South Dakota
Science Assessment

2022–2023

Volume 4:
Evidence of Reliability and Validity

south dakota
DEPARTMENT OF EDUCATION
Learning. Leadership. Service.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

1. Introduction and Overview of Reliability and Validity Evidence

The South Dakota Science Standards were adopted by the South Dakota Board of Education (BOE) in May of 2015. As a result, the South Dakota Science Assessment (SDSA) was administered to students in grades 5, 8, and 11 during the 2021–2022 school year in order to measure students' mastery of the new South Dakota Science Standards. The SDSA was administered online using an adaptive test design. Accommodated versions of the tests were available for each grade, including braille, Spanish-language versions, and Data Entry Interface (DEI) forms. Table 1 shows the complete list of summative tests that were delivered for the first year of operational administration in 2022–2023.

Table 1. 2022–2023 Operational Assessment Modes

Language/Format	Assessment Mode	Grade
English	Online	5, 8, and 11
Spanish	Online	5, 8, and 11
Braille	Online	5, 8, and 11
English Large Print /Data Entry Interface (DEI)	Paper	5, 8, and 11
Braille/Data Entry Interface (DEI)	Paper	5, 8, and 11

Given the intended uses of these tests, both reliability evidence and validity evidence were necessary to support appropriate inferences of student academic achievement based on the SDSA scores. The analyses to support reliability and validity evidence that are reported in this volume were based on reported test scores, including those for the online English-language version and the accommodated versions of the SDSA.

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

1. ***Reliability.*** The reliability estimates are presented by grade and demographic subgroups. This section also includes conditional standard error of measurement (CSEM), classification accuracy (CA), and classification consistency (CC) results by grade.
2. ***Content Validity.*** This section presents evidence showing that test forms were constructed to measure the three-dimensional South Dakota Science Standards with a sufficient number of items targeting each area of the test blueprint.
3. ***Internal Structure Validity.*** This section provides evidence regarding the internal relationships among the subscale scores to support their use and to justify the item response theory (IRT) measurement model. This type of evidence includes observed and disattenuated Pearson correlations among discipline scores per grade. The IRT model is a multi-dimensional model, with an overall dimension representing proficiency in science and nuisance dimensions that consider within-item local dependencies among scoring assertions (see Volume 1, Section 5.1, Annual Technical

Report). In this volume, the evidence is provided for 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 (i.e., unidimensional models) and models with a more elaborate internal structure.

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

1.1 RELIABILITY

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

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

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

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

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

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

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:

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

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

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

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

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

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

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

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

Because the TIF indicates the amount of information provided by the test at different points along the ability scale, its inverse indicates the lack of information at different points along the ability scale. This lack of information is the uncertainty, or the measurement error, of the cut score at various cut score points. For the derivation of heterogeneous measurement errors in IRT and how these errors are aggregated over the cut score distribution to obtain a single, marginal, IRT-based reliability coefficient, refer to Section 3, Reliability.

1.2 VALIDITY

Validity refers to the degree to which "evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014). Messick (1989) defines validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (p.13). Both definitions emphasize evidence and theory that support inferences and interpretations of test scores. The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) suggest five sources of validity evidence that can be used in evaluating a proposed interpretation of test scores. When validating test scores, these sources of evidence should be carefully considered.

The first source of evidence for validity is the relationship between the test content and the intended test construct (refer to Section 4, Evidence of Content Validity). For test score inferences to support a validity claim, the items should be representative of the content domain, and the content domain should be relevant to the proposed interpretation of test scores. To determine content representativeness, diverse panels of content experts conduct alignment studies, in which experts review individual items and rate them based on how well they match the test specifications or cognitive skills required for a construct (for details on the item development process, refer to Volume 2 of this technical report, Test Development).

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

The second source of validity evidence is based on "the fit between the construct and the detailed nature of performance or response actually engaged in by examinees" (AERA, APA, & NCME, 2014, p.12). This evidence is collected by surveying examinees about their performance strategies or responses to specific items. Because items are developed to measure specific constructs and intellectual processes, evidence that examinees have engaged in relevant performance strategies to correctly answer the items supports the validity of the test scores.

The third source of evidence for validity is based on *internal structure*, which is the degree to which the relationships among test items and test components relate to the construct on which the proposed test scores are interpreted. Dimensionality assessment, goodness-of-model-fit to data, and reliability analysis are possible analyses to examine internal structure (refer to Section 3, Reliability, and Section 4.2, Independent Alignment Study

Evidence of Internal-External Structure It is important to assess the degree to which the statistical relation between items and test components is invariant across groups. DIF analysis can be used to assess whether specific items function differently for subgroups of examinees (refer to Volume 1, Section 4.4, Annual Technical Report).

The fourth source of evidence for validity is the relationship of test scores to external variables. The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) divides this source of evidence into three parts: (1) convergent and discriminant evidence; (2) test-criterion relationships; and (3) validity generalization. Convergent evidence supports the relationship between the test and other measures intended to assess similar constructs. Conversely, discriminant evidence delineates the test from other measures intended to assess different constructs. To analyze both convergent and discriminant evidence, a multi-trait multi-method matrix (MTMM) can be used. Test-criterion relationships indicate how accurately test scores predict criterion performance. The degree of accuracy depends mainly on the test's purpose, such as classification, diagnosis, or selection. Test-criterion evidence is also used to investigate predictions of favoring different groups. Due to construct underrepresentation or construct-irrelevant components, the relation of test scores to a relevant criterion may differ from one group to another. Furthermore, validity generalization is related to whether the evidence is situation-specific or can be generalized across different settings and times. For example, sampling errors or range restrictions may need to be considered to determine whether the conclusions of a test can be assumed for the larger population. Convergent and discriminant validity evidence are discussed in Section 5.2, Convergent and Discriminant Validity.

The fifth source of validity evidence is the fact that intended and unintended consequences of test use should be included in the test-validation process. Determining the validity of the test should depend upon evidence directly related to the test; this process should not be influenced by external factors. For example, if an employer administers a test to determine hiring rates for different groups of people, an unequal distribution of skills related to the measurement construct does not necessarily imply a lack of validity for the test. However, if the unequal distribution of scores is

due to an unintended, confounding aspect of the test, this would interfere with the test's validity. As described in Volume 1, Annual Technical Report, and in this volume, test use should align with the intended purpose of the test.

Supporting a validity argument requires multiple sources of validity evidence. This enables one to evaluate whether sufficient evidence has been presented to support the intended uses and interpretations of the test scores. Thus, determining the validity of a test first requires an explicit statement regarding the intended uses of the test scores and, subsequently, evidence that the scores can be used to support these inferences.

2. Purpose of the South Dakota Science Assessment

The primary purpose of the SDSA is to yield accurate information on students' achievement of South Dakota's Science Standards. The SDSA measures the science knowledge and skills of South Dakota students in grades 5, 8, and 11. The South Dakota Department of Education (SDDOE) provides an overview of the SDSA at: https://doe.sd.gov/Assessment/science.aspx. Information about the South Dakota Science Standards is available at: https://doe.sd.gov/contentstandards/.

The SDSA supports instruction and student learning by measuring growth in student achievement. Assessments can be used as indicators to determine whether students in South Dakota have the knowledge and skills that are essential for college education and careers.

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

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

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

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

3. RELIABILITY

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

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

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

The marginal reliability of science for the overall sample is reported by grade in Table 2. The overall reliability ranges from 0.85 to 0.88. Due to the new structure of the test, Cambium Assessment, Inc. (CAI) explores the relationships between reliability and other important factors, such as the effect of nuisance dimensions (refer to Section 5 of Volume 1, Annual Technical Report). It is found that if the local dependencies among assertions pertaining to the same item are ignored, the marginal reliability will increase. Ignoring local dependencies can be achieved either by computing the maximum likelihood estimates (MLE) ability estimates under the unidimensional Rasch model, or by setting the variance parameters to zero for all item clusters when computing the marginal maximum likelihood estimation (MMLE) ability estimates under the one-parameter logistic (1PL) bifactor model (refer to Section 6 of Volume 1, Annual Technical Report). Therefore, by ignoring the local dependencies, which are substantial for many item clusters, the reliability coefficient is overestimating the true reliability of the test. Note, however, that local dependencies are also present to some degree in traditional assessments that make use of item groups (e.g., a set of items relating to the same reading passage). Traditional assessments typically do not account for local dependencies, and therefore, reported reliability coefficients may be overestimating, to some degree, the true reliability for these tests. The reliability coefficients are also reported for demographics subgroups and reporting categories in Appendix 4-A, Student Demographics and Reliability Coefficients.

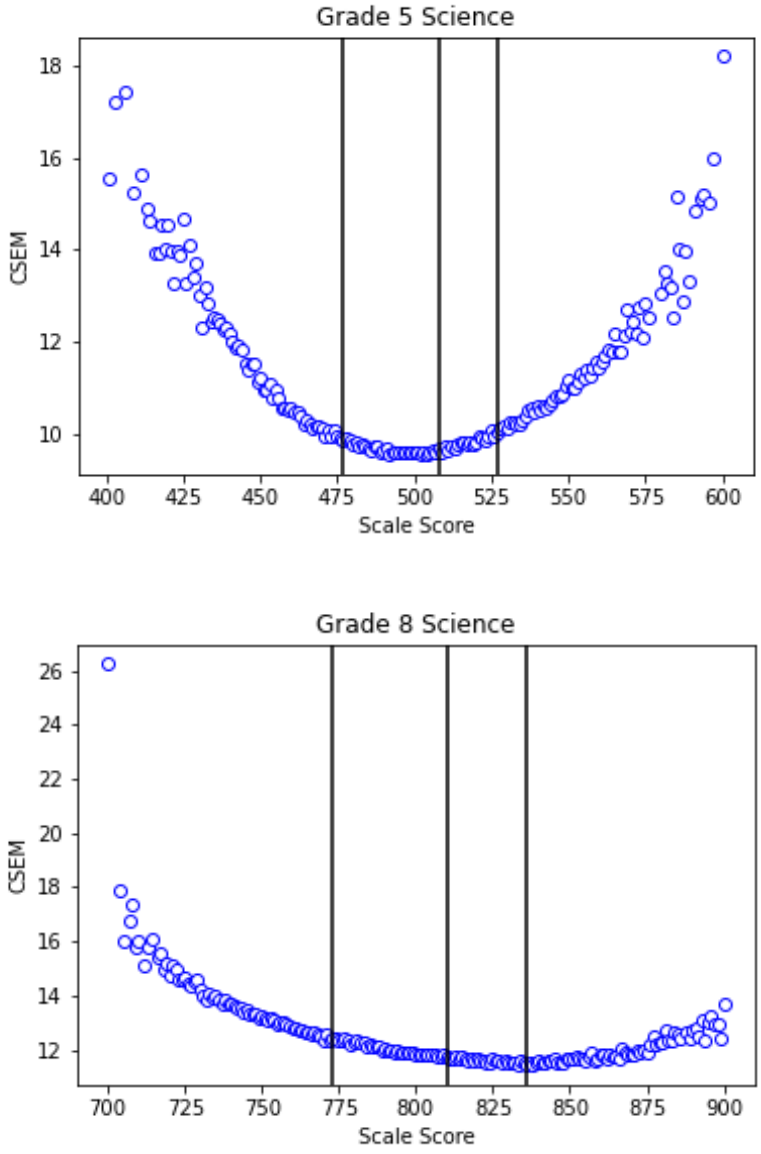
Table 2. Marginal Reliability Coefficients

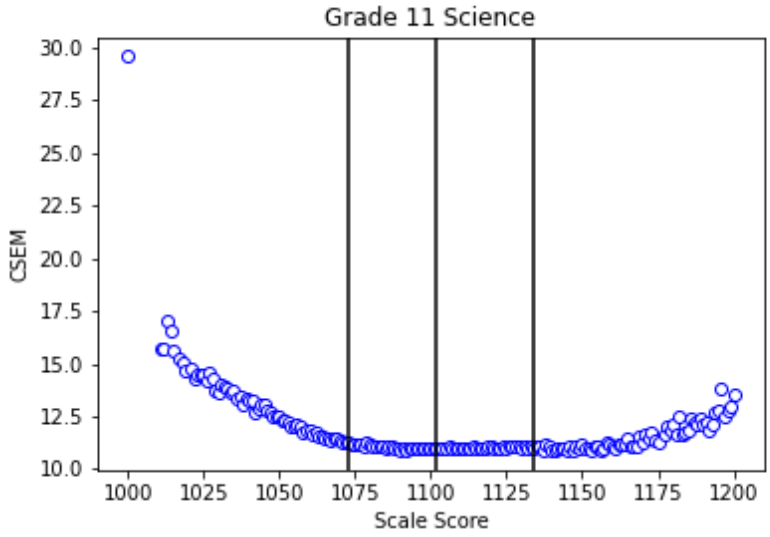
Grade	*N*	Reliability
5	10,605	0.88
8	10,744	0.87
11	9,474	0.85

3.1 Standard Error of Measurement

The computation method of CSEM is described in Section 6.4 of Volume 1, Annual Technical Report. Figure 1 presents the average CSEM for each scale score. The lowest standard errors are observed near the proficiency cut score (the middle vertical line) for grades 5 and 8, which is a desirable test property. The CSEM at each scale score is reported in Appendix 4-B, Conditional Standard Error of Measurement.

Figure 1. Conditional Standard Errors of Measurement





3.2 Reliability of Achievement Classification

When student achievement is reported in terms of achievement levels, the reliability of classifying students into a specific level can be computed in terms of the likelihood of accurate and consistent classification, as specified in Standard 2.16 in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014).

The reliability of achievement classification can be examined in relation to *classification accuracy* (CA) and *classification consistency* (CC). The first term (CA) refers to the agreement between classifications based on the test taken and classifications that would be made on the basis of the students' true scores if, hypothetically, they could be obtained. The second term (CC) refers to the agreement between classifications based on the test taken and classifications that would be made on the basis of an alternate, equivalently constructed test form.

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

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

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

where c_{Ll} and c_{Ul} denote the score corresponding to the lower and upper limits of 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 at level l (based on their true scores) given students from observed level k, and can be calculated as

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

where pl_j is the jth student's observed achievement level. The CA at level l is estimated by

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

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

The CA 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:

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

Table 3 provides overall the CA and the CA for the individual cut scores. The overall CA of the test ranges from 74.78% to 75.57%. The individual cut score accuracy rates are high across all grades, with the minimum value being 87.71% for grade 11. It denotes that we can accurately differentiate students above and below each cut score. The CA for demographic subgroups is presented in Appendix 4-C, Classification Accuracy and Consistency Indices by Subgroups.

Table 3. Classification Accuracy Index

Grade	Overall Accuracy (%)	Cut Score Accuracy (%)		
		Cut Score 1	*Cut Score 2*	*Cut Score 3*
5	75.57	92.44	90.22	92.69
8	75.25	92.50	88.46	94.18
11	74.78	92.82	87.71	94.20

3.2.2 Classification Consistency

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

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

where p_{jl} is the probability of the true score at 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 cuts (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 4 provides the overall CC and the CC for the cut scores. The overall CC of the test ranges from 65.04% to 66.70%. The individual cut score consistency rates are high across all grades, with the minimum value being 82.85% for grade 11. In all achievement levels, CA is slightly higher than CC. CC rates can be lower than CA rates; the consistency is based on two tests with measurement errors, but the accuracy is based on one test with a measurement error and the true score. The CC for demographic subgroups is presented in Appendix 4-C, Classification Accuracy and Consistency Indices by Subgroups.

Table 4. Classification Consistency Index

Grade	Overall Consistency (%)	Cut Score Consistency (%)		
		Cut Score 1	*Cut Score 2*	*Cut Score 3*
5	66.70	89.33	86.27	89.73
8	65.99	89.33	84.05	91.64
11	65.04	89.83	82.85	91.63

3.3 PRECISION AT CUT SCORES

Table 5 presents the mean CSEM at each achievement level by grade and includes achievement-level cut scores and associated CSEM. The CSEM at each scale score is reported in Appendix 4-B, Conditional Standard Error of Measurement.

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

Grade	Achievement Level	Mean CSEM	Cut Score (Scale Score)	CSEM at Cut Score
5	Level 1	10.67	-	-
	Level 2	9.65	477	9.85
	Level 3	9.79	508	9.67
	Level 4	10.66	527	10.04
8	Level 1	13.26	-	-
	Level 2	12.01	773	12.44
	Level 3	11.63	810	11.74
	Level 4	11.72	836	11.48
11	Level 1	12.06	-	-
	Level 2	11.04	1,073	11.31
	Level 3	11.01	1,102	11.03
	Level 4	11.18	1,134	11.01

4. Evidence of Content Validity

The knowledge and skills assessed by the SDSA are representative of the content standards of the larger knowledge domain. In this section, we describe the content standards for the SDSA and discuss the test development process and mapping SDSA 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

Test blueprints were developed to ensure that the test and the items were aligned to the South Dakota Science Standards that they were intended to measure. A complete description of the blueprint and test construction process can be found in Volume 2, Test Development.

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

Table 6. Number of Items for Each Discipline

Grade	Discipline	Item Clusters	Stand-Alone Items
5	Earth and Space Sciences (ESS)	18	31
	Life Sciences (LS)	13	29
	Physical Sciences (PS)	17	35
8	ESS	14	23
	LS	18	41
	PS	15	30
11	ESS	11	22
	LS	20	43
	PS	16	23

4.2 INDEPENDENT ALIGNMENT STUDY

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

The Wisconsin Center for Education Products and Services and WebbAlign served as an external evaluator and conducted an alignment study in June 2022. The purpose of the study was to examine the extent to which the SDSA item pool represented the South Dakota Science Content Standards as represented by the test blueprints in terms of range, complexity, depth, and breadth. The results of the alignment study are presented in Appendix 4-D, Independent Alignment Study Report.

In summary, study results suggest that the overall SDSA item bank for grade 5 had the capacity to fully meet all alignment criteria agreed upon and used in this study. For grades 8 and 11, study results suggest these item banks have the capacity to fully meet all alignment criteria except that items addressed at least 90% of standards for Range of Knowledge Correspondence–Population. It was concluded that the relative weaker Range of Knowledge (Population) for grades 8 and 11 item banks could be fully resolved with the addition of at least four items to the middle school item bank and six items to the high school item bank. SDDOE as well as CAI Content experts were informed the study results and have plans to expand the item bank for future administrations.

5. EVIDENCE OF INTERNAL-EXTERNAL STRUCTURE

In this section, the internal structure of the assessment is explored using the scores provided at the discipline level. The relationship between the discipline scores is just one indicator of the test dimensionality. The SDSA is calibrated with the Rasch testlet model (Wang & Wilson, 2005). The testlet 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 (CFA) is used to evaluate the fit of the IRT model and to compare the model to alternative models, including models with a simpler internal structure (i.e., unidimensional models without cluster effects) and models with a more elaborate internal structure (refer to Section 5.4, Confirmatory Factor Analysis).

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

5.1 Correlations Among Discipline Scores

Table 7 presents the observed and disattenuated correlation matrix of the discipline scores. The observed correlations range from 0.62 to 0.72, and disattenuated correlations range from 0.96 to 1.00.

In some instances, the observed correlations were lower than one might expect. However, as previously noted, the correlations were subject to a large amount of measurement error at the discipline level, given the limited number of items from which the scores were derived. Consequently, over-interpretation of these correlations as either high or low were made cautiously. After correcting for measurement error, the correlations between the discipline scores became very high. The dissattenuated correlations were close to 1, supporting the use of a psychometric model that did not include a separate dimension for each of the three disciplines.

Table 7. Correlations Among Disciplines

Grade	Discipline	Earth and Space Sciences (ESS)	Life Sciences (LS)	Physical Sciences (PS)
5	ESS	0.72*	0.96	0.96
	LS	0.68	0.70*	0.96
	PS	0.68	0.68	0.70*
8	ESS	0.65*	0.98	0.98
	LS	0.65	0.69*	1.00
	PS	0.66	0.69	0.69*
11	ESS	0.62*	1.00	1.00
	LS	0.67	0.69*	1.00
	PS	0.62	0.67	0.62*

Note. The values for cells shaded on the diagonal are marginal reliabilities for each discipline. Below the cells shaded on the diagonal are the observed correlations, and above the cells shaded on the diagonal are the disattenuated correlations. The disattenuated correlations larger than 1 were truncated to 1. * Indicates that the correlation is statistically significant at $p < .05$.

5.2 CONVERGENT AND DISCRIMINANT VALIDITY

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

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

Table 8. Correlations Across Subjects, Grade 5

Subject	Number of Students	Reporting Category	Science			ELA				Mathematics		
			ESS	*LS*	*PS*	*R*	*W*	*L*	*RES*	*CP*	*PMD*	*CR*
Science	10,546	Earth and Space Sciences (ESS)	0.72*	0.96	0.96	0.88	0.79	0.85	0.86	0.84	0.93	0.90
		Life Sciences (LS)	0.68	0.70*	0.96	0.89	0.81	0.87	0.88	0.82	0.91	0.88
		Physical Sciences (PS)	0.68	0.67	0.70*	0.88	0.79	0.87	0.85	0.84	0.92	0.88
English Language Arts (ELA)		Reading (R)	0.64	0.64	0.63	0.74*	0.86	0.91	0.93	0.79	0.88	0.86
		Writing (W)	0.57	0.58	0.56	0.63	0.73*	0.83	0.85	0.79	0.86	0.83
		Listening (L)	0.58	0.58	0.58	0.63	0.56	0.64*	0.89	0.78	0.86	0.84
		Research (RES)	0.62	0.63	0.61	0.68	0.62	0.61	0.73*	0.80	0.88	0.86
Mathematics		Concepts and Procedures (CP)	0.67	0.65	0.66	0.64	0.63	0.59	0.64	0.89*	0.99	0.97
		Problem Solving, Modeling, and Data Analysis (PMD)	0.65	0.63	0.64	0.63	0.61	0.57	0.62	0.78	0.69*	1.00
		Communicating and Reasoning (CR)	0.64	0.62	0.62	0.62	0.59	0.56	0.61	0.77	0.72	0.70*

Note. Cells shaded on the diagonal represent the reliability coefficient of the reporting category. Observed correlations are below the cells shaded on the diagonal; disattenuated correlations are above. The disattenuated correlations larger than 1 were truncated to 1. * Indicates that the correlation is statistically significant at $p < .05$.

Table 9. Correlations Across Subjects, Grade 8

Subject	Number of Students	Reporting Category	Science			ELA				Mathematics		
			ESS	*LS*	*PS*	*R*	*W*	*L*	*RES*	*CP*	*PMD*	*CR*
Science	10,634	Earth and Space Sciences (ESS)	0.65*	0.97	0.98	0.86	0.79	0.85	0.81	0.85	0.94	0.87
		Life Sciences (LS)	0.65	0.69*	0.99	0.89	0.81	0.87	0.85	0.85	0.94	0.87
		Physical Sciences (PS)	0.65	0.69	0.69*	0.89	0.81	0.85	0.86	0.86	0.94	0.88
English Language Arts (ELA)		Reading (R)	0.60	0.64	0.64	0.75*	0.90	0.93	0.92	0.81	0.91	0.84
		Writing (W)	0.54	0.57	0.57	0.66	0.72*	0.84	0.88	0.81	0.87	0.82
		Listening (L)	0.53	0.56	0.55	0.62	0.55	0.60*	0.89	0.80	0.88	0.83
		Research (RES)	0.54	0.58	0.59	0.66	0.62	0.57	0.69*	0.77	0.86	0.80
Mathematics		Concepts and Procedures (CP)	0.64	0.66	0.67	0.66	0.64	0.58	0.60	0.87*	1.00	0.99
		Problem Solving, Modeling, and Data Analysis (PMD)	0.62	0.64	0.64	0.65	0.60	0.56	0.58	0.78	0.67*	1.00
		Communicating and Reasoning (CR)	0.57	0.58	0.60	0.59	0.56	0.52	0.54	0.75	0.70	0.66*

Note. Cells shaded on the diagonal represent the reliability coefficient of the reporting category. Observed correlations are below the cells shaded on the diagonal; disattenuated correlations are above. The disattenuated correlations larger than 1 were truncated to 1. * Indicates that the correlation is statistically significant at $p < .05$.

Table 10. Correlations Across Subjects, Grade 11

Subject	Number of Students	Reporting Category	Science			ELA				Mathematics		
			ESS	*LS*	*PS*	*R*	*W*	*L*	*RES*	*CP*	*PMD*	*CR*
Science	9,404	Earth and Space Sciences (ESS)	0.62*	1.00	1.00	0.87	0.80	0.83	0.82	0.85	0.90	0.83
		Life Sciences (LS)	0.67	0.69*	1.00	0.89	0.82	0.85	0.84	0.87	0.91	0.85
		Physical Sciences (PS)	0.62	0.66	0.62*	0.86	0.81	0.83	0.82	0.87	0.90	0.86
English Language Arts (ELA)		Reading (R)	0.59	0.64	0.58	0.75*	0.88	0.93	0.94	0.80	0.83	0.80
		Writing (W)	0.54	0.58	0.55	0.66	0.74*	0.85	0.90	0.82	0.83	0.81
		Listening (L)	0.50	0.54	0.50	0.62	0.56	0.59*	0.91	0.78	0.81	0.77
		Research (RES)	0.54	0.58	0.54	0.67	0.64	0.58	0.69*	0.79	0.82	0.77
Mathematics		Concepts and Procedures (CP)	0.63	0.68	0.64	0.65	0.67	0.56	0.62	0.89*	0.95	0.94
		Problem Solving, Modeling, and Data Analysis (PMD)	0.59	0.63	0.59	0.60	0.59	0.52	0.57	0.75	0.70*	0.94
		Communicating and Reasoning (CR)	0.55	0.59	0.56	0.58	0.58	0.49	0.53	0.74	0.65	0.69*

Note. Cells shaded on the diagonal represent the reliability coefficient of the reporting category. Observed correlations are below the cells shaded on the diagonal; disattenuated correlations are above. The disattenuated correlations larger than 1 were truncated to 1. * Indicates that the correlation is statistically significant at $p < .05$.

Additionally, the correlation is computed among the overall scores for the three tested subjects: ELA, mathematics, and science as shown in Table 11.

Table 11. Correlations Across Spring 2023
English Language Arts, Mathematics, and Science Scores

Grade	*N*	ELA & Mathematics	ELA & Science	Mathematics & Science
5	10,546	0.79	0.81	0.79
8	10,634	0.78	0.78	0.78
11	9,404	0.78	0.76	0.78

5.3 CLUSTER EFFECTS

The SDSA is calibrated with the Rasch testlet model (Wang & Wilson, 2005). The testlet model is a high-dimensional model, incorporating a nuisance dimension for each item cluster in addition to an overall dimension representing the overall proficiency. Section 5 of Volume 1, Annual Technical Report, presents a detailed description of the IRT model. The internal (latent) structure of the model is presented in Figure 7. The validity evidence on the internal structure presented in this section relates to the presence of cluster effects (i.e., nuisance dimensions) and how substantial they are.

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

CAI examined the distribution of cluster variances obtained from the 2019 IRT calibrations for the entire Independent College and Career Readiness (ICCR) item bank.

For elementary school, the estimated value of the cluster variances of all operational, scored items ranged from 0 to 5.13, with a median value of 0.38 and a mean value of 0.78. As a comparison, the estimated variance parameter of the overall dimension for South Dakota elementary school in 2021 was $\widehat{\sigma}^2_{\theta SD} = 0.78$.

For middle school, the estimated value of the cluster variances of all operational, scored items ranged from 0 to 2.47, with a median value of 0.43 and a mean value of 0.57. The estimated variance parameter of the overall dimension for South Dakota middle school in 2021 was $\widehat{\sigma}^2_{\theta SD} = 0.47$.

For high school, the estimated value of the cluster variances of all operational, scored items ranged from 0.07 to 2.58, with a median value of 0.43 and a mean value of 0.52. The estimated variance parameter of the overall dimension for South Dakota high school in 2021 was $\widehat{\sigma}^2_{\theta SD} = 0.49$.

Figure 2 through Figure 4 present the histograms of the cluster variances expressed as the proportion of the systematic variance due to the cluster variance for each cluster (computed as $\eta_g = \frac{\hat{\sigma}_g^2}{\hat{\sigma}_{\theta_{SD}}^2 + \hat{\sigma}_g^2}$), where $\hat{\sigma}_{\theta_{SD}}^2$ is the variance estimate of the overall proficiency of South Dakota students. The variance proportion shows the relative magnitude of the variance of a cluster compared to the variance of the overall dimension. For instance, if the variance proportion of a cluster is larger than 0.5, then the cluster variance is larger than the overall variance; otherwise, the cluster variance is smaller than the overall variance. For all three grade bands, a wide range of cluster variances is observed. These results indicate that, for all grades, cluster effects can be substantial and provide evidence for the appropriateness of a psychometric model that explicitly takes local dependencies among the assertions of an item cluster into account.

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

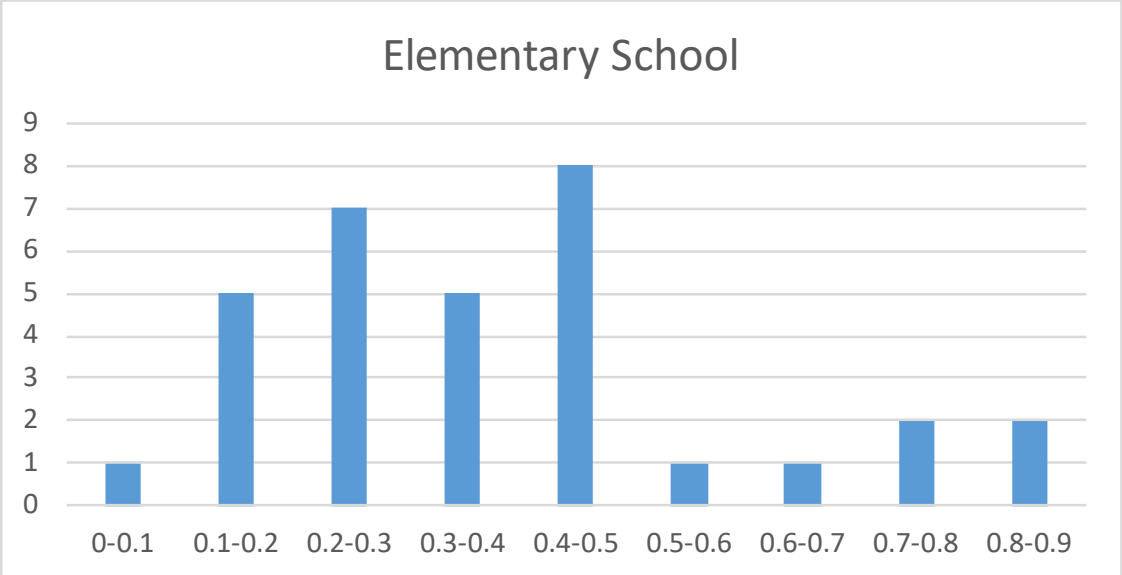


Figure 3. Cluster Variance Proportion for Operational Items in Middle School

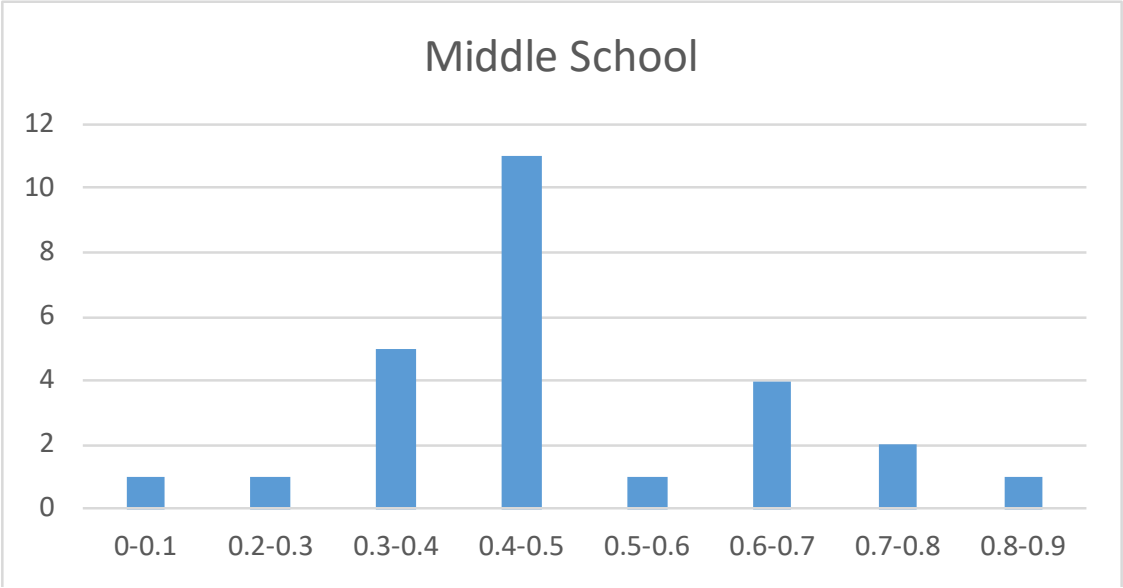
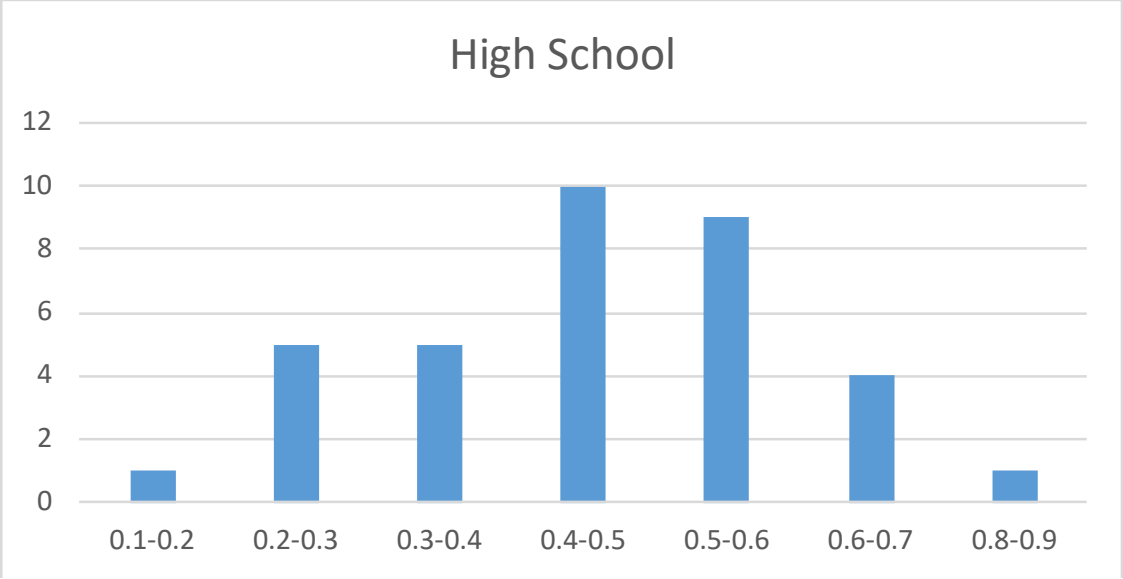


Figure 4. Cluster Variance Proportion for Operational Items in High School



5.4 CONFIRMATORY FACTOR ANALYSIS

In Section 5.3, Cluster Effects, evidence is presented for the existence of substantial cluster effects. In this section, the internal structure of the IRT model used for calibrating the item parameters is further evaluated using 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 CFA for discrete, observed variables are not well suited for incomplete data collection designs where each case has data only on a subset of the set of observed variables. The linear-on-the-fly (LOFT) test design results in sparse data matrices. Every student is only responding to a small number of items relative to the size of the item pool, so data are missing on most of the manifest variables for any given student. In 2018 and 2019, a LOFT test design was used for all operational science assessments inspired by a three-dimensional science framework, except for Utah's assessments. As a result, the student responses of these other states are not readily amenable for the application of CFA techniques.

