions to "determine what caused Austin's afternoon weather" were too open ended for these students.

- At least three students noted that the answer choices in Part B would have given them an idea of how to tackle the problem if they had read Part B before working with the simulator.

- Two students earned the most credits on Part A (4 score points) by (1) checking for air mass and movement, (2) choosing wind vane and anemometer when checking for movement, and (3) conducting one trial for air mass measured at 1 p.m. at locations 3, 4, and 5. One of these students said she was confused and overwhelmed when probed about this item.
    - "There was no way I could read this and understand it, I'll just look back and forth between [the chart and the table]." The student explained, "I've never been good with weather – it doesn't make sense to me how everything works . . . I didn't understand the table – like how it correlated with what I was putting in [Part A]. I was overwhelmed with eight measurements because it said, 'Do Part A and then Part B,' so I was thinking okay, I should do Part A and then Part B. But then after I did Part B, I realized that I should have looked at Part B first so I would know what eight measurements to take! I didn't know the difference in what would show up on the table if I chose air mass, or movement, or precipitation. I just didn't understand what difference it would make in each choice I had."

REASONING

The other student who earned 4 score points on the item had a somewhat better understanding of how to use the simulator to find out what caused Austin's afternoon weather.

> In her think-aloud, the student said that she was going to take measurements first at Location 3 because it's most central. She chose 3 p.m. because that's when the weather turned cold and wet in Austin. She then changed the measurement to Location 4 because "it's closest to Austin and what the chart pertains to." Said she would leave the time as 3 p.m. as that's when it was cold and wet. She said she would use the anemometer and the thermometer. She clicked *Take Measurement*. She said she would check for precipitation but didn't see any tools that pertained. She then chose movement at Location 3, using a wind vane and an anemometer, to see if the wind was going in that direction.

Item 1 (Part B)

SCORES

Only one student got credit for Part B, and this may have been by chance, given that the student only earned one of the eight possible points on Part A.

COMPREHENSION

At least three students did not realize that the numbers 1 through 8 on Part B were the eight measurements they were allowed to take in Part A, and that they were to pick measurements that showed evidence for the claim in column 1.

REASONING

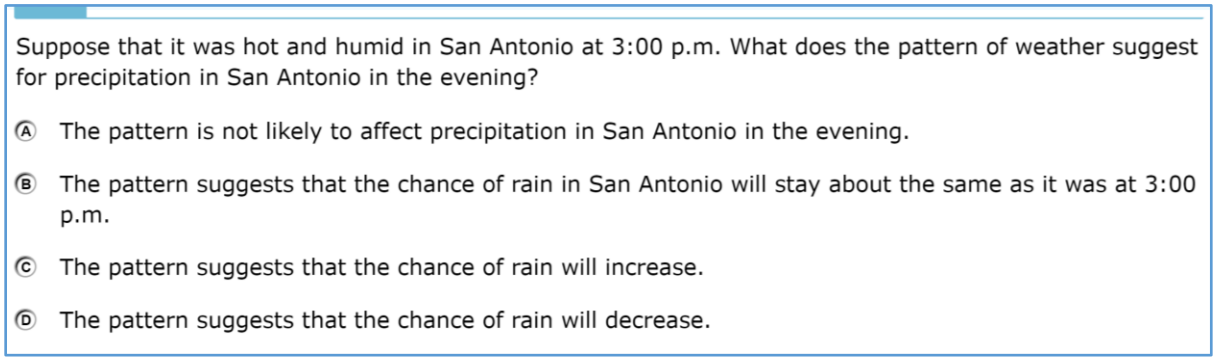
Given their performance on Part A, students had little to work with in Part B, even if they understood what they were supposed to do.

For example, one student said that she had to make her best guess in Part B because "none of my measurements in Part A told me anything because I took all the wrong measurements in Part A. Part B was truly kind of stressful for me."

Item 2

Item 2 of the Texas Weather cluster is shown in Figure 31.

Figure 31. Item 2: Texas Weather



SCORES

Four of the 10 students who attempted this item earned credit.

COMPREHENSION

Given performance on Item 1, it is unlikely that these students' scores actually reflected mastery of the content being assessed by the item.

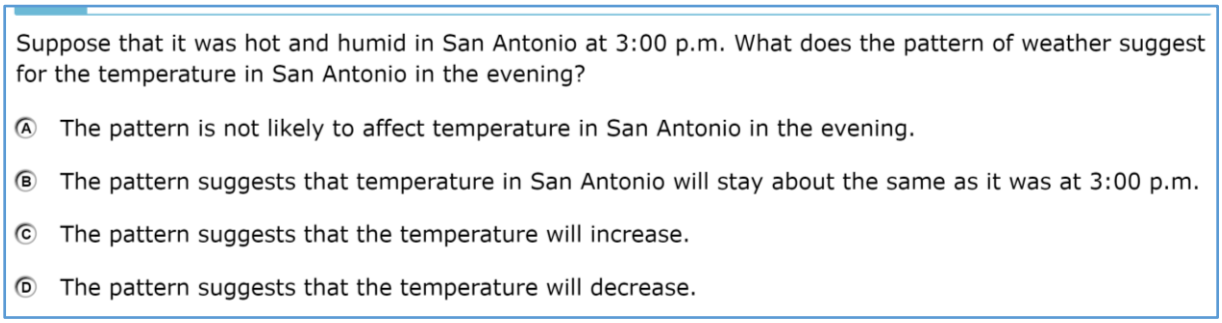
Some students understood "pattern of weather" as referring to the hour-by-hour weather report shown in the stimulus, and it's not clear that any of the students realized that the question pertained to a different location than the weather report (or Item 1).

For example, one student referred to the weather report table and said that the table indicates that the chance of rain will likely increase so he couldn't select decrease (pointing at both option A and option D). The student noted that option B suggests no change, but the table shows a very clear change in the chance of rain, therefore B could not be the answer. The student referred to the table again and said that the chance of rain was increasing, so C was the only possible answer that works.

Item 3

Item 3 of the Texas Weather cluster is shown in Figure 32.

Figure 32. Item 3: Texas Weather



SCORES

Six of the nine students who attempted this item earned credit.

COMPREHENSION

As with the other items in this cluster, students had, at best, a faulty understanding of this item. Consequently, as with Item 2, a correct response did not indicate mastery of the content being assessed.

> For example, one student said that, as soon as she read "temperature," she went to the weather report table, looked at the temperature at 3 p.m., and saw that the temperature was decreasing over time. The student then went back to the question and read through the options and noted that answer A was about no effect, that B was about staying the same, and C was about the temperature increasing. Since the temperature is decreasing, the student decided that answer D was the only one that matched the data.

3.4 DETAILED DISCUSSION BY CLUSTER: HIGH SCHOOL

3.4.1 Cluster 1: Blood Sugar Regulation

Performance Summary

The median time to complete the Blood Sugar Regulation cluster was 19 minutes. Table 29 and Table 30 indicate the number of students attaining cluster total scores and item scores within the specified ranges, respectively.

Table 29. Number of Students Attaining Cluster Total Scores in Specified Range: Blood Sugar Regulation

Score 7–6	Score 5–3	Score 2–1	Score 0
0	9	3	1

Note. Maximum score = 7; *n* = 13; two students ran out of time before completing this cluster.

Table 30. Number of Students Attaining Item Scores in Specified Range, by Item: Blood Sugar Regulation

	Maximum Item Score	Score 3	Score 2–1	Score 0
Item 1	3	8	4	1
Item 2	3	0	3	11

	Maximum Item Score	Score 2	Score 1	Score 0
Item 3	2	3	7	3

Note. *n* = 13; two students ran out of time before completing this cluster.

Task Demands

The following are task demands of the Blood Sugar Regulation cluster:

- Identify the outcome data that should be collected in an investigation to provide evidence that feedback mechanisms maintain homeostasis. This could include measurements and/or identifications of changes in the external environment, the response of the living system, stabilization/destabilization of the system's internal conditions, and/or the amount of systems for which data is collected.
- Make and/or record observations about the external factors affecting systems interacting to maintain homeostasis, responses of living systems to external conditions, and/or stabilization/destabilization of the system's internal conditions.

- Identify or describe the relationships, interactions, and/or processes that contribute to and/or participate in the feedback mechanisms maintaining homeostasis that lead to the observed data.
- Using the collected data, express or complete a causal chain explaining how the components of (a) mechanism(s) interact in response to a disturbance in equilibrium in order to maintain homeostasis. This may include indicating directions of causality in an incomplete model such as a flow chart or diagram or completing cause and effect chains.
- Evaluate the sufficiency and limitations of data collected to explain the cause and effect mechanism(s) maintaining homeostasis.

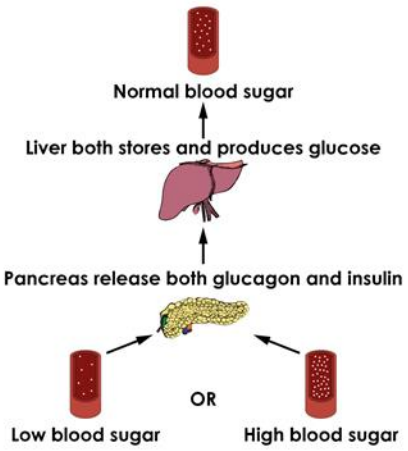
Stimulus

The stimulus for the Blood Sugar Regulation cluster is shown in Figure 33.

Figure 33. Stimulus: Blood Sugar Regulation

A hungry person eats a meal. Soon after the meal is completed, the person's blood sugar is elevated. After a while, the blood sugar levels return to their pre-meal levels.

Hunger is one of the body's symptoms of abnormal blood glucose levels, or blood sugar. Hunger alerts the body to eat, which almost immediately increases blood sugar. Both the pancreas and liver work together to maintain blood sugar concentrations in the range of 80-120 miligrams per deciliter (mg/dL). The pancreas helps regulate blood sugar by producing two types of hormones: glucagon and insulin. The normal range for blood glucagon levels is 60-200 picograms per mililiter (pg/mL) and the normal range for blood insulin levels is 65-200 picomole per liter (pmol/L). The liver both converts glucagon into glucose and stores glucose. The flowchart shows how the pancreas and liver participate in feedback mechanisms to help regulate blood sugar.



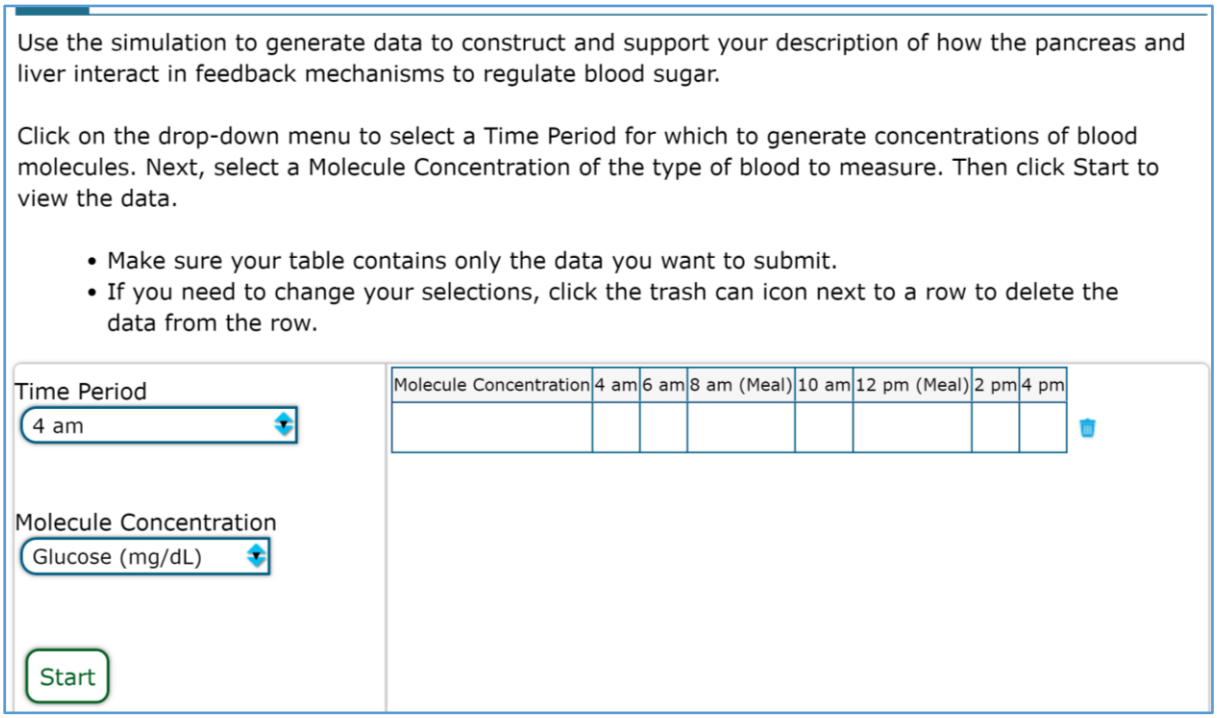
In the questions that follow, investigate and describe how the molecules produced and stored by the pancreas and liver interact in feedback mechanisms to regulate blood sugar.