The 2018 Utah operational field test for science made use of a set of fixed-form tests for each grade. Therefore, the data for each fixed-form test were complete, and the fixed-form tests were amenable to CFA. The Utah science standards, even though the standards are grade-specific for middle school, were developed under a framework similar to the one developed for the Next Generation Science Standards (NGSS), and a crosswalk was available between both sets of standards.

Utah is part of the Memorandum of Understanding (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 with respect to the internal structure of the SDSA.

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

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

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

The factor structure of a testlet model, which is the model used for calibration, is formally equivalent to a second-order model. Specifically, the testlet model is the model obtained after a Schmid Leiman transformation of the second-order model (Li, Bolt, & Fu, 2006; Rijmen, 2009; Yung, Thissen, & McLeod, 1999). In the corresponding second-order model, the group of assertions related to 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 CFA to evaluate the fit of the second-order model (described previously) to student data from spring 2018. Three additional structural models were included in the analysis as well. In the first model, there was only one factor representing overall science achievement. All assertions were indicators of this overall proficiency factor. The first model was a testlet model where all cluster variances were zero. In the second model, assertions were indicators of the corresponding science discipline, and each discipline was an indicator of the overall science achievement. This was a second-order model with science disciplines rather than clusters as first-order factors. This model did not take the cluster effects into account. In the last, most general model, assertions were indicators of the corresponding cluster, and clusters were indicators of the corresponding science discipline, with disciplines being indicators of the overall science achievement.

For the sake of simplicity, the models in the analysis are referred to as the following:

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

Figure 5 through

Figure 8 illustrate these four structural models. Model 1 is nested within Models 2, 3, and 4. Also, Models 2 and 3 are nested within Model 4. The paths from the factors to the assertions represent the first-order factor loadings. Note that all four models include factor loadings for the assertions,

which is different from the calibration model for which all the discrimination parameters of the assertions were set to 1.

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

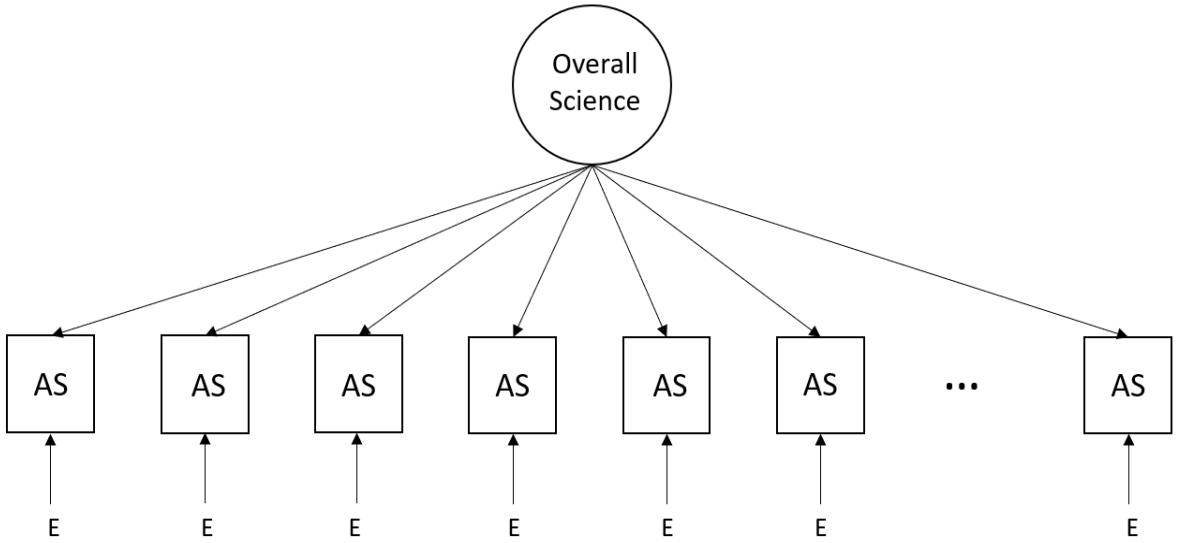


Figure 6. Second-Order Structural Model (Assertions-Disciplines-Overall Science): Model 2

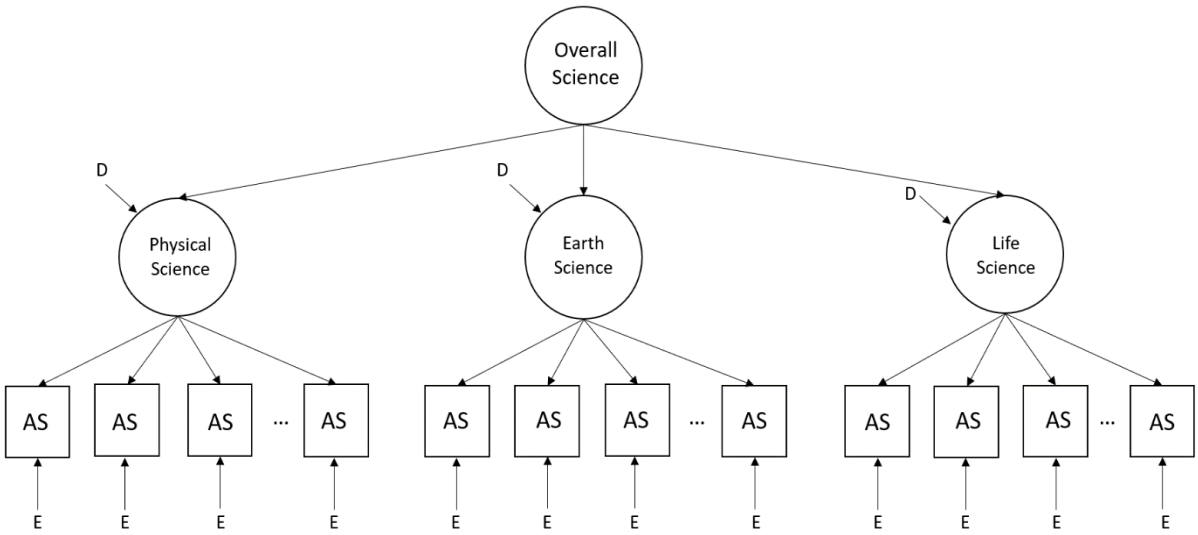


Figure 7. Second-Order Structural Model (Assertions-Clusters-Overall Science): Model 3

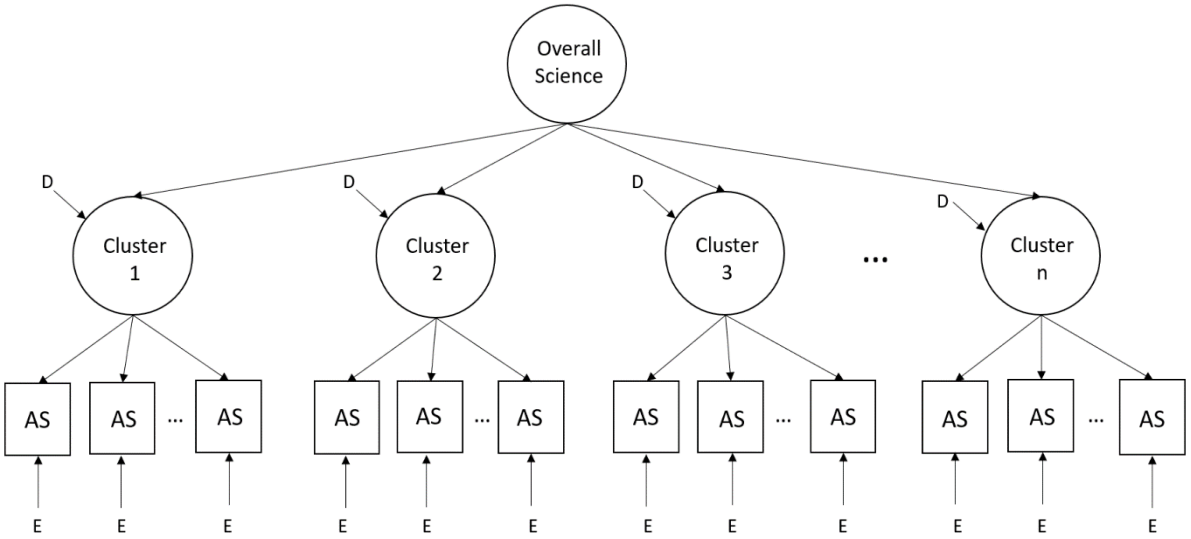
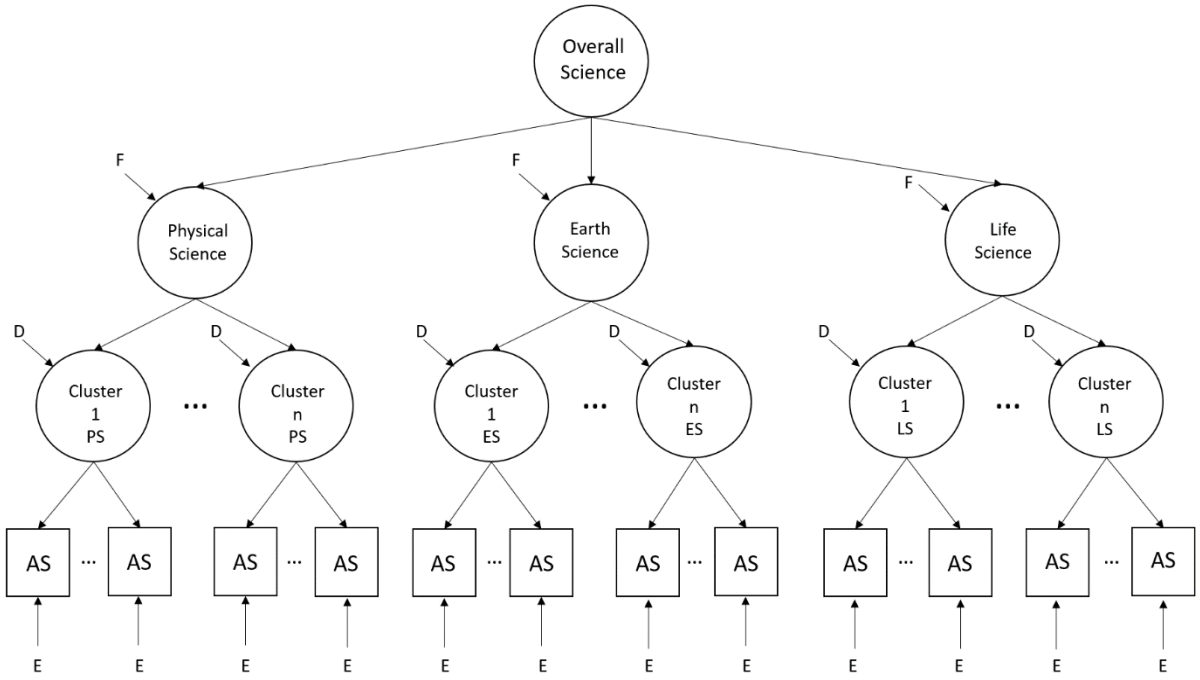


Figure 8. Third-Order Structural Model (Assertions-Clusters-Disciplines-Overall Science): Model 4



5.4.1 Results

For each test form, fit measures were computed for each of the four models. The fit measures used to evaluate goodness-of-fit were the comparative fit index (CFI), the Tucker-Lewis index (TLI), the root mean square error of approximation (RMSEA), and the standardized root mean residual (SRMR). CFI and TLI were relative fit indices, meaning they evaluated model fit by comparing the model of interest to a baseline model. RMSEA and SRMR were indices of absolute fit. Table 13 provides a list of these measures, along with the corresponding thresholds indicating a good fit.

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

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

*Brown, 2015; Hu & Bentler, 1999.

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

- Model 1 shows the most misfit across grades and forms.
- Across forms, Model 3 generally shows more improvement in model fit relative to Model 1 than Model 2 (i.e., higher values for CFI and TLI and lower values for RMSEA and SRMR). This means that accounting for the clusters results in a higher improvement in model fit over a single factor model than accounting for disciplines.
- Model 4 does not show improvement in model fit over Model 3. Fit measures remain 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 were taken into account) does not improve model fit.
- Overall model fit for Models 3 and 4 decreases with decreasing grades. For grade 8, all fit indices for Models 3 and 4 indicate good model fit for all three forms. For grade 7, all fit indices for Models 3 and 4 indicate good fit for two out of the six forms, and the degree of misfit for the other four forms is small. For grade 6, all three forms have fit indices above the threshold values for at least one of the absolute fit indices for Models 3 and 4.

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

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

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

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

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

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

Model	Form	CFI	TLI	RMSEA	SRMR
Model 1 Assertions-Overall Science (one-factor model)	**1**	**0.892**	**0.889**	0.060	0.074
	2	**0.938**	**0.936**	**0.083**	**0.109**
	3	**0.940**	**0.939**	0.052	0.065
	4	**0.937**	**0.936**	**0.068**	**0.114**
	5	**0.939**	**0.937**	**0.093**	**0.119**
	6	**0.898**	**0.895**	0.056	0.071
Model 2 Assertions-Disciplines-Overall Science (second-order model)	**1**	**0.908**	**0.906**	0.055	0.073
	2	0.962	0.961	**0.065**	**0.088**
	3	0.950	**0.949**	0.048	0.063
	4	0.955	0.954	0.058	**0.094**
	5	0.959	0.957	**0.077**	**0.103**
	6	**0.906**	**0.903**	0.054	0.070
Model 3 Assertions-Clusters-Overall Science (second-order model)	**1**	**0.938**	**0.937**	0.046	0.072
	2	0.974	0.973	0.054	**0.082**
	3	0.967	0.966	0.039	0.055
	4	0.977	0.976	0.041	0.072

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

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

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

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

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

For Models 3 and 4, grade 6 shows some degree of misfit across all three forms according to the measures of absolute model fit, especially for the SRMR. Further examination indicates that the lack of fit could be attributed to a single item that is common to all three grade-6 forms that are part of this factor analysis study. When this item is removed, only two forms have two or more clusters per discipline. The fit for both forms improves drastically in Models 3 and 4, with all fit measures excepting the SRMR for one form meeting the criteria for model fit. The SRMR value that exceeds the threshold value does so barely, with a value of 0.083. Table 17 shows

the fit measures for grade 6 after removal of the item causing misfit. Note that, unlike Models 3 and 4, Models 1 and 2 still do not meet the criteria of model fit after removing the item.[2]

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

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

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

Table 18 shows the estimated correlations among disciplines for Model 4 (third-order model). The correlations are all very high, ranging between 0.913 and 1.0. The high correlations between the disciplines in Model 4 indicate that, after considering the cluster effects, the disciplines do not add much to the model. This may explain why Model 4 does not show an improvement in fit compared to Model 3. Overall, the findings support the IRT model used for calibration.

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

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

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

Grade	Form	Discipline	Earth and Space Sciences (ESS)	Life Sciences (LS)
	4	Physical Sciences (PS)	0.938	0.913
		Earth and Space Sciences (ESS)	–	0.973
	5	Physical Sciences (PS)	0.931	0.944
		Earth and Space Sciences (ESS)	–	0.965
	6	Physical Sciences (PS)	0.941	0.928
		Earth and Space Sciences (ESS)	–	0.967
8	1	Physical Sciences (PS)	0.971	0.971
		Earth and Space Sciences (ESS)	–	0.970
	2	Physical Sciences (PS)	0.956	0.958
		Earth and Space Sciences (ESS)	–	0.935
	3	Physical Sciences (PS)	0.966	0.978
		Earth and Space Sciences (ESS)	–	0.988

5.4.2 Conclusion

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

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

6. FAIRNESS IN CONTENT

The principles of universal design (UD) provide guidance for developing tests that minimize the impact of construct-irrelevant factors on assessments of student achievement. UD enables access for the widest possible range of students. The following seven principles of UD 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 UD and apply them in the development of all test materials. In the review process, South Dakota educators and stakeholders verify adherence to the principles of UD. More details on how to reduce construct-irrelevant variance through universal design and on training on the principles of universal design are described in Section 2, Item Development Process That Supports Validity of Claims, as well as Appendix 2-C, Style Guide for Science Items, of Volume 2 of this technical report.

6.1 COGNITIVE LABORATORY STUDIES

In 2017, when the development of item clusters for the states that are part of the MOU started, cognitive lab studies were carried out to evaluate and refine the process of developing item clusters aligned to three-dimensional science standards. 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 was carried out in 2018 and 2019 to determine whether students using braille understood the task demands of selected accommodated three-dimensional science-aligned item clusters and were able to navigate the interactive features of these clusters in a manner that allowed them to fully display their knowledge and skills relative to the constructs being measured. In general, both the students who relied entirely on braille and/or Job Access With Speech (JAWS) and those who had enough vision to read on-screen text with magnification were able to find the information they needed to respond to the questions, navigate the various response formats, and finish within a reasonable amount of time. The clusters were clearly different from (and more complex than) other tests with which the students were familiar; however, the study recommended that students should be given adequate time to practice with at least one sample

cluster before taking the summative test. The study findings also proposed tool-specific recommendations on accessibility for visually impaired students. The reports of both sets of cognitive laboratory studies are presented in Appendix 4-E, Science Clusters Cognitive Lab Report, and Appendix 4-F, Braille Cognitive Lab Report.

6.2 STATISTICAL FAIRNESS IN ITEM STATISTICS

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

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.*** This result provides various measures of reliability at the aggregate and subgroup levels, showing that the reliability of all tests is in line with acceptable industry standards.
- ***Content Validity.*** This result provides evidence 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.*** This result provides evidence 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.*** This result provides evidence of convergent and discriminant validity to support the relationship between the test and other measures intended to assess similar and different constructs.
- ***Test Fairness.*** This result provides evidence that items are developed following the principles of universal design, which enables access for the widest possible range of students. Evidence of test fairness is provided statistically using DIF analysis in tandem with content review by specialists.

8. REFERENCES

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

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

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

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

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

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

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

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

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

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

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

Wang, W. C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement*, *29*, 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 4-A

Student Demographics and Reliability Coefficients

Student Demographics and Reliability Coefficients

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

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

*Subgroup is not reported due to small size ($N < 10$).

Table 1-A-2. Scale Score Summary by Reporting Category, Science Grade 5

Reporting Category	Mean	SD	Min	Max	Reliability	SEM
Physical Sciences	499.68	32.53	400.14	599.99	0.70	17.48
Earth and Space Sciences	499.97	33.77	400.14	599.99	0.72	17.50
Life Sciences	499.01	36.01	400.14	599.99	0.70	19.51

Table 1-A-3. Scale Score Summary by Reporting Category, Science Grade 8

Reporting Category	Mean	SD	Min	Max	Reliability	SEM
Physical Sciences	799.52	38.70	700.31	899.97	0.69	21.44
Earth and Space Sciences	798.50	36.58	700.31	899.97	0.65	21.42
Life Sciences	797.03	39.65	700.31	899.97	0.69	21.89

Table 1-A-4. Scale Score Summary by Reporting Category, Grade 11

Reporting Category	Mean	SD	Min	Max	Reliability	SEM
Physical Sciences	1,102.92	33.73	1,000.3	1,199.74	0.62	20.58
Earth and Space Sciences	1,099.77	31.54	1,000.3	1,199.74	0.62	19.38
Life Sciences	1,100.16	35.78	1,000.3	1,199.74	0.69	19.70

Appendix 4-B

Conditional Standard Error of Measurement

Conditional Standard Error of Measurement

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

Science Grade 5

Scale Score	Achievement Level	CSEM
401	1	15.53
403	1	17.22
406	1	17.45
409	1	15.26
411	1	15.63
413	1	14.89
414	1	14.61
416	1	13.93
417	1	13.91
418	1	14.54
419	1	14.03
420	1	14.54
421	1	13.98
422	1	13.26
423	1	13.95
424	1	13.87
425	1	14.66
426	1	13.26
427	1	14.12
428	1	13.38
429	1	13.70
430	1	13.01
431	1	12.30
432	1	13.18
433	1	12.82
434	1	12.43
435	1	12.53
436	1	12.49
437	1	12.38
438	1	12.26
439	1	12.30

Science Grade 5

Scale Score	Achievement Level	CSEM
440	1	12.17
441	1	12.02
442	1	11.86
443	1	11.93
444	1	11.84
445	1	11.50
446	1	11.39
447	1	11.52
448	1	11.50
449	1	11.11
450	1	11.19
451	1	10.96
452	1	10.95
453	1	11.10
454	1	10.75
455	1	10.95
456	1	10.79
457	1	10.57
458	1	10.57
459	1	10.49
460	1	10.55
461	1	10.46
462	1	10.45
463	1	10.38
464	1	10.21
465	1	10.30
466	1	10.19
467	1	10.11
468	1	10.15
469	1	10.16
470	1	10.11
471	1	9.96
472	1	10.08
473	1	9.96

Science Grade 5

Scale Score	Achievement Level	CSEM
474	1	10.05
475	1	9.92
476	1	9.89
477	2	9.85
478	2	9.88
479	2	9.84
480	2	9.76
481	2	9.81
482	2	9.74
483	2	9.75
484	2	9.72
485	2	9.66
486	2	9.64
487	2	9.70
488	2	9.71
489	2	9.60
490	2	9.63
491	2	9.66
492	2	9.55
493	2	9.59
494	2	9.57
495	2	9.59
496	2	9.59
497	2	9.59
498	2	9.58
499	2	9.58
500	2	9.58
501	2	9.58
502	2	9.54
503	2	9.60
504	2	9.55
505	2	9.59
506	2	9.64
507	2	9.59
508	3	9.67

Science Grade 5

Scale Score	Achievement Level	CSEM
509	3	9.58
510	3	9.74
511	3	9.64
512	3	9.74
513	3	9.67
514	3	9.78
515	3	9.81
516	3	9.82
517	3	9.76
518	3	9.80
519	3	9.77
520	3	9.79
521	3	9.92
522	3	9.89
523	3	9.87
524	3	9.95
525	3	10.05
526	3	9.93
527	4	10.04
528	4	10.10
529	4	10.15
530	4	10.12
531	4	10.23
532	4	10.23
533	4	10.22
534	4	10.22
535	4	10.27
536	4	10.37
537	4	10.51
538	4	10.54
539	4	10.46
540	4	10.60
541	4	10.50
542	4	10.59
543	4	10.56

Science Grade 5

Scale Score	Achievement Level	CSEM
544	4	10.65
545	4	10.74
546	4	10.80
547	4	10.83
548	4	10.88
549	4	11.03
550	4	11.15
551	4	11.00
552	4	10.99
553	4	11.11
554	4	11.28
555	4	11.21
556	4	11.38
557	4	11.25
558	4	11.45
559	4	11.55
560	4	11.41
561	4	11.58
562	4	11.71
563	4	11.81
564	4	11.76
565	4	12.16
566	4	11.78
567	4	11.80
568	4	12.12
569	4	12.69
570	4	12.22
571	4	12.44
572	4	12.19
573	4	12.73
574	4	12.10
575	4	12.82
576	4	12.53
580	4	13.03
581	4	13.51

Science Grade 5

Scale Score	Achievement Level	CSEM
582	4	13.26
583	4	13.20
584	4	12.51
585	4	15.14
586	4	14.00
587	4	12.87
588	4	13.99
589	4	13.33
591	4	14.86
593	4	15.11
594	4	15.20
596	4	15.01
597	4	15.98
600	4	18.20

Table 1-B-2. CSEM at Each Scale Score, Science Grade 8

Science Grade 8

Scale Score	Achievement Level	CSEM
700	1	26.24
704	1	17.87
705	1	15.99
707	1	16.74
708	1	17.33
709	1	15.80
710	1	16.03
712	1	15.11
713	1	15.77
714	1	16.10
716	1	15.44
717	1	15.57
718	1	15.00
719	1	15.21
720	1	14.76
721	1	15.10
722	1	15.00
723	1	14.56
724	1	14.59
725	1	14.70
726	1	14.47
727	1	14.34
728	1	14.53
729	1	14.60
730	1	14.24
731	1	14.00
732	1	13.84
733	1	14.05
734	1	13.90
735	1	14.03
736	1	13.87
737	1	13.73
738	1	13.81

Science Grade 8

Scale Score	Achievement Level	CSEM
739	1	13.69
740	1	13.73
741	1	13.63
742	1	13.52
743	1	13.56
744	1	13.41
745	1	13.47
746	1	13.34
747	1	13.35
748	1	13.29
749	1	13.15
750	1	13.22
751	1	13.19
752	1	13.13
753	1	13.14
754	1	13.06
755	1	12.93
756	1	13.03
757	1	12.99
758	1	12.91
759	1	12.87
760	1	12.83
761	1	12.80
762	1	12.70
763	1	12.74
764	1	12.66
765	1	12.67
766	1	12.57
767	1	12.67
768	1	12.55
769	1	12.47
770	1	12.38
771	1	12.54
772	1	12.38
773	2	12.44

Science Grade 8

Scale Score	Achievement Level	CSEM
774	2	12.45
775	2	12.37
776	2	12.35
777	2	12.39
778	2	12.32
779	2	12.22
780	2	12.28
781	2	12.31
782	2	12.27
783	2	12.25
784	2	12.11
785	2	12.17
786	2	12.09
787	2	12.13
788	2	12.13
789	2	12.01
790	2	12.02
791	2	11.95
792	2	11.99
793	2	11.92
794	2	12.01
795	2	11.93
796	2	11.93
797	2	11.92
798	2	11.92
799	2	11.90
800	2	11.85
801	2	11.86
802	2	11.84
803	2	11.79
804	2	11.81
805	2	11.84
806	2	11.79
807	2	11.77
808	2	11.78

Science Grade 8

Scale Score	Achievement Level	CSEM
809	2	11.82
810	3	11.74
811	3	11.71
812	3	11.69
813	3	11.72
814	3	11.66
815	3	11.75
816	3	11.67
817	3	11.61
818	3	11.71
819	3	11.63
820	3	11.64
821	3	11.62
822	3	11.60
823	3	11.54
824	3	11.52
825	3	11.66
826	3	11.62
827	3	11.59
828	3	11.55
829	3	11.57
830	3	11.55
831	3	11.51
832	3	11.50
833	3	11.44
834	3	11.59
835	3	11.53
836	4	11.48
837	4	11.52
838	4	11.44
839	4	11.54
840	4	11.60
841	4	11.49
842	4	11.56
843	4	11.61

Science Grade 8

Scale Score	Achievement Level	CSEM
844	4	11.63
845	4	11.68
846	4	11.55
847	4	11.54
848	4	11.55
849	4	11.65
850	4	11.69
851	4	11.71
852	4	11.77
853	4	11.70
854	4	11.67
855	4	11.60
856	4	11.73
857	4	11.90
858	4	11.59
859	4	11.66
860	4	11.80
861	4	11.85
862	4	11.64
863	4	11.82
864	4	11.77
865	4	11.78
866	4	11.70
867	4	12.02
868	4	11.90
869	4	11.83
870	4	11.81
871	4	11.83
872	4	12.01
873	4	11.93
874	4	11.95
875	4	11.91
876	4	12.18
877	4	12.53
878	4	12.17

Science Grade 8

Scale Score	Achievement Level	CSEM
879	4	12.25
880	4	12.24
881	4	12.74
882	4	12.32
883	4	12.33
884	4	12.63
885	4	12.56
886	4	12.43
878	4	12.17
879	4	12.25
880	4	12.24
881	4	12.74
882	4	12.32
883	4	12.33
884	4	12.63
885	4	12.56
886	4	12.43
888	4	12.66
889	4	12.40
890	4	12.73
891	4	12.76
892	4	12.48
893	4	13.06
894	4	12.35
895	4	13.01
896	4	13.23
897	4	12.91
898	4	12.91
899	4	12.39
900	4	13.71

Table 1-B-3. CSEM at Each Scale Score, Grade 11

Science Grade 11

Scale Score	Achievement Level	CSEM
1,000	1	29.56
1,011	1	15.69
1,012	1	15.68
1,013	1	17.02
1,014	1	16.53
1,015	1	15.57
1,017	1	15.26
1,018	1	15.03
1,019	1	14.66
1,021	1	14.73
1,022	1	14.34
1,023	1	14.49
1,024	1	14.46
1,025	1	14.51
1,026	1	14.21
1,027	1	14.57
1,028	1	14.26
1,029	1	13.76
1,030	1	13.63
1,031	1	14.04
1,032	1	13.88
1,033	1	13.83
1,034	1	13.62
1,035	1	13.72
1,036	1	13.38
1,037	1	13.40
1,038	1	13.11
1,039	1	13.35
1,040	1	13.21
1,041	1	13.27
1,042	1	12.69
1,043	1	12.85
1,044	1	13.05

Science Grade 11

Scale Score	Achievement Level	CSEM
1,045	1	13.02
1,046	1	12.79
1,047	1	12.72
1,048	1	12.49
1,049	1	12.54
1,050	1	12.47
1,051	1	12.35
1,052	1	12.33
1,053	1	12.19
1,054	1	12.00
1,055	1	12.02
1,056	1	12.08
1,057	1	12.03
1,058	1	11.74
1,059	1	11.85
1,060	1	11.85
1,061	1	11.62
1,062	1	11.72
1,063	1	11.55
1,064	1	11.59
1,065	1	11.48
1,066	1	11.42
1,067	1	11.37
1,068	1	11.43
1,069	1	11.48
1,070	1	11.36
1,071	1	11.31
1,072	1	11.30
1,073	2	11.31
1,074	2	11.18
1,075	2	11.18
1,076	2	11.18
1,077	2	11.19
1,078	2	11.10
1,079	2	11.24

Science Grade 11

Scale Score	Achievement Level	CSEM
1,080	2	11.14
1,081	2	11.05
1,082	2	11.06
1,083	2	11.10
1,084	2	11.09
1,085	2	11.01
1,086	2	11.04
1,087	2	11.08
1,088	2	11.01
1,089	2	10.98
1,090	2	10.92
1,091	2	11.02
1,092	2	10.92
1,093	2	11.01
1,094	2	10.97
1,095	2	10.94
1,096	2	10.98
1,097	2	10.97
1,098	2	10.97
1,099	2	10.98
1,100	2	10.97
1,101	2	10.95
1,102	3	11.03
1,103	3	10.95
1,104	3	11.01
1,105	3	10.96
1,106	3	11.06
1,107	3	10.95
1,108	3	10.97
1,109	3	10.94
1,110	3	11.00
1,111	3	10.96
1,112	3	11.02
1,113	3	11.00
1,114	3	11.05

Science Grade 11

Scale Score	Achievement Level	CSEM
1,115	3	11.03
1,116	3	11.02
1,117	3	10.97
1,118	3	11.01
1,119	3	11.06
1,120	3	11.06
1,121	3	10.97
1,122	3	11.02
1,123	3	10.96
1,124	3	11.10
1,125	3	11.00
1,126	3	11.09
1,127	3	11.07
1,128	3	11.04
1,129	3	11.04
1,130	3	11.02
1,131	3	11.07
1,132	3	10.99
1,133	3	11.10
1,134	4	11.01
1,135	4	11.08
1,136	4	11.05
1,137	4	10.88
1,138	4	11.13
1,139	4	11.09
1,140	4	10.93
1,141	4	10.90
1,142	4	10.94
1,143	4	10.95
1,144	4	10.96
1,145	4	10.90
1,146	4	11.12
1,147	4	10.93
1,148	4	11.08
1,149	4	11.02

Science Grade 11

Scale Score	Achievement Level	CSEM
1,150	4	11.13
1,151	4	10.99
1,152	4	11.03
1,153	4	10.92
1,154	4	11.05
1,155	4	11.10
1,156	4	10.93
1,157	4	10.97
1,158	4	11.25
1,159	4	11.13
1,160	4	11.06
1,161	4	11.00
1,162	4	11.15
1,163	4	11.16
1,164	4	11.15
1,165	4	11.44
1,166	4	11.12
1,167	4	11.07
1,168	4	11.12
1,169	4	11.57
1,170	4	11.29
1,171	4	11.61
1,172	4	11.50
1,173	4	11.76
1,174	4	11.40
1,175	4	11.31
1,177	4	11.65
1,178	4	11.98
1,179	4	11.79
1,180	4	12.10
1,181	4	11.66
1,182	4	12.49
1,183	4	11.69
1,184	4	11.74
1,185	4	11.87

Science Grade 11

Scale Score	Achievement Level	CSEM
1,186	4	12.44
1,187	4	12.20
1,188	4	12.08
1,189	4	12.38
1,190	4	12.12
1,191	4	12.20
1,192	4	11.86
1,193	4	12.10
1,194	4	12.67
1,195	4	12.80
1,196	4	13.81
1,197	4	12.48
1,198	4	12.80
1,199	4	13.01
1,200	4	13.55

Appendix 4-C

Classification Accuracy and Consistency Indices by Subgroups

Classification Accuracy and Consistency Indices by Subgroups

Table 1-C-1. Classification Accuracy by Demographic Subgroup

Group	N	Overall (%)	By Cut (%)			By Level (%)			
			Level 2 Cut	*Level 3 Cut*	*Level 4 Cut*	*Level 1*	*Level 2*	*Level 3*	*Level 4*
Grade 5									
All Students	10,598	75.57	92.44	90.22	92.69	85.91	75.61	59.82	81.54
Female	5,195	75.17	91.97	89.87	93.09	84.86	75.63	59.87	80.93
Male	5,403	75.96	92.89	90.55	92.30	86.89	75.59	59.78	82.03
African American	337	76.75	89.06	91.65	95.88	87.08	75.44	59.93	69.39
American Indian/Alaskan Native	1,228	80.01	88.25	94.12	97.53	88.16	74.30	59.89	75.50
Asian	200	73.35	91.89	89.83	91.40	79.38	76.80	60.04	79.25
Hispanic	855	77.53	90.54	91.73	95.09	87.08	76.71	59.82	78.06
Multi-Racial	641	75.32	91.47	89.62	93.98	86.57	73.46	59.42	80.30
Pacific Islander	15	76.33	91.81	86.76	97.51	75.50	77.27	54.20	85.27
White	7,322	74.63	93.62	89.39	91.36	84.37	75.85	59.83	82.11
Limited English Proficiency (LEP)	501	80.94	86.12	95.99	98.78	86.81	76.42	58.82	64.96
Non-LEP	10,097	75.31	92.75	89.93	92.38	85.79	75.57	59.83	81.56
Special Education (SPED)	1,746	81.30	89.20	94.91	97.07	88.68	75.02	59.39	82.20
Non-SPED	8,852	74.44	93.08	89.29	91.82	84.18	75.71	59.85	81.49
Grade 8									
All Students	10,694	75.25	92.50	88.46	94.18	85.48	74.24	64.48	81.68
Female	5,255	74.46	92.60	87.41	94.33	84.57	74.44	64.15	80.06
Male	5,439	76.01	92.40	89.47	94.05	86.23	74.02	64.82	82.91
African American	328	79.84	89.76	92.43	97.58	87.84	74.62	65.80	85.36
American Indian/Alaskan Native	1,108	80.74	88.71	93.63	98.35	87.65	75.46	64.22	81.58

Group	N	Overall (%)	By Cut (%)			By Level (%)			
			Level 2 Cut	*Level 3 Cut*	*Level 4 Cut*	*Level 1*	*Level 2*	*Level 3*	*Level 4*
Asian	142	75.97	92.30	90.16	93.41	84.19	76.81	65.74	75.86
Hispanic	815	77.80	90.33	90.43	96.96	87.94	74.42	62.96	82.12
Multi-Racial	559	75.71	90.84	89.67	95.10	84.66	74.36	64.66	81.36
Pacific Islander	16	73.76	87.87	87.58	98.20	83.55	68.77	60.90	-
White	7,726	73.95	93.52	87.22	93.09	83.73	74.00	64.54	81.74
Limited English Proficiency (LEP)	399	83.51	87.75	96.06	99.66	89.26	75.67	58.92	68.32
Non-LEP	10,295	74.93	92.68	88.16	93.97	85.05	74.19	64.50	81.71
Special Education (SPED)	1,253	81.95	87.95	95.26	98.70	88.64	73.41	62.63	81.59
Non-SPED	9,441	74.36	93.10	87.56	93.59	84.08	74.32	64.53	81.68
Grade 11									
All Students	9,468	74.78	92.82	87.71	94.20	83.09	69.76	73.14	82.54
Female	4,695	73.99	92.58	86.76	94.60	81.92	70.02	73.14	80.71
Male	4,773	75.56	93.05	88.65	93.81	84.00	69.45	73.14	83.66
African American	304	74.92	89.11	88.33	97.42	85.02	68.08	70.99	80.30
American Indian/Alaskan Native	742	75.87	86.46	90.75	98.61	84.56	69.35	70.62	79.46
Asian	163	75.42	92.75	87.62	95.00	85.05	69.40	72.48	86.90
Hispanic	627	76.55	90.88	89.66	95.96	85.32	70.96	73.28	78.52
Multi-Racial	437	74.63	91.53	87.17	95.89	79.52	70.04	74.34	86.05
Pacific Islander	15	72.43	85.08	90.53	96.79	84.37	65.21	63.02	-
White	7,180	74.50	93.90	87.23	93.33	82.15	69.79	73.25	82.55
Limited English Proficiency (LEP)	247	80.56	85.19	95.39	99.95	87.28	68.84	64.34	99.13
Non-LEP	9,221	74.62	93.02	87.51	94.05	82.62	69.78	73.17	82.53
Special Education (SPED)	868	78.76	85.87	93.95	98.89	86.26	68.98	72.61	85.43
Non-SPED	8,600	74.38	93.52	87.08	93.73	81.79	69.84	73.16	82.51

Table 1-C-2. Classification Consistency by Demographic Subgroup

Group	N	Overall (%)	By Cut (%)			By Level (%)			
			Level 2 Cut	*Level 3 Cut*	*Level 4 Cut*	*Level 1*	*Level 2*	*Level 3*	*Level 4*
Grade 5									
All Students	10,598	66.70	89.33	86.27	89.73	77.18	67.16	48.44	72.99
Female	5,195	66.16	88.69	85.87	90.25	75.55	67.57	48.36	71.83
Male	5,403	67.22	89.94	86.65	89.24	78.71	66.71	48.51	73.93
African American	337	68.20	84.86	88.08	94.19	79.85	67.14	48.45	53.96
American Indian/Alaskan Native	1,228	72.47	83.57	91.75	96.45	83.82	64.79	47.08	60.22
Asian	200	63.79	88.72	85.47	88.02	72.42	64.96	50.10	69.15
Hispanic	855	68.82	86.51	88.10	93.13	80.16	68.87	47.20	64.59
Multi-Racial	641	66.72	88.11	85.79	91.57	78.54	65.75	47.41	72.24
Pacific Islander	15	66.48	88.33	81.63	95.71	73.21	70.70	32.03	81.98
White	7,322	65.49	90.95	85.12	87.88	72.75	67.49	48.68	74.16
Limited English Proficiency (LEP)	501	73.43	80.65	94.06	98.23	82.28	66.84	44.37	36.90
Non-LEP	10,097	66.36	89.76	85.88	89.31	76.54	67.17	48.50	73.12
Special Education (SPED)	1,746	74.09	84.76	92.81	95.88	84.95	65.06	45.41	74.90
Non-SPED	8,852	65.24	90.23	84.98	88.52	72.86	67.50	48.68	72.84
Grade 8									
All Students	10,694	65.99	89.33	84.05	91.64	77.19	65.53	53.58	70.23
Female	5,255	64.98	89.43	82.70	91.84	75.52	66.01	53.21	67.44
Male	5,439	66.97	89.23	85.36	91.45	78.58	65.00	53.96	72.41
African American	328	71.39	85.22	89.28	96.27	82.91	65.65	51.15	70.36
American Indian/Alaskan Native	1,108	72.71	83.78	90.91	97.51	83.70	65.73	48.81	61.01
Asian	142	66.34	89.12	85.72	90.57	77.86	66.36	52.27	69.24
Hispanic	815	69.24	86.29	86.75	95.49	81.58	65.94	50.66	66.97

Group	N	Overall (%)	By Cut (%)			By Level (%)			
			Level 2 Cut	*Level 3 Cut*	*Level 4 Cut*	*Level 1*	*Level 2*	*Level 3*	*Level 4*
Multi-Racial	559	66.64	87.17	85.63	92.94	76.88	65.23	53.94	70.49
Pacific Islander	16	64.71	82.32	84.66	97.07	78.69	57.59	52.45	18.39
White	7,726	64.41	90.79	82.42	90.11	72.83	65.48	54.08	70.63
Limited English Proficiency (LEP)	399	76.42	82.48	94.19	99.48	85.46	66.67	34.14	57.86
Non-LEP	10,295	65.59	89.59	83.66	91.34	76.31	65.49	53.72	70.25
Special Education (SPED)	1,253	74.83	82.95	93.40	98.12	85.14	62.91	46.72	73.69
Non-SPED	9,441	64.82	90.17	82.81	90.78	73.99	65.81	53.84	70.13
Grade 11									
All Students	9,468	65.04	89.83	82.85	91.63	72.09	58.85	65.79	68.29
Female	4,695	63.91	89.49	81.52	92.13	68.76	59.68	65.80	63.40
Male	4,773	66.15	90.17	84.17	91.13	74.90	57.89	65.79	71.61
African American	304	65.87	84.67	84.43	96.12	77.09	58.88	60.46	67.47
American Indian/Alaskan Native	742	66.89	81.20	87.12	97.91	77.25	59.89	59.39	56.84
Asian	163	66.48	89.76	83.09	92.89	71.55	61.08	63.85	77.58
Hispanic	627	67.24	86.87	85.45	94.21	78.88	60.03	64.70	60.44
Multi-Racial	437	64.50	88.08	82.04	93.61	69.52	59.32	66.53	65.97
Pacific Islander	15	64.60	80.48	87.11	96.49	76.99	56.82	54.54	45.33
White	7,180	64.62	91.33	82.16	90.40	68.88	58.52	66.31	68.70
Limited English Proficiency (LEP)	247	72.96	79.13	93.48	99.89	83.31	57.35	44.63	88.15
Non-LEP	9,221	64.83	90.12	82.57	91.41	70.99	58.89	65.89	68.27
Special Education (SPED)	868	70.39	80.24	91.32	98.29	81.17	58.34	59.02	57.45
Non-SPED	8,600	64.50	90.80	82.00	90.95	68.83	58.90	66.01	68.44

Appendix 4-D

Independent Alignment Study Report

WebbAlign™

WCEPS
Wisconsin Center for
Education Products & Services

Alignment Analysis of South Dakota Science Assessment (SDSA) for Grades 5, 8, and 11 with Corresponding Grade Band South Dakota Science Standards

Sara C. Christopherson
October, 2022

Wisconsin Center for Educational Products and Services
Matt Messinger, Executive Director

Acknowledgements

Panelists:

Tim Cox	Panel Facilitator	Wisconsin
Melyssa Ferro	Panel Facilitator	Idaho
Ashley Bates		South Dakota
Carolyn Burns		South Dakota
Sharla Dowding		South Dakota
Andrew Koch		South Dakota
Austin Lopour		South Dakota
Susan Mercer		South Dakota
Amy Miller		South Dakota
Whitney Muller		South Dakota
Linda Pinz-Valdez		South Dakota
Patrick Skroch		South Dakota

The South Dakota Department of Education funded this independent, third-party content alignment analysis. Matt Gill, Chris Booth, and Beth Schiltz were the main contacts for communication.

Table of Contents

Executive Summary

A content alignment analysis was conducted in June 2022 to provide information about the degree of alignment of the South Dakota Science Assessment (SDSA) for Grades 5, 8, and 11 with the corresponding South Dakota Science Standards as pertains to fulfilling requirements as stated in Federal statute. The SDSA used a particular state-vetted subset of items that were part of a Shared Science Assessment Item Bank. The item bank is managed by Cambium Assessment (CAI) and is shared by multiple states. The initial group of participating states agreed to share items through a Memorandum of Understanding (MOU) that detailed a commitment to shared content, leadership, ideas, and methods.

The South Dakota Department of Education requested an item-level content analysis of the entire operational SDSA item bank as of the Spring, 2022 administration. A total of 321 items and item clusters were included in the analysis. Item-level data were already available for 150 of those items, which were reviewed in a 2019 content alignment analysis that included panelists from 10 of the MOU states. The remaining 171 items from the SDSA item bank were included in a content analysis conducted in June 2022 by panels of expert educators in Pierre, South Dakota. The results described in this report include alignment-related characteristics of the overall SDSA item bank followed by test-event-level findings. Alignment is reported according to nine criteria agreed upon by participating states, including South Dakota, to be used to evaluate alignment of the assessments with corresponding standards.

Study results suggest that the overall SDSA item bank for grade 5 had the capacity to fully meet all alignment criteria agreed upon and used in this study. The SDSA item bank for grades 8 and 11 weakly met South Dakota's expectation for inclusion of items that addressed at least 90% of standards (Range of Knowledge Correspondence - Population) but study results suggest these item banks have the capacity to fully meet all other alignment criteria. The weak Range of Knowledge (Population) for grades 8 and 11 item banks could be fully resolved with the addition of at least four items to the middle school item bank and six items to the high school item bank.

Item-level results were mapped onto actual SDSA test events sampled from each of grade 5, 8, and 11 to yield test-event-level content alignment results. The SDSA was adaptively administered in SY2021-2022. For each grade, test events were randomly sampled from at or near cut scores for Levels 2 (below proficiency), 3 (at proficiency), and 4 (above proficiency). This sampling allowed for information about alignment of test events generated across proficiency levels. All test events analyzed were found to be fully or acceptably aligned with corresponding standards.

The SDSA was found to have had the overall capacity to generate test forms that were fully or acceptably aligned with the corresponding grade band standards. This finding included consideration of the results of the item-level analyses of the overall item bank, sample SDSA test events, and SDSA blueprints. The evidence to support these findings includes:

- The SDSA blueprints identified South Dakota's intended sampling across reporting categories (as relates to Categorical Concurrence, Range of Knowledge (breadth) for individual test events, and Balance of Representation (emphasis)).
- Overall, the items within the SDSA item bank met South Dakota's expectations (as relates to Use of Phenomena, Dimensionality, Consistency of Cognitive Engagement, and relationships with scoring assertions).
- The SDSA item bank fully or weakly met the state's expectations for Range of Knowledge across the tested student population.
- An analysis of three sample test events from each of grades 5, 8, and 11 found that all test events were fully or acceptably aligned with corresponding standards, based on the criteria agreed upon by South Dakota and used in this analysis.

The SY2021-2022 Assessment Technical Report was not available at the time of this writing. If aggregate data from all administered test events within South Dakota show that the blueprints and item selection algorithm yielded test forms as expected, it would further strengthen the argument for the capacity of the item bank to generate fully or acceptably aligned test forms.

Panelists identified specific items that did not meet one or more alignment-related expectations, and warrant revisions or removal. Even for items that panelists agreed met alignment-related expectations, many editorial suggestions were made to correct errors found in text and graphics, improve clarity, and/or address scientific inaccuracy. This extent of editorial issues is typically not observed in a high-stakes operational assessment and included issues that could potentially affect student scores. Of the items flagged, most are stand-alone items associated with just one or two scoring assertions. Because of the relatively limited interactions, a single stand-alone item contributes proportionately minimally to a student's score. While these items are recommended for revision in ongoing item bank maintenance, they were generally not considered a significant threat to the alignment of test events.

Overall, however, panelists found that items and item clusters were meeting state expectations for assessment tasks to require integrated engagement with at least two (stand-alone items) or three (item clusters) dimensions specified in the targeted standard in order to make sense of a phenomenon. With just a very few exceptions, items required student cognitive engagement consistent with the expectations of the standards. Items were spread across the domains of Physical, Life, and Earth and Space Science, with no standard(s) overemphasized in the item bank. Overall, panelists found that the large majority of scoring assertions reasonably reflected inferences that could be made based on student interactions and corresponded to the expectations within the targeted standard.

Introduction and Methodology

The alignment of expectations for student learning with assessments for measuring students' attainment of these expectations is an essential attribute for an effective standards-based education system. The critical role of alignment in the success of Framework-influenced science standards was called out in the very first chapter of A Framework for K-12 Science Education:

> "The committee recognizes that the framework and subsequent standards will not lead to improvements in K-12 science education unless the other components of the system – curriculum, instruction, PD, and assessment – also change so they are aligned with the framework's vision." (NRC, 2012)

In the context of statewide summative assessments, content alignment is defined as the degree to which expectations (standards) and assessments are in agreement and serve in conjunction with one another to guide an education system toward students learning what they are expected to know and do (Webb, 1997). As such, content alignment is a quality of the relationship between expectations and assessments and not an attribute solely of either of these two system components. Content alignment describes the match between expectations and an assessment that can be legitimately improved by changing either student expectations or the assessments. As a relationship between two or more system components, content alignment is typically determined by using, at minimum, multiple criteria described in detail in a National Institute for Science Education (NISE) research monograph, *Criteria for Alignment of Expectations and Assessments in Mathematics and Science Education* (Webb, 1997). The corresponding Webb methodology used to evaluate content alignment has been refined and improved over the years, yielding a flexible, adaptable, effective, and efficient analytical approach. Some version of this alignment methodology has been used to analyze curriculum standards and assessments in nearly all states to satisfy or to prepare to satisfy the Title I compliance as required by the United States Department of Education (USDE). Modified and/or expanded versions of this alignment methodology have been used for studies of multi-dimensional assessments, computer adaptive tests (CATs), interim assessments, alternate assessments, for studies intended to inform vendor internal continuous improvement, and for other purposes. Evidence of content alignment is a critical component of a validity argument that student scores from an assessment can reasonably yield the intended inferences.

The study detailed in this report was conducted for the South Dakota Department of Education and was coordinated and facilitated by the WebbAlign program. WebbAlign operates out of the Wisconsin Center for Education Products and Services (WCEPS), a non-profit organization that strives to extend the reach of innovations developed at the University of Wisconsin Center for Education Research (WCER), including the Webb alignment methodology.

Overview of South Dakota Science Assessment

The South Dakota Science Assessment (SDSA) administered in SY2020-2021 was comprised of items drawn from a larger science item bank that was managed by Cambium Assessment (CAI). Some participating states agreed to share items through a Memorandum of Understanding (MOU) that detailed a commitment to share content, leadership, ideas, and methods. The overall shared science item bank includes items owned by CAI as well as items owned by particular states. Regardless of ownership, all items followed the same test

development and review processes. Each state uses its own assessment blueprint and a particular state-vetted subset of items from the shared item bank. All states' science standards are Framework-based but some include adjustments from the wording or scope of the NGSS PEs. The South Dakota Science Assessment used items from this same item bank but South Dakota participated independently and was not a part of the MOU as of spring 2022. The SDSA item bank included stand-alone items and item clusters for elementary, middle, and high school sciences that were grounded in NGSS performance expectations (PEs). Results as relates to South Dakota standards that differ from the NGSS are noted in Findings.

Item Structure Each stand-alone item and each item cluster within the item bank was designed to address a single standard (performance expectation). (Note that throughout this report, "item" may refer to both stand-alone items and to item clusters.) Both stand-alone items and item clusters were intended to be based on a specific real-world scenario and focused enough to require students' application of multiple dimensions of the standard in order to make sense of the phenomenon presented. Stand-alone items were intended to require application of two or three dimensions while item clusters were intended to require application of all three dimensions of a standard. Item clusters are multi-part items that include an extensive scenario, typically involving text, illustrations, data shown in a variety of formats, short animations and other features. Item clusters have between several and up to around 20 different student interactions. Item clusters are typically, but not always, presented on-screen via two panels. One panel provides the stimulus. The other panel contains the instructions, prompts, and answer spaces for the student interactions. Stand-alone items present a more concise scenario and include, at most, several student interactions. Items included many different types of interactions. All interactions were machine scorable.

Scoring Assertions Each stand-alone item and item cluster was associated with a set of binary (true/false) narrative scoring assertions, which constitute the scoring rationales for items. Each assertion is intended to describe a piece of content knowledge, skill, or ability (KSA) that is related to the targeted standard and that the student is expected to have demonstrated by successful interaction with the item. In general, an assertion states the student's action(s) within the item that provide(s) evidence for the corresponding inferences about student KSAs. Per assessment design, the number of scoring assertions for an item varies, depending on the evidence that an item can yield based on a student's response. The vast majority (89%) of SDSA stand-alone items had one or two scoring assertions, with up to five scoring assertions, maximum. SDSA item clusters had an average of nine scoring assertions and up to 18 scoring assertions, maximum.

Blueprints and Delivery SDSA blueprints separated items by the domains of Life Sciences, Physical Sciences, and Earth and Space Sciences. Each domain was further divided into sub-domains according to the Disciplinary Core Idea (DCI) arrangement of the standards. Blueprints specified the length of the test and the minimum and maximum number of items that could be included on a test event by DCI organization of the standards per domain. SDSA blueprints specified that each standard was represented on a test event by no more than one item cluster or stand-alone item. In general, blueprints specified that a test event could include no more than one item cluster or two stand-alone items that targeted standards within the same DCI sub-domain. Test events were administered online and used an adaptive item delivery in which the

item selection algorithm chooses items based on content value toward blueprint fulfillment as well as a match-to-ability based on student responses to previous items. In other words, each item is selected based on its contribution to meeting the blueprint specifications as well as student ability, given the items that have already been administered.

Study Design

The South Dakota Department of Education requested an item-level content analysis of the entire operational SDSA item bank as of the Spring, 2022 administration. A total of 321 items were included in the analysis. Item-level data were already available for 150 of those items, which were reviewed in a 2019 content alignment analysis that included panelists from 10 of the MOU states. The remaining 171 items from the SDSA item bank were included in a content analysis conducted in June, 2022 by panels of expert educators in Pierre, South Dakota. The full set of items included in this report is shown in **Table 1**. While both stand-alone items and item clusters may require multiple student interactions, the unit of analysis was the overall item or item cluster.

Table 1. SDSA Item Bank Operational Items, Spring 2022

Grade	Item Type	Total Operational Items SDSA, Spring 2022
Grade 5	Cluster	39
	Stand-Alone	79*
Grade 8	Cluster	25
	Stand-Alone	74
Grade 11	Cluster	34
	Stand-Alone	70
Total		321

*one item not included for review

Item-level results were mapped onto actual SDSA test events sampled from each of grades 5, 8, and 11 to yield test-event-level content alignment results. The SDSA was adaptively administered in SY2021-2022. For each grade, test events were randomly sampled from at or near cut scores for Levels 2 (below proficiency), 3 (at proficiency), and 4 (above proficiency). This sampling allowed for information about alignment of test events generated across proficiency levels.

The overall study was structured to answer four key research questions:

1. To what extent do the stand-alone items and item clusters satisfy the measurement target claims (standard and scoring assertions) identified in the CAI metadata?
2. What DOK - Category of Engagement (cognitive complexity) is required for successful completion of each interaction within a stand-alone item or item cluster and how does the DOK distribution within the SDSA item bank compare with the DOK distribution within the South Dakota Science Standards?
3. To what extent do the stand-alone items and item clusters satisfy the claim that the assessment is phenomenon-based?
4. To what extent was the SDSA program likely to generate test events that were aligned with corresponding grade-level academic standards, considering depth and breadth (specified in ESSA) as well as other alignment criteria agreed upon and used in this analysis?

The results reported here pertain only to the issue of alignment between the South Dakota Science Standards and South Dakota Science Assessment. Note that an alignment analysis of this nature does not serve as external verification of the general quality of the standards or assessments, but rather, the focus is on the degree of alignment.

Panelists

Twelve educators completed the content analysis of the grade 8 item bank and then split into two panels to complete the analysis of the grade 5 and grade 11 item banks. Information about participating panelists is provided in **Appendix F**. Along with the Study Director, WebbAlign brought two experienced group leaders to facilitate the SDSA panels. Per South Dakota specifications, state officials were responsible for recruiting qualified expert educators with content expertise as panelists for the in-person alignment institute. State officials reached out directly to a wide range of districts and individuals in efforts to recruit from diverse populations across the state, with consideration for race/ethnicity, socioeconomic, and regional factors (urban/suburban/rural). State officials also recruited carefully to ensure adequate content expertise across science disciplines and grades.

All panelists were expected to have the following qualifications:

- thorough knowledge of their discipline (be subject matter experts)
- thorough knowledge of the South Dakota Science Standards
- experience with Framework-based science assessment considerations
- experience in the appropriate grade band(s) science education based on South Dakota Science Standards
- experience with the SDSA test design or willingness to review the test design and released item samples in advance of the in-person work
- willingness to express professional opinions, and listen to the professional opinions of others; willingness to agree, disagree, persuade, and be persuaded; maintain collegial, respectful, and positive professional environment

A sequential account of the alignment study procedures is provided in the sections that follow.

Training and Coding

Appropriate training of the panelists at a content alignment institute is critical to the success of the project. A necessary outcome of training is for panelists to have a common, calibrated understanding of the DOK - Categories of Engagement language system, a shared understanding of the structure, including dimensionality, of the standards, and a shared understanding of the coding processes and associated evaluative steps.

During the morning of the first day of the content alignment institute, panelists received an overview of the assessment, the purpose of their work, the coding processes, the use of online interfaces to view items and record data, and general training on different evaluative steps, including calibration on the DOK - Categories of Engagement definitions for science. The general training at the alignment institute was crafted to contextualize the origins of DOK - Categories of Engagement (as an evaluative tool used to inform alignment studies of standards and assessments) and purpose (to support consistency in differentiating between and among categories of cognitive complexity through content analyses), and to highlight common misinterpretations and misconceptions to help panelists better understand and, therefore, consistently apply the language system.

In advance of the study, panelists were provided with pre-reading materials including descriptions of the evaluative steps in the coding processes and definitions of the four DOK - Categories of Engagement for science. Through interactive and participatory training on-site, panelists reviewed the definitions and worked toward a common understanding of the difference between and among each of the categories of complexity. Training was designed with consideration of core tenets of contemporary learning theory, including recognition of the critical importance of engaging prior understandings, as people's existing ideas greatly influence how they make sense of new ideas and construct knowledge (Posner, et al. 1982; NRC, 2000; NRC, 2005). As such, activities elicited panelists' ideas and presented opportunities for panelists to grapple with these existing ideas as well as consider if and how their existing ideas fit with (possibly new) ideas presented. Background information was shared about the overall conceptual model of complexity and the epistemology in which the model is grounded. Through facilitated activities, panelists also worked to differentiate concepts such as cognitive complexity, difficulty, and multidimensionality as well as the idea of sophistication of performance across the K-12 learning progression as described in the standards. As part of the training, panelists practiced assigning DOK - Categories of Engagement to sample tasks that were selected to foster important discussions to promote improved conceptual understanding of the tool. Explicit clarification was provided related to potential misinterpretations of the tool to evaluate complexity. Calibration on the concept of complexity is critical for alignment analysis work as evaluation of the extent to which an assessment addresses the "depth" of the standards (i.e. cognitive complexity) is a central expectation of the alignment evidence required per ESSA.

An alignment analysis of the South Dakota Science Assessment - Alternate (SDSAA) was conducted concurrently with the analysis of the SDSA, and much of the SDSA and SDSAA panelist orientation was conducted together. Panelists' responses from study evaluation forms suggest training was effective (**Appendix F)**. Panelists were asked to rate on a 1-10 scale the extent to which the in-person orientation along with pre-reading materials helped prepare them for the work. Fifteen out of the 16 panelists who completed the evaluation (94%) ranked their preparation between 7-10, with most ranking the preparation as 9-10 (63%).

After separating into SDSA and SDSAA panels, group leaders facilitated more extensive introductions and set the tone for a collaborative, respectful, professional work environment in which panelists were expected to share and adjudicate dissenting professional judgments. For panelists to make reliable judgements on the degree to which an assessment task measures student performance as relates to a particular assessment target, they must have a shared and thorough understanding of the assessment targets. Therefore, an analysis of assessment targets is a necessary component of any study that examines the degree of content alignment of assessments and expectations. This need is augmented, however, in the context of Framework-based and multidimensional standards, for which there is recognition of a lack of consensus for referents as pertains to alignment analyses (e.g. Fulmer, et al, 2018). In this case, individual standards were defined as the assessment targets. Panelists calibrated their interpretation of the standards as pertains to the complexity of each as well as what is (and is not) intended for a statewide summative assessment. The standards analysis by grade-band panels is a necessary component of a content alignment study but also, importantly, fosters thorough, nuanced, and calibrated understanding of the standards by panelists. Consensus DOK - Categories of Engagement were then entered into the online data collection system, the

WATv2. The consensus Category of Engagement values for all standards are summarized in the Findings section of this report and listed in **Appendix A**. Additional information about the tool itself is provided in **Appendix E.**

Panelists also calibrated their understanding of what would be considered an appropriate manifestation of the three-dimensional standards in the context of an on-demand summative assessment. Similarly, panelists worked to build a common understanding of other evaluative considerations, such as the expectation for students to *make sense of* a phenomenon in their work, i.e. that students were expected to figure something out rather than answer a question that simply uses a phenomenon as a context.

Next, panelists started into the analysis of the SDSA items. All SDSA panelists worked through the Grade 8 items and item clusters to calibrate their coding. Panelists first coded these items/clusters independently and then adjudicated as a large group, discussing any differences in interpretations. Group leaders facilitated discussions and adjudication if needed, and communicated any specific decision rules that arose. Group leaders also provided instruction and clarification on appropriate coding procedures and best practices for effective recording of comments in the WATv2. This initial calibration work was conducted to promote consistency in coding both within and between the two panels for each grade band.

Panelists were instructed to work through each stand-alone item and item cluster as if they were the student. Then, they were to determine what the item measured, i.e. what students needed to know or be able to do in order to successfully respond to the question. Panelists considered whether a student's correct response to the stand-alone item or item cluster would allow for a reasonable inference about the student's proficiency as related to one of the standards for the grade band. As panelists worked, no internal metadata were visible. After independently identifying a standard that they thought the item addressed, panelists then were instructed to compare their independent assignment with the standard as given in the internal (CAI) item metadata. If the internally coded standard was appropriate, they recorded it in the online data entry system. If panelists did not think that the internally coded standard was appropriate, they were to discontinue coding for that item or item cluster. It was considered a necessary condition that an item or item cluster reasonably address the internally coded standard. Only in that context could panelists complete coding, i.e., consider if the phenomenon was appropriate, decide if any parts of DCI element(s) were missing, evaluate the scoring assertions, etc.

Panelists also worked individually to assign a DOK - Category of Engagement to the stand-alone item or to an item cluster. Panelists were instructed to consider the Complexity of Engagement required by each student interaction and to record the highest Category of Engagement that was included to ensure that coding captured the full scope of the complexity of the interactions within a stand-alone item or item cluster (defined as the unit of analysis). Panelists responded to two additional questions about each item or item cluster: one about dimensionality and another about use of a phenomenon. The evaluative questions, details about each criterion, and notes on the recording of responses for these two questions were included within the coding instructions (see **Appendix F**).

Panelists were instructed to focus primarily on the content alignment between the standards and the assessment items and item clusters provided. However, panelists were able to provide qualitative input or feedback on the standards and on the assessment items and clusters by writing a note in the appropriate text box in the WATv2 data collection tool. Panelists could indicate whether there was a Source of Challenge issue with an item—i.e. a technical or content problem with the item that might cause the student who knows the material to give a wrong answer or enable someone who does not have the knowledge being tested to answer the item correctly. After a panelist completed coding all of the assessment items and item clusters within a batch, the WATv2 offered a set of debriefing questions to answer for each study. These questions solicited feedback from the panelists about assessment items as a whole and provided a space to record any topics that were not captured in the item-level coding data.

If needed and as time allowed, the results for each study were adjudicated after all of the panelists completed coding a batch of items. The adjudication process helped to ensure that the coding by panelists did not include spurious data and that the codes entered were those as intended. For example, adjudication can correct errors, such as if a panelist accidentally entered one standard but meant to enter another standard. Group leaders facilitated conversations about items or item clusters for which panelists differed significantly on data entry for one or more evaluative step(s). When these substantial differences in coding occur, it sometimes indicates a data entry error. If data are entered as intended, then it suggests that panelists are either interpreting some aspect of the evaluation process in very different ways or are interpreting the particular assessment task in very different ways. For standard assignment, only data entry issues were addressed in final adjudication as any clarifications or discussion of differences in perspective on standard selection were addressed as needed as panelists moved through an item batch. Panelists did not conduct adjudication specific to the evaluative prompts for Use of Phenomenon or for the evaluations of the relationships with Scoring Assertions, but sometimes discussed these codings in the context of overall discussion of an item. Overall, adjudication was conducted to foster full and appropriate interpretation of the assessment items/clusters and to ensure that panelists had coded the items/clusters as they intended. Panelists were not required to change their coding after the discussions. Panelist agreement statistics were computed after adjudication and are included in the Findings section of this report.

Data Analysis

Results of the item-level analysis are reported for the overall item bank as well as for sample test events and include suggestions for areas in which improvements are needed. For the analysis of item complexity, the final reported value was found by averaging the DOK - Categories of Engagement values across all panelists. Any variance among panelists was considered legitimate, for example, with the reported DOK - Category of Engagement for an item falling somewhere between the two or more assigned values. Such variation could signify differences in interpretation of an item or of the assessed content and/or a DOK - Category of Engagement that falls in between two of the four defined levels. Standard deviations are reported in the tables provided in **Appendix B**, which give one indication of the variance among panelists. Majority coding (at least 7 of 12 for Grade 8 and at least 4 out of 6 for Grades 5 and

11) was used to determine whether the criteria of Dimensionality and Use of Phenomena were met as well as for the two evaluative questions related to the Scoring Assertions.

The results from this study pertain specifically to the issue of alignment between the South Dakota Science Standards and SDSA program and sample test forms that were analyzed. While some feedback is provided on aspects of quality, the degree of alignment is the focus of the discussion in the results. The 24 items (out of 321) that were flagged for review, revision, or removal are tabulated in the Findings section of this report and identified individually in **Appendix C**.

Alignment Criteria Used for This Analysis

After input from and discussion with representatives from all MOU states as of 2018, the nine alignment criteria detailed in **Table 2** were agreed upon to be used to evaluate and report on the degree of alignment of standards with state assessments drawn from the overall shared science item bank. Anchored in the states' intended claims, specific cutoffs were assigned to each criterion using defined decision rules about what was considered acceptable. The rationales for all decision rules were provided to allow for a process that was as transparent as possible. Individual states could modify these levels of acceptable alignment as warranted for particular state circumstances. For example, adjustments were made for a state that used a high school biology assessment instead of a high school science assessment that sampled from across all domains. Similarly, decision rules and cutoffs for acceptable alignment must be appropriate to the context of the SDSA, to confirm that the assessment in some way measures student performance as intended by the South Dakota Science Standards and within the intentions of the state's assessment design. For example, breadth needs to be considered from the perspective of an individual student (by test form) as well as for an overall student population (by item bank capacity along with aggregate data from all administered assessments). In addition to depth and breadth (specified in ESSA), criteria were included that corresponded to the specific intents and claims of a Framework-based and multidimensional statewide summative assessment. For example, in the context of the SDSA, the structure of how students are to know, engage, and think about science is very relevant to how the measurement of knowledge should be designed. The degree to which the three-dimensional engagement as expressed in the standards was captured on the assessment is reflected in the Dimensionality/Structure of Knowledge Comparability criterion. A content analysis was also required to provide evidence of the Use of Phenomena as well as the relationship of scoring assertions to the standards and to the student interactions, relevant to the specifics of the assessment design. South Dakota officials reviewed and discussed the proposed decision rules for determination of acceptable cutoffs for each alignment criterion and approved each; no modifications were proposed for use in the evaluation of the SDSA in 2022 (**Table 2**). In the case of criteria which South Dakota expected to be met for **all** items/clusters, a 90% cutoff was used to allow some leeway for human error and differences in professional opinion.

Table 2. Consensus Alignment Criteria for SDSA with South Dakota Science Standards, 2022

Criterion	Intended Claim/Inference	Acceptable Cutoff
1. Use of Phenomena	Items/clusters require students to engage multiple dimensions of the standards ("use science") to make sense of phenomena. Each item/ cluster is grounded in a stimulus that meets the test development criteria for a phenomenon.	At least **90%** of items/clusters are considered phenomenon-based by a majority of panelists (e.g. at least 4 out of 6 panelists).
2. Dimensionality/ Structure of Knowledge Comparability	Item clusters require students to demonstrate integrated engagement with the three dimensions of SEPs, DCIs, and CCCs in the targeted standard. Stand-alone items require integrated engagement with two **or** three of the dimensions specified in the targeted standard.	At least **90%** of clusters are considered three-dimensional by a majority of panelists; at least **90%** of stand-alone items are considered multi-dimensional by a majority of panelists.
3. Categorical Concurrence*	Test events have the potential to yield sufficient evidence to make inferences about student knowledge, skills, and abilities (KSAs) as relates to each reporting category.	A test form will include at least **six (6)** opportunities to respond to items that target the standards within each reporting category, two (2) of which are item clusters (per SDSA blueprint.) For the item bank, content categories concur with the standards' categories.
4. Consistency of Cognitive Engagement	The assessment elicits work that is as cognitively demanding as the expectations in the standards.	While some interactions may be DOK Category 1, no items/clusters should include only DOK 1 interactions. Proportions of items/clusters with DOK 2 and 3 opportunities reflect grade band standards. Some aspects of DOK Category 4 Standards will be assessed but the full scope of DOK 4 standards is expected to be assessed in the classroom.
5. Range of Knowledge Correspondence (Population)*	State-specific claims will be considered against aggregate data from all administered test events in the state in conjunction with a comparison of independent assignment of standard with internal vendor metadata.	At least **90%** of Standards have the potential to be assessed across the student population.
6. Range of Knowledge Correspondence (Individual)*	Test events assess an appropriate breadth of the standards, as defined by the SDSA blueprint. Assessed standards are sampled across topics within each reporting category for individual students.	Test forms analyzed meet blueprint specifications for Range. Blueprints are expected to specify sampling across topics (or other sublevels for each reporting category).

Table 2 Cont'd. Consensus Alignment Criteria for SDSA with South Dakota Science Standards, 2022

Criterion	Intended Claim/Inference	Acceptable Cutoff
7. Balance of Representation*	No standard is targeted more than once on any test event.	A standard should not be targeted more than once on a test event; each stand-alone item and item cluster should target a different standard.
8. Relationship of Scoring Assertions with Student Interactions	In aggregate, the Scoring Assertions for an item/item cluster appropriately represent the inferences about student knowledge, skills, and abilities that can be made based on successful interactions with an item/cluster.	For at least **90%** of all items/clusters, a majority of panelists consider a large majority of the Scoring Assertions (at least ~75%) to appropriately represent the inferences about student KSAs that can be made based on successful interactions with an item/cluster.
9. Relationship of Scoring Assertions with Standards	In aggregate, the Scoring Assertions for an item/item cluster appropriately represent the three-dimensional expectations of the targeted standard. (At least two of the three dimensions for stand-alone items.)	For at least **90%** of all items/clusters, a majority of panelists consider a large majority of the Scoring Assertions (at least ~75%) to appropriately represent the expectations within the corresponding standard.