Details by Item

Item 1

Item 1 of the Blood Sugar Regulation cluster is shown in Figure 34.

Figure 34. Item 1: Blood Sugar Regulation



SCORES

Student scores on this item are as follows:

- Eight students earned 3 score points (full credit).
- Three students earned 2 score points.
- Two students earned 1 score point.

COMPREHENSION

Seven students expressed some confusion in figuring out how to generate data in the simulation. For example, one student was confused by the layout of the item and by the term "simulation" because she was not sure whether she should test all the options or provide her own answer. At this point she skipped ahead to look at the next items to see if they would provide any clues as to how she should proceed on Item 1 but did not find that helpful. She was very unsure what to do next and seemed overwhelmed by the options. After some flipping back and forth, she decided to measure all three values for each of the times offered.

At least three students went back to Item 1 and re-generated the data in the simulation once they knew that they had to create three graphs in Item 2.

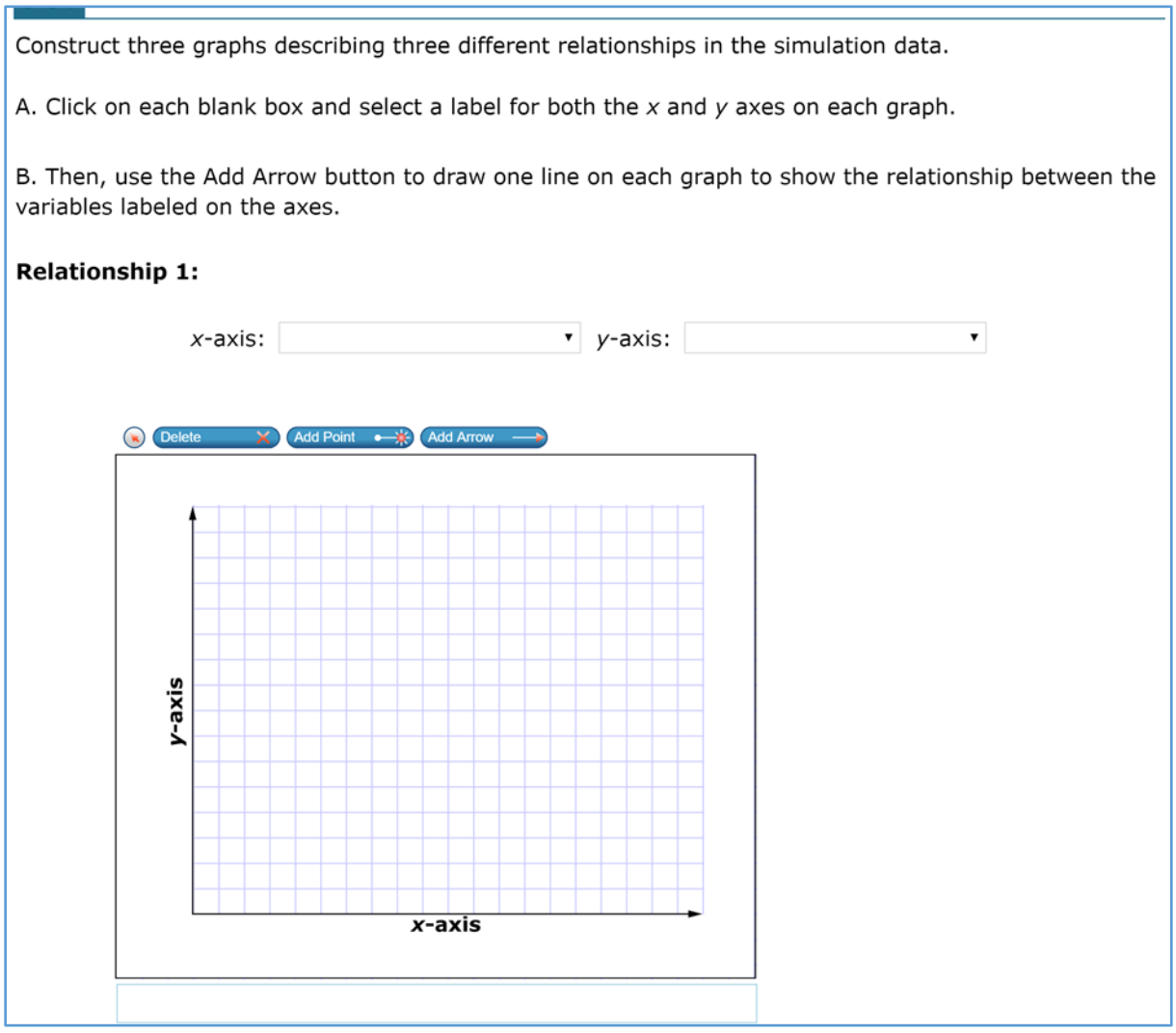
REASONING

Students used the simulations as a learning experience. For example, when asked how he decided how many simulations to do, one student said, "Well, I knew that there was three different substances (glucose, glucagon, and insulin). I wasn't really sure how it worked, and then once I did it, I was like 'OK well that's when you have a meal,' so I knew from the reading that's when your blood sugar spikes."

Item 2

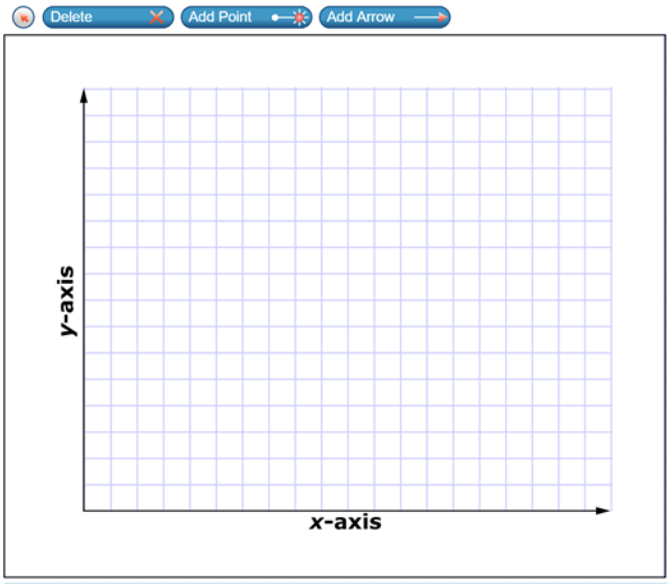
Item 2 of the Blood Sugar Regulation cluster is shown in Figure 35.

Figure 35. Item 2: Blood Sugar Regulation



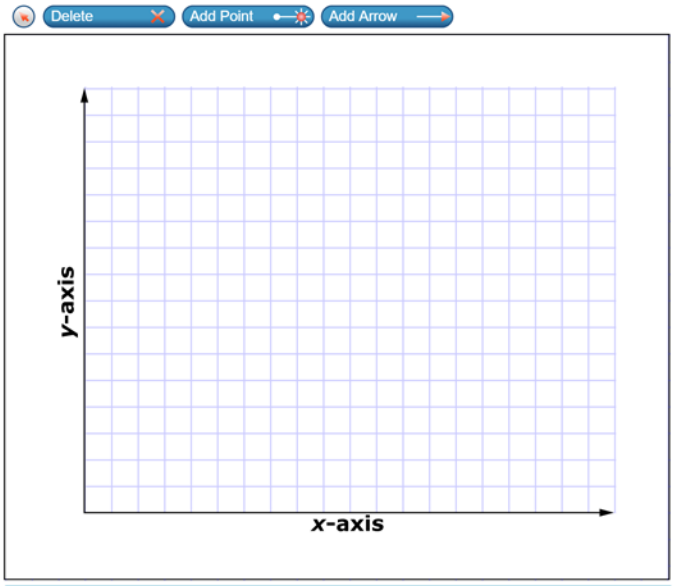
Relationship 2:

x-axis: ▾ *y*-axis: ▾



Relationship 3:

x-axis ▾ *y*-axis ▾



SCORES

Student scores on this item are as follows:

- No students earned 3 score points (full credit).
- Two students earned 2 score points.
- One student earned 1 score point.

COMPREHENSION

Eight students expressed some confusion as to how to construct the graphs of the simulation data. For example, one student was "kind of confused" about where to draw the second and third graphs. Initially she did not see the answer grids for the second and third graphs, but even after she noticed the additional answer grids, some confusion lingered.

At least five students were not sure how to represent the units or values on the graphs, and two students did not draw any graphs for that reason. For example, for the first relationship, one student chose glucose versus time for the first relationship, but he was not sure which value to put on which axis: "I've never looked at the concentration of molecules and tried to graph it, and I feel like there are a lot of things I'm missing to help me figure out what to do. I think I may be overcomplicating it to myself."

REASONING

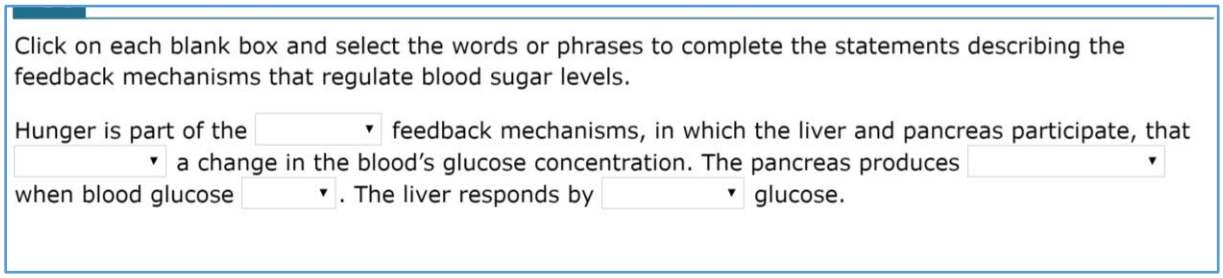
The following is an example of how one student reasoned through the construction of one of the graphs.

> The student said that he was going to place concentration on the x-axis and time on the y-axis because "in sciences you usually do time on the y-axis and concentration and stuff on the x-axis. I don't know why, it's what I've always known." He selected *Glucose Concentration* for the x-axis and *Time Passed after Eating* for the y-axis. He used the numbers for the glucose concentrations from the simulation in Item 1 to plot points on the graph. He said, "I feel like it spikes up like 5 times so I'll put it a decent amount, 6, 8 and then 10, and it kind of stays pretty high but not as high, so like right there, and then it drops a little bit again, and then it spikes up in a big lunge, and then it drops back down again to here, but it kind of stayed, and then it spiked the highest peak at dinner." He then started to connect the points, and said, "I don't know what the point of the arrows are, I'm just going to connect them all to show their relationship. That's my best guess to show what happened each hour."

Item 3

Item 3 of the Blood Sugar Regulation cluster is shown in Figure 36.

Figure 36. Item 3: Blood Sugar Regulation



SCORES

Student scores on this item are as follows:

- Three students earned 2 score points (full credit).
- Seven students earned 1 score point.
- Among these 10 students,
    - four earned a point for correctly filling the blanks in the statement about hunger; and
    - seven earned a point for correctly filling the blanks in the statement about the roles of the pancreas and the liver.

COMPREHENSION

No students expressed confusion about this item.

REASONING

In responding to Item 3, five students referred to the stimulus, and two students referred to the simulation results in Item 1.

3.4.2 Cluster 2: Saving the Tuna

Performance Summary

The median time to complete the Saving the Tuna cluster was 14 minutes. Table 31 and Table 32 indicate the number of students attaining cluster total scores and items scores within the specified ranges, respectively.