Student scores on the SDSA were not used to make specific claims about:

- Engineering Design (ETS) standards or engineering.
- Science, Technology, Society, and the Environment Connections
- Nature of Science

Details on the criteria used for determining the degree of content alignment between standards and assessments are provided on the following pages. For each criterion, the cutoffs for acceptability are defined. If overall results met these defined cutoffs, the criterion was considered to be met. For criteria related to Use of Phenomena, Dimensionality, Range, and Scoring Assertions, a criterion was considered to be "weakly met" if results fell within 10% of the expected cutoff.

Reporting Categories
Study results for each of Grades 5, 8, and 11 are reported by the domains of Physical Science, Life Science, and Earth and Space Science. Consensus DOK - Category of Engagement values for all standards are given in **Appendix A**. All standards were included in the analysis with the exception of South Dakota Science Standard HS-LS4-7. This standard was not included in SDSA blueprints for the SY2021-2022 administration.

Total number of standards by grade:

- Grade 5 science standards: 42
- Grade 8 science standards: 52
- Grade 11 science standards: 64

In the descriptions below, the term “standards” may be used as an umbrella term, to refer to expectations in general.

1. Use of Phenomena The SDSA was intended to be phenomenon-based, meaning that items/clusters required students to engage multiple dimensions of the standards (“use science”) to make sense of a phenomenon. Per test development criteria, a phenomenon was expected to be based on a specific real-world scenario, reflect grade-appropriate content and complexity, and be focused enough to require students' application of a SEP in the context of a DCI and CCC. While stand-alone items could be two *or* three dimensional, they were still expected to require students to use multiple dimensions of a standard to make sense of a phenomenon. To meet this expectation at the item level, a majority of the reviewers on a panel (at least 7 of 12 for Grade 8 and at least 4 out of 6 for Grades 5 and 11) must have considered the item or item cluster to have met the test development criteria for a phenomenon, as indicated in their independent coding. To meet this criterion at the item bank level and test event level, 90% of items must have been coded affirmatively by a majority of panelists. A 90% cutoff was used to allow some leeway for human error and differences in professional opinion.

2. Dimensionality / Structure of Knowledge Comparability All SDSA assessment items were intended to be multi-dimensional, meaning that items/clusters required a student to engage with and interweave two or three dimensions of the standards to make sense of phenomena. For an item cluster, successful completion of the task was expected to require students to engage with the specific three dimensions identified in the corresponding standard, at minimum. (NOTE that the particular DCI, SEP, and CCC of a standard are not expected to exist in isolation of other DCIs, SEPs, and/or CCCs in the context of a task and consequently, items and item clusters may have included multiple SEPs or CCCs, etc.) For a stand-alone item, successful completion of the task was expected to require students to engage with at least two of the specific three dimensions identified in the corresponding standard. To meet this expectation at an item level, a majority of the reviewers on a panel (at least 7 of 12 for Grade 8 and at least 4 out of 6 for Grades 5 and 11) must have indicated in their independent coding that an item cluster required student engagement with all three dimensions of the standard or that a stand-alone item required student engagement with two or three of the dimensions. To meet this criterion at the item bank level and test event level, 90% of items must have been coded affirmatively by a majority of panelists. A 90% cutoff was used to allow some leeway for human error and differences in professional opinion.

3. Categorical Concurrence The South Dakota Science Standards were organized by the content categories of Physical, Life, and Earth and Space Sciences. Each of these categories was further divided by content (DCI organization). An important aspect of alignment between standards and assessments is whether both address the same content categories. The Categorical Concurrence criterion provides a very general indication of alignment if both documents incorporate the same content. The criterion of Categorical Concurrence between standards and assessments is met if the same or consistent categories of content appear in both documents. Overall Categorical Concurrence at the item bank level is reported for informational purposes. For a particular test event, this criterion was judged by determining whether the assessment included items targeting standards from each reporting category. Grounded in calculations based on a procedure developed by Subkoviak (1988), it is typically assumed that an assessment would have to have at least six items for measuring content from a reporting category for a minimum acceptable level of Categorical Concurrence to exist between the domain and the assessment (Webb, 1999). The number of items (six) is based on estimating the number of items that could produce a reasonably reliable score for estimating students' mastery of content on that subscale. Of course, many factors must be considered in determining what a reasonable number is, including the reliability of the subscale, the mean score, and cutoff score for determining mastery. A cutoff of six items per reporting category was consistent with South Dakota expectations as reflected in the SDSA blueprints and was used in this analysis. Because both stand-alone items and item clusters included multiple student interactions, actual Categorical Concurrence within each domain is greater than the item count.

4. Consistency of Cognitive Engagement (DOK - Category of Engagement) A Framework for K-12 Science Education and the resulting NGSS both emphasize a conceptual shift in science standards, related to the complexity of student engagement with science concepts and scientific thinking (NGSS Appendix A, Conceptual Shift #4). As a central conceptual shift, attention must be given to determine if and in what ways different types of student cognitive engagement (i.e. cognitive complexity) are being interpreted both in the expectations and the assessment. Consistency of Cognitive Engagement between content standards and an assessment indicates alignment if what is elicited from students on the assessment is as demanding cognitively as what students are expected to know and do as stated in the corresponding standards. The DOK - Categories of Engagement tool is used to guide content analysis for the purpose of differentiating between and among the different types of complexity of cognitive engagement required by learning expectations and tasks. For consistency to exist between the assessment and the reporting categories, as judged in this analysis, two conditions applied. First, no items or item clusters should require only DOK - Category 1 Cognitive Engagement. While it was considered acceptable for some interactions within an item or item cluster to be Category 1, successful completion of an item or item cluster could not require only Category 1 work, per South Dakota expectations and consistent with the intent of Framework-based standards. Second, the proportion of items and item clusters with Category 2 and Category 3 opportunities should reflect the proportion of DOK - Category 2 and Category 3 expectations in the Performance Expectations. Category 4 expectations, which are complex tasks that require extended time (such as the "sustained investigations" expected by the Framework) are not expected to be appropriately or authentically assessed in an on-demand context. All of the items and item clusters in the SDSA Item Bank, therefore, were expected to

provide opportunities for cognitive engagement within DOK - Category 2 and Category 3. Although this was expected for **all** items/clusters, a 90% cutoff was used to allow some leeway for human error and differences in professional opinion. To meet this criterion at the test event level, no items were to be DOK 1, and the items within each domain should include at least one opportunity to engage at DOK 3.

DOK – Category of Engagement for Science The Category of Engagement descriptions help to clarify how different types of complexity are represented in the sciences and are summarized below. Full descriptions for science as well as an explanation for the relationship of each Category with the expectations of Framework-based standards are included in **Appendix E**.

Category 1 includes tasks such as recalling facts and terms, recognizing structures or properties, reproducing standard scientific representations, or performing routine procedures. The Framework and NGSS documentation specify that Category 1 type expectations are not intended as assessment targets. For example, NGSS Appendix C calls out "…a huge transition, from a focus on knowledge itself to a focus on putting…knowledge to use—a transition that in and of itself necessitates a corresponding leap in rigor" and notes that new standards "focus on understanding rather than memorization" (NGSS Appendix C, 2013). Because "[p]erformance expectations are the assessable statements of what students should know and be able to do" and are intended "to make clear the intent of the assessments" (NGSS, 2013) it can be inferred that no standard should be considered to expect only Category 1 type work. While an explicit goal of Framework-based standards is to promote a shift away from assessing students on Category 1 types of tasks, resulting standards recognize that students will indeed need to engage with Category 1 tasks in the context of broader work to make sense of a phenomenon. For example, although "[n]o part of the NGSS specifies the student outcome of defining a gene – it is…implicit that in order to demonstrate proficiency on MS-LS3-1, students will have to be introduced to the concept of a gene through curriculum and instruction" (NGSS Appendix B, 2013). Similarly, students will need to use particular tools and protocols, and learn new terms. Overall, students may need to develop fluency with Category 1 expectations but they are not appropriate as overall summative assessment targets, per Framework and NGSS. Because of this clear expectation within the standards, it is critical that educators and assessment developers can consistently differentiate between Category 1 and Category 2 tasks.

Category 2 tasks require students to connect ideas and make sense of relationships and interactions between and among concepts and ideas, anchored in evidence-based thinking. The conceptual understanding emphasized by Category 2 expectations are reflected in multiple places in Framework and NGSS documentation. For example, Appendix A conceptual shift number four states that "[t]he NGSS focus on deeper understanding of content as well as application of content" (NGSS Appendix A, 2013). Appendix C also underscores this key shift, noting that "the NGSS focus [is] on understanding rather than memorization" (NGSS Appendix C, 2013). This, in turn, reflects the Framework committee's intent to "give time for students to…achieve depth of understanding of the core ideas" (NRC, 2012). A core overall goal of Framework-influenced standards, including NGSS, is for students to demonstrate knowledge-in-use as they make sense of phenomena, consistent with many Category 2 types of expectations.

Category 3 tasks involve abstract, analytical, hypothetical, critical, evaluative, original (to the student), and innovative thinking, including crafting reasoned scientific arguments based on

evidence. Category 3 expectations are reflected in the Framework committee's intent to "give time for students to engage in scientific…argumentation" (NRC, 2012) and goal of supporting students as they "discove[r] new knowledge, solv[e] challenging problems, and generat[e] innovations" including addressing "problems not previously encountered" (NGSS Appendix C, 2013).

Category 4 tasks expect at least the complexity of Category 3 but require extended and iterative sensemaking, corresponding to the "expectation…that students generate and interpret evidence and develop explanations of the natural world through sustained investigations" or that students "carry out empirical investigations in order to develop or evaluate knowledge claims" (NRC, 2013). While subcomponents of Category 4 tasks may be represented in an on-demand assessment, they are more appropriately and authentically assessed in the classroom.

5. Range of Knowledge Correspondence (Population) In the context of the SDSA, the criterion of Range must be considered for the overall tested population as well as for the individual student. The item-level analysis was used to determine the degree to which the claims within the item metadata could be substantiated. These findings can be considered alongside aggregate data from all administered test events in the state (when available; at the time of this writing the SDSA Technical Report for SY2021-2022 was not available). South Dakota expected at least 90% of the standards within each domain to have the potential to be assessed across the student population.

6. Range of Knowledge Correspondence (Individual) Traditionally, Range of Knowledge at the test event level is calculated against the full scope of a set of assessed academic standards, and a test form is expected to sample the knowledge, skills, and abilities from at least half of the full set of standards. In the context of the South Dakota Science Standards, it is not reasonable to consider the full range of grade-band standards across multiple disciplines as the referent, both because of the vast scope of the standards as well as because the standards are intended to foster deep engagement with science versus broad coverage of topics. Instead, state stakeholders defined what was appropriate and reasonable for assessment on an individual test event. The intended range was then codified in the test blueprint, which served as the referent. For SDSA reporting categories and assessments to be aligned, the breadth of knowledge expected on the test blueprint should be comparable to the breadth of knowledge sampled on a test form. In other words, the span of knowledge expected of students by a reporting category (as defined by a test blueprint) should correspond to the span of knowledge that students need to correctly answer the assessment items. Because the test blueprint served as the referent, fidelity to blueprint specifications can serve as evidence for meeting this alignment criterion, interpreted in the context of the results of the item-level content analysis (if they provide independent verification of internal metadata for targeted standard). Test blueprints were organized by domain and by DCI, and specified the expected range of sampling within each DCI as well as across all DCIs within the domain.

7. Balance of Representation In addition to comparable depth and breadth of knowledge, aligned reporting categories and assessments require that knowledge be distributed in the intended proportions. The Balance of Representation criterion, as applies to the test-event level, specifies that no standard is targeted more than once on any single SDSA test event.

8. Relationship of Scoring Assertions with Student Interactions Each stand-alone item and item cluster was scored with a set of binary (true/false) narrative Scoring Assertions, which constituted the scoring rationales for items. Each Assertion was intended to describe a piece of content knowledge, skill, or ability (KSA) that was related to the targeted standard and that the student was inferred to have demonstrated by successful interaction with the item. In general, an Assertion stated the student's action(s) within the item that provided evidence for the corresponding inference about student KSAs. Panelists were instructed to carefully read through each individual Scoring Assertion and then consider whether or not the Assertions, in aggregate, adequately reflected reasonable inferences about student knowledge, skills, and abilities based on their work on the assessment item or item cluster. Panelists could find that one or more of the Assertions slightly misstated, overstated, or understated the inferences that could be made but to code this criterion affirmatively ("Yes"), panelists needed to agree that a large majority (~75%) of the Scoring Assertions described a direct inference that could be made from the student's correct responses. To meet this criterion at the item bank and test event level, at least 90% of all items/clusters within each domain must be coded affirmatively as represented in panelists' independent coding. Although South Dakota expected this criterion to be met for **all** items/clusters, a 90% cutoff was used to allow some leeway for human error and differences in professional opinion.

9. Relationship of Scoring Assertions with Standards As scoring rationales, the Assertions were expected to appropriately reflect the assessment targets (i.e., the standards). This evaluative point was intended as a cross-check to "close the loop" on the measurement chain of reasoning for each item or item cluster. If an item (or cluster) adequately targeted a particular standard, and the Scoring Assertions appropriately reflected inferences about a student's successful work on the item (or cluster), then it would be expected that the Scoring Assertions circled back to the standard, and adequately reflected at least two (for stand-alone items) or all three (for item clusters) of the three-dimensional expectations therein. To meet this expectation at an item level, a majority of panelists must have indicated in their independent coding that the Scoring Assertions, in aggregate, represented the expectations explicit within the corresponding standard. To meet this criterion at the item bank and test event level, at least 90% of all items/item clusters must be coded affirmatively. Although South Dakota expected this criterion to be met for **all** items/clusters, a 90% cutoff was used to allow some leeway for human error and differences in professional opinion.

Source of Challenge and Panelist Comments The Source of Challenge criterion is used to identify items with issues that can cause a student to answer the item correctly or incorrectly for the wrong reason. Bias and sensitivity issues, as well as technical issues and errors, could all be considered a Source of Challenge problem. These types of issues are uncommon on high-quality operational assessments as they are typically addressed during test development. Panelists were instructed to document any Source of Challenge issue. Panelists could also leave comments about each item. After coding each item batch, panelists were asked to respond to debriefing questions. Responses to these questions provide qualitative and holistic feedback about the item bank and the alignment relationships between the standards and the items.

Summary Findings: Standards and SDSA Item Bank Characteristics by Alignment Criterion

The results and a discussion of both the standards analysis and the item-level analysis of the overall item bank are presented in this section. In order for student scores on an assessment to support the intended inferences about student achievement (as represented within the Scoring Assertions) as relates to the South Dakota Science Standards, there must be a close underlying relationship between and among the standards, the assessment items, and the Scoring Assertions. The results of the standards analyses therefore provide context that can support interpretation of the results of the item-level analysis.

In general, the SDSA items for all grade bands met the state's expectations although specific items (four item clusters and 20 stand-alone items) were flagged for Source of Challenge, for not meeting one or more alignment expectations, and/or for needing editorial corrections (**Appendix C**). These items should be closely reviewed and some unquestionably warrant revisions or removal. Overall, however, panelists found that items met states' expectations (detailed in the previous section).

Overall, the item-level analysis found that the SDSA item bank for each grade band showed the capacity to generate aligned test events as summarized in **Table 3** below. In **Table 3**, a "YES" indicates that the cutoff for the criterion was met (as specified in **Table 2** and described in the previous section of this report). If results fell within 10% of the cutoff, the criterion was reported as "WEAKLY" met. For all grades, cutoffs for criteria were met or weakly met.

Table 3. Overall Results by Alignment Criterion and Grade, SDSA Item Bank as of Spring 2022

	Was the criterion met for each grade?		
Criterion	**Grade 5**	**Grade 8**	**Grade 11**
Use of Phenomenon	YES	YES	YES
Dimensionality/Structure of Knowledge	YES	YES	YES
Categorical Concurrence	YES	YES	YES
Consistency of Cognitive Engagement	YES	YES	YES
Range of Knowledge (Population)	YES	WEAKLY*	WEAKLY**
Balance of Representation	YES	YES	YES
Relationship of Scoring Assertions with Student Interactions	YES	YES	YES
Relationship of Scoring Assertions with Standards	YES	YES	YES

*For the middle school Life Science domain, this criterion was unmet. If considering the full set of middle school standards, this criterion was weakly met.
**For the high school Physical Science domain, this criterion was unmet. If considering the full set of high school standards, this criterion was weakly met.

The grades 8 and 11 item banks need to be supplemented if South Dakota expects them to have the capacity to assess at least 90% of the standards. This issue could be fully resolved with the addition of at least four items to the grade 8 item bank and six items to the grade 11 item bank. Item bank weaknesses for Range are not considered an alignment concern early in program development, but rather can be a focus for ongoing improvement.

Item Bank Characteristics by Alignment Criterion

1. Use of Phenomena Based on the item-level analysis of the overall SDSA item bank, and with the exception of several items flagged for revision or removal because all interactions were DOK - Category 1 (**Appendix C**), all items were considered by a majority of panelists to meet South Dakota's expectations for being phenomenon-based. (If an item is coded as DOK - Category 1 in its entirety, it indicates that the item did not require a student to *interact* with the phenomenon presented, and therefore means that even if a specific real-world scenario was presented, it did not meet the full set of expectations for Use of Phenomena.)

2. Dimensionality / Structure of Knowledge Comparability Panelists considered dimensionality from several perspectives, as dimensionality of items was related to their assignment of standard, their evaluation of items' Use of Phenomena, and their evaluation of relationships with Scoring Assertions. With the exception of items that did not meet one or more other criteria (e.g. were considered Category 1) all items were considered by a majority of panelists to meet South Dakota's expectations for dimensionality: item clusters were found to require students to demonstrate integrated engagement with the three dimensions specified in the targeted standards. Stand-alone items were found to require students to demonstrate integrated engagement with two or three of the dimensions specified in the targeted standards.

3. Categorical Concurrence Overall results as pertains to Categorical Concurrence of the item banks by grade and domain are summarized in **Table 4**. For each grade and domain, the total number of items is given.

Table 4. Number of Shared Science Assessment Items Included in this Report by Domain and Grade Band Based on Item-Level Analysis of Overall Operational Item Bank, Spring 2022

Number of SDSA Items by Grade and Reporting Category			
	Grade 5*	**Grade 8****	**Grade 11*****
PS	41	30	25
LS	36	40	59
ESS	40	29	20
TOTAL	**117**	**99**	**104**

*For grade 5, item 367 was listed but not found in item bank. One PS item and one ESS item were flagged for revision or removal.
**For grade 8, one PS item and one LS item were flagged for revision or removal.
***For grade 11, five LS items and one ESS item were flagged for revision or removal.

The vast majority of standards were represented by items within the overall item bank even when taking into account any items flagged by panelists for review. South Dakota Science Standard HS-LS4-7 was not included in SDSA blueprints for the SY2021-2022 administration. All standards with no corresponding items are listed in **Table 5**. If a standard had one corresponding item but that item was flagged by panelists because it did not meet one or more of the evaluative criteria, then the standard was listed as unrepresented.

Table 5. Unrepresented Standards in SDSA Operational Item Bank (as of Spring, 2022) by Grade and Domain Based on Item-Level Analysis of Item Bank (321 Items)

Standards Not Represented in the SDSA Item Bank by Grade and Domain			
Domain	**Grade 5**	**Grade 8**	**Grade 11**
PS	All standards represented	MS-PS1-1 MS-PS2-4 MS-PS2-5	HS-PS1-8 HS-PS2-3 HS-PS2-5 HS-PS3-2 HS-PS4-2 HS-PS4-3 HS-PS4-4
LS	All standards represented	MS-LS1-1 MS-LS1-5 MS-LS1-7 MS-LS4-2	HS-LS1-3 HS-LS4-6
ESS	5-ESS2-1	MS-ESS3-3	HS-ESS2-3 HS-ESS2-4 HS-ESS3-3

4. Consistency of Cognitive Engagement (DOK - Category of Engagement) The last two columns of **Table 6** show the distribution of standards at each DOK - Category of Engagement next to the distribution of assessment items (both stand-alone and item clusters) within the item bank at each DOK - Category of Engagement by grade band. This allows for a broad-stroke look at the overall complexity of the items within the item bank for each grade in relation to the overall complexity of the standards. All standards were judged to have a complexity Category of 2, 3, or 4. Across grade bands, the vast majority of standards were considered DOK - Category 2 (70% to 76%). Between 16% and 23% of the standards for each grade band were considered DOK - Category 3. The remaining standards were considered DOK - Category 4 (5% in elementary; 6% in middle school; 14% in high school).

Although no items or item clusters were expected to include only DOK - Category 1 interactions, a very small percentage of grades 5 and 8 items (4 items total; <2%) were identified as such. While this information can be used for ongoing improvements to the item bank, it is not considered a threat to the alignment of test events with standards. Items and item clusters were found to be DOK - Categories 2 and 3, corresponding proportionately, overall, to the complexity of the grade-band standards. The item cluster structure used this assessment, along with the multiple types of information provided within the stimulus and the multiple types of student interactions possible, provided opportunities for students to engage in a wide variety of complex tasks. The results of the item-level analysis suggest that the item bank has the capacity to generate test events that meet the criterion of Consistency of Cognitive Engagement, which expects the assessment to elicit work that is as cognitively demanding as the expectations in the standards.

Table 6. Standards by DOK - Category of Engagement Compared with DOK - Category of Engagement Distribution of Items within the Overall SDSA Item Bank, Spring 2022

SDSA Grade	Total Number of Standards	DOK - Category of Cognitive Engagement	Number of Standards by Level	Standards % Category of Engagement Distribution by Level	Items % Category of Engagement Distribution by level
Grade 5	42	1 2 3 4	0 32 8 2	0 76 19 5	2 90 8 0
Grade 8	52	1 2 3 4	0 37 12 3	0 71 23 6	2 72 26 0
Grade 11	64	1 2 3 4	0 45 10 9	0 70 16 14	0 84 16 0

5. Range of Knowledge Correspondence (Population) Range of Knowledge for the student population shows the breadth of standards represented within an item bank. The total set of standards was identified by the SDSA blueprints for each grade. **Table 7** shows the total number of standards by grade and domain next to the number of standards not represented. The rightmost column shows the percentage of standards with one or more corresponding items by grade and domain. South Dakota expected at least 90% of the standards to have the potential to be assessed across the student population.

Table 7. SDSA Item Bank Range of Knowledge Correspondence (Population) by Grade and Domain, Spring 2022

Domain by Grade	Total Number of Standards	Number of Standards Not Represented	Percentage of Standards Targeted by Items Within SDSA Item Bank
Grade 5			
PS	17	0	100%
LS	12	0	100%
ESS	13	1	92%
Grade 8			
PS	19	3	84%
LS	19	4	79%
ESS	14	1	93%
Grade 11			
PS	24	7	71%
LS	24*	2	96%
ESS	16	3	81%

*included in SY2022-2023 assessment; 25 LS standards total

For the Grade 5 item bank, all but one standard was represented within the item bank. For Grade 8, three Physical Science standards, four Life Science standards, and one Earth and Space Science standard were unrepresented in the item bank. For Grade 11, seven Physical Science standards were unrepresented in the item bank along with two Life Science and three Earth and Space Science standards.

Based on the results of the item-level analysis, the SDSA item bank for Grade 5 met state expectations to include items that address at least 90% of the corresponding standards. The SDSA item bank for grades 8 and 11 weakly met this expectation overall, but did not meet the expectation for Grade 8 Life Science domain nor for Grade 11 Physical Science domain. This issue could be fully resolved with the addition of at least four items to the middle school item bank and six items to the high school item bank. Two of the new or added middle school items would need to address unrepresented Physical Science standards, and the other two items would need to address unrepresented Life Science standards. For high school, five of the new or added items would need to address unrepresented standards within the Physical Science domain and one item would need to address an unrepresented standard within the Earth and Space Science domain. Item bank weaknesses for Range are not considered an alignment concern early in program development, but rather can be a focus for ongoing improvement.

6. Range of Knowledge (Individual) Range of Knowledge at the individual student level is addressed in the test-event level findings.

7. Balance of Representation In addition to comparable depth and breadth of knowledge, aligned reporting categories and assessments require that knowledge be distributed in the intended proportions. Results of the item bank analysis show that items were reasonably distributed among the targeted standards. Between one and nine items were found to correspond to each of the represented standards; no standard was overemphasized in the item bank. Overemphasis applies only in circumstances when one particular standard is represented in excess of all others.

8. Relationship of Scoring Assertions with Student Interactions Panelists' independent coding met expectations for this criterion: with a few exceptions, a majority of panelists agreed that the Scoring Assertions reasonably described the inferences that could be made based on successful student interactions with the assessment for over 90% of all items. For just one Grade 5 item, two Grade 8 items, and three Grade 11 items (<2% of items overall), independent coding did not yield a panel majority agreement with the Scoring Assertions as related to student interactions.

9. Relationship of Scoring Assertions with Standards Panelists' independent coding met expectations for this criterion: with some exceptions, a majority of panelists agreed that the Scoring Assertions reasonably reflected the expectations of the corresponding standard. For four Grade 5, five Grade 8 items, and 12 Grade 11 items, (<7% of items overall) independent coding did not yield a panel majority agreement with the Scoring Assertions as related to the targeted standard. However, some of the disagreement was due to panelists marking the Scoring Assertions as not reflecting the standard if not all three dimensions of the standard were addressed, instead of considering if the Scoring Assertions reflected the two dimensions that were addressed (for two-dimensional items) as per coding protocol.

Source of Challenge and Item-Level Comments The Source of Challenge criterion is used to identify items with issues that could cause a student to answer the item correctly or incorrectly for the wrong reason. Bias and sensitivity issues, as well as technical issues and errors, could all be considered a Source of Challenge problem. Across grades, 16 items were flagged with Source of Challenge issues (Grade 5 item 379, Grade 8 items 126, 173, 577, 678, 682, and Grade 11 items 199, 216, 359, 476, 492, 497, 559, 564, 565, 668). Issues identified included inclusion of content outside of a standard's assessment boundaries, unclear directions, graphics, or response modes, and errors in graphics or text that could affect at least some component of student responses. Panelists included some comments related to Source of Challenge in their notes as well. For some issues and topics, multiple panelists left similar comments. All comments should be reviewed and considered, including those made by an individual panelist, as one person may have noticed something that others did not. Many of the issues identified have straightforward resolutions, including slight adjustments and corrections to errors, after which the items would be expected to be appropriate and viable. Some of these issues require larger-scale reconsideration of the item or item cluster. Panelists also wrote notes about many items. Some notes included actionable suggestions for item improvements. Panelist notes also contain comments and feedback, including many commendations. These notes may be helpful to identify exemplar items that can be used as models for future item development.

Items Flagged for Review and Revision or Removal All items flagged for Source of Challenge, along with eight other items that did not meet one or more of the alignment expectations are itemized in **Appendix C**. Items that panelists identified as having weak connections between Scoring Assertions and the full scope of a standard were not included in this list, provided that no other issues were identified and that the item was considered to address a core component of the standard. This was consistent with state expectations as it was considered acceptable for an item to address only two of the three dimensions of a standard, for example. It is recommended that South Dakota and CAI review and revise or remove the items flagged and included in **Appendix C**. Inaccurate graphics and other science issues should be prioritized for resolution. However, the total flagged item count does not exceed the cutoffs established for this study. In general, states agreed that when an expectation was intended for all items, a 90% cutoff would be used to allow some leeway for human error and differences in professional opinion. When aggregating items that did not meet one or more expectation(s), the proportion of items still falls under this threshold. Panelist comments can be also be used to inform revisions that could help limit excess difficulty in items. The full text of panelist comments was provided to states and to CAI but redacted for public release.

Summary Findings: Alignment of SDSA Test Events with Corresponding South Dakota Science Standards

Test-event-level alignment findings are given on the pages that follow. Across forms, any alignment weaknesses affected only a small proportion of the scoring assertions (<10%), and the overall test forms were still considered acceptably aligned. The consistency in test-event-level findings suggests that the SDSA Item Bank was operating as intended.

Cutoffs for Each Alignment Criterion for Individual Test Events:
For individual test events, acceptable alignment for Categorical Concurrence, Range of Knowledge (Individual), and Balance of Representation was defined by the blueprint. Categorical Concurrence and Range of Knowledge (Individual) were considered met if the test form was consistent with the cutoffs used in this analysis (**Table 2**), weakly met if the test form fell short by no more than one item cluster or two stand-alone items per criterion and domain, and unmet if the test form fell short by more than one item or two item clusters per criterion and domain. The Balance of Representation criterion for test events is binary and was either met or unmet.

For individual test events, a reporting category was considered to have met the criterion of Consistency of Cognitive Engagement if the domain had no stand-alone items or item clusters with only DOK - Category 1 interactions and reflected the distribution of complexity as expressed in the standards, meaning the domain included at least one item with one or more interactions at DOK Category 3. This expectation was considered weakly met if results fell short by no more than one item per domain. Weakly met indicates that the criterion was nearly met, within a margin that could simply be due to error or reasonable variation in reviewer coding. The criterion was considered unmet if results fell short by two or more items per domain.

For Use of Phenomenon, Relationship of Scoring Assertions with Student Interactions, and Relationship of Scoring Assertions with Standards, test events met the criterion if at least 90% of items were coded affirmatively (see **Table 2**), weakly met the criterion if results were within 10% of this cutoff (i.e. at least 80% of items met expectations).

Cutoffs for Overall Alignment of Test Events with Standards:
Typically, a summative assessment test form has been considered fully aligned with corresponding standards if no changes were needed and acceptably aligned if it needed between one and five items revised or replaced. This widely accepted decision rule was grounded in the context of a typical multiple-choice test form of around 50 items that were generally equally weighted. Five items therefore constituted approximately 10% of the test form. If between six and 10 items (more than 10% and up to 20% of items) needed revision or replacement, the test form was considered to need slight adjustments. If a test form needed over 10 items (greater than 20% of items) revised or replaced, it was considered to need major adjustment. These decision rules did not apply in the context of the SDSA, which included multi-part items and item clusters that vary in the number of associated Scoring Assertions. Most SDSA stand-alone items had one or two Scoring Assertions, with up to five Scoring Assertions, maximum. SDSA item clusters had an average of nine Scoring Assertions and up to 18 Scoring Assertions, maximum. Because items vary in their contribution to a student's score, the approximate percentage of Scoring Assertions affected by unmet alignment expectations was

used to categorize the degree of alignment for a test event. The same typical decision rules were applied, as described on the previous page, but in the context of the Scoring Assertions. Therefore, a test form was considered fully aligned if no changes were needed and acceptably aligned if it needed revisions or replacements corresponding to up to 10% of the overall Scoring Assertions for the test form. A test form was considered to need slight adjustments if it needed revisions or replacements corresponding to between 10% and 20% of the overall Scoring Assertions for the test form, and to need major adjustments if it needed revisions or replacements corresponding to over 20% of the overall Scoring Assertions for the test form.

To determine the overall percentage of affected Scoring Assertions, the exact number of Scoring Assertions was used for any specific item(s) that required revision or replacement. For test forms that needed the addition of one or more items that offered DOK Category 3 cognitive engagement, a per-item estimate was used of 4% of the total Scoring Assertions for the test form. This estimate was based on an average item cluster, comprised of three parts and associated with nine Scoring Assertions, and an average test form, associated with 71 Scoring Assertions (SDSA, Grade 5), 84 Scoring Assertions (SDSA, Grade 8), or 83 Scoring Assertions (SDSA, Grade 11). When an item cluster was rated a DOK - Category 3, the opportunity for Category 3 cognitive engagement generally corresponded to one part of a multi-part item. Therefore, inclusion of a Category 3 opportunity can be considered equivalent to at least around three Scoring Assertions, approximately 4% of the total scoring assertions for a test for each grade. These decision rules allow for an overall categorization of the degree of alignment that takes into account the varying contribution of items to a student's score. For example, a test form would be found to need slight improvements if it needed replacement or revision of three items, each of which was associated with multiple Scoring Assertions such that the total number of Assertions was >10% of the overall set of Assertions for the test form. Another test form, however, would be found to be acceptably aligned if it needed replacement or revision of three items, each of which were associated with a single Scoring Assertion, as in that case the overall proportion of affected Scoring Assertions would only be approximately 3-5%.

Alignment Results: Sample Test Events

Alignment findings are reported for three SDSA sample test events for each of Grades 5, 8, and 11. For all grades, test events included six item clusters, two per domain, and 12 stand-alone items, four per domain. In order to provide information about alignment of test events generated for students across a range of achievement, test events were randomly sampled from at or near cut scores for Levels 2 (below proficiency), 3 (at proficiency), and 4 (above proficiency) for each grade.

Overall test-event-level alignment results are summarized in **Table 8**. Based on the cutoffs for the alignment criteria agreed upon and used in this study, all SDSA test forms analyzed would be considered fully or acceptably aligned with corresponding South Dakota Science Standards. The approximate numbers of replaced or revised items necessary for full alignment are provided for each test form. However, because items vary in their contribution to a student's score, the approximate percentage of affected Scoring Assertions was used to categorize the degree of alignment.

Table 8. Overall Alignment Findings for SDSA Grades 5, 8, and 11 Sample Test Forms with Corresponding Standards

Test Form	Findings	Approx. Number of Items that Need Revision/ Replacement for Full Alignment	Approx. % of Total Assertions that Need Revision/ Replacement for Full Alignment*
SDSA Grade 5 Form 1 (Level 2 – below proficiency)	Acceptably Aligned*	2 items	8%
SDSA Grade 5 Form 2 (Level 3 – at proficiency)	Acceptably Aligned	2 items	6%
SDSA Grade 5 Form 3 (Level 4 – above proficiency)	Acceptably Aligned	2 items	5%
SDSA Grade 8 Form 1 (Level 2 – below proficiency)	Acceptably Aligned	1 item	4%
SDSA Grade 8 Form 2 (Level 3 – at proficiency)	Fully Aligned	--	--
SDSA Grade 8 Form 3 (Level 4 – above proficiency)	Acceptably Aligned	1 item	4%
SDSA Grade 11 Form 1 (Level 2 – below proficiency)	Acceptably Aligned	1 item	4%
SDSA Grade 11 Form 2 (Level 3 – at proficiency)	Acceptably Aligned	2 items	6%
SDSA Grade 11 Form 3 (Level 4 – above proficiency)	Acceptably Aligned	1 item	10%

*Item 379 was included on this test form but was not found within the items provided for review. This item has only one scoring assertion and so even if it were included, it would not affect the overall finding of acceptable alignment.

The distribution of items by DOK - Category of Cognitive Engagement is shown in **Tables 9 - 11**. Any alignment weaknesses identified for test forms are described. Because each test form addresses a different set of standards, and because the SDSA blueprints do not select for cognitive engagement (i.e. "cognitive complexity" or "depth"), the distribution of items at DOK - Category 2 and Category 3 cognitive engagement was expected to vary to some extent between and among test events. Therefore, this variation was not considered an alignment issue based on state expectations. The findings related to the distribution of items by Category of Engagement at the test event level are reported here for informational purposes. Revision or removal of the two Grade 5 items and two Grade 8 items flagged as DOK 1 would resolve some

weakness for Consistency of Cognitive Engagement. To ensure that all test forms provide DOK 3 interaction for all domains, items would need to be selected for complexity or adjustments to items would need to be made, for example, to ensure that all item clusters included DOK 3 interactions.