Table 31. Number of Students Attaining Cluster Total Scores in Specified Range: Saving The Tuna

Score 7–6	Score 5–3	Score 2–1	Score 0
1	2	5	4

Note. Maximum score = 7; *n* = 12; three students ran out of time before completing this cluster.

Table 32. Number of Students Attaining Item Scores in Specified Range, by Item: Saving the Tuna

	Maximum Item Score	Score 3	Score 2–1	Score 0
Item 1 (Part A)	3	0	6	6

	Maximum Item Score	Score 1	Score 0
Item 1 (Part B)	1	6	6
Item 1 (Part C)	1	1	11

	Maximum Item Score	Score 2	Score 1	Score 0
Item 2 (Part A and B)	2	3	0	9

Note. n = 12; three students ran out of time before completing this cluster.

Task Demands

The following are task demands of the Saving the Tuna cluster:

- Articulate, describe, illustrate, or select the relationships, interactions, and/or processes to be explained. This may entail sorting relevant from irrelevant information or features.
- Express or complete a causal chain explaining how human activity impacts the environment. This may include indicating directions of causality in an incomplete model such as a flow chart or diagram or completing cause and effect chains.
- Identify evidence supporting the inference of causation that is expressed in a causal chain.

- Use an explanation to predict the environmental outcome given a change in the design of human technology.
- Describe, identify, and/or select information needed to support an explanation.

Stimulus

The stimulus for the Saving the Tuna cluster is shown in Figure 37.

Figure 37. Stimulus: Saving the Tuna

Saving the Tuna

North Atlantic bluefin tuna are one of the most prized fish in danger of overfishing. One 342 kilogram (kg) tuna sold for close to $400,000 dollars at a fish market in Tokyo.

Bluefin tuna are the apex predators in their ecosystem. They hunt, travel, and live within schools, or large groups, of other bluefin tuna individuals. Bluefins start out as extremely tiny larvae, no more than a few millimeters long, and weigh only a few hundredths of a gram. Within three to five years, sexually mature adults can reach lengths of three feet (about one meter) and can weigh over 600 kg. As adults, they can dive as deep as 914 meters and can swim very long distances in the open ocean during migration season. Their migration season spans from approximately May to June, during which they spawn near the Gulf of Mexico.

Because bluefin are prized fish that vary greatly in size and can be found in schools, or groups, within a wide range of water depths, netting fishing methods are commonly used to target and catch these individuals. However, fishing nets often catch bycatch individuals, or non-tuna individuals. The table summarizes several netting fishing methods and the relative amounts of targeted tuna and bycatch individuals caught at one time by each method.

Summary of Netting Fishing Methods

Method	Description	Type of Targetted Catch	Total Number of Individuals Caught at a Time	Percent of Total Catch that is Bycatch (%)	Types of Bycatch Caught
Purse Seining	Large wall of netting that herds fish together and then envelops them when the net is pulled by a drawstring	Schooling or spawning fish	Hundreds to thousands	35 - 70	Sea turtles, dolphins, and other fish
Cast Netting	Small-meshed netting cast from shore or canoes that expands a relatively small area	Groups of small fish	Up to a hundred	10 - 30	Other small fish
Gilnetting	Large curtains of netting suspended by a system of floats and weights that can either be anchored to the seafloor or allowed to float at the surface	All types of fish	Hundreds to thousands	40 - 75	Sea birds, sea turtles, octopi, shark, dolphins, other fish, and crustacea
Midwater Trawling	Gigantic nets that span the size of five football fields pulled by large industrial ships through the open ocean, catching entire schools of fish	All types of open-ocean fish	Thousands to tens of thousands	30 - 75	Sea turtles, shark, dolphins, and other fish
Seine Netting	Small-meshed netting suspended vertically by floats and weights from the surface of intertidal water to enclose and concentrate fish	Crustacea and shell fish	Less than a hundred	10 - 30	Sea birds and other small fish

Your task is to design, evaluate, and refine solutions for reducing the impacts of human fishing on the population of tuna and other native species in the Northern Atlantic Ocean.

Details by Item

Item 1

Item 1 of the Saving the Tuna cluster is shown in Figure 38.

Figure 38. Item 1: Saving the Tuna

The following question has three parts. First, answer part A. Next, answer part B. Then, answer part C.

Part A

Select the boxes to evaluate the tradeoff considerations of each fishing method.

- You may select more than one method per column.

	Likely to Catch the Greatest Number of Tuna Individuals	Likely to Catch the Least Number of Tuna Individuals	Likely to be the Best at Targeting Tuna Individuals	Likely to be the Worst at Targeting Tuna Individuals	Likely to be the Best at Protecting Biodiversity of Ecosystem	Likely to be the Worst at Protecting Biodiversity of Ecosystem
Purse seining	☐	☐	☐	☐	☐	☐
Cast netting	☐	☐	☐	☐	☐	☐
Gilnetting	☐	☐	☐	☐	☐	☐
Midwater trawling	☐	☐	☐	☐	☐	☐
Seine netting	☐	☐	☐	☐	☐	☐

Part B

Based on the evaluation of tradeoff considerations in part A, which fishing method best limits the negative effects of human fishing on non-tuna populations in the Northern Atlantic?

Ⓐ purse seining

Ⓑ cast netting

Ⓒ gilnetting

Ⓓ midwater trawling

Ⓔ seine netting

Part C

Click on each blank box and select a word or phrase to complete a statement describing a change that can be made to decrease the amount of bycatch for the method identified as the worst in targeting tuna individuals in part A.

▾ the ▾ will improve the targeting of bluefin tuna.

Item 1 (Part A)

SCORES

Student scores on this item are as follows:

- No students earned 3 score points (full credit).
- Two students earned 2 score points.
- Four students earned 1 score point.
- Six students earned no score points.

COMPREHENSION

Several students expressed confusion with different aspects of this sub-question including

- completely missing two of the columns in the *Summary of Netting Fishing Methods* table, which was a critical reference for this sub-question; and
- confusion with the response-entry table, including overlooking the instructions stating that it was permissible to select more than one method for each column.

REASONING

All students methodically navigated through the response-entry table and used the *Summary of Netting Fishing Methods* chart in the stimulus to figure out their responses. For example:

- One student first lined up the *Summary of Netting Fishing Methods* chart next to the response-entry table so that he could read the descriptions easily and fill out the table. For the first column (*Likely to Catch the Greatest Number of Tuna Individuals*), the student said, "The first one I will cancel out will be *cast netting* because it says up to 100, and also *seine netting* because that's less than 100. I would say *gillnetting* and *purse* [are] the two top because it says they catch up to 100s to 1,000s for both of those. Wait; sorry, I was reading that wrong. Okay, *midwater trawling* was 1,000s to 10,000s because that's what I was thinking instead of 100s to 2,000s, so *midwater trawling* will be my answer." The student continued in the same manner for each of the six columns.
- Not all the student's conclusions from the *Summary of Netting Fishing Methods* chart were correct, however, probably because of deficiencies in the student's knowledge about ecology. For example, for column 5 (*Likely to be the Best at Protecting Biodiversity of Ecosystem*), the student said, "I would say both *gillnetting* and *midwater trawling* because they both take all types of fish, they are not going after specific fish, which means that they're not taking one species of fish out of the water; they're taking multiple, so there's less chance of one fish being taken out of the ecosystem."

Item 1 (Part B)

SCORES

Six students earned credit on this sub-item.

COMPREHENSION

One student was confused, saying that she did not understand the question and she did not know about each type of net.

REASONING

In responding to this sub-item, four students referred to their responses in Part A, and four students referred to the *Summary of Netting Fishing Methods* chart.

Item 1 (Part C)

SCORES

One student earned credit on this sub-item.

COMPREHENSION

Several students clearly did not understand the sub-item and guessed on questionable grounds.

> For example, one student read out loud all of the options under the second drop-down menu and said that he did not really understand the question: "I'm confused because in re-reading the question, it makes it seems like it was asking which net would decrease the chance of getting a tuna, but re-reading the answer choices, it's not asking that as much as I thought it would be. So, I'm going to go with *decreasing* instead of *increasing* because it says decrease in the sentence, and then something about negatives."

Another student indicated that she initially thought the sub-item was looking for a change in any of the methods that would decrease the amount of tuna by catch. Later she realized that the sub-item was referencing something specific in Part A. She went through all the drop-down options and hesitated a lot over her answer, changing it several times.

REASONING

In responding to this sub-item, five students referred to their responses in Part A, and six students referred to the *Summary of Netting Fishing Methods* chart.

Item 2

Item 2 of the Saving the Tuna cluster is shown in Figure 39.

Figure 39. Item 2: Saving the Tuna

The following question has two parts. First answer part A. Then, answer part B.

Three solutions proposed by scientific and environmental organizations to protect and restore the Northern Atlantic bluefin tuna population are shown in the table.

Solutions to Protect and Restore the Bluefin Tuna Populations

Solution	Description
1	Completely restricting the catching of juvenile bluefin
2	Limiting the total number of adult bluefin that can be caught
3	Removing juvenile bluefin from the Northern Atlantic to raise in captivity

Part A

Which Bluefin characteristic serves as the criteria on which all three solutions are based?

Ⓐ body mass

Ⓑ body length

Ⓒ ability to reproduce

Ⓓ ability to dive for prey

Part B

Select the **two** netting characteristics that are most important to consider when designing fishing nets for use in implementing the three solutions.

☐ mesh size of the net

☐ overall size of the net

☐ ability of the net to move

☐ depth of the net's location within the water column

SCORES

Student scores on this item are as follows:

- Three students earned 2 score points (full credit).
- No students earned 1 score point.
- Nine students earned no score points.

- Part A contributed one-third of the weight to the total item score, and 11 students selected the correct response for Part A.
- Part B contributed two-thirds of the weight to the total item score. Students only received credit for Part B if they correctly identified two netting characteristics that are important to consider when designing fishing nets for use in implementing the three solutions. While only three students correctly selected both characteristics, seven other students correctly selected one of the characteristics (four selected the *depth of the net's location in the water* column, and three selected the *mesh size of the net* column).

COMPREHENSION

One student did not understand the term "mesh size." She understood mesh as a verb, e.g., "meshing things together."

REASONING

When responding to Part B, only one student referred to the *Solutions to Protect and Restore the Bluefin Tuna Populations* table included with the item; four students referred to the *Summary of Netting Fishing Methods* chart in the cluster stimulus, and two students referred to the text in the cluster stimulus.

The following is an example of how one student used the reference materials to draw two conclusions about how to design the net to protect and restore the tuna population. Rather than considering any of the solution strategies proposed in the cluster stimulus, the student seemed to focus on supporting a method that would selectively catch adult tuna rather than juveniles, but one of the net characteristics he identified (*depth of the net's location within the water colum*n) counted as correct.

> The student looked at the fishing method characteristics and said, "They're going to want to increase the depth of the net's location within the water column because the adults can dive as deep as 914 meters and can swim very long distances, so they're going to want to increase the depth and the overall size of the net to catch them." When asked where the student got the information to answer the question, the student said, "I looked at the top of the article where it says that they dive as deep as 914 meters and can swim very long distances in the open ocean. So, I said increase the overall size to make the catch wider so they can't swim outside of the range of the net and also increase the depth since they can go pretty low."

3.4.3 Cluster 3: Tomcods

Performance Summary

The median time to complete the Tomcods cluster was 17 minutes. Table 33 and Table 34 indicate the number of students attaining cluster total scores and item scores within the specified ranges, respectively.

Table 33. Number of Students Attaining Cluster Total Scores in Specified Range: Tomcods

Score 8–6	Score 5–4	Score 3–1	Score 0
0	1	9	4

Note. Maximum score = 8; n = 14; one student ran out of time before completing this cluster.

Table 34. Number of Students Achieving Item Scores in Specified Range, by Item: Tomcods

	Maximum Item Score	Score 5–4	Score 3–1	Score 0
Item 1 (Parts A–C)	5	0	2	12

	Maximum Item Score	Score 1	Score 0
Item 2 (Part A)	1	6	8
Item 2 (Part B)	1	0	14
Item 3	1	10	4

Note. n = 14; one student ran out of time before completing this cluster.

Task Demands

The following are task demands of the Tomcods cluster:

- Based on the provided data, identify, describe, or construct a claim regarding the effect of changes to the environment on (1) the increases in the number of individuals of some species, (2) the emergence of new species over time, and (3) the extinction of other species.
- Sort inferences about the effect of changes to the environment on (1) the increases in the number of individuals of some species, (2) the emergence of new species over time, and (3) the extinction of other species into those that are supported by the data, contradicted by the data, outliers in the data, or neither, or some similar classification.
- Identify patterns of information/evidence in the data that support correlative/causative inferences about the effect of changes to the environment on (1) the increases in the number of individuals of some species, (2) the emergence of new species over time, and (3) the extinction of other species.

- Construct an argument using scientific reasoning drawing on credible evidence to explain the effect of changes to the environment on (1) the increases in the number of individuals of some species, (2) the emergence of new species over time, and (3) the extinction of other species.
- Identify additional evidence that would help clarify, support, or contradict a claim or causal argument regarding the effect of changes to the environment on (1) the increases in the number of individuals of some species, (2) the emergence of new species over time, and (3) the extinction of other species.
- Identify, summarize, or organize given data or other information to support or refute a claim regarding the effect of changes to the environment on (1) the increases in the number of individuals of some species, (2) the emergence of new species over time, and (3) the extinction of other species.

Stimulus

The stimulus for the Tomcods cluster is shown in Figure 40.

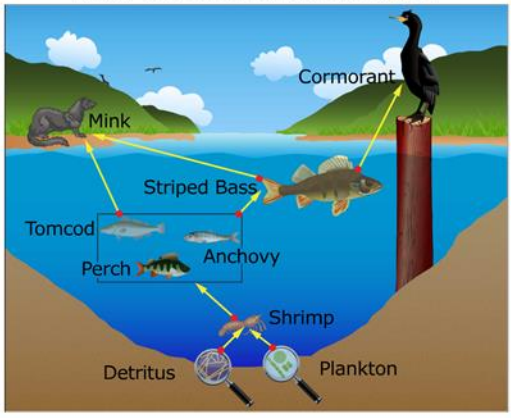
Figure 40. Stimulus: Tomcods

Atlantic Tomcod Thrive in Contaminated Hudson River

Polychlorinated biphenyls (PCBs) are chemicals that were produced from 1929 to 1979 for industrial and commercial uses. One electric company released 1.3 million pounds of PCBs into the Hudson River from 1947 to 1976. In 1979, PCBs were banned. However, the Hudson River still has high levels of PCBs today because they settle into sediments on the bottom and do not break down. When most fish embryos are exposed to PCBs, the immune system of the embryo is disrupted, causing the fish to develop smaller hearts that do not function properly, resulting in death. Many fish populations declined or disappeared from the Hudson River because of PCB exposure. However, one fish population, the Atlantic Tomcod, does not have this reaction to PCBs and thrives.

The picture shows a food web for the Hudson River. The liver of several aquatic species were tested for the presence of PCBs. The levels of PCBs in the livers of the tomcod were among the highest reported. Both striped bass and mink populations have also been found to have high levels of PCBs.

Food Web of the Hudson River



Tomcod were captured from the Hudson River and from rivers not contaminated by PCBs. The tomcod were tested for the AHR2 protein, which is responsible for regulating the toxic effects of PCB. The percentage of tomcod that contained the AHR2 protein mutation is shown in the table.

Percentage of Tomcod with AHR2 Protein Mutation

River	Percentage of Tomcod with Mutation
Hudson River, New York	99
Hackensack River, New Jersey	92
Niantic River, Connecticut	6
Shinnecock Bay, New York	5

Following are two hypotheses about the success of the tomcod in the contaminated Hudson River.

Hypothesis 1: The tomcod population did not decrease in response to PCB exposure because tomcod do not take in as many PCBs as other fish species through their food consumption or absorption from the water.

Hypothesis 2: The tomcod population did not decrease in response to PCB exposure because they have evolved resistance to the effects of PCBs through natural selection.

As you work through the questions, evaluate the evidence to determine which hypothesis of how the tomcods are able to overcome exposure to deadly PCBs is **best** supported.

Reference: Isaac Wirgin, et al. "....Atlantic Tomcod from the Hudson River." *Science* 331 (2011):1322–1325.

Details by Item

Item 1

Item 1 of the Tomcods cluster is shown in Figure 41.

Figure 41. Item 1: Tomcods

The following question has three parts. First, answer Part A. Next, answer part B. Then, answer part C.

Part A

Select the boxes to indicate whether each statement supports or refutes Hypothesis 1 or Hypothesis 2. You can select more than one box for each statement.

	Supports Hypothesis 1	Refutes Hypothesis 1	Supports Hypothesis 2	Refutes Hypothesis 2
There is a higher percentage of AHR2 protein mutations in the Hudson River than in rivers not contaminated by PCBs.	☐	☐	☐	☐
PCBs accumulate in striped bass and mink as a result of food consumption.	☐	☐	☐	☐
There is a high level of PCBs in the liver of tomcod in the Hudson River.	☐	☐	☐	☐
The tomcod population thrives in the PCB-contaminated Hudson River.	☐	☐	☐	☐
Tomcod feed on small PCB-contaminated bottom feeders but do not show any effects of PCB-exposure.	☐	☐	☐	☐

Part B

Click on each box to select the word or phrase that **best** completes the statement.

▾ is most probable because ▾ the evidence supports this hypothesis and ▾ the evidence refutes this hypothesis.

Part C

Select additional evidence to support the hypothesis selected in part B.

- ☐ The Hudson River shrimp and plankton do not take in as much PCB as the fish species.
- ☐ DNA evidence shows changes to the gene for AHR2 in the tomcod of the Hudson River.
- ☐ Changes to the AHR2 protein are acquired in response to environmental cues and are not genetic.
- ☐ The Hackensack River shares an estuary with the Hudson River, allowing fish to pass genes back and forth.

SCORES

Student scores on this item are as follows:

- No students earned 5 score points (full credit) on this item.
- The highest score earned was 2 points, and this was achieved by two students, who each earned 1 point for Part A and 1 point for Part B. No one achieved any points for Part C.
- The remaining 12 students earned no credit.

COMPREHENSION

It is hard to extract any detailed information on students' comprehension or reasoning because students floundered so badly on this question.

REASONING

In Part A, most students did conscientiously work their way through the list of evidence and try to determine which supported or refuted each hypothesis, but their reasoning was substantially flawed, perhaps because they did not understand the applicable content knowledge.

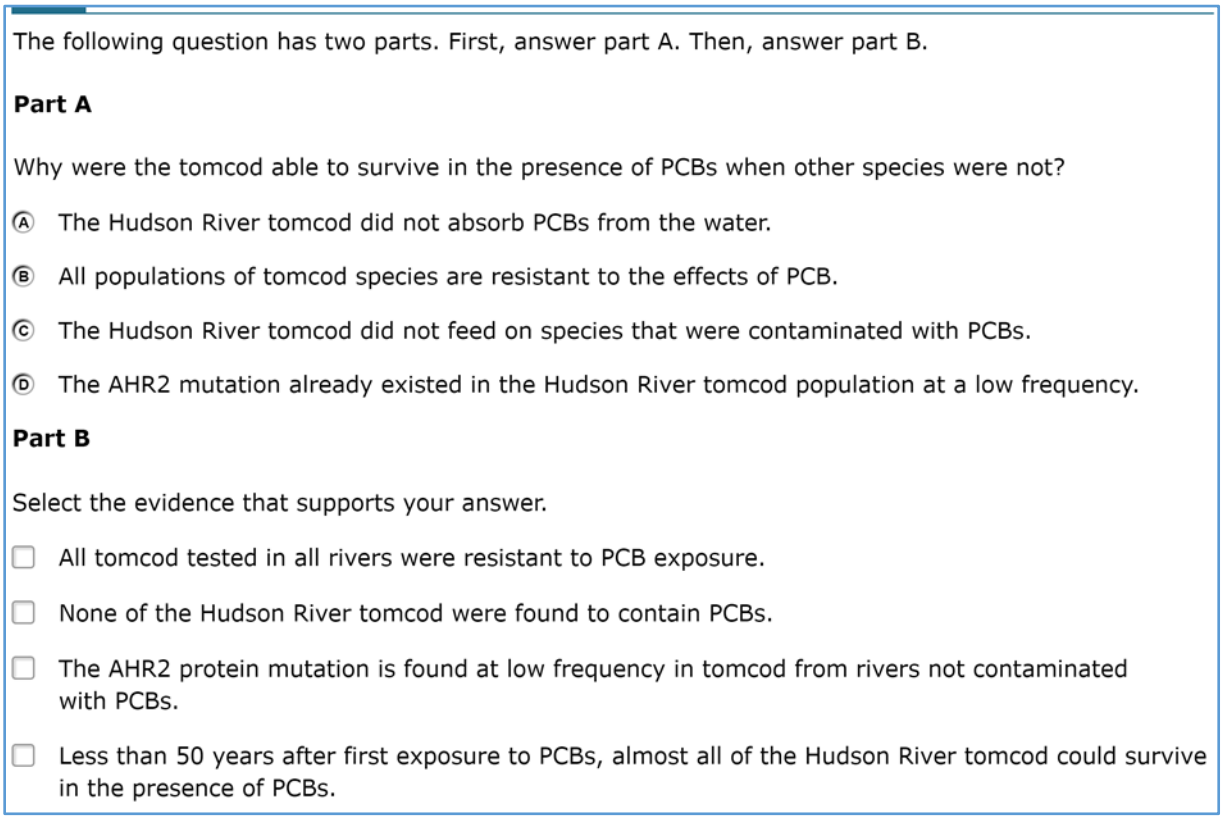
> For example, one student read out loud Hypothesis 1 and 2 in the introduction. She said, "So there's a higher percentage in the Hudson River than in rivers not contaminated," and selected Supports Hypothesis 1 for line 1 "because it's talking about how this one is saying that it's from the water and not from the fish." She read out loud part of line 2, looked quickly at the table in the introduction, and said that it's "actually going against it [refutes Hypothesis] because this one is talking about how it's because of the water not because of the fish, because of the food they are consuming, and they are not talking about the actual fish," then clicked Refutes Hypothesis 1. She read out loud line 3. She said she was going to select Refutes Hypothesis 1 because "it's the same as the first one, because it's saying how the species through the food, not the fish itself." She read out loud line 4 and immediately said that it supports Hypothesis 2 because "it's talking about how it is contained in the actual river, not the fish's fault, but the river's fault." She read out loud line 5 and said immediately that line 5 also supports Hypothesis 2 because, "of the natural selection."

Students who did not have good comprehension of Part A had even less chance of reasoning their way through Parts B or C, both of which built on conclusions from Part A.

Item 2

Item 2 of the Tomcods cluster is shown in Figure 42.

Figure 42. Item 2: Tomcods



Scores

Student scores on this item are as follows:

- Six students earned credit on Part A by choosing the correct explanation for why Tomcods can survive in the presence of PCBs.
- Three of those students also selected one of the pieces of evidence that supported their explanation, but they received no credit for Part B because they did not select both the applicable pieces of evidence.
- Three other students also selected one piece of "correct" evidence, but they had not chosen the right explanation in Part A, so it was unclear exactly what they were supporting.

Comprehension

Although it was hardly the only reason why students had difficulty with this item, students were clearly challenged by having to pick more than one right answer in Part B, perhaps because they are not familiar with multi-select items and just stopped looking after they had made one selection. It might have helped to cue the students if the stem had specified that they had to select ALL the evidence that supported their explanation.