Panelists identified issues and specified concerns with individual items. This information was provided to South Dakota Department of Education and to CAI. Specific items that did not meet one or more expectations and that were included on the sample test events were taken into consideration in the overall alignment results reported here. In other words, if a test event included one of the items flagged for removal in the overall item bank analysis, that item was considered to need revision or replacement for full alignment as reported in the test-event-level results. All qualitative feedback collected was provided to the state and to CAI.

SDSA Grade 5 Sample Test Events The three SDSA Grade 5 sample test events analyzed were found to be acceptably aligned with South Dakota Science Standards based on the criteria used in this analysis. One form included item 367, which was not located within the set of reviewed items. However, this unknown item was associated with only a single scoring assertion, constituting only ~1% of the overall scoring assertions on the test form and its inclusion would not affect overall findings. Disregarding that item, Form 1 additionally needed the revision or replacement of one Life Science and one Earth and Space Science item to ensure the opportunity for Category 3 cognitive engagement within those domains to fully meet state expectations. Form 1 also included an item (379) which persisted in the item bank without revisions but was flagged in 2019 for content-related issues and because scoring assertions were found not to adequately reflect the expectations of the targeted standard. This item was associated with just two scoring assertions, constituting <3% of the overall scoring assertions for the test form. To fully meet state expectations, both Forms 2 and 3 were found to need one item revised or replaced to ensure the opportunity for DOK - Category 3 cognitive engagement within the Life Science domain. Both Forms 2 and 3 included another item (444) which persisted in the item bank without revisions but was flagged in 2019 as DOK Category 1 and recommended for revision or removal to ensure no items with only Category 1 interactions. The item required students to recall a specific fact and did not offer students the opportunity to engage with a phenomenon to make sense of the science involved. This item was associated with just one scoring assertion, constituting <2% of the overall scoring assertions on each test form. Form 3 also needed the revision or replacement of one Physical Science item and one Earth and Space Science item to ensure the opportunity for Category 3 cognitive engagement within each domain. Panelists included qualitative feedback and suggestions for many items across forms that merit consideration. Overall, however, for all forms analyzed, all domains met or weakly met Consistency of Cognitive Engagement and fully met all other criteria.

Table 9. Distribution of Items by DOK - Category of Cognitive Engagement, SDSA Grade 5

Grade 5 Form 1	DOK - Cognitive Engagement by Domain		
Reporting Categories	% Category 1	% Category 2	% Category 3
Physical Science	0%	67%	33%
Life Science*	0%	100%	0%
Earth and Space Science	0%	100%	0%
Grade 5 Form 2	**DOK - Cognitive Engagement by Domain**		
Reporting Categories	% Category 1	% Category 2	% Category 3
Physical Science	0%	83%	17%
Life Science	0%	100%	0%
Earth and Space Science	17%	66%	17%
Grade 5 Form 3	**DOK - Cognitive Engagement by Domain**		
Reporting Categories	% Category 1	% Category 2	% Category 3
Physical Science	0%	67%	33%
Life Science	0%	100%	0%
Earth and Space Science	17%	67%	17%

*One Life Science item on this form was not included in review

Grade 8 Sample Test Events The three SDSA Grade 8 sample test events analyzed were found to be fully or acceptably aligned with South Dakota Science Standards based on the criteria used in this analysis. To fully meet state expectations, Forms 1 and 3 were found to need one item revised or replaced to ensure no items with only DOK - Category 1 interactions. Both forms contained item 395, which was flagged as DOK - Category 1 in 2019 and recommended for revision or removal but persisted in the item bank unchanged. This item required students to recall specific inputs and outputs of a chemical reaction. In addition to the recall nature of the task, reviewers also noted that the assessment boundary for the targeted standard specifies that an assessment of the standard should not include details about the chemical reactions referenced in the standard. However, this item constituted only ~3-4% of the overall scoring assertions for the test forms and all other alignment criteria were met for all domains. Form 2 fully met all alignment criteria for all domains. Panelists included qualitative feedback for several items across forms that merit consideration.

Table 10. Distribution of Items by DOK - Category of Cognitive Engagement, SDSA Grade 8

Grade 8 Form 1	DOK - Cognitive Engagement by Domain		
Reporting Categories	% Category 1	% Category 2	% Category 3
Physical Science	0%	67%	33%
Life Science	17%	66%	17%
Earth and Space Science	0%	50%	50%
Grade 8 Form 2	**DOK - Cognitive Engagement by Domain**		
Reporting Categories	% Category 1	% Category 2	% Category 3
Physical Science	0%	50%	50%
Life Science	0%	67%	33%
Earth and Space Science	0%	67%	33%
Grade 8 Form 3	**DOK - Cognitive Engagement by Domain**		
Reporting Categories	% Category 1	% Category 2	% Category 3
Physical Science	0%	67%	33%
Life Science	17%	50%	33%
Earth and Space Science	0%	67%	33%

Grade 11 Sample Test Events The three SDSA grade 11 sample test events analyzed were all found to be acceptably aligned with South Dakota Science Standards based on the criteria used in this analysis. To fully meet state expectations, Form 1 needed the revision or replacement of one Life Science item and Form 2 needed the revision or replacement of one Earth and Space Science item to ensure the opportunity for DOK - Category 3 cognitive engagement within each reporting category. For Form 3, panelists flagged item 668 for removal or replacement. This item required students to draw on content knowledge outside of the high school Life Science standards (speciation by polyploidy). Without incorporation of outside knowledge of this content, the provided information could not be interpreted to successfully engage with most of the interactions within the item. This item would need to be replaced with another item that provided opportunity for DOK - Category 3 cognitive engagement. Panelists included qualitative feedback and suggestions for several additional items across forms that merit consideration. Panelists also commented on particular interactions, such as in items that allowed students to conduct trials, that they thought offered great opportunities for students to engage with complex tasks that were consistent with the multidimensional expectations of the standards.

Table 11. Distribution of Items by DOK - Category of Cognitive Engagement, SDSA Grade 11

Grade 11 Form 1	**DOK - Cognitive Engagement by Domain**		
Reporting Categories	% Category 1	% Category 2	% Category 3
Physical Science	0%	67%	33%
Life Science	0%	100%	0%
Earth and Space Science	0%	83%	17%
Grade 11 Form 2	**DOK - Cognitive Engagement by Domain**		
Reporting Categories	% Category 1	% Category 2	% Category 3
Physical Science	0%	83%	17%
Life Science	0%	83%	17%
Earth and Space Science	0%	100%	0%
Grade 11 Form 3	**DOK - Cognitive Engagement by Domain**		
Reporting Categories	% Category 1	% Category 2	% Category 3
Physical Science	0%	83%	17%
Life Science	0%	83%	17%*
Earth and Space Science	0%	67%	33%

*item flagged for removal

Reliability Among Panelists

Panelists engaged in limited adjudication of their data after all panelists finished their coding for the item batches. These discussions were used to identify any mistakes in coding and ensure that the data were entered as intended. Panelists were not required to change their coding after discussion unless they found a compelling reason. The agreement statistics shown in **Table 12** were computed after adjudication for the analyses completed in June 2022. An intraclass correlation value greater than 0.8 generally indicates a high level of agreement among the reviewers. A pairwise comparison value greater than 0.7 generally indicates a high level of agreement among the reviewers. Pairwise comparison for DOK assignment was adequate to high for all item batches (average 0.83). Intraclass correlation is not as meaningful when there is little variation in DOK and reviewers generally agree (e.g. in **Table 12** see Batch 49, in which all but one item was coded as DOK 2) and in these cases pairwise comparison can provide more information.

Table 12. Intraclass and Pairwise Comparisons for Assignment of DOK - Category of Cognitive Engagement by Grade and Batch of Items

Grade 5		
Batch	**Intraclass Correlation**	**Pairwise Comparison**
48	0.64	0.89
49	0.39	0.81
50	0.88	0.90
Grade 8		
Batch	**Intraclass Correlation**	**Pairwise Comparison**
51	0.71	0.86
52	0.87	0.70
53	0.92	0.88
Grade 11		
Batch	**Intraclass Correlation**	**Pairwise Comparison**
54	0.67	0.75
55	0.88	0.77
56	0.95	0.87

Panelists independently recorded a “Yes” or “No” response to each prompt about the relationship of the Scoring Assertions, in aggregate, with the actual student interactions and with the corresponding standard. Panelists also independently recorded a “Yes” or “No” response to a prompt about whether the item/cluster met the expectations for Use of Phenomenon. Panelists did not conduct adjudication specific to these evaluative prompts but sometimes discussed these codings in the context of overall discussion about an item. Panelist agreement for their codings of Use of Phenomenon and Scoring Assertions’ relationship to student interactions and to standards is shown in **Table 13.** For each table, the first column shows the percentage of items in each grade band item bank for which all panelists coded the same way, either all coding “Yes” or all coding “No” in response to each evaluative prompt. The middle column shows the percentage of items for which the vast majority of panelists agreed (all-but-one for 6-person coding and all-but-one-or-two for 12-person coding). The rightmost column shows the percentage of items for which there was greater disagreement among panelists.

Panelists were very consistent in their coding for Use of Phenomenon, with consensus agreement on over 90% of items and >80% agreement on nearly all items across grades. Panelists were also very consistent in their coding for the evaluations of the Scoring Assertions with >80% agreement on between 78% to 98% of items reviewed. The greatest variation in coding was for the evaluation of the relationship of Scoring Assertions to the standard. Panelists approached this evaluative step in different ways. Additional training and clarification may have helped to improve consistency in coding. While the state expectations allowed for stand-alone items to be two-dimensional, some panelists coded these items as not meeting the expectation for the Scoring Assertions to reflect the content of the standard in cases where the Scoring Assertions did not address the full three-dimensionality of the standard. Additionally, when panelists took issue with some qualitative aspect of the item, they often communicated this through a negative coding related to the Scoring Assertions. This may have also contributed to the greater variation in coding for the evaluative components shown in **Table 13.** The greater variation did not interfere with interpretation of findings.

Table 13. SDSA Panelist Agreement for Rating of Use of Phenomenon and Rating of Scoring Assertions to Student Interactions and to Standard by Grade

Batch	Use of Phenomenon		
	% of items with 100% panelist agreement	% of items with > 80% panelist agreement	% of items with < ~80% panelist agreement
Grade 5	98%	100%	-
Grades 8	93%	100%	-
Grades 11	93%	98%	2%
Batch	**Scoring Assertion to Student Interaction**		
	% of items with 100% panelist agreement	% of items with > ~80% panelist agreement	% of items with < ~80% panelist agreement
Grade 5	97%	98%	2%
Grades 8	27%	78%	22%
Grades 11	49%	80%	20%
Batch	**Scoring Assertion to Standard**		
	% of items with 100% panelist agreement	% of items with > ~80% panelist agreement	% of items with < ~80% panelist agreement
Grade 5	49%	89%	11%
Grades 8	20%	53%	47%
Grades 11	29%	65%	35%

Summary Findings by Research Question:
The research questions used to guide the study design and execution are presented on the following pages, along with the corresponding study findings.

Research Question 1: To what extent do the stand-alone items and item clusters satisfy the measurement target claims (standard and Scoring Assertions) identified in the CAI metadata?

Study results found that the vast majority of items satisfied the measurement target claims identified in the CAI metadata. Panelists disagreed with aspects of Scoring Assertions for multiple items, but their independent coding met expectations for this criterion: a panel majority agreed with the standard and Scoring Assertion metadata for >90% of items.

Research Question 2: What DOK-Category of Cognitive Engagement is required for successful completion of each interaction within a stand-alone item or item cluster and how does the DOK distribution within the SDSA item bank compare with the DOK distribution within the South Dakota Science Standards?

Of the 321 items included in the analysis, only four (<2%) were flagged for revision or removal with the primary issue identified related to the Category of Engagement. These items were found to include only Category 1 interactions (requiring recall of information only). While this information can be used for ongoing improvements to the item bank, it was not considered a threat to alignment. Aside from these few items, all items and item clusters were found to be Categories 2 and 3, corresponding proportionately, overall, to the complexity of the grade-band standards. The vast majority of items considered Category 3 were item clusters. The item cluster structure of this assessment, along with the multiple types of information provided within the stimulus and the multiple types of student interactions possible, allowed opportunities for a wide variety of tasks that were considered to require Category 3 cognitive engagement. Panelists noted that most student interactions within an item cluster were typically Category 2, but that sometimes at least one of the interactions required students to interweave the components of the tasks such that there was at least one Category 3 interaction. Panelists coded an item cluster (defined as the unit of analysis) to the highest Category of Engagement that it included to ensure that coding captured the full scope of the complexity of an item cluster.

Research Question 3: To what extent do the stand-alone items and item clusters satisfy the claim that the assessment is phenomenon-based?

The South Dakota Science Assessment was intended to require students to engage multiple dimensions of the standards ("use science") to make sense of phenomena. With the exception of the four items flagged for revision or removal due to requiring only Category 1 interactions (and, therefore, not requiring students to make sense of a phenomenon) all items were considered to meet expectations for being phenomenon-based.

Research Question 4: To what extent was the SDSA likely to generate test events that were aligned with corresponding grade-level academic standards, considering depth and breadth (specified in ESSA) as well as other alignment criteria agreed upon and used in this analysis?

Item-level results were mapped onto actual SDSA test events sampled from each of grade 5, 8, and 11 to yield test-event-level content alignment results. Because the SDSA was adaptively administered in SY2021-2022, test events used in the analysis were randomly sampled from at or near cut scores for Levels 2 (below proficiency), 3 (at proficiency), and 4 (above proficiency) for each grade. This sampling allowed for information about alignment of test events generated for students across achievement levels. Based on the cutoffs for the alignment criteria agreed upon and used in this study, all nine SDSA test forms analyzed were considered fully or acceptably aligned with corresponding South Dakota Science Standards. Across forms, any alignment weaknesses affected only a small proportion of the scoring assertions (<10%), and the overall test forms were still considered acceptably aligned. The consistency in test-event-level findings suggests that the SDSA Item Bank was operating as intended.

Conclusion

This report summarizes the results of a content alignment analysis of the South Dakota Science Assessment (SDSA) for Grades 5, 8, and 11 with corresponding South Dakota Science Standards as pertains to fulfilling requirements as stated in Federal statute. The SDSA used a particular state-vetted subset of items that were part of an item bank managed and owned in part by Cambium Assessment (CAI) and shared by multiple states. A total of 321 items and item clusters were included in the analysis. The results described in this report include alignment-related characteristics of the overall SDSA item bank as well as test-event-level findings. Alignment is reported according to nine criteria agreed upon by participating states, including South Dakota, to be used to evaluate alignment of the assessments with corresponding standards:

1. **Use of Phenomena:** Stand-alone items and item clusters were expected to be grounded in a stimulus that met the test development criteria for a phenomenon. Items/clusters were expected to require students to engage multiple dimensions of the standards ("use science") to make sense of those phenomena.
2. **Categorical Concurrence:** Test events were expected to yield sufficient evidence to make inferences about student knowledge, skills, and abilities (KSAs) as relates to each reporting category.
3. **Dimensionality (Structure of Knowledge):** Item clusters were expected to require students to demonstrate integrated engagement with the three dimensions of Science and Engineering Practices (SEPs), Disciplinary Core Ideas (DCIs), and Crosscutting Concepts (CCCs) specified in the targeted standard. Stand-alone items were expected to require students to demonstrate integrated engagement with two or three of the dimensions specified in the targeted standard.
4. **Consistency of Cognitive Engagement:** The assessment was expected to elicit work that is as cognitively demanding as the expectations in the standards.
5. **Range of Knowledge Correspondence (Individual):** Test events were expected to assess an appropriate breadth of the standards, as defined by the test blueprint. For individual students, assessed standards were expected to be sampled across topics within each reporting category.
6. **Range of Knowledge Correspondence (Population):** At least 90% of standards within a grade band were expected to have the potential to be assessed across the student population. State-specific claims were expected to be consistent with aggregate data from all administered test events in the state in conjunction with results from an independent analysis of vendor metadata.
7. **Balance of Representation:** No standard should be targeted more than once on any single test event.
8. **Relationship of Scoring Assertions with Student Interactions:** In aggregate, the scoring assertions for an item/item cluster were expected to appropriately represent the inferences about student knowledge, skills, and abilities that could be made based on successful interactions with an item/cluster.
9. **Relationship of Scoring Assertions with PEs:** In aggregate, the scoring assertions for an item/item cluster were expected to appropriately represent the three-dimensional expectations of the targeted PE.

Cutoffs for acceptability are given in **Table 2** and detailed within the report section **Alignment Criteria Used for This Analysis** (p. 11).

Overall, the SDSA program was found to have the capacity to generate test forms that were fully or acceptably aligned with the corresponding grade band standards. This finding included consideration of the results of the item-level analyses of the overall item bank, sample SDSA test events, and SDSA blueprints. The evidence to support these findings includes:

- The SDSA blueprints identified South Dakota's intended sampling across reporting categories (as relates to Categorical Concurrence, Range of Knowledge (breadth) for individual test events, and Balance of Representation (emphasis)).
- Overall, the items within the SDSA item bank met South Dakota's expectations (as relates to Use of Phenomena, Dimensionality, Consistency of Cognitive Engagement, and relationships with scoring assertions).
- The SDSA item bank fully (grade 5) or weakly (grades 8 and 11) met the state's expectations for Range of Knowledge across the tested student population. The weak Range of Knowledge (Population) for grades 8 and 11 item banks could be fully resolved with the addition of at least four items to the middle school item bank and six items to the high school item bank.
- An analysis of three sample test events from each of grades 5, 8, and 11 found that all test events were fully or acceptably aligned with corresponding standards, based on the criteria agreed upon by states and used in this analysis.

At the test event level, some variation in the Consistency of Cognitive Engagement was expected between and among test forms because the blueprint did not specify any distribution for the complexity of items. Nearly all instances of DOK - Category 3 cognitive engagement were found within item clusters. Therefore, the variation in distribution of DOK - Category 2 and 3 tasks between and among test forms will depend almost entirely on the particular item clusters assigned for each domain on a test event. Over half of the test forms analyzed did not include one or more item cluster(s) with at least one opportunity for DOK - Category 3 cognitive engagement for all domains. However, the overall distribution of complexity of items in the item bank was found to be appropriate (DOK - Category 2 and 3) in relation to the distribution of complexity in the Performance Expectations. As such, Consistency of Cognitive Engagement of the assessment with standards can be expected across the tested student population. If South Dakota wishes to have greater consistency in the distribution of items by Category of Cognitive Engagement between and among test forms, adjustments would need to be made to the item bank and/or to the test blueprints. Adjustments could also help ensure that all test events included at least one item per domain that required DOK - Category 3 cognitive engagement.

Panelists identified specific items that did not meet one or more alignment-related expectations, and warrant revisions or removal. Even for items that panelists agreed met alignment-related expectations, many editorial suggestions were made to correct errors found in text and graphics, improve clarity, and/or address scientific inaccuracy. This extent of editorial issues is typically not observed in a high-stakes operational assessment and included issues that could potentially affect student scores. Of the items flagged, most are stand-alone items associated with just one or two scoring assertions. Because of the relatively limited interactions, a single stand-alone item contributes proportionately minimally to a student's score. While these items were not considered a significant threat to the alignment of test events, it is suggested that South Dakota / CAI consider revision or removal of all items flagged in the item-level analysis as

well as consider panelist feedback to support ongoing maintenance of and improvement to the item bank.

Overall, panelists found that items and item clusters were meeting state expectations for assessment tasks to require integrated engagement with at least two (stand-alone items) or three (item clusters) dimensions specified in the targeted standard in order to make sense of a phenomenon. With just a very few exceptions, items required student cognitive engagement consistent with the expectations of the standards. Items were spread across the domains of Physical, Life, and Earth and Space Science, with no standard(s) overemphasized in the item bank. Overall, panelists found that the large majority of scoring assertions reasonably reflected inferences that could be made based on student interactions and corresponded to the expectations within the targeted standard. At the test event level, all nine SDSA test forms analyzed were considered fully or acceptably aligned with corresponding South Dakota Science Standards. The consistency in test-event-level findings, across proficiency levels and grades, suggests that the SDSA Item Bank was operating as intended.

Bibliography and References

Achieve, Inc.; Stanford/SCALE (2019) Reconceptualizing Alignment for NGSS Assessments. NARST Annual Meeting. Symposium paper available (01/2020) at https://snapgse.stanford.edu/sites/g/files/sbiybj10126/f/narst_symposium_2019_3.31.19.pdf

Badrinarayan, A., Christopherson, S. and Harris, C. (2017) The Alignment Challenge. Reidy Interactive Lecture Series: Assessing Student Learning of the Next Generation Science Standards. The National Center for the Improvement of Educational Assessment

Badrinarayan, A, Christopherson, S., Gong, B., McCrae, A. (2018) Developing a Common Language to Understand Content Complexity for Alignment Studies of the NGSS. National Conference on Student Assessment

Badrinarayan, A., Christopherson, S., Davis-Becker, S., Everson, H., and Forte, E. (2019). Representing Cognitive Complexity in Test Design and Evaluation. ATP Innovations in Testing

National Research Council. 2012. A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas. Washington, DC: The National Academies Press. https://doi.org/10.17226/13165.

National Research Council. (2005). How Students Learn: History, Mathematics, and Science in the Classroom. Committee on How People Learn, A Targeted Report for Teachers, M.S. Donovan and J.D. Bransford, Editors. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

NGSS Lead States. 2013. Next Generation Science Standards: For States, By States. Washington, DC: The National Academies Press.

NGSS Lead States. 2013. NGSS Release: How to Read the Next Generation Science Standards (NGSS). Washington, DC: The National Academies Press.

Subkoviak, M. J. (1988). A practitioner's guide to computation and interpretation of reliability indices for mastery tests. Journal of Educational Measurement, 25(1), 47-55.

Valencia, S. W., & Wixson, K. K. (2000). Policy-oriented research on literary standards and assessment. In M. L. Kamil, P. B. Mosenthal, P. D. Pearson, & R. Barr (Eds.), Handbook of reading research: Vol. III. Mahwah, NJ: Lawrence Erlbaum.

Webb, N. L. (1997). Criteria for alignment of expectations and assessments in mathematics and science education. Council of Chief State School Officers and National Institute for Mathematics Education Research Monograph No. 6. Madison: University of Wisconsin, Wisconsin Center for Education Research.

Webb, N. L. (1999). Alignment of science and mathematics standards and assessments in four states. Council of Chief State School Officers and National Institute for Science Education Research Monograph No. 18. Madison, WI: University of Wisconsin.

Webb, N. L. (2003). Alignment study in language arts, mathematics, science, and social studies of state standards and assessments in four states. Washington, D. C.: Council of Chief State School officers.

Wise, S. L., Kingsbury, G. G., and Webb, N. L. (2015). Evaluating content alignment in computerized adaptive testing. Educational Measurement: Issues and Practices, Winter, 34, 4, pp. 41-48.

WebbAlign™

WCEPS
Wisconsin Center for
Education Products & Services

Appendix A

Group Consensus Category of Engagement - DOK Values for South Dakota Science Standards Grades 3-12

January, 2023

	South Dakota Science Standard	DOK
PS	Elementary School (Grades 3-5) Physical Science	
PS2	Motion and Stability: Forces and Interactions	
3-PS2-1	Plan and carry out an investigation to provide evidence of the effects of balanced and unbalanced forces on the motion of an object. (SEP: 3; DCI: PS2.A, PS2.B; CCC: Cause/Effect)	4
3-PS2-2	Make observations and/or measurements of an object's motion to provide evidence for how a pattern can be used to predict future motion. (SEP: 3; DCI: PS2.A; CCC: Patterns)	3
3-PS2-3	Ask questions about cause and effect relationships of electric or magnetic interactions between two objects not in contact with each other. (SEP: 1; DCI: PS2.B; CCC: Cause/Effect)	2
3-PS2-4	Define a simple design problem that can be solved by applying scientific ideas about magnets.* (SEP: 1; DCI: PS2.B; CCC: Technology)	3
5-PS2-1	Support an argument that the gravitational force exerted by Earth on objects is directed down. (SEP: 7; DCI: PS2.B; CCC: Cause/Effect)	2
PS3	Energy	
4-PS3-1	Use evidence to construct an explanation relating the speed of an object to the energy of that object. (SEP: 6; DCI: PS3.A ; CCC: Energy/Matter)	2
4-PS3-2	Make observations to provide evidence for how energy can be transferred from place to place by sound, light, heat, and electric currents. (SEP: 3; DCI: PS3.A, PS3.B; CCC: Energy/Matter)	2
4-PS3-3	Ask questions and predict outcomes about the changes in energy that occur when objects collide. (SEP: 1; DCI: PS3.A, PS3.B, PS3.C; CCC: Energy/Matter)	2
4-PS3-4	Design, test, and refine a device that converts energy from one form to another.* (SEP: 6; DCI: PS3.B, PS3.D, ETS1.A ; CCC: Energy/Matter)	4
5-PS3-1	Use models to describe that energy in animals' food (used for body repair, growth, motion, and to maintain body warmth) was once energy from the sun. (SEP: 2; DCI: PS3.D, LSI.C ; CCC: Energy/Matter)	2
PS4	Waves and Their Applications in Technologies for Information Transfer	
4-PS4-1	Develop a model of waves to describe patterns in terms of amplitude and wavelength and to provide evidence that waves can cause objects to move. (SEP: 2 ; DCI: PS4.A; CCC: Patterns)	2
4-PS4-2	Develop a model to describe how light reflecting from objects and entering the eye allows objects to be seen. (SEP: 2 ; DCI: PS4.B; CCC: Cause/Effect)	2
4-PS4-3	Create and compare multiple solutions that use patterns to transfer information.* (SEP: 6; DCI: PS4.C, ETS1.C; CCC: Patterns, Technology)	3

PS1	Matter and Its Interactions	
5-PS1-1	Develop a model to describe that matter is made of particles too small to be seen. (SEP: 2; DCI: PS1.A; CCC: Scale/Prop.)	2
5-PS1-2	Measure and graph quantities to provide evidence that regardless of the type of change that occurs when heating, cooling, or mixing substances, the total weight of matter is conserved. (SEP: 5; DCI: PS1.A, PS1.B; CCC: Scale/Prop.)	2
5-PS1-3	Make observations and measurements to identify materials based on their properties. (SEP: 3; DCI: PS1.A; CCC: Scale/Prop.)	2
5-PS1-4	Conduct an investigation to determine whether the mixing of two or more substances results in new substances. (SEP: 3; DCI: PS1.B; CCC: Cause/Effect)	3
LS	Elementary School (Grades 3-5) Life Science	
LS1	From Molecules to Organisms: Structures and Processes	
3-LS1-1	Develop models to describe that organisms have unique and diverse life cycles but all have in common birth, growth, reproduction, and death. (SEP: 1 ; DCI: LS1.B; CCC: Patterns)	2
4-LS1-1	Construct an argument that plants and animals have internal and external structures that function to support survival, growth, behavior, and reproduction. (SEP: 7; DCI: LS1.A; CCC: Systems)	2
4-LS1-2	Use a model to describe that animals receive different types of information through their senses, process the information in their brain, and respond to the information in different ways. (SEP: 2; DCI: LS1.D; CCC: Systems)	2
5-LS1-1	Support an argument that plants get the materials they need for growth chiefly from air and water. (SEP: 7; DCI: LS1.C; CCC: Energy/Matter)	2
LS2	Ecosystems: Interactions, Energy, and Dynamics	
3-LS2-1	Construct an argument that some animals form groups that help members survive. (SEP: 7; DCI: LS2.D; CCC: Cause/Effect)	2
5-LS2-1	Develop a model to describe the movement of matter and energy among producers, consumers, decomposers, and the environment. (SEP: 2; DCI:LS2.A, LS2.B ; CCC: Systems)	2

LS3	Heredity: Inheritance and Variation of Traits	
3-LS3-1	Analyze and interpret data to provide evidence that plants and animals have traits inherited from parents and that variations of these traits exist in a group of similar organisms. (SEP: 4; DCI: LS3.A, LS3.B; CCC: Patterns)	2
3-LS3-2	Use evidence and reasoning to support the explanation that traits can be influenced by the environment. (SEP: 6; DCI: LS3.A, LS3.B; CCC: Cause/Effect)	2
LS4	Biological Unity and Diversity	
3-LS4-1	Analyze and interpret data from fossils to provide evidence of the organisms and the environments in which they lived long ago. (SEP: 4; DCI: LS4.A; CCC: Scale/Prop.)	2
3-LS4-2	Use evidence and reasoning to construct an explanation for how the variations in characteristics among individuals of the same species may provide advantages in surviving, finding mates, and reproducing. (SEP: 6; DCI: LS4.B; CCC: Cause/Effect)	2
3-LS4-3	Construct an argument with evidence how some organisms thrive, some struggle to survive, and some cannot survive in a particular habitat. (SEP: 7; DCI: LS4.C; CCC: Cause/Effect)	2
3-LS4-4	Make a claim about the merit of a solution to a problem caused when the environment changes and the types of plants and animals that live there may change.* (SEP: 7; DCI: LS2.C, LS4.D; CCC: Systems, Technology	3
ESS	Elementary School (Grades 3-5) Earth and Space Science	
ESS2	Earth's Systems	
3-ESS2-1	Represent data in tables and graphical displays to describe typical weather conditions expected during a particular season. (SEP: 4; DCI: ESS2.D; CCC: Patterns)	2
3-ESS2-2	Obtain and combine information to describe climates in different regions of the world. (SEP: 8; DCI: ESS2.D ; CCC: Patterns)	2
4-ESS2-1	Make observations and/or measurements to provide evidence of the effects of weathering or the rate of erosion by water, ice, wind, or vegetation. (SEP: 3; DCI: ESS2.A, ESS2.E; CCC: Cause/Effect)	2
4-ESS2-2	Analyze and interpret data from maps to describe patterns of Earth's features. (SEP: 4; DCI: ESS2.B; CCC: Patterns)	2
5-ESS2-1	Develop a model to describe the interaction of geosphere, biosphere, hydrosphere, and/or atmosphere. (SEP: 2; DCI: ESS2.A; CCC: Systems)	2
5-ESS2-2	Describe and graph the amounts and percentages of water and fresh water in various reservoirs to provide evidence about the distribution of water on Earth. (SEP: 5; DCI: ESS2.C; CCC: Scale/Prop.)	2

ESS3	Earth and Human Activity	
3-ESS3-1	Make a claim about the merit of a design solution that reduces the impacts of a weather-related hazard.* (SEP: 7; DCI: ESS3.B ; CCC: Cause/Effect, Technology)	3
4-ESS3-1	Obtain and combine information to describe that energy and fuels are derived from natural resources and their uses affect the environment. (SEP: 8; DCI: ESS3.A; CCC: Cause/Effect, Technology)	2
4-ESS3-2	Generate and compare multiple solutions to reduce the impacts of natural Earth processes on humans. (SEP: 6; DCI: ESS3.B, ETS1.B; CCC: Cause/Effect, Technology)	3
5-ESS3-1	Obtain and combine information about ways individual communities use science ideas to protect the Earth's resources and environment. (SEP:8; DCI: ESS3.C; CCC: Systems)	3
ESS1	Earth's Place in the Universe	
4-ESS1-1	Identify evidence from patterns in rock formations and fossils in rock layers to support an explanation for changes in a landscape over time. (SEP: 6; DCI: ESS1.C ; CCC: Patterns)	2
5-ESS1-1	Support an argument that differences in the apparent brightness of the sun compared to other stars is due to distances from the Earth. (SEP: 7; DCI: ESS1.A; CCC: Scale/Prop.)	2
5-ESS1-2	Represent data in graphical displays to reveal patterns of daily changes in length and direction of shadows, day and night, and the seasonal appearance of some stars in the night sky. (SEP: 4; DCI: ESS1.B ; CCC: Patterns)	2

	South Dakota Science Standard	DOK
MS-PS	Middle School (Grades 6-8) Physical Science	
MS-PS1	Matter and Its Interactions	
MS-PS1-1	Develop models to describe the atomic composition of simple molecules and extended structures. (SEP:2 ; DCI: PS1.A; CCC: Scale/Prop.)	2
MS-PS1-2	Analyze and interpret data on the properties of substances before and after the substances interact to determine if a chemical reaction has occurred. (SEP: 8; DCI: PS1.A, PS1.B; CCC: Patterns)	2
MS-PS1-3	Obtain and evaluate information to describe that synthetic materials come from natural resources and impact society. (SEP: 8; DCI: PS1.A, PS1.B; CCC: Structure/Function, Technology	3
MS-PS1-4	Develop a model that predicts and describes changes in particle motion, temperature, and state of a pure substance when thermal energy is added or removed. (SEP: 2; DCI: PS1.A, PS3.A; CCC: Cause/Effect)	2
MS-PS1-5	Develop and use a model to describe how the total number of atoms does not change in a chemical reaction and thus mass is conserved. (SEP: 2 ; DCI: PS1.B; CCC: Energy/Matter)	2
MS-PS1-6	Design, construct, test, and modify a device that either releases or absorbs thermal energy by chemical processes.* (SEP: 6; DCI: PS1.B, ETS1.B, ETS1.C; CCC: Energy/Matter)	4
MS-PS2	Motion and Stability: Forces and Interactions	
MS-PS2-1	Design a solution to a problem involving the motion of two colliding objects that illustrates Newton's Third Law.* (SEP: 6; DCI: PS2.A; CCC: Systems, Technology)	3
MS-PS2-2	Plan an investigation to provide evidence that the change in an object's motion depends on the sum of the forces on the object and the mass of the object. (SEP: 3; DCI: PS2.A; CCC: Stability/Change)	3
MS-PS2-3	Ask questions about data to determine the factors that affect the strength of electric and magnetic forces. (SEP: 1; DCI: PS2.B; CCC: Cause/Effect)	2
MS-PS2-4	Construct and present arguments using evidence to support the claim that gravitational interactions are attractive and depend on the masses of interacting objects. (SEP: 7; DCI: PS2.B; CCC: Systems)	3
MS-PS2-5	Conduct an investigation and evaluate the experimental design to provide evidence that fields exist between objects exerting forces on each other even though the objects are not in contact. (SEP: 3; DCI: PS2.B; CCC: Cause/Effect)	4

MS-PS3	Energy	
MS-PS3-1	Construct and analyze graphical displays of data to describe the relationships of kinetic energy to the mass of an object and to the speed of an object. (SEP: 4; DCI: PS3.A ; CCC: Scale/Prop.)	2
MS-PS3-2	Develop a model to describe that when the arrangement of objects interacting at a distance changes, different amounts of potential energy are stored in the system. (SEP: 2; DCI: PS3.A, PS3.C; CCC: Systems)	2
MS-PS3-3	Design, construct, and test a device that either minimizes or maximizes thermal energy transfer.* (SEP: 6; DCI: PS3.A, PS3.B, ETS1.A, ETS1.B, ; CCC: Energy/Matter)	4
MS-PS3-4	Plan an investigation to determine the relationships among the energy transferred, the type of matter, the mass, and the change in the average kinetic energy of the particles as measured by the temperature of the sample. (SEP: 3; DCI: PS3.A, PS3.B; CCC: Scale/Prop.)	3
MS-PS3-5	Engage in argument from evidence to support the claim that when the kinetic energy of an object changes, energy is transferred to or from the object. (SEP: 7; DCI: PS3.B; CCC: Energy/Matter)	2
MS-PS4	Waves and Their Applications in Technologies for Information Transfer	
MS-PS4-1	Use mathematical representations to describe a simple model for waves that includes how the amplitude of a wave is related to the energy in a wave. (SEP: 5; DCI: PS4.A; CCC: Patterns)	2
MS-PS4-2	Develop and use a model to describe how waves are reflected, absorbed, or transmitted through various materials. (SEP: 2; DCI: PS4.A, PS4.B; CCC: Structure)	2
MS-PS4-3	Obtain, evaluate and communicate information to support the claim that digitized signals are a more reliable way to encode and transmit information than analog signals. (SEP: 8; DCI: PS4.C; CCC: Structure, Technology)	2

MS-LS	Middle School (Grades 6-8) Life Science	
MS-LS1	From Molecules to Organisms: Structures and Processes	
MS-LS1-1	Plan and carry out an investigation to provide evidence that living things are made of cells; either one cell or many different types and numbers of cells. (SEP: 3; DCI: LS1.A; CCC: Scale/Prop., Technology)	2
MS-LS1-2	Develop and use a model to describe the function of a cell as a whole and ways parts of cells contribute to the function. (SEP: 2; DCI: LS1.A; CCC: Structure/Function)	2
MS-LS1-3	Construct an argument supported by evidence for how the body is a system of interacting subsystems composed of groups of cells. (SEP: 7; DCI: LS1.A; CCC: Systems)	2
MS-LS1-4	Construct an argument based on empirical evidence and scientific reasoning to support an explanation for how characteristic animal behaviors and specialized plant structures affect the probability of successful reproduction of animals and plants respectively. (SEP: 7; DCI: LS1.B; CCC: Cause/Effect)	2
MS-LS1-5	Construct a scientific explanation based on evidence for how environmental and genetic factors influence the growth of organisms. (SEP: 6; DCI: LS1.B; CCC: Cause/Effect)	2
MS-LS1-6	Construct a scientific explanation based on evidence for the role of photosynthesis in the cycling of matter and flow of energy into and out of organisms. (SEP: 6, Nature Science/Empirical Evidence; DCI: LS1.C, PS3.D; CCC: Energy/Matter)	2
MS-LS1-7	Develop a model to describe how food is rearranged through chemical reactions forming new molecules that support growth and/or release energy as this matter moves through an organism. (SEP: 2; DCI: LS1.C, PS3.D; CCC: Energy/Matter)	2
MS-LS2	Ecosystems: Interactions, Energy, and Dynamics	
MS-LS2-1	Analyze and interpret data to provide evidence for the effects of resource availability on organisms and populations of organisms in an ecosystem. (SEP: 4; DCI: LS2.A; CCC: Cause/Effect)	2
MS-LS2-2	Construct an explanation that predicts patterns of interactions among organisms across multiple ecosystems. (SEP: 6; DCI: LS2.A; CCC: Patterns)	2
MS-LS2-3	Develop a model to describe the cycling of matter and flow of energy among living and nonliving parts of an ecosystem. (SEP: 2; DCI: LS2.B; CCC: Energy/Matter)	2

MS-LS2-4	Construct an argument supported by empirical evidence that changes to physical or biological components of an ecosystem affect populations. (SEP: 7; DCI: LS2.C ; CCC: Stability/Change)	2
MS-LS2-5	Evaluate competing design solutions for maintaining biodiversity and ecosystem services.* (SEP: 7; DCI: LS2.C, LS4.D, ETS1.B ; CCC: Stability/Change, Technology)	3
MS-LS3	Heredity: Inheritance and Variation of Traits	
MS-LS3-1	Develop and use a model to describe why structural changes to genes (mutations) located on chromosomes may affect proteins and may result in harmful, beneficial, or neutral effects to the structure and function of the organism. (SEP:2; DCI: LS3.A, LS3.B; CCC: Structure/Function)	2
MS-LS3-2	Develop and use a model to describe why asexual reproduction results in offspring with identical genetic information and sexual reproduction results in offspring with genetic variation. (SEP: 2; DCI: LS1.B, LS3.A, LS3.B; CCC: Cause/Effect)	2
MS-LS4	Biological Unity and Diversity	
MS-LS4-1	Analyze and interpret data for patterns in the fossil record that document the existence, diversity, extinction, and change of life forms throughout the history of life on Earth. (SEP: 4; DCI: LS4.A; CCC: Patterns)	2
MS-LS4-2	Apply scientific ideas to construct an explanation for similarities and differences among modern organisms and between modern and fossil organisms to infer evolutionary relationships. (SEP: 6; DCI: LS4.A; CCC: Patterns)	2
MS-LS4-4	Construct an explanation based on evidence that describes how genetic variations of traits in a population increase some individuals' probability of surviving and reproducing in a specific environment. (SEP: 6; DCI: LS4.B; CCC: Cause/Effect)	2
MS-LS4-5	Obtain, evaluate, and communicate information about how technological advances have changed the way humans influence the inheritance of desired traits in organisms. * (SEP: 8; DCI: LS4.B; CCC: Cause/Effect, Technology)	3
MS-LS4-6	Use mathematical representations to support explanations of how natural selection may lead to increases and decreases of specific traits in populations over time. (SEP: 5; DCI: LS4.C; CCC: Cause/Effect)	2

MS-ESS	Middle School (Grades 6-8) Earth and Space Science	
MS-ESS1	Earth's Place in the Universe	
MS-ESS1-1	Develop and use a model of the Earth-sun-moon system to describe the cyclic patterns of lunar phases, eclipses of the sun and moon, and seasons. (SEP: 2; DCI: ESS1.A, ESS1.B; CCC: Patterns)	2
MS-ESS1-2	Develop and use a model to describe the role of gravity in the motions within galaxies and the solar system. (SEP: 2; DCI: ESS1.A, ESS1.B; CCC: Systems)	2
MS-ESS1-3	Analyze and interpret data to determine scale properties of objects in the solar system. (SEP: 4; DCI: ESS1.B; CCC: Scale/Prop., Technology)	2
MS-ESS2	Earth's Systems	
MS-ESS2-1	Develop a model to describe the cycling of Earth's materials and the flow of energy that drives this process. (SEP: 2; DCI: ESS2.A; CCC: Stability/Change)	2
MS-ESS2-2	Construct an explanation based on evidence for how geoscience processes have changed Earth's surface at varying time and spatial scales. (SEP: 6; DCI: ESS2.A, ESS2.C; CCC: Scale/Prop.)	2
MS-ESS2-3	Analyze and interpret data on the age of the Earth, distribution of fossils and rocks, continental shapes, and seafloor structures to provide evidence of the past plate motions. (SEP: 4; DCI: ESS2.B, ESS1.C; CCC: Patterns)	2
MS-ESS2-4	Develop a model to describe the cycling of water through Earth's systems driven by energy from the sun and the force of gravity. (SEP: 2; DCI: ESS2.C; CCC: Energy/Matter)	2
MS-ESS2-5	Collect data to provide evidence for how the motions and complex interactions of air masses results in changes in weather conditions. (SEP: 3; DCI: ESS2.C, ESS2.D; CCC: Cause/Effect)	3
MS-ESS2-6	Develop and use a model to describe how unequal heating and rotation of the Earth cause patterns of atmospheric and oceanic circulation that determine regional climates. (SEP: 2; DCI: ESS2.C, ESS2.D; CCC: Systems)	2

MS-ESS3	Earth and Human Activity	
MS-ESS3-1	Construct a scientific explanation based on evidence for how the uneven distributions of Earth's mineral, energy, and groundwater resources are the result of past and current geoscience processes. (SEP: 6; DCI: ESS3.A ; CCC: Cause/Effect , Technology)	2
MS-ESS3-2	Analyze and interpret data on natural hazards to forecast future catastrophic events and inform the development of technologies to mitigate their effects. (SEP: 4; DCI: ESS3.B; CCC: Patterns, Technology)	3
MS-ESS3-3	Apply scientific principles to design a method for monitoring and minimizing a human impact on the environment.* (SEP: 6 ; DCI: ESS3.C; CCC: Cause/Effect, Technology)	3
MS-ESS3-4	Construct an argument supported by evidence for how increases in human population and per- capita consumption of natural resources impact Earth's systems. (SEP: 7; DCI: ESS3.C; CCC: Cause/Effect, Technology, Nature Science/Consequence-Actions)	3
MS-ESS3-5	Ask questions to clarify evidence of the factors that may have caused a change in global temperatures over the past century. (SEP: 1; DCI: ESS3.D; CCC: Stability/Change)	3

	South Dakota Science Standard	DOK
HS-PS	High School (Grades 9-12) Physical Science	
HS-PS1	Matter and Its Interactions	
HS-PS1-1	Use the periodic table as a model to predict the relative properties of elements based on the patterns of electrons in the outermost energy level of atoms. (SEP: 2; DCI: PS1.A, PS2.B; CCC: Patterns)	2
HS-PS1-2	Construct and revise an explanation for the outcome of a simple chemical reaction based on the outermost electron states of atoms, trends in the periodic table, and knowledge of the patterns of chemical properties. (SEP: 6; DCI: PS1.A, PS1.B; CCC: Patterns)	2
HS-PS1-3	Plan and carry out an investigation to gather evidence to compare the structure of substances at the bulk scale to infer the strength of electrical forces between particles. (SEP: 3; DCI: PS1.A, PS2.B; CCC: Patterns)	4
HS-PS1-4	Develop a model to illustrate that the release or absorption of energy from a chemical reaction system depends upon the changes in total bond energy. (SEP: 2; DCI: PS1.A, PS1.B; CCC: Energy/Matter)	2
HS-PS1-5	Construct an explanation based on evidence about the effects of changing the temperature or concentration of the reacting particles on the rate at which a reaction occurs. (SEP: 6; DCI: PS1.B; CCC: Patterns)	2
HS-PS1-6	Refine the design of a chemical system by specifying a change in conditions that would produce increased amounts of products at equilibrium.* (SEP: 6; DCI: PS1.B, ETS1.C; CCC: Stability/Change)	2
HS-PS1-7	Use mathematical representations to support the claim that atoms, and therefore mass, are conserved during a chemical reaction. (SEP: 5; DCI: PS1.B; CCC: Energy/Matter, Nature of Science/Consistency)	2
HS-PS1-8	Develop models to illustrate the changes in the composition of the nucleus of the atom and the energy released during the processes of fission, fusion, and radioactive decay. (SEP: 2; DCI: PS1.C; CCC: Energy/Matter)	2
HS-PS2	Motion and Stability: Forces and Interactions	
HS-PS2-1	Analyze data to support the claim that Newton's Second Law of motion describes the mathematical relationship among the net force on a macroscopic object, its mass, and its acceleration. (SEP: 4; DCI: PS2.A; CCC: Cause/Effect)	2
HS-PS2-2	Use mathematical representations to support the claim that the total momentum of a system of objects is conserved when there is no net force on the system. (SEP: 5; DCI: PS2.A ; CCC: Systems)	2

HS-PS2-3	Design, evaluate, and refine a device that minimizes the force on a macroscopic object during a collision.* (SEP: 6; DCI: PS2.A, ETS1.A, ETS1.C; CCC: Cause/Effect)	4
HS-PS2-4	Use mathematical representations of Newton's Law of Gravitation and Coulomb's Law to describe and predict the gravitational and electrostatic forces between objects. (SEP: 5; DCI: PS2.B; CCC: Patterns)	2
HS-PS2-5	Plan and carry out an investigation to provide evidence that an electric current can produce a magnetic field and that a changing magnetic field can produce an electric current. (SEP: 3; DCI: PS2.B, PS3.A; CCC: Cause/Effect)	4
HS-PS2-6	Communicate scientific and technical information about why the molecular-level structure is important in the functioning of designed materials.* (SEP: 8; DCI: PS1.A, PS2.B; CCC: Structure/Function)	2
HS-PS3	Energy	
HS-PS3-1	Create a computational model to calculate the change in the energy of one component in a system when the change in energy of the other component(s) and energy flows in and out of the system are known. (SEP: 5; DCI: PS3.A, PS3.B ; CCC: Systems)	2
HS-PS3-2	Develop and use models to illustrate that energy at the macroscopic scale can be accounted for as a combination of energy associated with the motions of particles (objects) and energy associated with the relative position of particles (objects). (SEP: 2 ; DCI: PS3.A; CCC: Energy/Matter)	2
HS-PS3-3	Design, build, and refine a device that works within given constraints to convert one form of energy into another form of energy. (SEP: 6; DCI: PS3.A, PS3.D, ETS1.A; CCC: Energy/Matter, Technology)	4
HS-PS3-4	Plan and carry out an investigation to provide evidence that the transfer of thermal energy when two components of different temperature are combined within a closed system results in a more uniform energy distribution among the components in the system (Second Law of Thermodynamics). (SEP: 3; DCI: PS3.B, PS3.D; CCC: Systems)	4
HS-PS3-5	Develop and use a model of two objects interacting through electric or magnetic fields to illustrate the forces between objects and the changes in energy of the objects due to the interaction. (SEP: 2; DCI: PS3.C; CCC: Cause/Effect)	2

HS-PS4	Waves and Their Applications in Technologies for Information Transfer	
HS-PS4-1	Use mathematical representations to support a claim regarding relationships among the frequency, wavelength, and speed of waves traveling in various media. (SEP: 5; DCI: PS4.A; CCC: Cause/Effect)	2
HS-PS4-2	Evaluate questions about the advantages of using a digital transmission and storage of information. (SEP: 1; DCI: PS4.A; CCC: Stability/Change, Technology)	2
HS-PS4-3	Evaluate the claims, evidence, and reasoning behind the idea that electromagnetic radiation can be described either by a wave model or a particle model, and that for some situations one model is more useful than the other. (SEP: 7; DCI: PS4.A, PS4.B; CCC: Systems)	3
HS-PS4-4	Evaluate the validity and reliability of claims in published materials of the effects that different frequencies of electromagnetic radiation have when absorbed by matter. (SEP: 8; DCI: PS4.B; CCC: Cause/Effect)	3
HS-PS4-5	Communicate technical information about how some technological devices use the principles of wave behavior and wave interactions with matter to transmit and capture information and energy.* (SEP: 8; DCI: PS3.D, PS4.A, PS4.B, PS4.C; CCC: Cause/Effect, Technology)	2
HS-LS	High School (Grades 9-12) Life Science	
HS-LS1	From Molecules to Organisms: Structures and Processes	
HS-LS1-1	Construct an explanation based on evidence for how the structure of DNA determines the structure of proteins which carry out the essential functions of life through systems of specialized cells. (SEP: 6; DCI: LS1.A; CCC: Structure/Function)	2
HS-LS1-2	Develop and use a model to illustrate the hierarchical organization of interacting systems that provide specific functions within multicellular organisms. (SEP: 2; DCI: LS1.A; CCC: Systems)	2
HS-LS1-3	Plan and carry out an investigation to provide evidence that feedback mechanisms maintain homeostasis. (SEP: 3; DCI: LS1.A; CCC: Stability/Change)	4
HS-LS1-4	Use a model to illustrate the role of cellular division (mitosis) and differentiation in producing and maintaining complex organisms. (SEP: 2; DCI: LS1.B; CCC: Systems)	2
HS-LS1-5	Use a model to illustrate how photosynthesis transforms light energy into stored chemical energy. (SEP: 2; DCI: LS1.C; CCC: Systems, Energy/Matter)	2

HS-LS1-6	Construct and revise an explanation based on evidence for how carbon, hydrogen, and oxygen from sugar molecules may combine with other elements to form amino acids and/or other large carbon-based molecules. (SEP: 6; DCI: LS1.C; CCC: Energy/Matter)	2
HS-LS1-7	Use a model of the major inputs and outputs of cellular respiration (aerobic and anaerobic) to exemplify the chemical process in which the bonds of food molecules are broken, the bonds of new compounds are formed, and a net transfer of energy results. (SEP: 2; DCI: LS1.C; CCC: Energy/Matter)	2
HS-LS2	Ecosystems: Interactions, Energy, and Dynamics	
HS-LS2-1	Use mathematical and/or computational representations to support explanations of factors that affect carrying capacity of ecosystems at different scales. (SEP: 5; DCI: LS2.A; CCC: Scale/Prop.)	2
HS-LS2-2	Use mathematical representations to support and revise explanations based on evidence about factors affecting biodiversity and populations in ecosystems of different scales. (SEP: 5; DCI: LS2.A, LS2.C; CCC: Scale/Prop.)	2
HS-LS2-3	Construct and revise an explanation based on evidence for the cycling of matter and flow of energy in aerobic and anaerobic conditions. (SEP:6; DCI: LS2.B; CCC: Energy/Matter)	2
HS-LS2-4	Use mathematical representations to support claims for the cycling of matter and flow of energy among organisms in an ecosystem. (SEP: 5; DCI: LS2.B; CCC: Energy/Matter)	2
HS-LS2-5	Develop a model to illustrate the role of photosynthesis and cellular respiration in the cycling of carbon among the biosphere, atmosphere, hydrosphere, and geosphere. (SEP: 2; DCI: LS2.B, PS3.D; CCC: Systems)	2
HS-LS2-6	Evaluate the claims, evidence, and reasoning that the complex interactions in ecosystems maintain relatively consistent numbers and types of organisms under stable conditions; however, moderate to extreme fluctuations in conditions may result in new ecosystems. (SEP: 7; DCI: LS2.C; CCC: Stability/Change)	3
HS-LS2-7	Design, evaluate, and refine a solution for reducing the impacts of human activities on the environment and biodiversity.* (SEP: 6; DCI: LS2.C, LS4.D, ETS1.B; CCC: Stability/Change)	4
HS-LS2-8	Evaluate the evidence for the role of group behavior on individual and species' chances to survive and reproduce. (SEP: 7; DCI: LS2.D; CCC: Cause/Effect)	3

HS-LS3	Heredity: Inheritance and Variation of Traits	
HS-LS3-1	Ask questions to clarify relationships about the role of DNA and chromosomes in coding the instructions for characteristic traits passed from parents to offspring. (SEP: 1; DCI: LS1.A, LS3.A; CCC: Cause/Effect)	2
HS-LS3-2	Make and defend a claim based on evidence that 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. (SEP: 7; DCI: LS3.B; CCC: Cause/Effect)	2
HS-LS3-3	Apply concepts of statistics and probability to explain the variation and distribution of expressed traits in a population. (SEP: 4; DCI: LS3.B; CCC: Scale/Prop.)	2
HS-LS4	Biological Unity and Diversity	
HS-LS4-1	Communicate scientific information that common ancestry and biological evolution are supported by multiple lines of empirical evidence. (SEP: 8; DCI: LS4.A; CCC: Patterns)	2
HS-LS4-2	Construct an explanation based on evidence that the process of evolution primarily results from four factors: (1) the potential for a species to increase in number, (2) the heritable genetic variation of individuals in a species due to mutation and sexual reproduction, (3) competition for limited resources, and (4) the proliferation of those organisms that are better able to survive and reproduce in the environment. (SEP: 6; DCI: LS4.B, LS4.C; CCC: Cause/Effect)	2
HS-LS4-3	Apply concepts of statistics and probability to support explanations that organisms with an advantageous heritable trait tend to increase in proportion to organisms lacking this trait. (SEP: 4; DCI: LS4.B, LS4.C; CCC: Patterns)	2
HS-LS4-4	Construct an explanation based on evidence for how natural selection leads to adaptation of populations. (SEP: 6; DCI: LS4.C ; CCC: Cause/Effect)	2
HS-LS4-5	Evaluate the evidence supporting claims that changes in environmental conditions may result in: (1) increases in the number of individuals of some species, (2) the emergence of new species over time, and (3) the extinction of other species. (SEP: 7; DCI: LS4.C; CCC: Cause/Effect)	3
HS-LS4-6	Use a simulation to research and analyze possible solutions for the adverse impacts of human activity on biodiversity . (SEP: 5; DCI: LS4.C, LS4.D, ETS1.B; CCC: Cause/Effect)	3
HS-LS4-7	Analyze displays of pictorial data to compare patterns of similarities in the embryological development across multiple species to identify relationships not evident in the fully formed anatomy. (SEP: 4; DCI: LS4.A ; CCC: Patterns)	2

HS-ESS	High School (Grades 9-12) Earth and Space Science	
HS-ESS1	Earth's Place in the Universe	
HS-ESS1-1	Develop a model based on evidence to illustrate the life span of the sun and the role of nuclear fusion in the sun's core to release energy that eventually reaches Earth in the form of radiation. (SEP: 2; DCI: ESS1.A, PS3.D; CCC: Scale/Prop)	2
HS-ESS1-2	Construct an explanation of the Big Bang Theory based on astronomical evidence of light spectra, motion of distant galaxies, and composition of matter in the universe. (SEP: 6; DCI: PS4.B, ESS1.A; CCC: Energy/Matter, Technology)	2
HS-ESS1-3	Communicate scientific ideas about the way stars, over their life cycle, produce elements. (SEP: 8; DCI: ESS1.A; CCC: Energy/Matter)	2
HS-ESS1-4	Use mathematical or computational representations to predict the motion of orbiting objects in the solar system. (SEP: 5; DCI: ESS1.B, ESS1.A; CCC: Scale/Prop, Technology)	2
HS-ESS1-5	Evaluate evidence of the past and current movements of continental and oceanic crust and the theory of plate tectonics to explain the ages of crustal rocks. (SEP: 7; DCI: ESS1.C, ESS2.B, PS1.C; CCC: Patterns)	2
HS-ESS1-6	Apply scientific reasoning and evidence from ancient Earth materials, meteorites, and other planetary surfaces to construct an account of Earth's formation and early history. (SEP: 6; DCI: ESS1.C, PS1.C; CCC: Stability/Change)	2
HS-ESS2	Earth's Systems	
HS-ESS2-1	Analyze geoscience data to make the claim that one change to Earth's surface can create feedback that cause changes to other Earth systems. (SEP: 2; DCI: ESS2.A, ESS2.B; CCC: Stability/Change)	3
HS-ESS2-2	Develop a model based on evidence of Earth's interior to describe the cycling of matter by thermal convection. (SEP: 4; DCI: ESS2.A, ESS2.D; CCC: Stability/Change, Technology)	2
HS-ESS2-3	Use a model to describe how variations in the flow of energy into and out of Earth's systems result in changes in climate. (SEP: 2; DCI: ESS2.A, ESS2.B, PS4.A; CCC: Energy/Matter, Technology)	2
HS-ESS2-4	Plan and carry out an investigation of the properties of water and its effects on Earth materials and surface processes. (SEP: 2; DCI: ESS1.B, ESS2.A, ESS2.D; CCC: Cause/Effect)	4

HS-ESS3	Earth and Human Activity	
HS-ESS3-1	Construct an explanation based on evidence for how the availability of natural resources, occurrence of natural hazards, and changes in climate have influenced human activity. (SEP: 6; DCI: ESS3.A, ESS3.B ; CCC: Cause/Effect, Technology)	2
HS-ESS3-2	Evaluate competing design solutions for developing, managing, and utilizing energy and mineral resources based on cost-benefit ratios.* (SEP: 7; DCI: ESS3.A, ETS1.B; CCC: Technology)	3
HS-ESS3-3	Create a computational simulation to illustrate the relationships among management of natural resources, the sustainability of human populations, and biodiversity. (SEP: 5; DCI: ESS3.C; CCC: Stability/Change, Technology)	3
HS-ESS3-4	Evaluate or refine a technological solution that reduces impacts of human activities on natural systems.* (SEP: 6; DCI: ESS3.C; ETS1.B; CCC: Stability/Change, Technology)	3
HS-ESS3-5	Analyze geoscience data and the results from global climate models to make an evidence-based forecast of the current rate of global or regional climate change and associated future impacts to Earth systems. (SEP: 4; DCI: ESS3.D; CCC: Stability/Change)	4
HS-ESS3-6	Use a computational representation to illustrate the relationships among Earth systems and how those relationships are being modified due to human activity. (SEP: 5; DCI: ESS2.D, ESS3.D; CCC: Systems)	2

WebbAlign™

WCEPS
Wisconsin Center for
Education Products & Services

Appendix B

Item-Level Content Analysis Data for SDSA Items Reviewed in June 2022 by Batch

January, 2023

Brief Explanation of Data Included in Appendix B

Table 1 (SDSA Grade x Batch x)
The DOK value for each assessment item given by each reviewer. The intraclass correlation for the group of reviewers is given on the last row.

Table 2 (SDSA Grade x Batch x)
The DOK level and standard code assigned by each reviewer for each item.

Table 3 (SDSA Grade x Batch x)
This lists for each standard all of the items coded by the group of reviewers as corresponding to the standard. The number of reviewers who coded the item is given in parentheses.

Table 4 (SDSA Grade x Batch x)
This list for each item all of the standards coded by the group of reviewers as corresponding to the item. The number of reviewers who coded the standard is given in after the colon.

Table 5 (SDSA Grade x Batch x)
This table can be used to compare approximately the DOK level of a standard to the average DOK level of the items reviewers assigned to the standard. This table is helpful to identify items with a lower DOK level that should be replaced by an item with a higher DOK level to improve the Depth-of-Knowledge Consistency. The DOK listed in the table for each item is generally the mode DOK for that item.

Grade 5 Batch 48

Table 1 (SDSA Grade 5 Batch 48). *Depth-of-Knowledge Levels by Item and Reviewers' Intraclass Correlation*

Item	Reviewer 1	Reviewer 2	Reviewer 3	Reviewer 4	Reviewer 5	Reviewer 6
1	2	2	2	2	2	2
2	2	2	2	2	2	2
3	2	2	3	2	2	2
4	2	2	3	2	2	2
5	2	3	2	2	2	2
6	2	3	2	2	2	2
7	2	2	2	2	2	2
8	2	2	2	2	2	2
9	2	2	2	2	2	2
10	2	2	2	2	2	2
11	2	2	2	2	2	2
12	2	2	2	2	2	2
13	2	2	2	2	2	2
14	2	2	2	2	2	2
15	2	2	2	2	2	2
16	2	3	2	2	2	2
17	2	2	3	2	2	2
18	2	2	2	2	2	2
19	2	2	2	2	2	2

Intraclass correlation - -.0641
Pairwise Comparison - 0.89

Table 2 (SDSA Grade 5 Batch 48). *DOK Levels and Objectives Code by Each Reviewer*

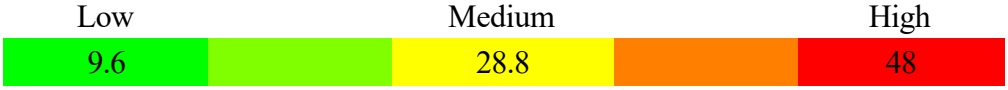
Item	DOK	Obj	S1 Obj	S2 Obj	DOK	Obj	S1 Obj	S2 Obj	DOK	Obj	S1 Obj	S2 Obj	DOK	Obj	S1 Obj	S2 Obj	DOK	Obj	S1 Obj	S2 Obj	DOK	Obj	S1 Obj	S2 Obj
1	2	5-LS2-1			2	5-LS2-1			2	5-LS2-1			2	5-LS2-1			2	5-LS2-1			2	5-LS2-1		
2	2	3-LS1-1			2	3-LS1-1			2	3-LS1-1			2	3-LS1-1			2	3-LS1-1			2	3-LS1-1		
3	2	3-LS1-1			2	3-LS1-1			3	3-LS1-1			2	3-LS1-1			2	3-LS1-1			2	3-LS1-1		
4	2	3-LS4-4			2	3-LS4-4			3	3-LS4-4			2	3-LS4-4			2	3-LS4-4			2	3-LS4-4		
5	2	3-LS4-3			3	3-LS4-3			2	3-LS4-3			2	3-LS4-3			2	3-LS4-3			2	3-LS4-3		
6	2	3-LS4-2			3	3-LS4-2			2	3-LS4-2			2	3-LS4-2			2	3-LS4-2			2	3-LS4-2		
7	2	4-LS1-1			2	4-LS1-1			2	4-LS1-1			2	4-LS1-1			2	4-LS1-1			2	4-LS1-1		
8	2	4-LS1-1			2	4-LS1-1			2	4-LS1-1			2	4-LS1-1			2	4-LS1-1			2	4-LS1-1		
9	2	4-LS1-1			2	4-LS1-1			2	4-LS1-1			2	4-LS1-1			2	4-LS1-1			2	4-LS1-1		
10	2	3-LS4-2			2	3-LS4-2			2	3-LS4-2			2	3-LS4-2			2	3-LS4-2			2	3-LS4-2		
11	2	4-LS1-1			2	4-LS1-1			2	4-LS1-1			2	4-LS1-1			2	4-LS1-1			2	4-LS1-1		
12	2	4-LS1-1			2	4-LS1-1			2	4-LS1-1			2	4-LS1-1			2	4-LS1-1			2	4-LS1-1		
13	2	3-LS4-2			2	3-LS4-2			2	3-LS4-2			2	3-LS4-2			2	3-LS4-2			2	3-LS4-2		
14	2	4-LS1-1			2	4-LS1-1			2	4-LS1-1			2	4-LS1-1			2	4-LS1-1			2	4-LS1-1		
15	2	3-LS4-2			2	3-LS4-2			2	3-LS4-2			2	3-LS4-2			2	3-LS4-2			2	3-LS4-2		
16	2	3-LS2-1			3	3-LS2-1			2	3-LS2-1			2	3-LS2-1			2	3-LS2-1			2	3-LS2-1		
17	2	3-LS4-4			2	3-LS4-4			3	3-LS4-4			2	3-LS4-4			2	3-LS4-4			2	3-LS4-4		
18	2	3-LS3-1			2	3-LS3-1			2	3-LS3-1			2	3-LS3-1			2	3-LS3-1			2	3-LS3-1		
19	2	4-LS1-1			2	4-LS1-1			2	4-LS1-1			2	4-LS1-1			2	4-LS1-1			2	4-LS1-1		
Objective Pairwise Comparison: 1																								
Standard Pairwise Comparison: 1																								

Table 3 (SDSA Grade 5 Batch 48). *Number of Reviewers Coding an Item by Objective (Item Number: Number of Reviewers)*

Low			Medium				High
0			28.8				48
LS							
LS1							
3-LS1-1	2(12)	3(6)					
4-LS1-1	7(12)	8(12)	9(6)	11(12)	12(6)	14(12)	19(12)
4-LS1-2							
5-LS1-1							
LS2							
3-LS2-1	16(48*)						
5-LS2-1	1(6)						
LS3							
3-LS3-1	18(6)						
3-LS3-2							
LS4							
3-LS4-1							
3-LS4-2	6(48)	10(6)	13(6)	15(6)			
3-LS4-3	5(36)						
3-LS4-4	4(6)	17(6)					

**GLOBAL: high numbers reflect weighting of items*

Table 4 (SDSA Grade 5 Batch 48). *Number of Reviewers Coding an Objective by Item (Objective: Number of Reviewers)*



1 37-	5-LS2-1:6
2 618-3013	3-LS1-1:12
3 519-3508	3-LS1-1:6
4 522-17404	3-LS4-4:6
5 544-17410	3-LS4-3:36
6 660-17445	3-LS4-2:48
7 453-17474	4-LS1-1:12
8 456-17613	4-LS1-1:12
9 528-17614	4-LS1-1:6
10 457-17615	3-LS4-2:6
11 483-17743	4-LS1-1:12
12 499-17824	4-LS1-1:6
13 500-17829	3-LS4-2:6
14 680-17831	4-LS1-1:12
15 504-17838	3-LS4-2:6
16 540-18027	3-LS2-1:48
17 636-18088	3-LS4-4:6
18 651-18374	3-LS3-1:6
19 653-18428	4-LS1-1:12

Table 5 (SDSA Grade 5 Batch 48). *Assessment Item DOK vs Consensus DOK (Item Number: Number of Reviewers [Average DOK])*

Low DOK		Matched DOK		High DOK

LS							
LS1							
3-LS1-1: [2]	2:(12)[2]	3:(6)[2]					
4-LS1-1: [2]	7:(12)[2]	8:(12)[2]	9:(6)[2]	11:(12)[2]	12:(6)[2]	14:(12)[2]	19:(12)[2]
4-LS1-2							
5-LS1-1							
LS2							
3-LS2-1: [2]	16:(48)[2]						
5-LS2-1: [2]	1:(6)[2]						
LS3							
3-LS3-1: [2]	18:(6)[2]						
3-LS3-2							
LS4							
3-LS4-1							
3-LS4-2: [2]	6:(48)[2]	10:(6)[2]	13:(6)[2]	15:(6)[2]			
3-LS4-3: [2]	5:(36)[2]						
3-LS4-4: [3]	4:(6)[2]	17:(6)[2]					

Grade 5 Batch 49

Table 1 (SDSA Grade 5 Batch 49). *Depth-of-Knowledge Levels by Item and Reviewers' Intraclass Correlation*

Item	Reviewer 1	Reviewer 2	Reviewer 3	Reviewer 4	Reviewer 5	Reviewer 6
1	2	2	2	2	2	2
2	3	3	2	2	2	2
3	2	2	2	2	2	2
4	2	2	2	2	2	2
5	2	3	2	2	2	2
6	2	2	2	2	2	2
7	2	2	3	2	2	2
8	2	3	3	2	2	2
9	2	2	3	2	2	2
10	2	2	2	2	2	2
11	2	2	3	3	3	2
12	2	2	2	2	3	2
13	2	2	2	2	2	2
14	2	2	2	2	2	2
15	2	3	2	2	2	2
16	2	2	2	2	2	2
17	2	2	3	2	2	2
18	2	2	2	2	2	2
19	2	2	2	2	2	2
20	2	2	2	2	2	2
21	2	3	2	2	2	2

Intraclass correlation - .3939
Pairwise Comparison - 0.81

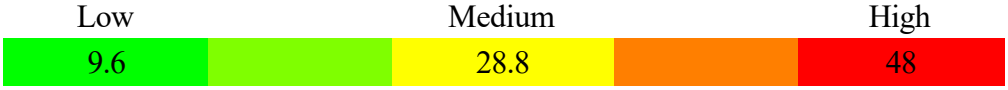
Table 2 (SDSA Grade 5 Batch 49). *DOK Levels and Objectives Code by Each Reviewer*

Item	DOK	Obj	S1 Obj	S2 Obj	DOK	Obj	S1 Obj	S2 Obj	DOK	Obj	S1 Obj	S2 Obj	DOK	Obj	S1 Obj	S2 Obj	DOK	Obj	S1 Obj	S2 Obj	DOK	Obj	S1 Obj	S2 Obj
1	2	4-PS3-3			2	4-PS3-3			2	4-PS3-3			2	4-PS3-3			2	4-PS3-3			2	4-PS3-3		
2	3	5-PS1-1			3	5-PS1-1			2	5-PS1-1			2	5-PS1-1			2	5-PS1-1			2	5-PS1-1		
3	2	3-PS2-1			2	3-PS2-1			2	3-PS2-1			2	3-PS2-1			2	3-PS2-1			2	3-PS2-1		
4	2	5-PS1-1			2	5-PS1-1			2	5-PS1-1			2	5-PS1-1			2	5-PS1-1			2	5-PS1-1		
5	2	5-PS1-3			3	5-PS1-3			2	5-PS1-3			2	5-PS1-3			2	5-PS1-3			2	5-PS1-3		
6	2	5-PS2-1			2	5-PS2-1			2	5-PS2-1			2	5-PS2-1			2	5-PS2-1			2	5-PS2-1		
7	2	3-PS2-4			2	3-PS2-4			3	3-PS2-4			2	3-PS2-4			2	3-PS2-4			2	3-PS2-4		
8	2	5-PS1-4			3	5-PS1-4			3	5-PS1-4			2	5-PS1-4			2	5-PS1-4			2	5-PS1-4		
9	2	4-PS4-3			2	4-PS4-3			3	4-PS4-3			2	4-PS4-3			2	4-PS4-3			2	4-PS4-3		
10	2	4-PS3-2			2	4-PS3-2			2	4-PS3-2			2	4-PS3-2			2	4-PS3-2			2	4-PS3-2		
11	2	3-PS2-1			2	3-PS2-1			3	3-PS2-1			3	3-PS2-1			3	3-PS2-1			2	3-PS2-1		
12	2	4-PS3-2			2	4-PS3-2			2	4-PS3-2			2	4-PS3-2			3	4-PS3-2			2	4-PS3-2		
13	2	4-PS3-1			2	4-PS3-1			2	4-PS3-1			2	4-PS3-1			2	4-PS3-1			2	4-PS3-1		
14	2	4-PS3-4			2	4-PS3-4			2	4-PS3-4			2	4-PS3-4			2	4-PS3-4			2	4-PS3-4		
15	2	4-PS4-1			3	4-PS4-1			2	4-PS4-1			2	4-PS4-1			2	4-PS4-1			2	4-PS4-1		
16	2	4-PS3-2			2	4-PS3-2			2	4-PS3-2			2	4-PS3-2			2	4-PS3-2			2	4-PS3-2		
17	2	4-PS3-1			2	4-PS3-1			3	4-PS3-1			2	4-PS3-1			2	4-PS3-1			2	4-PS3-1		
18	2	3-PS2-1			2	3-PS2-1			2	3-PS2-1			2	3-PS2-1			2	3-PS2-1			2	3-PS2-1		
19	2	5-PS1-3			2	5-PS1-3			2	5-PS1-3			2	5-PS1-3			2	5-PS1-3			2	5-PS1-3		
20	2	4-PS3-4			2	4-PS3-4			2	4-PS3-4			2	4-PS3-4			2	4-PS3-4			2	4-PS3-4		
21	2	4-PS3-4			3	4-PS3-4			2	4-PS3-4			2	4-PS3-4			2	4-PS3-4			2	4-PS3-4		
Objective Pairwise Comparison: 1																								
Standard Pairwise Comparison: 1																								