REASONING

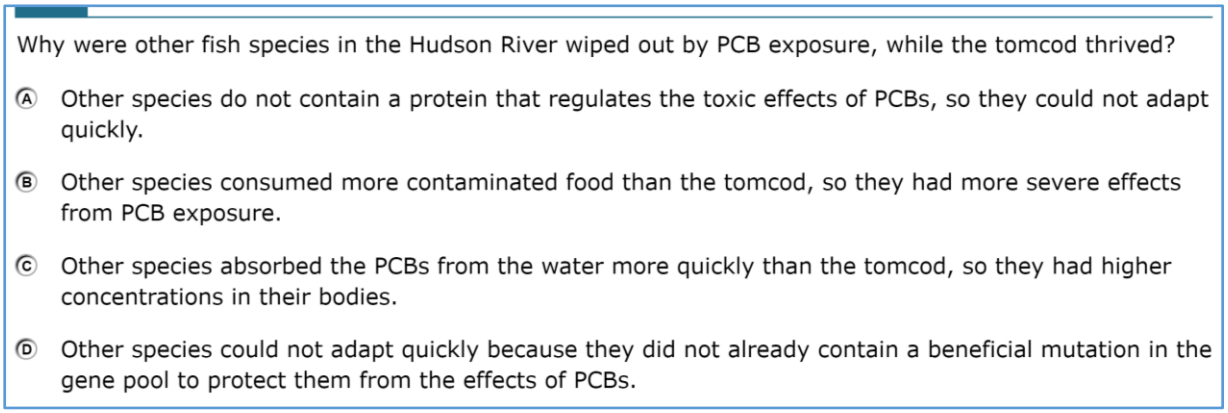
The following is an example of the reasoning of one of the students who correctly identified option D as the reason why Tomcod survived in Part A,

> The student read option A out loud and said, "That's a lie! Because it says up there tomcod have a bunch of it, so that's definitely a lie." The student read option B out loud, saying, "I'm going to say No, because, in the [student looked back to the table on the left] Niantic River and the Shinnecock Bay, they did not have that mutation. So, I'm going to say B is wrong." The student read option C out loud, saying, "OK wrong, because they eat the plankton and the shrimp, and they said earlier that they eat bottom feeders that have it." Student read option D out loud and said, "Yes, because then they would have made it and had a bunch with that mutation."

Item 3

Item 3 of the Tomcods cluster is shown in Figure 43.

Figure 43. Item 3: Tomcods



SCORES

Students did the best on this item; 10 students earned credit.

COMPREHENSION

No features of this item appeared to confuse students.

REASONING

Students who chose the right answer demonstrated plausible reasoning that supported the inference that the students had mastered the concept being tested.

> For example, one student read out loud response option A and said, "That's a good one, that might be the one." He read out loud response option B and said, "That one does not make any sense because all fish, I'm assuming. [are] about the same size will eat about the same, and I know that goldfish don't fill their stomach. I believe they go for all fish, they are all eating like crazy, so I would not click that one." He read out loud response option C twice and said, "Again, that's the same explanation for C as B, I would not click it." He

read out loud response option D and said, "That's the one I'm going to click, because that one is exactly referring to natural selection and . . . it's like a gene, something in their mutation that they could protect themselves from the effects of it, but it's in the gene pool and it's referring to natural selection and the crossing of two species to get your genes and I would go with D, and A would be a close choice."

3.4.4 Cluster 4: Tuberculosis

Performance Summary

The median time to complete the Tuberculosis cluster was 10 minutes. Table 35 and Table 36 indicate the number of students attaining cluster total scores and items scores within the specified ranges, respectively.

Table 35. Number of Students Attaining Cluster Total Scores in Specified Range: Tuberculosis

Score 5–4	Score 3–1	Score 0
1	9	4

Note. Maximum score = 5; *n* = 14; one student ran out of time before completing this cluster.

Table 36. Number of Students Attaining Item Scores in Specified Range, by Item: Tuberculosis

	Maximum Item Score	Score 3	Score 2–1	Score 0
Item 1	3	1	5	8

	Maximum Item Score	Score 1	Score 0
Item 2 (Part A)	1	6	8
Item 2 (Part B)	1	1	13

Note. n = 14; one student ran out of time before completing this cluster.

Task Demands

The following are task demands of the Tuberculosis cluster:

- Based on the provided data, make or construct a claim regarding inheritable genetic variations that may result from: (1) new genetic combinations through meiosis, (2) viable errors occurring during replication, and/or (3) mutations caused by environmental factors. This does not include selecting a claim from a list.
- Sort inferences about inheritable genetic variation into those that are supported by the data, contradicted by the data, outliers in the data, or neither, or some similar classification.
- Identify patterns of information/evidence in the data that support correlative/causative inferences about inheritable genetic variation.
- Construct an argument using scientific reasoning drawing on credible evidence to explain inheritable genetic variations may result from: (1) new genetic combinations through meiosis, (2) viable errors occurring during replication, and/or (3) mutations caused by environmental factors (handscored constructed response).

- Identify additional evidence that would help clarify, support, or contradict a claim or causal argument.
- Identify, describe, and/or construct alternate explanations or claims and cite the data needed to distinguish among them.
- Predict outcomes of genetic variations, given the cause and effect relationships of inheritance.

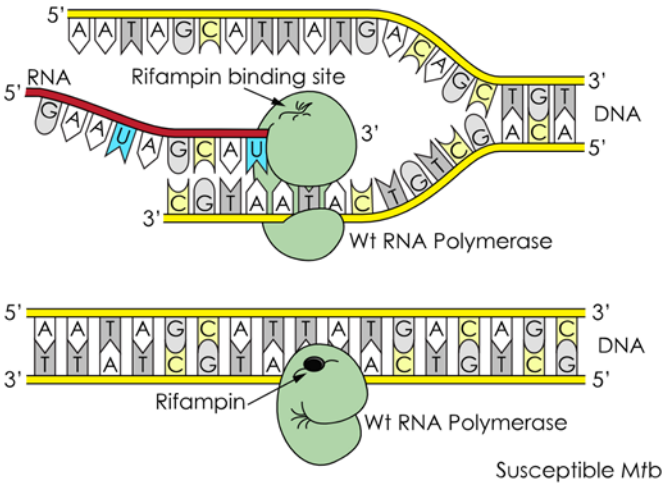
Stimulus

The stimulus for the Tuberculosis cluster is shown in Figure 44.

Figure 44. Stimulus: Tuberculosis

Antibiotic Resistant Tuberculosis

Antibiotic-resistant bacteria present a growing health care problem. The bacteria *Mycobacterium tuberculosis (Mtb)* causes the disease tuberculosis. One antibiotic used to treat tuberculosis is rifampin. Rifampin works by binding to amino acids 36-67 of the RNA polymerase protein of *Mycobacterium tuberculosis.* This binding makes the RNA polymerase protein inactive and the cell dies. This is illustrated below:



However, when treated with the antibiotic rifampin, some *Mycobacterium tuberculosis* bacteria are killed, but others survive. The bacteria that are killed are called "susceptible" to the antibiotic.

Scientists grow 3 mutant strains of *Mycobacterium tuberculosis* bacteria in a lab and sequence their DNA to compare to the wild-type strain that is not resistant to rifampin. Review the information provided.

Comparison of Mutant *Mycobacterium Tuberculosis* Bacteria to Wild-Type

Strain	DNA Sequence Change	Amino Acid Position	Amino Acid Change
Mutant 1	G to A substitution mutation	30	Alanine to Threonine
Mutant 2	C to A substitution mutation	51	No change
Mutant 3	G to T substitution mutation	46	Aspartic Acid to Tyrosine

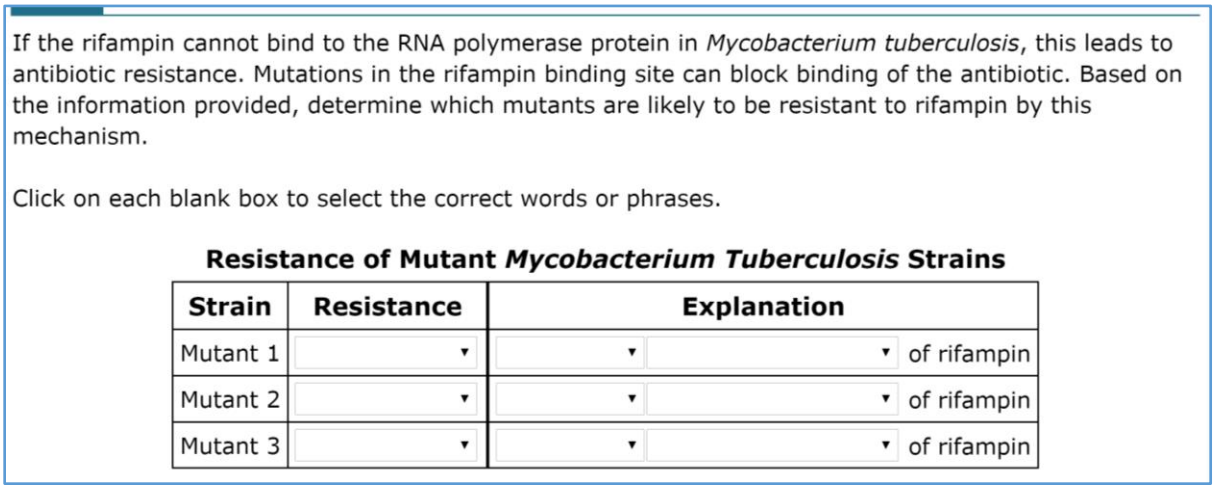
As you work through the questions, evaluate the evidence to identify the source of genetic variation for antibiotic resistance in *Mycobacterium tuberculosis*.

Details by Item

Item 1

Item 1 in the Tuberculosis cluster is shown in Figure 45.

Figure 45. Item 1: Tuberculosis



SCORES

One student earned 3 score points (full credit), and she was the only one to earn a point for correctly determining and explaining the resistance status of Mutant 3.

Five other students each earned 1 score point. Three of these students earned their point for correctly determining and explaining the resistance status of Mutant 2, and two earned their point for Mutant 1.

COMPREHENSION

Four students reported that they found this item confusing and did not understand how to derive the necessary information from the stimulus.

> For example, one student said that Item 1 was confusing and that it was not really addressed [in the stimulus]. He said he was doing a lot of "assuming" because "it's talking about 'resistant,' and he only saw the word once." He also said that "it seemed weird that all three of them would be not resistant," although it is not clear on what basis he concluded that all three mutant strains were not resistant.

Four students reported using things they learned in science classes at school to help them respond to this item. For example,

- one student said that she knew about the amino acid from Biology in freshman year, and
- another student said that he learned about the topic in a biotech class two weeks prior to the interview.

REASONING

All but two of the students referred to the comparison table in the stimulus when responding to this item; four students referred to the diagram.

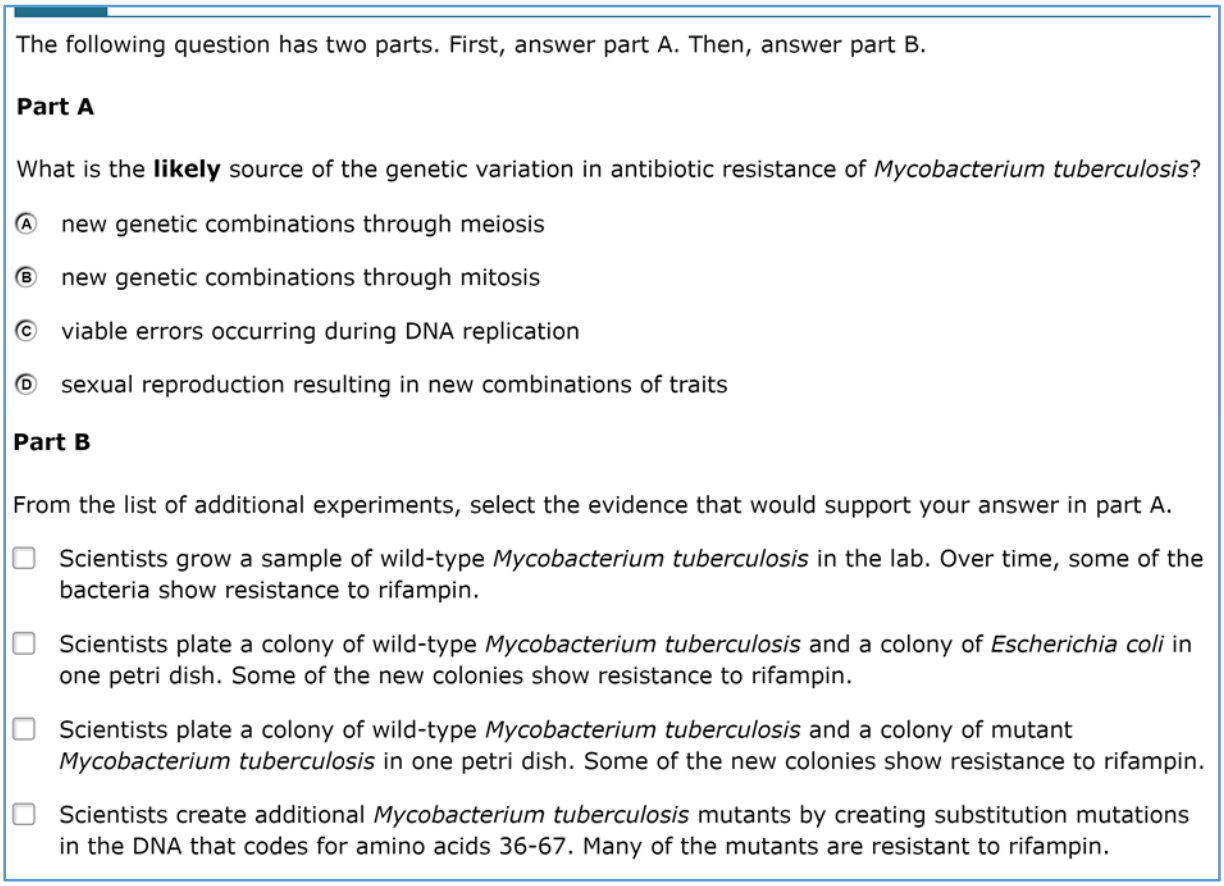
Although only one student had the correct responses for all three of the mutant strains, several used the stimulus materials in the intended manner to reason through the problem.