Table 3 (SDSA Grade 5 Batch 49). *Number of Reviewers Coding an Item by Objective (Item Number: Number of Reviewers)*

Low		Medium		High
0		28.8		48
PS				
PS2				
3-PS2-1	3(6)	11(12)	18(6)	
3-PS2-2				
3-PS2-3				
3-PS2-4	7(6)			
5-PS2-1	6(6)			
PS3				
4-PS3-1	13(12)	17(12)		
4-PS3-2	10(12)	12(6)	16(6)	
4-PS3-3	1(6)			
4-PS3-4	14(12)	20(6)	21(18)	
5-PS3-1				
PS4				
4-PS4-1	15(48)			
4-PS4-2				
4-PS4-3	9(6)			
PS1				
5-PS1-1	2(42)	4(6)		
5-PS1-2				
5-PS1-3	5(48)	19(6)		
5-PS1-4	8(18)			

Table 4 (SDSA Grade 5 Batch 49). *Number of Reviewers Coding an Objective by Item (Objective: Number of Reviewers)*



1 28-	4-PS3-3:6
2 657-1122	5-PS1-1:42
3 520-3520	3-PS2-1:6
4 521-17401	5-PS1-1:6
5 523-17415	5-PS1-3:48
6 524-17456	5-PS2-1:6
7 525-17461	3-PS2-4:6
8 526-17462	5-PS1-4:18
9 527-17463	4-PS4-3:6
10 455-17612	4-PS3-2:12
11 529-17639	3-PS2-1:12
12 464-17648	4-PS3-2:6
13 467-17662	4-PS3-1:12
14 677-17677	4-PS3-4:12
15 531-17712	4-PS4-1:48
16 498-17815	4-PS3-2:6
17 507-17844	4-PS3-1:12
18 511-17851	3-PS2-1:6
19 534-17902	5-PS1-3:6
20 634-18066	4-PS3-4:6
21 675-18268	4-PS3-4:18

Table 5 (SDSA Grade 5 Batch 49). *Assessment Item DOK vs Consensus DOK (Item Number: Number of Reviewers [Average DOK])*

Low DOK		Matched DOK		High DOK

PS			
PS2			
3-PS2-1: [4]	3:(6)[2]	11:(12)[2]	18:(6)[2]
3-PS2-2			
3-PS2-3			
3-PS2-4: [3]	7:(6)[2]		
5-PS2-1: [2]	6:(6)[2]		
PS3			
4-PS3-1: [2]	13:(12)[2]	17:(12)[2]	
4-PS3-2: [2]	10:(12)[2]	12:(6)[2]	16:(6)[2]
4-PS3-3: [2]	1:(6)[2]		
4-PS3-4: [4]	14:(12)[2]	20:(6)[2]	21:(18)[2]
5-PS3-1			
PS4			
4-PS4-1: [2]	15:(48)[2]		
4-PS4-2			
4-PS4-3: [3]	9:(6)[2]		
PS1			
5-PS1-1: [2]	2:(42)[2]	4:(6)[2]	
5-PS1-2			
5-PS1-3: [2]	5:(48)[2]	19:(6)[2]	
5-PS1-4: [3]	8:(18)[2]		

Grade 5 Batch 50

Table 1 (SDSA Grade 5 Batch 50). *Depth-of-Knowledge Levels by Item and Reviewers' Intraclass Correlation*

Item	Reviewer 1	Reviewer 2	Reviewer 3	Reviewer 4	Reviewer 5	Reviewer 6
1	2	2	2	2	2	2
2	2	2	2	2	2	2
3	2	2	2	2	2	2
4	2	3	2	2	2	2
5	2	2	2	2	2	2
6	2	2	3	2	2	2
7	2	2	2	2	2	2
8	2	2	2	2	2	2
9	2	2	2	2	2	2
10	2	2	3	2	2	2
11	2	2	2	2	2	2
12	2	2	2	2	2	2
13	2	2	2	2	2	2
14	2	2	3	2	2	2
15	2	2	2	2	2	2
16	3	3	3	3	3	3
17	2	2	3	2	2	2
18	2	2	2	2	2	2
19	2	2	2	2	2	2
20	2	2	3	2	2	2
21	2	2	2	2	2	2

Intraclass correlation - .8848
Pairwise Comparison - 0.9

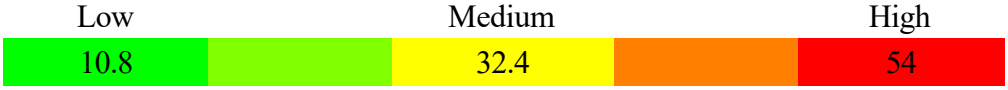
Table 2 (SDSA Grade 5 Batch 50). *DOK Levels and Objectives Code by Each Reviewer*

Item	*DOK*	*Obj*	*S1 Obj*	*S2 Obj*	*DOK*	*Obj*	*S1 Obj*	*S2 Obj*	*DOK*	*Obj*	*S1 Obj*	*S2 Obj*	*DOK*	*Obj*	*S1 Obj*	*S2 Obj*	*DOK*	*Obj*	*S1 Obj*	*S2 Obj*	*DOK*	*Obj*	*S1 Obj*	*S2 Obj*
1	2	4-ESS2-2			2	4-ESS2-2			2	4-ESS2-2			2	4-ESS2-2			2	4-ESS2-2			2	4-ESS2-2		
2	2	3-ESS2-2			2	3-ESS2-2			2	3-ESS2-2			2	3-ESS2-2			2	3-ESS2-2			2	3-ESS2-2		
3	2	3-ESS2-2			2	3-ESS2-2			2	3-ESS2-2			2	3-ESS2-2			2	3-ESS2-2			2	3-ESS2-2		
4	2	5-ESS1-1			3	5-ESS1-1			2	5-ESS1-1			2	5-ESS1-1			2	5-ESS1-1			2	5-ESS1-1		
5	2	4-ESS2-2			2	4-ESS2-2			2	4-ESS2-2			2	4-ESS2-2			2	4-ESS2-2			2	4-ESS2-2		
6	2	5-ESS3-1			2	5-ESS3-1			3	5-ESS3-1			2	5-ESS3-1			2	5-ESS3-1			2	5-ESS3-1		
7	2	4-ESS3-1			2	4-ESS3-1			2	4-ESS3-1			2	4-ESS3-1			2	4-ESS3-1			2	4-ESS3-1		
8	2	5-ESS1-2			2	5-ESS1-2			2	5-ESS1-2			2	5-ESS1-2			2	5-ESS1-2			2	5-ESS1-2		
9	2	5-ESS1-2			2	5-ESS1-2			2	5-ESS1-2			2	5-ESS1-2			2	5-ESS1-2			2	5-ESS1-2		
10	2	4-ESS3-2			2	4-ESS3-2			3	4-ESS3-2			2	4-ESS3-2			2	4-ESS3-2			2	4-ESS3-2		
11	2	4-ESS3-1			2	4-ESS3-1			2	4-ESS3-1			2	4-ESS3-1			2	4-ESS3-1			2	4-ESS3-1		
12	2	4-ESS3-1			2	4-ESS3-1			2	4-ESS3-1			2	4-ESS3-1			2	4-ESS3-1			2	4-ESS3-1		
13	2	4-ESS3-1			2	4-ESS3-1			2	4-ESS3-1			2	4-ESS3-1			2	4-ESS3-1			2	4-ESS3-1		
14	2	4-ESS3-2			2	4-ESS3-2			3	4-ESS3-2			2	4-ESS3-2			2	4-ESS3-2			2	4-ESS3-2		
15	2	5-ESS1-2			2	5-ESS1-2			2	5-ESS1-2			2	5-ESS1-2			2	5-ESS1-2			2	5-ESS1-2		
16	3	3-ESS3-1			3	3-ESS3-1			3	3-ESS3-1			3	3-ESS3-1			3	3-ESS3-1			3	3-ESS3-1		
17	2	4-ESS3-2			2	4-ESS3-2			3	4-ESS3-2			2	4-ESS3-2			2	4-ESS3-2			2	4-ESS3-2		
18	2	5-ESS1-2			2	5-ESS1-2			2	5-ESS1-2			2	5-ESS1-2			2	5-ESS1-2			2	5-ESS1-2		
19	2	3-ESS3-1			2	3-ESS3-1			2	3-ESS3-1			2	3-ESS3-1			2	3-ESS3-1			2	3-ESS3-1		
20	2	4-ESS3-2			2	4-ESS3-2			3	4-ESS3-2			2	4-ESS3-2			2	4-ESS3-2			2	4-ESS3-2		
21	2	4-ESS2-1			2	4-ESS2-1			2	4-ESS2-1			2	4-ESS2-1			2	4-ESS2-1			2	4-ESS2-1		
Objective Pairwise Comparison: 1																								
Standard Pairwise Comparison: 1																								

Table 3 (SDSA Grade 5 Batch 50). *Number of Reviewers Coding an Item by Objective (Item Number: Number of Reviewers)*

Low		Medium		High
0		32.4		54
ESS				
ESS2				
3-ESS2-1				
3-ESS2-2	2(6)	3(12)		
4-ESS2-1	21(6)			
4-ESS2-2	1(6)	5(6)		
5-ESS2-1				
5-ESS2-2				
ESS3				
3-ESS3-1	16(54)	19(12)		
4-ESS3-1	7(6)	11(12)	12(6)	13(6)
4-ESS3-2	10(6)	14(6)	17(6)	20(6)
5-ESS3-1	6(12)			
ESS1				
4-ESS1-1				
5-ESS1-1	4(42)			
5-ESS1-2	8(6)	9(6)	15(6)	18(6)

Table 4 (SDSA Grade 5 Batch 50). *Number of Reviewers Coding an Objective by Item (Objective: Number of Reviewers)*



1 26-	4-ESS2-2:6
2 35-	3-ESS2-2:6
3 60-	3-ESS2-2:12
4 541-1184	5-ESS1-1:42
5 515-2761	4-ESS2-2:6
6 517-3051	5-ESS3-1:12
7 518-3102	4-ESS3-1:6
8 459-17629	5-ESS1-2:6
9 460-17630	5-ESS1-2:6
10 530-17679	4-ESS3-2:6
11 681-17835	4-ESS3-1:12
12 510-17850	4-ESS3-1:6
13 512-17861	4-ESS3-1:6
14 545-17873	4-ESS3-2:6
15 533-17877	5-ESS1-2:6
16 536-17922	3-ESS3-1:54
17 635-18079	4-ESS3-2:6
18 674-18080	5-ESS1-2:6
19 684-18087	3-ESS3-1:12
20 685-18107	4-ESS3-2:6
21 640-18170	4-ESS2-1:6

Table 5 (SDSA Grade 5 Batch 50). *Assessment Item DOK vs Consensus DOK (Item Number: Number of Reviewers [Average DOK])*

Low DOK		Matched DOK		High DOK

ESS				
ESS2				
3-ESS2-1				
3-ESS2-2: [2]	2:(6)[2]	3:(12)[2]		
4-ESS2-1: [2]	21:(6)[2]			
4-ESS2-2: [2]	1:(6)[2]	5:(6)[2]		
5-ESS2-1				
5-ESS2-2				
ESS3				
3-ESS3-1: [3]	16:(54)[3]	19:(12)[2]		
4-ESS3-1: [2]	7:(6)[2]	11:(12)[2]	12:(6)[2]	13:(6)[2]
4-ESS3-2: [3]	10:(6)[2]	14:(6)[2]	17:(6)[2]	20:(6)[2]
5-ESS3-1: [3]	6:(12)[2]			
ESS1				
4-ESS1-1				
5-ESS1-1: [2]	4:(42)[2]			
5-ESS1-2: [2]	8:(6)[2]	9:(6)[2]	15:(6)[2]	18:(6)[2]

Grade 8 Batch 51

Table 1 (SDSA Grade 8 Batch 51). *Depth-of-Knowledge Levels by Item and Reviewers Intraclass Correlation*

Item	Reviewer 1	Reviewer 2	Reviewer 3	Reviewer 4	Reviewer 5	Reviewer 6	Reviewer 7	Reviewer 8	Reviewer 9	Reviewer 10	Reviewer 11	Reviewer 12
1	2	2	2	2	2	2	2	2	2	2	2	2
2	2	2	2	2	2	2	2	2	2	2	3	2
3	2	2	2	2	2	2	2	2	3	2	2	2
4	3	2	2	2	3	2	3	2	2	2	3	2
5	2	2	2	2	2	2	2	2	2	2	2	2
6	2	2	2	2	2	2	2	2	2	2	2	2
7	2	2	2	2	2	2	2	2	2	2	2	2
8	2	2	2	2	2	2	3	2	2	2	2	2
9	2	3	3	3	2	2	1	2	2	2	2	2
10	2	2	2	2	2	2	3	2	2	2	2	2
11	2	2	2	2	2	2	2	2	2	2	2	2
12	2	2	2	2	2	2	2	2	2	2	2	2
13	2	2	2	2	2	2	3	2	2	2	2	2
14	2	2	2	2	2	2	2	2	2	2	2	2
15	2	2	2	2	2	2	2	2	2	2	2	2
16	2	2	2	2	2	2	2	2	2	2	2	2
17	2	2	2	2	2	2	2	2	2	2	2	2
18	2	2	2	2	2	2	3	2	2	2	2	2
19	2	2	2	2	2	2	2	2	2	2	2	2
20	3	2	3	2	2	2	2	2	3	2	3	3
21	2	2	2	2	2	2	2	2	2	2	2	2
22	2	2	2	2	2	2	3	2	2	2	2	2
23	2	3	2	2	2	2	2	2	2	2	2	2
24	3	2	3	3	2	3	2	2	2	2	3	3

Intraclass correlation - .7054
Pairwise Comparison - 0.86

Table 2 (SDSA Grade 8 Batch 51). *DOK Levels and Objectives Code by Each Reviewer*

SDSA Grade 8 Batch 51

Item	DOK	Obj	S1 Obj	S2 Obj	DOK	Obj	S1 Obj	S2 Obj	DOK	Obj	S1 Obj	S2 Obj	DOK	Obj	S1 Obj	S2 Obj	DOK	Obj	S1 Obj	S2 Obj	DOK	Obj	S1 Obj	S2 Obj	DOK	Obj	S1 Obj	S2 Obj	DOK	Obj	S1 Obj	S2 Obj	DOK	Obj	S1 Obj	S2 Obj	DOK	Obj	S1 Obj	S2 Obj	DOK	Obj	S1 Obj	S2 Obj	DOK	Obj	S1 Obj	S2 Obj	DOK	Obj	S1 Obj	S2 Obj
1	2	MS-LS1-3			2	MS-LS1-3			2	MS-LS1-3			2	MS-LS1-3			2	MS-LS1-3			2	MS-LS1-3			2	MS-LS1-3			2	MS-LS1-3			2	MS-LS1-3			2	MS-LS1-3			2	MS-LS1-3			2	MS-LS1-3			2	MS-LS1-3			2	MS-LS1-3		
2	2	MS-LS1-6			2	MS-LS1-6			2	MS-LS1-6			2	MS-LS1-6			2	MS-LS1-6			2	MS-LS1-6			2	MS-LS1-6			2	MS-LS1-6			2	MS-LS1-6			3	MS-LS1-6			2	MS-LS1-6			2	MS-LS1-6			2	MS-LS1-6		
3	2	MS-LS2-2			2	MS-LS2-2			2	MS-LS2-2			2	MS-LS2-2			2	MS-LS2-2			2	MS-LS2-2			2	MS-LS2-2			2	MS-LS2-2			3	MS-LS2-2			2	MS-LS2-2			2	MS-LS2-2			2	MS-LS2-2			2	MS-LS2-2		
4	3	MS-LS4-6			2	MS-LS4-6			2	MS-LS4-6			2	MS-LS4-6			3	MS-LS4-6			2	MS-LS4-6			3	MS-LS4-6			2	MS-LS4-6			2	MS-LS4-6			3	MS-LS4-6			2	MS-LS4-6			2	MS-LS4-6			2	MS-LS4-6		
5	2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4		
6	2	MS-LS4-1			2	MS-LS4-1			2	MS-LS4-1			2	MS-LS4-1			2	MS-LS4-1			2	MS-LS4-1			2	MS-LS4-1			2	MS-LS4-1			2	MS-LS4-1			2	MS-LS4-1			2	MS-LS4-1			2	MS-LS4-1			2	MS-LS4-1		
7	2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4		
8	2	MS-LS1-6			2	MS-LS1-6			2	MS-LS1-6			2	MS-LS1-6			2	MS-LS1-6			2	MS-LS2-3			3	MS-LS1-6			2	MS-LS1-6			2	MS-LS1-6			2	MS-LS2-3			2	MS-LS1-6			2	MS-LS1-6			2	MS-LS1-6		
9	2	MS-LS4-5			3	MS-LS4-5			3	MS-LS4-5			3	MS-LS4-5			2	MS-LS4-5			2	MS-LS4-5			1	MS-LS4-5			2	MS-LS4-5			2	MS-LS4-5			2	MS-LS4-5			2	MS-LS4-5			2	MS-LS4-5			2	MS-LS4-5		
10	2	MS-LS1-2			2	MS-LS1-2			2	MS-LS1-2			2	MS-LS1-2			2	MS-LS1-2			2	MS-LS1-2			3	MS-LS1-2			2	MS-LS1-2			2	MS-LS1-2			2	MS-LS1-2			2	MS-LS1-2			2	MS-LS1-2			2	MS-LS1-2		
11	2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-1			2	MS-LS2-1			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4		
12	2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4		
13	2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			3	MS-LS2-2			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4		
14	2	MS-LS4-1			2	MS-LS4-1			2	MS-LS4-1			2	MS-LS4-1			2	MS-LS4-1			2	MS-LS4-1			2	MS-LS2-4			2	MS-LS4-1			2	MS-LS4-1			2	MS-LS4-1			2	MS-LS4-1			2	MS-LS4-1			2	MS-LS4-1		
15	2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4		
16	2	MS-LS2-3			2	MS-LS2-3			2	MS-LS2-3			2	MS-LS2-3			2	MS-LS2-3			2	MS-LS2-3			2	MS-LS2-3			2	MS-LS2-3			2	MS-LS2-3			2	MS-LS2-3			2	MS-LS2-3			2	MS-LS2-3			2	MS-LS2-3		
17	2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-1			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4			2	MS-LS2-4		
18	2	MS-LS2-2			2	MS-LS2-2			2	MS-LS2-2			2	MS-LS2-2			2	MS-LS2-2			2	MS-LS2-2			3	MS-LS2-2			2	MS-LS2-2			2	MS-LS2-2			2	MS-LS2-2			2	MS-LS2-2			2	MS-LS2-2			2	MS-LS2-2		
19	2	MS-LS4-1			2	MS-LS4-1			2	MS-LS4-1			2	MS-LS4-1			2	MS-LS4-1			2	MS-LS4-1			2	MS-LS4-1			2	MS-LS4-1			2	MS-LS4-1			2	MS-LS4-1			2	MS-LS4-1			2	MS-LS4-1			2	MS-LS4-1		
20	3	MS-LS2-5			2	MS-LS2-5			3	MS-LS2-5			2	MS-LS2-5			2	MS-LS2-5			2	MS-LS2-5			2	MS-LS2-5			2	MS-LS2-5			3	MS-LS2-5			2	MS-LS2-5			3	MS-LS2-5			3	MS-LS2-5			3	MS-LS2-5		
21	2	MS-LS1-6			2	MS-LS1-6			2	MS-LS1-6			2	MS-LS1-6			2	MS-LS1-6			2	MS-LS1-6			2	MS-LS1-6			2	MS-LS1-6			2	MS-LS1-6			2	MS-LS1-6			2	MS-LS1-6			2	MS-LS1-6			2	MS-LS1-6		

22	2	MS-LS3-1			2	MS-LS3-1			2	MS-LS3-1			2	MS-LS3-1			2	MS-LS3-1			2	MS-LS3-1			3	MS-LS3-1			2	MS-LS3-1			2	MS-LS3-1			2	MS-LS3-1			2	MS-LS3-1			2	MS-LS3-1		
23	2	MS-LS3-2			3	MS-LS3-2			2	MS-LS3-2			2	MS-LS3-2			2	MS-LS3-2			2	MS-LS3-2			2	MS-LS3-2			2	MS-LS3-2			2	MS-LS3-2			2	MS-LS3-2			2	MS-LS3-2			2	MS-LS3-2		
24	3	MS-LS2-5			2	MS-LS2-5			3	MS-LS2-5			3	MS-LS2-5			2	MS-LS2-5			3	MS-LS2-5			2	MS-LS2-5			2	MS-LS2-5			2	MS-LS2-5			2	MS-LS2-5			3	MS-LS2-5			3	MS-LS2-5		

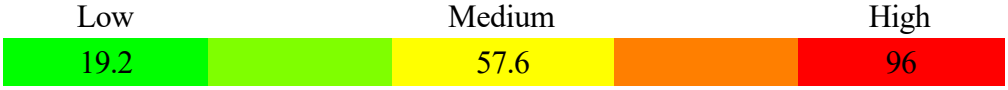
Objective Pairwise Comparison: 0.95

Standard Pairwise Comparison: 1

Table 3 (SDSA Grade 8 Batch 51). *Number of Reviewers Coding an Item by Objective (Item Number: Number of Reviewers)*

Low			Medium				High	
0			57.6				96	
MS-LS								
MS-LS1								
MS-LS1-1								
MS-LS1-2	10(24)							
MS-LS1-3	1(24)							
MS-LS1-4								
MS-LS1-5								
MS-LS1-6	2(12)	8(20)	21(12)					
MS-LS1-7								
MS-LS2								
MS-LS2-1	17(1)	11(4)						
MS-LS2-2	13(1)	3(12)	18(24)					
MS-LS2-3	8(4)	16(24)						
MS-LS2-4	14(1)	5(24)	7(24)	11(20)	12(12)	13(11)	15(12)	17(11)
MS-LS2-5	20(12)	24(24)						
MS-LS3								
MS-LS3-1	22(36)							
MS-LS3-2	23(36)							
MS-LS4								
MS-LS4-1	6(12)	14(11)	19(12)					
MS-LS4-2								
MS-LS4-4								
MS-LS4-5	9(12)							
MS-LS4-6	4(96)							

Table 4 (SDSA Grade 8 Batch 51). *Number of Reviewers Coding an Objective by Item (Objective: Number of Reviewers)*



1 126-	MS-LS1-3:24	
2 173-	MS-LS1-6:12	
3 571-2970	MS-LS2-2:12	
4 577-3514	MS-LS4-6:96	
5 579-17413	MS-LS2-4:24	
6 581-17430	MS-LS4-1:12	
7 590-17432	MS-LS2-4:24	
8 621-17454	MS-LS1-6:20	MS-LS2-3:4
9 622-17459	MS-LS4-5:12	
10 661-17601	MS-LS1-2:24	
11 454-17607	MS-LS2-1:4	MS-LS2-4:20
12 462-17635	MS-LS2-4:12	
13 463-17644	MS-LS2-2:1	MS-LS2-4:11
14 475-17718	MS-LS2-4:1	MS-LS4-1:11
15 486-17785	MS-LS2-4:12	
16 678-17790	MS-LS2-3:24	
17 490-17794	MS-LS2-1:1	MS-LS2-4:11
18 491-17798	MS-LS2-2:24	
19 494-17806	MS-LS4-1:12	
20 625-17865	MS-LS2-5:12	
21 629-18017	MS-LS1-6:12	
22 671-18040	MS-LS3-1:36	
23 644-18194	MS-LS3-2:36	
24 650-18365	MS-LS2-5:24	

Table 5 (SDSA Grade 8 Batch 51). *Assessment Item DOK vs Consensus DOK (Item Number: Number of Reviewers [Average DOK])*

Low DOK		Matched DOK		High DOK

MS-LS								
MS-LS1								
MS-LS1-1								
MS-LS1-2: [2]	10:(24)[2]							
MS-LS1-3: [2]	1:(24)[2]							
MS-LS1-4								
MS-LS1-5								
MS-LS1-6: [2]	2:(12)[2]	8:(20)[2]	21:(12)[2]					
MS-LS1-7								
MS-LS2								
MS-LS2-1: [2]	11:(4)[2]	17:(1)[2]						
MS-LS2-2: [2]	3:(12)[2]	13:(1)[3]	18:(24)[2]					
MS-LS2-3: [2]	8:(4)[2]	16:(24)[2]						
MS-LS2-4: [2]	5:(24)[2]	7:(24)[2]	11:(20)[2]	12:(12)[2]	13:(11)[2]	14:(1)[2]	15:(12)[2]	17:(11)[2]
MS-LS2-5: [3]	20:(12)[2]	24:(24)[2]						
MS-LS3								
MS-LS3-1: [2]	22:(36)[2]							
MS-LS3-2: [2]	23:(36)[2]							
MS-LS4								
MS-LS4-1: [2]	6:(12)[2]	14:(11)[2]	19:(12)[2]					
MS-LS4-2								
MS-LS4-4								
MS-LS4-5: [3]	9:(12)[2]							
MS-LS4-6: [2]	4:(96)[2]							

Grade 8 Batch 52

Table 1 (SDSA Grade 8 Batch 52). *Depth-of-Knowledge Levels by Item and Reviewers Intraclass Correlation*

Item	Reviewer 1	Reviewer 2	Reviewer 3	Reviewer 4	Reviewer 5	Reviewer 6	Reviewer 7	Reviewer 8	Reviewer 9	Reviewer 10	Reviewer 11	Reviewer 12
1	2	2	3	2	3	2	3	2	2	2	3	3
2	2	2	2	2	2	2	2	2	2	2	2	2
3	2	2	2	2	2	2	2	2	2	2	2	2
4	1	2	2	2	2	2	2	1	2	2	2	1
5	2	2	2	2	2	2	2	2	2	2	2	2
6	2	2	2	2	2	2	3	3	2	2	2	2
7	2	2	2	2	2	2	2	2	2	2	2	2
8	2	2	3	2	2	2	3	2	2	2	2	2
9	2	3	2	2	2	2	3	2	2	2	2	2
10	3	3	4	2	3	3	3	3	2	2	3	2
11	2	2	4	2	2	2	2	2	2	2	2	3
12	2	2	2	2	2	2	2	2	2	2	2	2
13	3	2	3	2	2	2	3	3	3	2	2	2
14	2	2	4	2	2	3	3	2	3	3	2	2
15	3	2	3	2	3	3	3	2	3	2	2	2
16	3	3	3	3	3	2	3	3	3	3	3	2

Intraclass correlation - .872
Pairwise Comparison - 0.7

Table 2 (SDSA Grade 8 Batch 52). *DOK Levels and Objectives Code by Each Reviewer*

Item	DOK	Obj	S1 Obj	S2 Obj	DOK	Obj	S1 Obj	S2 Obj	DOK	Obj	S1 Obj	S2 Obj	DOK	Obj	S1 Obj	S2 Obj	DOK	Obj	S1 Obj	S2 Obj	DOK	Obj	S1 Obj	S2 Obj	DOK	Obj	S1 Obj	S2 Obj	DOK	Obj	S1 Obj	S2 Obj	DOK	Obj	S1 Obj	S2 Obj	DOK	Obj	S1 Obj	S2 Obj	DOK	Obj	S1 Obj	S2 Obj	DOK	Obj	S1 Obj	S2 Obj
1	2	MS-PS2-2			2	MS-PS2-2			3	MS-PS2-2			2	MS-PS2-2			3	MS-PS2-2			2	MS-PS2-2			3	MS-PS2-2			2	MS-PS2-2			2	MS-PS2-2			2	MS-PS2-2			3	MS-PS2-2			3	MS-PS2-2		
2	2	MS-PS4-1			2	MS-PS4-1			2	MS-PS4-1			2	MS-PS4-1			2	MS-PS4-1			2	MS-PS4-1			2	MS-PS4-1			2	MS-PS4-1			2	MS-PS4-1			2	MS-PS4-1			2	MS-PS4-1			2	MS-PS4-1		
3	2	MS-PS3-1			2	MS-PS3-1			2	MS-PS3-1			2	MS-PS3-1			2	MS-PS3-1			2	MS-PS3-1			2	MS-PS3-1			2	MS-PS3-1			2	MS-PS3-1			2	MS-PS3-1			2	MS-PS3-1			2	MS-PS3-1		
4	1	MS-PS1-5			2	MS-PS1-5			2	MS-PS1-5			2	MS-PS1-5			2	MS-PS1-5			2	MS-PS1-5			2	MS-PS1-5			1	MS-PS1-5			2	MS-PS1-5			2	MS-PS1-5			2	MS-PS1-5			1	MS-PS1-5		
5	2	MS-PS1-4			2	MS-PS1-4			2	MS-PS1-4			2	MS-PS1-4			2	MS-PS1-4			2	MS-PS1-4			2	MS-PS1-4			2	MS-PS1-4			2	MS-PS1-4			2	MS-PS1-4			2	MS-PS1-4			2	MS-PS1-4		
6	2	MS-PS4-2			2	MS-PS4-2			2	MS-PS4-2			2	MS-PS4-2			2	MS-PS4-2			2	MS-PS4-2			3	MS-PS4-2			3	MS-PS4-2			2	MS-PS4-2			2	MS-PS4-2			2	MS-PS4-2			2	MS-PS4-2		
7	2	MS-PS4-2			2	MS-PS4-2			2	MS-PS4-2			2	MS-PS4-2			2	MS-PS4-2			2	MS-PS4-2			2	MS-PS4-2			2	MS-PS4-2			2	MS-PS4-2			2	MS-PS4-2			2	MS-PS4-2			2	MS-PS4-2		
8	2	MS-PS1-3			2	MS-PS1-3			3	MS-PS1-3			2	MS-PS1-3			2	MS-PS1-3			2	MS-PS1-3			3	MS-PS1-3			2	MS-PS1-3			2	MS-PS1-3			2	MS-PS1-3			2	MS-PS1-3			2	MS-PS1-3		
9	2	MS-PS4-2			3	MS-PS4-2			2	MS-PS4-2			2	MS-PS4-2			2	MS-PS4-2			2	MS-PS4-2			3	MS-PS4-2			2	MS-PS4-2			2	MS-PS4-2			2	MS-PS4-2			2	MS-PS4-2			2	MS-PS4-2		
10	3	MS-PS3-3			3	MS-PS3-3			4	MS-PS3-3			2	MS-PS3-3			3	MS-PS3-3			3	MS-PS3-3			3	MS-PS3-3			3	MS-PS3-3			2	MS-PS3-3			2	MS-PS3-3			3	MS-PS3-3			2	MS-PS3-3		
11	2	MS-PS3-3			2	MS-PS3-3			4	MS-PS3-3			2	MS-PS3-3			2	MS-PS3-3			2	MS-PS3-3			2	MS-PS3-3			2	MS-PS3-3			2	MS-PS3-3			2	MS-PS3-3			2	MS-PS3-3			3	MS-PS3-3		
12	2	MS-PS4-3			2	MS-PS4-3			2	MS-PS4-3			2	MS-PS4-3			2	MS-PS4-3			2	MS-PS4-3			2	MS-PS4-3			2	MS-PS4-3			2	MS-PS4-3			2	MS-PS4-3			2	MS-PS4-3			2	MS-PS4-3		
13	3	MS-PS3-5			2	MS-PS3-4			3	MS-PS3-4			2	MS-PS3-4			2	MS-PS3-4			2	MS-PS3-4			3	MS-PS3-4			3	MS-PS3-4			3	MS-PS3-3			2	MS-PS3-4			2	MS-PS3-3			2	MS-PS3-4		
14	2	MS-PS3-3			2	MS-PS3-3			4	MS-PS3-3			2	MS-PS3-3			2	MS-PS3-3			3	MS-PS3-3			3	MS-PS3-3			2	MS-PS3-3			3	MS-PS3-3			3	MS-PS3-3			2	MS-PS3-3			2	MS-PS3-3		
15	3	MS-PS2-2			2	MS-PS2-2			3	MS-PS2-2			2	MS-PS2-2			3	MS-PS2-2			3	MS-PS2-2			3	MS-PS2-2			2	MS-PS2-2			3	MS-PS2-2			2	MS-PS2-2			2	MS-PS2-2			2	MS-PS2-2		
16	3	MS-PS4-1			3	MS-PS4-1			3	MS-PS4-1			3	MS-PS4-1			3	MS-PS4-1			2	MS-PS4-1			3	MS-PS4-1			3	MS-PS4-1			3	MS-PS4-1			3	MS-PS4-1			3	MS-PS4-1			2	MS-PS4-1		

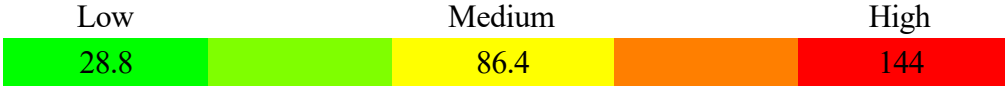
Objective Pairwise Comparison: 0.97

Standard Pairwise Comparison: 1

Table 3 (SDSA Grade 8 Batch 52). *Number of Reviewers Coding an Item by Objective (Item Number: Number of Reviewers)*

Low		Medium		High
0		86.4		144
MS-PS				
MS-PS1				
MS-PS1-1				
MS-PS1-2				
MS-PS1-3	8(12)			
MS-PS1-4	5(12)			
MS-PS1-5	4(12)			
MS-PS1-6				
MS-PS2				
MS-PS2-1				
MS-PS2-2	1(24)	15(12)		
MS-PS2-3				
MS-PS2-4				
MS-PS2-5				
MS-PS3				
MS-PS3-1	3(12)			
MS-PS3-2				
MS-PS3-3	13(2)	10(24)	11(12)	14(12)
MS-PS3-4	13(9)			
MS-PS3-5	13(1)			
MS-PS4				
MS-PS4-1	2(12)	16(144)		
MS-PS4-2	6(12)	7(12)	9(36)	
MS-PS4-3	12(12)			

Table 4 (SDSA Grade 8 Batch 52). *Number of Reviewers Coding an Objective by Item (Objective: Number of Reviewers)*