> For example, one student looked at the comparison table in the stimulus and said, "It says that the Rifampin works by binding to amino acids 36-67 of the RNA. And then it says down here that, because of the G to A substitution mutation, the amino acid positions at number 30, and then . . . it is resistant because it changed it from 36 to 30, so then the Rifampin can't bind to it…So I would say it's resistant, but there's no change of rifampin—oh yeah, change to the—outside of the binding site." "Mutant 2 changed it C to A. Mutant 2 changes the amino acid to 51, so there's no change, so I'm going to mark *Not Resistant* because it's still within 36-67, so I'm going to say no change inside the binding site." "And Mutant 3 is a G to T substitution to 46. And 46 is still within 36-67, so I'm going to say *Not Resistant*, because there is a change from aspartic acid to tyrosine, Inside the binding site."

Item 2

Item 2 of the Tuberculosis cluster is shown in Figure 46.

Figure 46. Item 2: Tuberculosis



Item 2 (Part A)

Scores

Half of the students (seven students) earned credit on this sub-item.

Comprehension

No features of this item appeared to confuse students.

Reasoning

Three students looked back to one or more parts of the stimulus while working on this sub-item.

Four students said they used, or tried to use, material learned in school to help them respond to this sub-item. For example,

- one student said, "I am trying to go back to my knowledge of mitosis and meiosis and DNA replications," and

- another student said, "Usually errors that occur during DNA replication can be bad, and I remember back from when I was a freshman that it's not hereditary."

Some students used test-wise strategies to make plausible guesses, so a correct answer did not necessarily represent full mastery.

> For example, one student (who correctly selected C, *viable errors occurring during DNA replication*) said in his think aloud, "All this right now has to do with DNA . . .I don't see anything about meiosis and mitosis on the chart." When asked how he came up with his answer, he said, "I didn't think it was A or B cause it's talking about meiosis and mitosis, which was not discussed in the article, and then same with D. I did the viable errors because it's talking about DNA strands, so that's why I chose C."

Item 2 (Part B)

SCORES

Only one student earned credit for this sub-item. In part, the difficulty resulted from an incorrect interpretation of the sub-item, as explained further in the Comprehension section below.

Of the two correct options, five students selected *Scientists grow a sample of wild-type Mycobacterium tuberculosis in the lab . . .* and seven students selected *Scientists create additional Mycobacterium tuberculosis mutants by creating substitution mutations in the DNA . . .*

COMPREHENSION

To earn credit for this item, students had to select both the experiments that could provide evidence to support the conclusion they selected in Part A. However, this is not clearly stated in the instructions, so most students stopped after they thought they had found one relevant experiment. Only three students marked two options, and two students said that they thought that they were only allowed to choose one option.

One student expressed confusion with the second response option. He did not know what *Escherichia coli* was and the relationship might be between it and *Mycobacterium tuberculosis*.

REASONING

At least four students referred to the text, diagram, and/or comparison table when responding to this sub-item.

3.5 STUDENTS' OVERALL PERCEPTIONS OF THE TEST

3.5.1 Topics Studied

Elementary School (n=18)

- Eleven students reported that they had studied topics related to the Desert Plants cluster, such as the life cycle of a plant and how plants survive in a desert habitat.
- Ten students had studied topics related to the Grand Canyon cluster, although not all of them learned about fossils or contemporary animals that can be found in the canyon. One student learned about fossils and rock formations as part of the history of Utah.
- Nine students had studied topics related to the Terrarium Matter Cycle cluster, such as "plants have carbon dioxide, but a whole plant needs water, soil, and sun," and some had conducted an experiment in which one group of students tried to grow plants in a dark environment and another group tried to grow plants in the sunlight.
- Although no students were familiar with topics related to the German Pyramid Candle cluster, five students had studied heat transfer.

Generally, each of the Utah students had studied more of these topics than the California students, and their lessons were more closely aligned with the topics of the science clusters. One of the Utah students said he had studied all four of the topics:

> "At the beginning of the year we studied the heat one and how we can help make a motor turn something on, like a light bulb. I thought of that. Maybe it was just backwards, the light was helping the fan to spin. The light was turning or making it spin by the energy it was producing. I remember last year in 4th grade we studied the Grand Canyon and the animals, and we did a little bit this year, and the animals that were living in the walls like trilobite and some others like starfish. We saw this video of this hole that was in Arizona, and there were tons of fossils in it. I think we studied a little bit on the terrarium one . . . We studied a little bit about [the desert plants]. About how each plant could survive."

Middle School (n = 12)

- Nine of the 11 students who responded to the Galilean Moons cluster question reported that they had studied related topics, such as moons, the solar system, space, and the planets, although their studies were not as in-depth as the animation and the data table.
- Only three students had studied the water cycle or how it applied to fog.
- Four students had studied some aspects of weather, including warm and cold fronts, but not as in-depth as the Texas Weather cluster.
- Eight students had studied animals and the types of relationships between animals, although not necessarily about hippos.

High School (n = 15)

- Thirteen students reported that they had studied topics related to the Tuberculosis cluster, such as DNA, mutations, mitosis, meiosis, and amino acids.
- Seven students had studied topics related to the Blood Sugar Regulation cluster, although not as in-depth as these questions. In referring to the Blood Sugar Regulation cluster, one student said that they had reviewed molecule concentrations but never discussed meals or "not that in-depth, more gone over these and what they do for the body." Another student said she had studied feedback loops and homeostasis.
- Five students had studied topics related to the Tomcods cluster, such as the food web, ecology, and PCBs.
- Only two students said that they had studied topics related to the Saving the Tuna cluster, but they did not provide any information about which specific topics.

3.5.2 Use of Similar Online Tests and Tools

Elementary School (n=18)

All but one student had previously taken online tests; the subjects of the tests varied and included science, mathematics, reading, and/or "grammar." The online tests they had used included Galileo, SALT, ATI, and, for the Utah students, SAGE.

All but one of the students said that they had used similar online tools, including being able to expand the screen from left to right and vice versa; videos; dictionaries; navigation buttons such as arrows, a scroll bar, Back, Next, and Zoom in/Zoom out buttons; and drop-down menus. One student said that her previous experience with online tests involved individual questions rather than clusters, and another student said that there were "more pictures to move around" on the other online test.

Middle School (n = 12)

All 11 students who responded to this question had previously taken online tests; the subjects varied and included science, mathematics, and/or English language arts.

All but two of the students said that they had used similar online tools (including the Connect Line tool and Graphing tool for plotting points), animations, videos, and navigation buttons such as the Next, Back, Pause, and Zoom in/Zoom out buttons. One student said that he previously had to draw lines, but only straight lines, nothing like the graphs she had to draw in the Morning Fog cluster. Another student mentioned that layout of the items was familiar, including having the stimulus on the left side of the screen and the questions on the right side.

High School (n = 15)

All but two students had previously taken online tests; the test subjects varied and included science, mathematics, and English.

All but one of the students said that they had used similar online tools including at least one of the following: graphs, diagrams, the Connect Line tool, checkboxes, and a layout that presented a stimulus on one side of the screen and the associated questions on the other side. One student said that a standardized test he took the previous day was exactly the same, "the interface is the same," although he was not able to expand the screen on the standardized test. One student mentioned two other functionalities that he had used on other tests: the Highlighting tool and the ability to add a note to a paragraph and view it later.

3.6 OVERALL THOUGHTS ABOUT TEST DIFFICULTY

Elementary School (n=18)

Nine students felt that the test had both easy and hard parts and described the overall difficulty as "in between." Examples include the following:

- One student said, "I think the test was in between those because some of it I got confused on and some other pieces like this [referring to Item 1 of the Redwall Limestone cluster] was easy since it gave us these maps about where it lived and the rest was kind of simple. For this one [referring to Item 2 of the Redwall Limestone cluster], it was simple."
- One student said, "Some of them were hard, some of them were confusing, some of them were easy – that's how I feel about this test. The hardest part was [the Terrarium Matter Cycle cluster], question two, Part A [of the Terrarium Matter Cycle cluster] because "I didn't understand what they meant about X, Y, and Z – I had to think about what they mean."
- Another student thought the test was "right in the middle, good. It wasn't too easy or too difficult." The student did not find any of it particularly confusing.
- Five students described only one of the items as being difficult, and four of the five students said the hard item was Item 2 Part A in the Terrarium Matter Cycle cluster. Examples include the following:
    - One student said, "There was one I skipped. I didn't really like that. Because there was too much going on," referring to Item 2 in the Terrarium Matter Cycle cluster.
    - One student felt that the hardest question was on "the terrarium with the diagram and the X, Y, and Z stuff. The others you just had to think about, and you could solve them."
    - Another student said, "Overall, I think it's really good. I found the terrarium a little confusing. It is a good test to have about things you need to know." When asked if the questions were hard or easy, the student said they were easy except for the terrarium question. He said he got confused on the circle of energy.

By contrast, four students expressed that the test was easy. Examples include the following:

- One student did not feel like any of it was confusing, and he was not nervous. He thought the questions were very specific. It was easy for him to navigate through the tools and figure out how to answer the questions.
- One student said, "It took some time for me to think of the answers, but I thought it was pretty easy."

Middle School (n = 12)

All 12 students responded to the end-of-test question on what they thought of the test. Seven of the students felt that the test was not too hard. For example:

- One student thought that the questions were reasonably easy but were hard for someone who hadn't learned a lot of this material. She said that, in general, she is well educated in science, but a lot of these topics are "very random." The student felt like she could have told the interviewer about the water cycle, but not how it works in this specific scenario.

- One student said that the test "was good, yeah. It wasn't hard." The student said that Item 3 of the Galilean Moon cluster was hard.

- Another student thought the questions got harder as she went along, and the hardest problem was the Texas Weather cluster. She had to reread some of the questions, but overall, she thought they were clear.

By contrast, five students expressed that the test was difficult or challenging. For example:

- One student thought that the test was good, but kind of difficult. She mentioned that students like her brother, who is dyslexic, would find it helpful to have the questions read out loud to them. She also said some of the questions were harder because she hadn't gone over the content yet and didn't know what some of the moons were.

- Another student thought the test was "pretty difficult." It was confusing for the student because she had to go back and reread items to understand the process and how to figure it out.

- A student said it was definitely "more challenging" than tests he had taken.

- A student said, "I thought it was kind of confusing. We've studied the moon one a bit, the hippos for sure, and then the water cycle and the temperature we haven't, so for doing all of those for my first time, I couldn't quite make it out. I was totally lost on the Morning Fog in the Valley."

High School (n = 15)

All 15 students responded to the end of the test question on what they thought of the test, although three students did not comment on whether the test was easy or difficult. (One of these latter students described it as "pretty interesting" and "different." Another said he liked the multiple-choice items, the diagrams, tables, and having multiple parts to a question.)

Ten students felt that the test was in the "middle range" of difficulty, with some questions being clearer than others. Four students felt that the Tomcods cluster was confusing, and three students felt that the Blood Sugar Regulation cluster was confusing.

Two students described the test as being difficult. One of these students said the test did not relate to his past studies, but he thought it would be a good test for students who were studying these topics. He also said the types of questions were different than he was used to: – "it's not like normal standardized testing kinds of questions." The student noted that he had not studied these topics even though he was an Advanced Placement (AP) Biology student. Consequently, he was unsure who the target audience of the test might be. The other student mentioned that she found the questions "kinda hard" because there were so many parts to each question. The reading parts were clear, but the structure of the questions could be confusing, according to the student.

APPENDIX 1: CHARACTERISTICS OF SAMPLE, BY CLUSTER GRADE LEVEL AND STUDENT

Table 1-A. Elementary School Sample

Student	Location	Grade	Gender	Lunch Program	Ethnicity	Language at Home	IEP (Disability)	Science Grades
1	California	5	Male	No	Asian	English	No (N/A)	Mostly A's
2	California	5	Male	No	Caucasian	English	No (N/A)	Mostly A's
3	California	5	Male	No	Asian	English	No (N/A)	Mostly A's
4	California	5	Male	No	Caucasian	English	No (N/A)	Mostly A's
5	California	5	Male	No	African American	English	No (N/A)	Mostly B's
6	California	5	Male	No	Caucasian	English	No (N/A)	Mostly A's
7	California	5	Female	Yes	Other	English	No (N/A)	Mostly B's
8	California	5	Male	Yes	Caucasian	English	No (N/A)	Mostly A's
9	California	5	Male	Yes	Hispanic	English	No (N/A)	Mostly A's
10	California	5	Male	No	Caucasian	English	No (N/A)	Mostly B's
11	California	5	Female	No	Caucasian	English	No (N/A)	Mostly B's
12	California	5	Female	No	Caucasian	English	No (N/A)	Mostly B's
13	Utah	6	Male	–	Caucasian	–	–	–
14	Utah	6	Male	–	Caucasian	–	–	–
15	Utah	5	Male	–	Caucasian	–	–	–
16	Utah	6	Female	–	Caucasian	–	–	–
17	Utah	5	Male	–	Caucasian	–	–	–
18	Utah	5	Female	–	Caucasian	–	–	–

Note. –: Missing data

Table 1-B. Middle School Sample

Student	Location	Grade	Gender	Lunch Program	Ethnicity	Language at Home	IEP (Disability)	Honors/ Advanced Classes	Science Grades
1	California	9	Female	No	Other	English	No (N/A)	Math	Mostly A's
2	California	9	Male	No	African American	English	No (N/A)	None	Mostly B's
3	California	9	Female	No	Caucasian	English	No (N/A)	None	Mostly A's
4	California	8	Female	No	Caucasian	N/A	No (N/A)	None	Mostly A's
5	California	9	Female	No	Asian	English	No (N/A)	Math, Science, Reading	Mostly A's
6	California	8	Female	No	Caucasian	English	No (N/A)	Math	Mostly A's
7	California	9	Male	Yes	Caucasian	English	Yes (Specific Learning Disability)	None	Mostly A's
8	California	8	Male	Yes	Hispanic	English	No (N/A)	None	Mostly A's
9	California	8	Male	Yes	Caucasian	English	No (N/A)	None	Mostly A's
10	California	8	Male	No	African American	English	No (N/A)	None	Mostly A's
11	California	8	Male	No	Asian	English	No (N/A)	Math, Science, Reading	Mostly A's
12	California	8	Female	No	Asian	English	No (N/A)	None	Mostly A's

Table 1-C. High School Sample

Student	Location	Grade	Gender	Lunch Program	Ethnicity	Language at Home	IEP (Disability)	Honors/ Advanced Classes	Science Grades/ Achievement*
1	California	11	Female	No	Caucasian	English	No (N/A)	None	Mostly A's
2	California	11	Female	No	Hispanic	English	No (N/A)	None	Mostly A's
3	California	11	Female	No	Other	English	No (N/A)	None	Mostly A's
4	California	11	Female	No	Caucasian	English	No (N/A)	AP Chemistry	Mostly A's
5	California	11	Female	Yes	Hispanic	English	No (N/A)	IB Honors Science	Mostly A's
6	California	11	Female	No	Hispanic	English	No (N/A)	None	Mostly B's
7	California	11	Female	No	Caucasian	English	Yes (ADHD)	None	Mostly A's
8	California	11	Male	No	Asian	English	No (N/A)	IB Biology, Chemistry	Mostly A's
9	California	11	Male	Yes	Hispanic	English	No (N/A)	None	Mostly B's
10	California	11	Female	No	Caucasian	English	No (N/A)	Chemistry	Mostly B's
11	California	11	Male	Yes	Prefer not to answer	English	No (N/A)	None	Mostly B's
12	California	11	Male	No	Caucasian	English	No (N/A)	None	Mostly B's
13	Connecticut	10	Female	–	African American	–	–	–	High Achieving
14	Connecticut	11	Male	–	Caucasian	–	–	–	High Achieving
15	Connecticut	12	Female	–	Hispanic	–	–	–	High Achieving

Note. *Parent report of science grades or teacher estimate of achievement level.
–: Missing data

Appendix E

Braille Cognitive Lab Report

Cognitive Lab Study: Accessibility of Science Clusters for Braille Readers

Fran Stancavage

Susan Cole

April 2019

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

This set of cognitive labs was designed to determine if students using braille can understand the task demands of selected interactive Next Generation Science Standards (NGSS)-aligned science clusters and 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. The clusters for the study were sampled from those that had already been selected for braille translation. The cognitive labs were designed to address the following three research questions:

1. Can students using braille provide responses to the selected interactive NGSS-aligned science clusters that are consistent with their knowledge and skills relative to the constructs of interest?
2. Within the selected clusters, can students successfully navigate all the included interaction types, or are further modifications needed to make the clusters fully accessible?
3. How much time do students using braille require to work their way through the selected clusters, and what strategies can be recommended to enable students using braille to complete clusters within a single testing session (to improve continuity)?

Although the Cambium Assessment, Inc (CAI) team was able to collect relevant data for this cognitive lab study, there were some limitations to the analysis. Most importantly, there were far fewer eligible visually-impaired students willing to participate in the study than anticipated, and some of them, although technically readers of braille, did not use braille while responding to the science questions in the cognitive labs. In addition, in several of the cognitive lab sessions, students' interactions with the clusters was hampered by technical issues with the Job Access With Speech (JAWS) screen-reading software and/or the Refreshable Braille Display (RDB) supplied locally, as well as by text-to-speech (TTS) tagging or braille embossing problems that arose in the beta-version materials. The latter were used in the cognitive labs due to the timing of the study.

2. METHODS

2.1 STUDY DESIGN

Two science clusters were sampled for each grade band (i.e., elementary, middle, and high school), and tailored protocols were developed for each cluster. The original design called for a minimum of six cognitive labs at each grade level, but due to recruitment challenges (discussed further in this section), labs were only conducted with ten students in total. The cognitive labs were held in Oregon and West Virginia between October 2018 and January 2019. The interviews lasted two hours, and each student was presented with one or both clusters for their grade band, depending on how much time the student took to complete the first cluster.

As part of the cognitive lab introductory activities, students were trained in the concurrent think-aloud technique. Using an elementary-level science cluster, which was not one of the clusters evaluated in the study, the interviewer first modeled the technique in Part A (first scored question) and then had the student practice in Part B (second scored question).

Students then moved on to their first assigned cluster. They were encouraged to think out loud as they worked through the cluster, and interviewers were instructed to use follow-up probes to clarify and expand on what the student said (or what the student was observed doing). Probes, which were tailored to the specifics of the cluster, focused on whether the student was able to find all the information needed to respond to the questions, what the student thought about the ways in which they had to enter answers to questions (for questions with innovative response formats), and if they would change anything about the way the information was presented to make it easier to work on the questions. A final probe allowed the student to report on anything else they found notable about the questions or introductory material in the cluster.

Students who were able to complete the first cluster by the 1.5-hour mark (out of the scheduled 2-hour lab) were moved on to the second cluster for their grade band. Probes were only administered after the student had completed all the questions in a given cluster in order to ensure that probing on the earlier questions would not influence the student's interactions with the later questions.[1]

Interviewers brought embossed braille forms to the cognitive labs. The site was responsible for providing other resources, such as JAWS and an RBD. CAI requested that a teacher of the visually impaired (TVI) or a teacher assistant be present in the room during the cognitive lab and assist the student as they would during an actual test. In most cases, prior to the interview, the interviewer briefly discussed with the TVI/teacher assistant what resources the student used to navigate online tests and how frequently/in what ways the TVI/teacher assistant typically assisted the student during testing. This information helped the interviewer to further tailor their probes and observations.

2.2 INTERVIEWER TRAINING

The project leads provided a 4-hour training for the interviewers who would be conducting the cognitive labs. Because all the interviewers were experienced in the cognitive interview technique, the training primarily focused on reviewing the content of the clusters and familiarizing the interviewers with the test platform and the specifics of the cognitive lab protocols. An assessment program manager was present at the training to provide an overview of the test platform and to respond to any technical questions.

2.3 STUDY SAMPLE

Permission to recruit students for the study was secured from four states. In each state, the project manager and project director worked with relevant school and district personnel to recruit eligible students and coordinate logistics. Ultimately, only two states, Oregon and West Virginia, were able to provide students for the study.

The recruitment materials specified a need for students in grades 6, 7, 9, 10, or 12 who use braille, and all the recruited students were in fact able to use braille to some degree; however, an unanticipated complication was that some of the students who were partially sighted chose to use other resources (e.g., the Zoom tool) to navigate the clusters. Given that there were so few students

[1]To stay within the agreed-upon 2-hour time limit, the interviewer sometimes stopped the student before they finished the second cluster in order to leave sufficient time for probing.

available, the CAI team took whomever was recruited. The characteristics of the sample, by student, are shown in Table 1 below.

Students in grades 6 and 7 were administered the elementary-school-level clusters, students in grades 9 and 10 were administered the middle-school-level clusters, and students in grade 12 were administered the high-school-level clusters.

Table 1. Characteristics of Sample, by Student

Student	Grade	Gender	Resources Used in the Cognitive Lab
1	6	Male	JAWS, RBD, braille*
2	6	Female	Zoom, larger cursor
3	9	Male	Zoom, larger cursor, JAWS, braille
4	9	Male	Zoom
5	9	Male	JAWS, RBD
6	10	Male	JAWS, RBD, braille
7	10	Female	Braille, ChromeVox**
8	10	Female	Zoom
9	12	Female	Zoom, JAWS, braille
10	12	Male	Inverse colors, zoom

Note. **Braille* refers to the embossed braille forms
**ChromeVox is an alternative TTS reader.

3. FINDINGS AND RECOMMENDATIONS

3.1 RESOURCES USED

The students used the available resources in a variety of ways during the cognitive labs. It was common for the students to switch between resources (e.g., moving between embossed braille, JAWS [sometimes coupled with an RBD], the Zoom tool [where relevant]). Some of the partially-sighted students chose to use only zoom, citing reasons such as having only "beginner" level braille skills or feeling that navigation using braille took longer; others switched between the Zoom tool and other resources. One TVI reported that the partially-sighted student they were assisting switched based on "eye fatigue and lighting conditions." At least two students used the embossed braille forms almost exclusively to read the questions and reference the introductory materials, but switched to JAWS to enter their answers. One of these students reported that they used the embossed braille forms because it was easier than scrolling up and down the page using JAWS. Another partially-sighted student used the embossed braille forms and a screen reader similar to JAWS, but they also looked very closely at the screen to see where to place the cursor when responding to the questions.

Two students, one assigned to a middle school cluster and the other assigned to a high school cluster, reported that they would normally be offered a Perkins Brailler (also called Perkins Braille Writer) to take notes during testing. The CAI team did not anticipate or provide this resource,

which is the equivalent to scratch paper for a braille user and is a standard accommodation for visually-impaired students in testing situations. It can also be used by the student to type the answers in braille, after which the TVI/teacher assistant can transcribe the answers and enter them into the test system.

3.1.1 Hardware and Software Resources

As mentioned previously, there were technical issues with some of the locally-supplied resources used in the cognitive labs. In both states, JAWS often did not work smoothly, and there were instances in which the RBD did not operate at all. As a result, some of the students struggled more with navigation than they usually would. In a couple of cases, these students reported depending more on the TVI/teacher assistant and embossed braille forms than they normally would have.

One TVI noted that every difficulty that their student encountered had come up in a real testing situation—problems with the RBD crashing, unpredictable behavior with JAWS, and "bad" embossed braille forms. The TVI said that, even when everything is tested in advance (as the RBD is), resources still do not necessarily work inside CAI's test delivery system (TDS).

3.1.2 Embossed Braille Forms

Students were generally taken aback when they first realized the number of pages in the embossed braille forms, and, with no prior exposure to the science clusters, they had not anticipated or prepared for the need to keep track of information across multiple pages. Most of the other challenges that students experienced with this resource arose from inadvertent errors in the beta-version forms. Some of these errors were fixed after the first cognitive lab, but others persisted. In a normal cognitive lab study with a larger subject pool, all protocols would be pilot tested, which would have offered an opportunity to fix problems like this before the materials were used in the actual study.

However, some students also reported encountering graphical elements that—as rendered—were difficult to discriminate on the embossed forms. For example, one student reported that it was hard to differentiate between the two graph lines that, in the print version, were distinguished by different tones of grey. Another student indicated that it was difficult to discern the overall layout of a map of the United States, in which some states were highlighted for sharing a common characteristic, because the state lines, the line marking the boundary of the United States, and the lines outlining the Great Lakes were all too similar.

Regardless of these various issues, most students felt that the braille forms were easier to work with than using JAWS.

3.1.3 JAWS and Other Online Navigation Issues

There were significant problems with JAWS that prolonged the time it took students to work through the clusters. Some of these problems were caused by TTS-formatting configuration errors that were not caught in advance, but others had to do with the way in which JAWS was set up by the TVI/teacher assistant. An example of the latter was an instance in which JAWS was accidentally set to read all the navigation marks and not just the substance of the text. Proper settings are covered in the *Braille Requirements and Testing Manual*, but were not discussed with the TVIs/teacher assistants who were preparing for the cognitive labs.

Other challenges were caused by conventions with which the students were not familiar. In particular, students often appeared confused when JAWS skipped over a table or figure that had been judged as too complex to be read successfully by JAWS. It might have been helpful if the TTS tagging had included embedded text that instructed students to switch to the screen image or the embossed braille forms in order to see the contents of the table or figure.

For tables that were read by JAWS, at least one student noted that it would be helpful for JAWS to indicate when the table was entered and exited, rather than just reading "table of checkboxes" multiple times as it progressed through the table; however, it was not clear whether the student had JAWS set up correctly.

Several students had difficulties using the Tab key effectively, repeatedly finding themselves in some other location than they expected when they tabbed forward or back. There seemed to be some interaction between problems with tabbing and the students' confusion about JAWS not reading the tables and figures (however, it should be noted that one student, who did not have any problems navigating with JAWS, said that it would have been very helpful to be able to easily tab between the question stem and the response fields so that students could quickly review the question—potentially multiple times—as they considered their response).

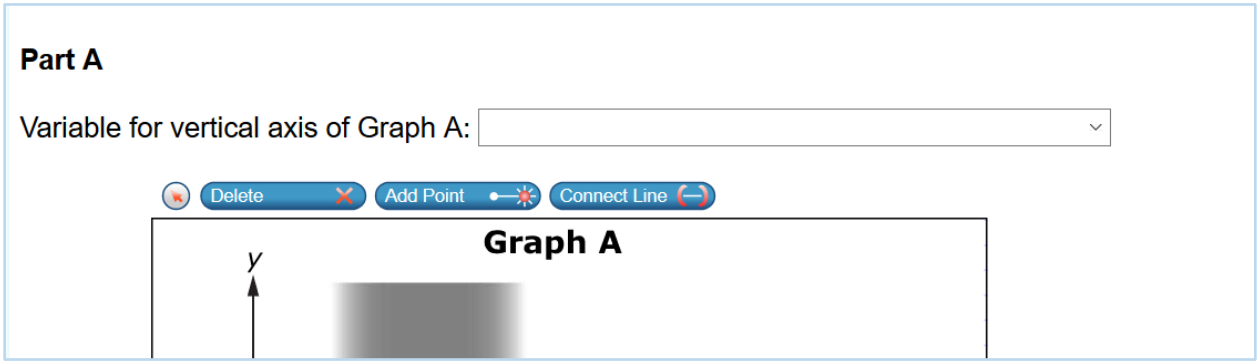
Finally, there were issues associated with the way in which drop-down boxes were handled by JAWS. Some students were not familiar with the term "combo boxes," which was used to describe these boxes, and many students were confused by the ways in which JAWS handled the response options for these boxes. In some cases, it appeared that JAWS did not read these choices at all (which was consistent with the current TXX business rules), while in other cases JAWS read the options, but only after a response was selected. Finally, the tagging may have been inadequate, as at least one student didn't understand what JAWS was reading until the TVI showed them where the various parts of the question were, especially the text in the drop-down boxes.

3.1.4 Zoom Tool

Students who used the Zoom tool did not encounter many problems applying this tool to the science clusters, although one student failed to discern at least one drop down box as they moved through the text. These students did, however, suggest several modifications that they felt would improve their experience, including the following:

- Enable the user to change the size of tables or images on all sides rather than just two sides to avoid having to scroll sideways.
- Add additional spacing in the text; at x3 or greater zoom, the spacing is too tight.
- Make the sizing of the answer buttons consistent when zoomed in—currently the answer buttons on the multiple-choice questions stayed small, whereas other answer buttons got larger when zoomed in.
- To help with viewing the drop-down boxes (see example in Figure 1), format the boxes with high contrast or a thicker line.

Figure 1. Example Drop-Down Box



3.1.5 Assistance from the TVI/Teacher Assistant

The level of TVI/teacher assistance varied in relation to the student's fluency with the other resources. An added factor in the level of assistance provided to students in the cognitive labs was the failure of the RBDs in some sessions. Without the RBD, students who could not see the computer screen required assistance to enter their responses.

The most facile student in our sample, who was very comfortable using both the embossed braille forms and JAWS, still asked for some assistance from the TVI, particularly with online navigation. At the other end of the scale, the following vignette illustrates how one TVI worked with a student who needed considerable support.

Example of a TVI assisting a student who was not very facile with the other resources available.

One student began by letting JAWS read through the entire introduction and most of the questions before asking if they could pause it. The TVI gave the student the instructions to do so. The student said that they were being hit with too much information at once, so they asked for the embossed braille form. The TVI found the first page and directed the student through most of the content, reading a lot of it out loud. The TVI noted that this was an official accommodation that the student was allowed to use during tests. The student had difficulty reading the braille out loud–stumbling over words and parts of words and asked the TVI for a lot of help with the figures. When the student had trouble reading Table 1 (included in the introduction) on the braille form, they decided to go back to JAWS. JAWS jumped ahead to Table 2 (part of the first scorable question), and it took some effort for the student to go back to Table 1. The TVI helped the student find Table 1, and the student followed along on the braille form as JAWS read the text preceding Table 1 out loud; however, JAWS did not read Table 1, instead skipping to the next paragraph of text. The student wanted to try typing on the keyboard to see if it would help bring up the table, but the TVI explained that there was no text box to type anything into. The TVI suggested that the student tab forward. The TVI said that in a real test situation, she would offer to read the table at this point. The student said this would be helpful, and the interviewer indicated that this was acceptable, so the TVI read the table out loud while the student followed along on the braille form.

3.2 General Accessibility Issues

An accessibility issue that, although it primarily affects the embossed braille forms, also has implications for screen layout, has to do with the inconsistent locations in which cluster components (e.g., questions, tables and figures, other text) appear on the page. Without the ability to quickly discern the overall layout of each page or screen, it was much harder for students in the study to process the information being conveyed. One student mentioned that it would be helpful if question stems consistently appeared on the top of the page, as in some cases the display that follows the item identifier (e.g., Part A) starts with a table or other graphic, with the text of the item stem following. Given the student feedback, it would be better to position the table/graphic below the item stem. Another student was observed to completely overlook a short paragraph of text that appeared between two large graphics in the introduction. Moreover, there were no sufficient cues to alert the student to the fact that they had missed an element. When blocks are being prepared for braille readers and other visually impaired students, it would be helpful to take these considerations into account and modify the page and screen layouts accordingly.

Similarly, one student's thoughts about how they would use the various resources to efficiently work through the science clusters (see graphic below), suggest another modification that would help maximize accessibility.

Thoughts from a student on how to best use resources to work through the science clusters.

Both the student and their TVI noted that working with the embossed braille forms for the science clusters was a departure from their usual testing experience because most traditional test questions can be rendered on a single page. Upon reflection, the student said that the strategy that would work best for them would be to

- first read through the whole cluster using the embossed braille form; and then
- navigate the questions with JAWS and an RBD, referring back to text passages as needed using these tools; however, where there was a need to refer back to a figure or chart, use the embossed braille.

The student indicated that to successfully carry out this strategy, they would need a better system for keeping all the braille pages organized so as to be able to quickly access the necessary graphics. Providing an index, or some form of page headers, might help with this problem.

3.3 Timing and Continuity

One of the goals at the beginning of the study was to determine whether students could complete an entire cluster during a single testing session; the results suggest that timing will not be a major issue, so long as schools are able to provide uninterrupted 1-hour testing sessions, if necessary. Despite the technical issues with JAWS, the RBD, and the braille forms, all but two of the students were able to complete at least one of the clusters during the cognitive labs, and one of the students who failed to complete the cluster was not focused or motivated to respond to the questions. The labs were approximately 1.5 hours long, not including the introduction and think-aloud modeling

and practice. Given that they involved thinking aloud and probing, as well as working the questions, 1-hour testing sessions should be sufficient for actual administrations.

4. CONCLUSIONS

In general, both the students who relied entirely on braille and/or JAWS and those who had some vision and were able to read the screen with the Zoom tool 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. To varying degrees, assistance from the TVI/teacher assistant was necessary, but this was most likely not qualitatively different from the assistance that would be provided on a more traditional test.

However, the clusters were clearly different from (and more complex than) other tests with which the students were familiar, and students should be given adequate time to practice with at least one sample cluster before taking the state test. It would also be helpful for students to work with their TVIs/teacher assistants in advance to develop a strategy for organizing and using the information required to answer the test questions. For example, students might want to take notes on a Perkins Brailler as they work. Given that the challenges of the science clusters are not unlike the challenges that students are likely to encounter under curricula based on NGSS or Common Core State Standards (CCSS) or their equivalent, students could be expected to become more fluent in the requisite skills as such curricula become more widespread.

Because of the large numbers of substantively important figures and tables in the clusters, we judge the embossed braille forms to be essential for any student who cannot see the material on the screen with magnification. Embossing is already set to "automatic" on all CAI science tests; however, in the case of the science clusters, test administrators (TAs) should be instructed to have the forms available before the student begins work on a given cluster, as the embossing would otherwise be very disruptive.

A major challenge that we observed in the cognitive labs—which would apply to more conventional tests, as well—was the temperamental functioning of JAWS and the RBDs. There were multiple instances of these resources failing during the cognitive labs, even when they had been tested in advance. This might be avoided with more rigorous user acceptance testing (UAT) of items using JAWS, but it also might require changes at the local level, such as better training for TVIs/teacher assistants or better maintenance of the devices.

Among the innovative response formats encountered in the science clusters that were used in the cognitive labs, the drop-down boxes proved to be the most problematic (specifically for students who were trying to navigate the science clusters using JAWS), since the drop-down options were not tagged to be read by JAWS. CAI should consider changes to the business rules in order to allow the drop-down options to be read.

The following recaps the tool-specific recommendations offered in the report.

For braille forms,

- make sure that graphic elements, such as graph or map lines, are bold enough or sufficiently contrasted to be easily discriminated;

- consider reformatting so that page layout is more predictable (e.g., always keeping text together rather than interspersing it with large graphics); and/or
- consider adding an index or page headers to make it easier for students to keep track of information across multiple sheets of embossed braille.

For JAWS,

- provide more cues when a student needs to switch to the braille form or the screen image to view a table or figure that JAWS will skip over;
- add navigation markers to indicate when the reader is entering or exiting a table if tables are tagged to be read by JAWS; and/or
- provide a way for the student to readily tab between the question stem and the response field(s).

For the Zoom tool,

- enable the user to change the size of tables or images on all sides rather than just two sides to avoid having to scroll sideways;
- add additional spacing in the text; at x3 or greater zoom, the spacing is too tight;
- make the sizing of the answer button consistent when zoomed in—as currently configured, the answer buttons on the multiple-choice questions stay small, whereas other buttons get larger when zoomed in; and/or
- format the boxes with high contrast to help with viewing the drop-down boxes.