1 72-	MS-PS2-2:24		
2 129-	MS-PS4-1:12		
3 658-2947	MS-PS3-1:12		
4 572-3031	MS-PS1-5:12		
5 578-3516	MS-PS1-4:12		
6 466-17658	MS-PS4-2:12		
7 468-17666	MS-PS4-2:12		
8 584-17716	MS-PS1-3:12		
9 484-17744	MS-PS4-2:36		
10 626-17870	MS-PS3-3:24		
11 627-17999	MS-PS3-3:12		
12 665-18001	MS-PS4-3:12		
13 628-18016	MS-PS3-3:2	MS-PS3-4:9	MS-PS3-5:1
14 633-18062	MS-PS3-3:12		
15 641-18178	MS-PS2-2:12		
16 686-18866	MS-PS4-1:144		

Table 5 (SDSA Grade 8 Batch 52). *Assessment Item DOK vs Consensus DOK (Item Number: Number of Reviewers [Average DOK])*

Low DOK		Matched DOK		High DOK

MS-PS				
MS-PS1				
MS-PS1-1				
MS-PS1-2				
MS-PS1-3: [3]	8:(12)[2]			
MS-PS1-4: [2]	5:(12)[2]			
MS-PS1-5: [2]	4:(12)[2]			
MS-PS1-6				
MS-PS2				
MS-PS2-1				
MS-PS2-2: [3]	1:(24)[2]	15:(12)[2]		
MS-PS2-3				
MS-PS2-4				
MS-PS2-5				
MS-PS3				
MS-PS3-1: [2]	3:(12)[2]			
MS-PS3-2				
MS-PS3-3: [4]	10:(24)[3]	11:(12)[2]	13:(2)[2]	14:(12)[2]
MS-PS3-4: [3]	13:(9)[2]			
MS-PS3-5: [2]	13:(1)[3]			
MS-PS4				
MS-PS4-1: [2]	2:(12)[2]	16:(144)[3]		
MS-PS4-2: [2]	6:(12)[2]	7:(12)[2]	9:(36)[2]	
MS-PS4-3: [2]	12:(12)[2]			

Grade 8 Batch 53

Table 1 (SDSA Grade 8 Batch 53). *Depth-of-Knowledge Levels by Item and Reviewers Intraclass Correlation*

Item	Reviewer 1	Reviewer 2	Reviewer 3	Reviewer 4	Reviewer 5	Reviewer 6	Reviewer 7	Reviewer 8	Reviewer 9	Reviewer 10	Reviewer 11	Reviewer 12
1	2	2	3	2	2	2	2	2	2	2	2	2
2	2	2	2	2	2	2	2	2	2	2	2	2
3	2	3	3	2	2	2	3	2	2	3	2	2
4	3	3	3	2	2	3	3	3	3	3	3	3
5	2	2	2	2	2	2	3	2	2	2	2	2
6	2	2	2	2	2	2	3	2	2	2	2	2
7	2	2	2	2	2	2	3	2	2	2	2	2
8	2	2	2	2	2	2	2	2	2	2	2	2
9	2	2	2	2	2	2	2	2	2	2	2	2
10	2	2	2	2	2	2	3	2	2	2	2	2
11	2	2	2	2	2	2	2	2	2	2	2	2
12	2	2	2	2	2	2	2	2	2	2	2	2
13	2	2	2	2	2	2	3	2	2	2	2	2
14	2	2	2	2	2	2	2	2	2	2	2	2
15	2	2	2	2	2	2	2	2	2	2	2	2

Intraclass correlation - .9239
Pairwise Comparison - 0.88

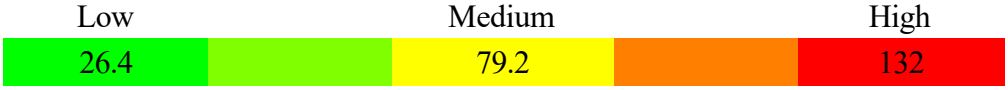
Table 2 (SDSA Grade 8 Batch 53). *DOK Levels and Objectives Code by Each Reviewer*

Item	DOK	Obj	S1 Obj	S2 Obj	DOK	Obj	S1 Obj	S2 Obj	DOK	Obj	S1 Obj	S2 Obj	DOK	Obj	S1 Obj	S2 Obj	DOK	Obj	S1 Obj	S2 Obj	DOK	Obj	1 bj	2 bj	OK	bj	j	j	DK	j	j	j	K			K											
1	2	MS-ESS3-4			2	MS-ESS3-4			3	MS-ESS3-4			2	MS-ESS3-4			2	MS-ESS3-4				MS-ESS3-				MS-ESS3-				S-S3-				-S3-				3-				3-							
2	2	MS-ESS3-1			2	MS-ESS3-1			2	MS-ESS3-1			2	MS-ESS3-1			2	MS-ESS3-				MS-ESS3-				S-SS3-				S-S3-				-S3-				3-				3-							
3	2	MS-ESS2-6			3	MS-ESS2-6			3	MS-ESS2-6			2	MS-ESS2-6			2	MS-ESS2-				MS-ESS2-				S-SS2-				S-S2-				-S2-				2-				2-							
4	3	MS-ESS3-4			3	MS-ESS3-4			3	MS-ESS3-4			2	MS-ESS3-4			2	MS-ESS3-				MS-ESS3-				S-SS3-				S-S3-				-S3-				3-				3-							
5	2	MS-ESS1-1			2	MS-ESS1-1			2	MS-ESS1-1			2	MS-ESS1-1			2	MS-ESS1-				MS-ESS1-				S-SS1-				S-S1-				-S1-				1-				1-							
6	2	MS-ESS1-1			2	MS-ESS1-1			2	MS-ESS1-1			2	MS-ESS1-1			2	MS-ESS1-				MS-ESS1-				S-SS1-				S-S1-				-S1-				1-				1-							
7	2	MS-ESS1-2			2	MS-ESS1-2			2	MS-ESS1-2			2	MS-ESS1-2			2	MS-ESS1-				MS-ESS1-				S-SS1-				S-S1-				-S1-				1-				1-							
8	2	MS-ESS1-2			2	MS-ESS1-2			2	MS-ESS1-2			2	MS-ESS1-2			2	MS-ESS1-				MS-ESS1-				S-SS1-				S-S1-				-S1-				1-				1-							
9	2	MS-ESS1-2			2	MS-ESS1-2			2	MS-ESS1-2			2	MS-ESS1-2			2	MS-ESS1-				MS-ESS1-				S-SS1-				S-S1-				-S1-				1-				1-							
10	2	MS-ESS2-2			2	MS-ESS2-2			2	MS-ESS2-2			2	MS-ESS2-2			2	MS-ESS2-				MS-ESS2-				S-SS2-				S-S2-				-S2-				2-				2-							
11	2	MS-ESS1-1			2	MS-ESS1-1			2	MS-ESS1-1			2	MS-ESS1-1			2	MS-ESS1-				MS-ESS1-				S-SS1-				S-S1-				-S1-				1-				1-							
12	2	MS-ESS1-3			2	MS-ESS1-3			2	MS-ESS1-3			2	MS-ESS1-3			2	MS-ESS1-				MS-ESS1-				S-SS1-				S-S1-				-S1-				1-				1-							
13	2	MS-ESS1-1			2	MS-ESS1-1			2	MS-ESS1-1			2	MS-ESS1-1			2	MS-ESS1-				MS-ESS1-				S-SS1-				S-S1-				-S1-				1-				1-							
14	2	MS-ESS1-1			2	MS-ESS1-1			2	MS-ESS1-1			2	MS-ESS1-1			2	MS-ESS1-				MS-ESS1-				S-SS1-				S-S1-				-S1-				1-				1-							
15	2	MS-ESS2-1			2	MS-ESS2-1			2	MS-ESS2-1			2	MS-ESS2-1			2	MS-ESS2-				MS-ESS2-				S-SS2-				S-S2-				-S2-				2-				2-							
Objective Pairwise Comparison: 1																																																
Standard Pairwise Comparison: 1																																																

Table 3 (SDSA Grade 8 Batch 53). *Number of Reviewers Coding an Item by Objective (Item Number: Number of Reviewers)*

Low		Medium			High
0		79.2			132
MS-ESS					
MS-ESS1					
MS-ESS1-1	5(12)	6(36)	11(12)	13(12)	14(12)
MS-ESS1-2	7(12)	8(12)	9(24)		
MS-ESS1-3	12(24)				
MS-ESS2					
MS-ESS2-1	15(12)				
MS-ESS2-2	10(36)				
MS-ESS2-3					
MS-ESS2-4					
MS-ESS2-5					
MS-ESS2-6	3(24)				
MS-ESS3					
MS-ESS3-1	2(24)				
MS-ESS3-2					
MS-ESS3-3					
MS-ESS3-4	1(12)	4(132)			
MS-ESS3-5					

Table 4 (SDSA Grade 8 Batch 53). *Number of Reviewers Coding an Objective by Item (Objective: Number of Reviewers)*



1 139-	MS-ESS3-4:12
2 385-2965	MS-ESS3-1:24
3 573-3049	MS-ESS2-6:24
4 659-3084	MS-ESS3-4:132
5 469-17668	MS-ESS1-1:12
6 592-17671	MS-ESS1-1:36
7 472-17709	MS-ESS1-2:12
8 473-17710	MS-ESS1-2:12
9 474-17711	MS-ESS1-2:24
10 583-17715	MS-ESS2-2:36
11 482-17741	MS-ESS1-1:12
12 487-17787	MS-ESS1-3:24
13 502-17836	MS-ESS1-1:12
14 682-17878	MS-ESS1-1:12
15 672-18055	MS-ESS2-1:12

Table 5 (SDSA Grade 8 Batch 53). *Assessment Item DOK vs Consensus DOK (Item Number: Number of Reviewers [Average DOK])*

Low DOK		Matched DOK		High DOK

MS-ESS					
MS-ESS1					
MS-ESS1-1: [2]	5:(12)[2]	6:(36)[2]	11:(12)[2]	13:(12)[2]	14:(12)[2]
MS-ESS1-2: [2]	7:(12)[2]	8:(12)[2]	9:(24)[2]		
MS-ESS1-3: [2]	12:(24)[2]				
MS-ESS2					
MS-ESS2-1: [2]	15:(12)[2]				
MS-ESS2-2: [2]	10:(36)[2]				
MS-ESS2-3					
MS-ESS2-4					
MS-ESS2-5					
MS-ESS2-6: [2]	3:(24)[2]				
MS-ESS3					
MS-ESS3-1: [2]	2:(24)[2]				
MS-ESS3-2					
MS-ESS3-3					
MS-ESS3-4: [3]	1:(12)[2]	4:(132)[3]			
MS-ESS3-5					

Grade 11 Batch 54

Table 1 (SDSA Grade 11 Batch 54). *Depth-of-Knowledge Levels by Item and Reviewers Intraclass Correlation*

Item	Reviewer 1	Reviewer 2	Reviewer 3	Reviewer 4	Reviewer 5	Reviewer 6
1	2	2	2	2	2	2
2	2	3	2	2	2	2
3	2	2	2	2	2	2
4	2	2	2	2	2	2
5	2	2	2	2	2	2
6	2	2	2	2	2	2
7	2	3	2	2	2	2
8	2	3	2	2	2	2
9	2	2	2	2	2	2
10	2	3	2	2	2	2
11	2	3	2	2	2	2
12	1	2	2	2	2	2
13	2	2	2	2	2	2
14	2	2	2	2	2	2
15	2	3	2	2	2	2
16	2	3	2	2	2	2
17	2	3	2	3	2	2
18	2	3	2	3	2	2
19	2	3	2	2	2	2
20	2	2	2	3	2	2
21	2	2	2	2	2	2
22	3	2	2	2	2	2
23	2	3	2	2	2	2
24	2	3	2	2	2	2
25	3	3	2	2	2	2
26	2	3	2	2	2	2
27	2	2	2	3	3	2
28	2	2	2	2	2	2
29	2	3	2	2	2	2
30	2	2	2	3	2	1
31	2	2	2	2	2	2
32	2	2	2	3	2	2
33	2	3	2	2	2	2
34	2	3	3	3	3	3
35	2	3	3	3	3	3
36	2	3	2	2	2	2
37	2	2	2	2	2	2

Intraclass correlation - .6718
Pairwise Comparison - 0.75

Table 2 (SDSA Grade 11 Batch 54). *DOK Levels and Objectives Code by Each Reviewer*

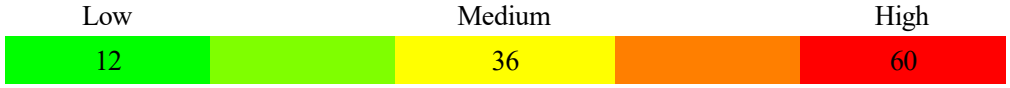
Item	DOK	Obj	S1 Obj	S2 Obj	DOK	Obj	S1 Obj	S2 Obj	DOK	Obj	S1 Obj	S2 Obj	DOK	Obj	S1 Obj	S2 Obj	DOK	Obj	S1 Obj	S2 Obj	DOK	Obj	S1 Obj	S2 Obj
1	2	HS-LS2-2			2	HS-LS2-2			2	HS-LS2-2			2	HS-LS2-2			2	HS-LS2-2			2	HS-LS2-2		
2	2	HS-LS1-4			3	HS-LS1-4			2	HS-LS1-4			2	HS-LS1-4			2	HS-LS1-4			2	HS-LS1-4		
3	2	HS-LS3-1			2	HS-LS3-1			2	HS-LS3-1			2	HS-LS3-1			2	HS-LS3-1			2	HS-LS3-1		
4	2	HS-LS1-5			2	HS-LS1-5			2	HS-LS1-5			2	HS-LS1-5			2	HS-LS1-5			2	HS-LS1-5		
5	2	HS-LS3-2			2	HS-LS3-2			2	HS-LS3-2			2	HS-LS3-2			2	HS-LS3-2			2	HS-LS3-2		
6	2	HS-LS1-1			2	HS-LS1-1			2	HS-LS1-1			2	HS-LS1-1			2	HS-LS1-1			2	HS-LS1-1		
7	2	HS-LS1-2			3	HS-LS1-2			2	HS-LS1-2			2	HS-LS1-2			2	HS-LS1-2			2	HS-LS1-2		
8	2	HS-LS3-3			3	HS-LS3-3			2	HS-LS3-3			2	HS-LS3-3			2	HS-LS3-3			2	HS-LS3-3		
9	2	HS-LS1-1			2	HS-LS1-1			2	HS-LS1-1			2	HS-LS1-1			2	HS-LS1-1			2	HS-LS1-1		
10	2	HS-LS4-5			3	HS-LS4-5			2	HS-LS4-5			2	HS-LS4-5			2	HS-LS4-5			2	HS-LS4-5		
11	2	HS-LS1-7			3	HS-LS1-7			2	HS-LS1-7			2	HS-LS1-7			2	HS-LS1-7			2	HS-LS1-7		
12	1	HS-LS2-4			2	HS-LS2-4			2	HS-LS2-4			2	HS-LS2-4			2	HS-LS2-4			2	HS-LS2-4		
13	2	HS-LS1-2			2	HS-LS1-2			2	HS-LS1-2			2	HS-LS1-2			2	HS-LS1-2			2	HS-LS1-2		
14	2	HS-LS1-6			2	HS-LS1-6			2	HS-LS1-6			2	HS-LS1-6			2	HS-LS1-6			2	HS-LS1-6		
15	2	HS-LS4-3			3	HS-LS4-3			2	HS-LS4-3			2	HS-LS4-3			2	HS-LS4-3			2	HS-LS4-3		
16	2	HS-LS4-1			3	HS-LS4-1			2	HS-LS4-1			2	HS-LS4-1			2	HS-LS4-1			2	HS-LS4-1		
17	2	HS-LS4-5			3	HS-LS4-5			2	HS-LS4-5			3	HS-LS4-5			2	HS-LS4-5			2	HS-LS4-3		
18	2	HS-LS1-4			3	HS-LS1-4			2	HS-LS1-4			3	HS-LS1-4			2	HS-LS1-4			2	HS-LS1-4		
19	2	HS-LS4-1			3	HS-LS4-1			2	HS-LS1-1			2	HS-LS4-1			2	HS-LS4-1			2	HS-LS4-1		
20	2	HS-LS1-4			2	HS-LS1-4			2	HS-LS1-4			3	HS-LS1-4			2	HS-LS1-4			2	HS-LS1-4		
21	2	HS-LS2-7			2	HS-LS2-7			2	HS-LS2-7			2	HS-LS2-7			2	HS-LS2-7			2	HS-LS2-7		
22	3	HS-LS2-7			2	HS-LS2-7			2	HS-LS2-7			2	HS-LS2-7			2	HS-LS2-7			2	HS-LS2-7		
23	2	HS-LS4-1			3	HS-LS4-1			2	HS-LS4-1			2	HS-LS4-1			2	HS-LS4-1			2	HS-LS4-1		
24	2	HS-LS3-2			3	HS-LS3-2			2	HS-LS3-2			2	HS-LS3-2			2	HS-LS3-2			2	HS-LS3-2		
25	3	HS-LS2-7			3	HS-LS2-7			2	HS-LS2-7			2	HS-LS2-7			2	HS-LS2-7			2	HS-LS2-7		
26	2	HS-LS2-6			3	HS-LS2-6			2	HS-LS2-6			2	HS-LS2-6			2	HS-LS2-6			2	HS-LS2-2		
27	2	HS-LS4-5			2	HS-LS4-5			2	HS-LS4-5			3	HS-LS4-5			3	HS-LS4-5			2	HS-LS2-1		
28	2	HS-LS4-1			2	HS-LS4-1			2	HS-LS4-1			2	HS-LS4-1			2	HS-LS4-1			2	HS-LS4-1		
29	2	HS-LS4-5			3	HS-LS4-5			2	HS-LS4-5			2	HS-LS4-5			2	HS-LS4-5			2	HS-LS4-5		
30	2	HS-LS2-4			2	HS-LS2-4			2	HS-LS2-4			3	HS-LS2-4			2	HS-LS2-4			1	HS-LS2-4		
31	2	HS-LS1-7			2	HS-LS1-7			2	HS-LS1-7			2	HS-LS1-7			2	HS-LS1-7			2	HS-LS1-7		
32	2	HS-LS2-4			2	HS-LS2-4			2	HS-LS2-4			3	HS-LS2-4			2	HS-LS2-4			2	HS-LS2-4		
33	2	HS-LS2-5			3	HS-LS2-5			2	HS-LS2-5			2	HS-LS2-5			2	HS-LS2-5			2	HS-LS2-5	HS-LS2-5	
34	2	HS-LS2-5			3	HS-LS2-5			3	HS-LS2-5			3	HS-LS2-5			3	HS-LS2-5			3	HS-LS2-5		
35	2	HS-LS3-2			3	HS-LS3-2			3	HS-LS3-2			3	HS-LS3-2			3	HS-LS3-2			3	HS-LS3-2		
36	2	HS-LS1-2			3	HS-LS1-2			2	HS-LS1-2			2	HS-LS1-2			2	HS-LS1-2			2	HS-LS1-2		
37	2	HS-LS2-2			2	HS-LS2-2			2	HS-LS2-2			2	HS-LS2-2			2	HS-LS2-2			2	HS-LS2-2		
Objective Pairwise Comparison: 0.96																								
Standard Pairwise Comparison: 1																								

Table 3 (SDSA Grade 11 Batch 54). *Number of Reviewers Coding an Item by Objective (Item Number: Number of Reviewers)*

Low		Medium		High
0		36		60

HS-LS				
HS-LS1				
HS-LS1-1	19(1)	6(6)	9(12)	
HS-LS1-2	7(6)	13(6)	36(24)	
HS-LS1-3				
HS-LS1-4	2(60)	18(24)	20(18)	
HS-LS1-5	4(6)			
HS-LS1-6	14(6)			
HS-LS1-7	11(24)	31(6)		
HS-LS2				
HS-LS2-1	27(1)			
HS-LS2-2	26(2)	1(12)	37(12)	
HS-LS2-3				
HS-LS2-4	12(12)	30(12)	32(12)	
HS-LS2-5	34(54)	33(28)		
HS-LS2-6	26(10)			
HS-LS2-7	21(6)	22(6)	25(12)	
HS-LS2-8				
HS-LS3				
HS-LS3-1	3(6)			
HS-LS3-2	5(6)	24(12)	35(60)	
HS-LS3-3	8(6)			
HS-LS4				
HS-LS4-1	16(12)	19(5)	23(12)	28(6)
HS-LS4-2				
HS-LS4-3	17(1)	15(6)		
HS-LS4-4				
HS-LS4-5	10(6)	17(5)	27(5)	29(12)
HS-LS4-6				

Table 4 (SDSA Grade 11 Batch 54). *Number of Reviewers Coding an Objective by Item (Objective: Number of Reviewers)*



1 201-	HS-LS2-2:12	
2 563-3034	HS-LS1-4:60	
3 566-3064	HS-LS3-1:6	
4 548-3070	HS-LS1-5:6	
5 451-17421	HS-LS3-2:6	
6 569-17424	HS-LS1-1:6	
7 551-17425	HS-LS1-2:6	
8 452-17431	HS-LS3-3:6	
9 552-17440	HS-LS1-1:12	
10 553-17451	HS-LS4-5:6	
11 458-17618	HS-LS1-7:24	
12 465-17654	HS-LS2-4:12	
13 470-17696	HS-LS1-2:6	
14 476-17721	HS-LS1-6:6	
15 477-17726	HS-LS4-3:6	
16 478-17727	HS-LS4-1:12	
17 480-17729	HS-LS4-3:1	HS-LS4-5:5
18 485-17748	HS-LS1-4:24	
19 488-17788	HS-LS1-1:1	HS-LS4-1:5
20 489-17789	HS-LS1-4:18	
21 492-17800	HS-LS2-7:6	
22 594-17804	HS-LS2-7:6	
23 495-17807	HS-LS4-1:12	
24 496-17811	HS-LS3-2:12	
25 497-17814	HS-LS2-7:12	
26 624-17822	HS-LS2-2:2	HS-LS2-6:10
27 501-17833	HS-LS2-1:1	HS-LS4-5:5
28 503-17837	HS-LS4-1:6	
29 505-17839	HS-LS4-5:12	
30 506-17842	HS-LS2-4:12	
31 508-17845	HS-LS1-7:6	
32 509-17847	HS-LS2-4:12	
33 557-17909	HS-LS2-5:24	
34 559-17932	HS-LS2-5:54	
35 668-18034	HS-LS3-2:60	
36 631-18036	HS-LS1-2:24	
37 638-18140	HS-LS2-2:12	

Table 5 (SDSA Grade 11 Batch 54). *Assessment Item DOK vs Consensus DOK (Item Number: Number of Reviewers [Average DOK])*

Low DOK		Matched DOK		High DOK

HS-LS				
HS-LS1				
HS-LS1-1: [2]	6:(6)[2]	9:(12)[2]	19:(1)[2]	
HS-LS1-2: [2]	7:(6)[2]	13:(6)[2]	36:(24)[2]	
HS-LS1-3				
HS-LS1-4: [2]	2:(60)[2]	18:(24)[2]	20:(18)[2]	
HS-LS1-5: [2]	4:(6)[2]			
HS-LS1-6: [2]	14:(6)[2]			
HS-LS1-7: [2]	11:(24)[2]	31:(6)[2]		
HS-LS2				
HS-LS2-1: [2]	27:(1)[2]			
HS-LS2-2: [2]	1:(12)[2]	26:(2)[2]	37:(12)[2]	
HS-LS2-3				
HS-LS2-4: [2]	12:(12)[2]	30:(12)[2]	32:(12)[2]	
HS-LS2-5: [2]	33:(24)[2]	34:(54)[3]		
HS-LS2-6: [3]	26:(10)[2]			
HS-LS2-7: [4]	21:(6)[2]	22:(6)[2]	25:(12)[2]	
HS-LS2-8				
HS-LS3				
HS-LS3-1: [2]	3:(6)[2]			
HS-LS3-2: [2]	5:(6)[2]	24:(12)[2]	35:(60)[3]	
HS-LS3-3: [2]	8:(6)[2]			
HS-LS4				
HS-LS4-1: [2]	16:(12)[2]	19:(5)[2]	23:(12)[2]	28:(6)[2]
HS-LS4-2				
HS-LS4-3: [2]	15:(6)[2]	17:(1)[2]		
HS-LS4-4				
HS-LS4-5: [3]	10:(6)[2]	17:(5)[2]	27:(5)[2]	29:(12)[2]
HS-LS4-6				

Grade 11 Batch 55

Table 1 (SDSA Grade 11 Batch 55). *Depth-of-Knowledge Levels by Item and Reviewers Intraclass Correlation*

Item	Reviewer 1	Reviewer 2	Reviewer 3	Reviewer 4	Reviewer 5	Reviewer 6
1	2	2	2	2	2	2
2	2	2	2	2	2	2
3	2	3	2	2	2	2
4	3	3	3	3	3	3
5	2	3	3	3	3	3
6	2	3	2	2	3	2
7	2	3	3	3	3	2
8	2	2	3	2	2	2
9	2	2	2	2	2	2
10	2	2	2	2	2	2
11	2	2	3	2	2	2
12	2	2	2	2	2	2
13	2	2	3	3	2	2

Intraclass correlation - .8818
Pairwise Comparison - 0.77

Table 2 (SDSA Grade 11 Batch 55). *DOK Levels and Objectives Code by Each Reviewer*

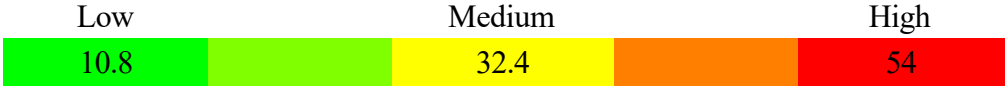
Item	DOK	Obj	S1 Obj	S2 Obj	DOK	Obj	S1 Obj	S2 Obj	DOK	Obj	S1 Obj	S2 Obj	DOK	Obj	S1 Obj	S2 Obj	DOK	Obj	S1 Obj	S2 Obj	DOK	Obj	S1 Obj	S2 Obj
1	2	HS-PS1-6			2	HS-PS1-6			2	HS-PS1-6			2	HS-PS1-6			2	HS-PS1-6			2	HS-PS1-6		
2	2	HS-PS2-4			2	HS-PS2-4			2	HS-PS2-4			2	HS-PS2-4			2	HS-PS2-4			2	HS-PS2-4		
3	2	HS-PS1-4			3	HS-PS1-4			2	HS-PS1-4			2	HS-PS1-4			2	HS-PS1-4			2	HS-PS1-4		
4	3	HS-PS3-3			3	HS-PS3-3			3	HS-PS3-3			3	HS-PS3-3			3	HS-PS3-3			3	HS-PS3-3		
5	2	HS-PS4-5	HS-PS4-5		3	HS-PS4-5			3	HS-PS4-5			3	HS-PS4-5			3	HS-PS4-5			3	HS-PS4-5		
6	2	HS-PS1-1			3	HS-PS1-1			2	HS-PS1-1			2	HS-PS1-1			3	HS-PS1-1			2	HS-PS1-1		
7	2	HS-PS1-5			3	HS-PS1-5			3	HS-PS1-5			3	HS-PS1-5			3	HS-PS1-5			2	HS-PS1-5		
8	2	HS-PS4-5			2	HS-PS4-5			3	HS-PS4-5			2	HS-PS4-5			2	HS-PS4-5			2	HS-PS4-5		
9	2	HS-PS3-4			2	HS-PS3-4			2	HS-PS3-4			2	HS-PS3-4			2	HS-PS3-4			2	HS-PS3-4		
10	2	HS-PS1-2			2	HS-PS1-2			2	HS-PS1-2			2	HS-PS1-2			2	HS-PS1-2			2	HS-PS1-2		
11	2	HS-PS4-1			2	HS-PS4-1			3	HS-PS4-1			2	HS-PS4-1			2	HS-PS4-1			2	HS-PS4-1		
12	2	HS-PS2-6			2	HS-PS2-6			2	HS-PS2-6			2	HS-PS2-6			2	HS-PS2-6			2	HS-PS2-6		
13	2	HS-PS4-1			2	HS-PS4-1			3	HS-PS4-1			3	HS-PS4-1			2	HS-PS4-1			2	HS-PS4-1		
Objective Pairwise Comparison: 1																								
Standard Pairwise Comparison: 1																								

Table 3 (SDSA Grade 11 Batch 55). *Number of Reviewers Coding an Item by Objective (Item Number: Number of Reviewers)*

Low		Medium		High
0		32.4		54

HS-PS		
HS-PS1		
HS-PS1-1	6(48)	
HS-PS1-2	10(6)	
HS-PS1-3		
HS-PS1-4	3(6)	
HS-PS1-5	7(42)	
HS-PS1-6	1(12)	
HS-PS1-7		
HS-PS1-8		
HS-PS2		
HS-PS2-1		
HS-PS2-2		
HS-PS2-3		
HS-PS2-4	2(6)	
HS-PS2-5		
HS-PS2-6	12(6)	
HS-PS3		
HS-PS3-1		
HS-PS3-2		
HS-PS3-3	4(54)	
HS-PS3-4	9(6)	
HS-PS3-5		
HS-PS4		
HS-PS4-1	11(6)	13(12)
HS-PS4-2		
HS-PS4-3		
HS-PS4-4		
HS-PS4-5	8(18)	5(49)

Table 4 (SDSA Grade 11 Batch 55). *Number of Reviewers Coding an Objective by Item (Objective: Number of Reviewers)*



1 180-	HS-PS1-6:12
2 199-	HS-PS2-4:6
3 547-2981	HS-PS1-4:6
4 564-3036	HS-PS3-3:54
5 567-3103	HS-PS4-5:42
6 620-17452	HS-PS1-1:48
7 558-17919	HS-PS1-5:42
8 670-18039	HS-PS4-5:18
9 632-18045	HS-PS3-4:6
10 643-18186	HS-PS1-2:6
11 645-18206	HS-PS4-1:6
12 646-18216	HS-PS2-6:6
13 654-18429	HS-PS4-1:12

Table 5 (SDSA Grade 11 Batch 55). *Assessment Item DOK vs Consensus DOK (Item Number: Number of Reviewers [Average DOK])*

Low DOK		Matched DOK		High DOK

HS-PS		
HS-PS1		
HS-PS1-1: [2]	6:(48)[2]	
HS-PS1-2: [2]	10:(6)[2]	
HS-PS1-3		
HS-PS1-4: [2]	3:(6)[2]	
HS-PS1-5: [2]	7:(42)[3]	
HS-PS1-6: [2]	1:(12)[2]	
HS-PS1-7		
HS-PS1-8		
HS-PS2		
HS-PS2-1		
HS-PS2-2		
HS-PS2-3		
HS-PS2-4: [2]	2:(6)[2]	
HS-PS2-5		
HS-PS2-6: [2]	12:(6)[2]	
HS-PS3		
HS-PS3-1		
HS-PS3-2		
HS-PS3-3: [4]	4:(54)[3]	
HS-PS3-4: [4]	9:(6)[2]	
HS-PS3-5		
HS-PS4		
HS-PS4-1: [2]	11:(6)[2]	13:(12)[2]
HS-PS4-2		
HS-PS4-3		
HS-PS4-4		
HS-PS4-5: [2]	5:(42)[3]	8:(18)[2]

Grade 11 Batch 56

Table 1 (SDSA Grade 11 Batch 56). *Depth-of-Knowledge Levels by Item and Reviewers Intraclass Correlation*

Item	Reviewer 1	Reviewer 2	Reviewer 3	Reviewer 4	Reviewer 5	Reviewer 6
1	2	2	2	2	2	2
2	2	2	2	2	2	2
3	2	2	2	2	2	2
4	2	3	3	3	3	3
5	3	2	3	3	3	3

Intraclass correlation - .9533
Pairwise Comparison - 0.87

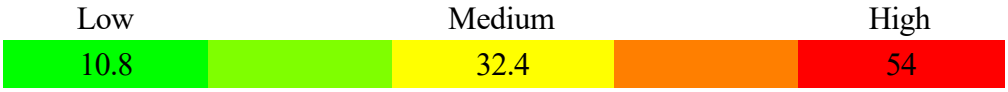
Table 2 (SDSA Grade 11 Batch 56). *DOK Levels and Objectives Code by Each Reviewer*

Item	DOK	Obj	S1 Obj	S2 Obj	DOK	Obj	S1 Obj	S2 Obj	DOK	Obj	S1 Obj	S2 Obj	DOK	Obj	S1 Obj	S2 Obj	DOK	Obj	S1 Obj	S2 Obj	DOK	Obj	S1 Obj	S2 Obj
1	2	HS-ESS3-6			2	HS-ESS3-6			2	HS-ESS3-6			2	HS-ESS3-6			2	HS-ESS3-6			2	HS-ESS3-6		
2	2	HS-ESS2-2			2	HS-ESS2-2			2	HS-ESS2-2			2	HS-ESS2-2			2	HS-ESS2-2			2	HS-ESS2-2		
3	2	HS-ESS2-3			2	HS-ESS2-3			2	HS-ESS2-3			2	HS-ESS2-3			2	HS-ESS2-3			2	HS-ESS2-3		
4	2	HS-ESS1-2			3	HS-ESS1-2			3	HS-ESS1-2			3	HS-ESS1-2			3	HS-ESS1-2			3	HS-ESS1-2		
5	3	HS-ESS3-4			2	HS-ESS3-4			3	HS-ESS3-4			3	HS-ESS3-4			3	HS-ESS3-4			3	HS-ESS3-4		
Objective Pairwise Comparison: 1																								
Standard Pairwise Comparison: 1																								

Table 3 (SDSA Grade 11 Batch 56). *Number of Reviewers Coding an Item by Objective (Item Number: Number of Reviewers)*

Low		Medium		High
0		32.4		54
HS-ESS				
HS-ESS1				
HS-ESS1-1				
HS-ESS1-2				4(48)
HS-ESS1-3				
HS-ESS1-4				
HS-ESS1-5				
HS-ESS1-6				
HS-ESS2				
HS-ESS2-1				
HS-ESS2-2				2(6)
HS-ESS2-3				3(6)
HS-ESS2-4				
HS-ESS3				
HS-ESS3-1				
HS-ESS3-2				
HS-ESS3-3				
HS-ESS3-4				5(54)
HS-ESS3-5				
HS-ESS3-6				1(6)

Table 4 (SDSA Grade 11 Batch 56). *Number of Reviewers Coding an Objective by Item (Objective: Number of Reviewers)*



1 216-	HS-ESS3-6:6
2 219-	HS-ESS2-2:6
3 565-3057	HS-ESS2-3:6
4 556-17897	HS-ESS1-2:48
5 666-18008	HS-ESS3-4:54

Table 5 (SDSA Grade 11 Batch 56). *Assessment Item DOK vs Consensus DOK (Item Number: Number of Reviewers [Average DOK])*

Low DOK		Matched DOK		High DOK

HS-ESS	
HS-ESS1	
HS-ESS1-1	
HS-ESS1-2: [2]	4:(48)[3]
HS-ESS1-3	
HS-ESS1-4	
HS-ESS1-5	
HS-ESS1-6	
HS-ESS2	
HS-ESS2-1	
HS-ESS2-2: [2]	2:(6)[2]
HS-ESS2-3: [2]	3:(6)[2]
HS-ESS2-4	
HS-ESS3	
HS-ESS3-1	
HS-ESS3-2	
HS-ESS3-3	
HS-ESS3-4: [3]	5:(54)[3]
HS-ESS3-5	
HS-ESS3-6: [2]	1:(6)[2]

WebbAlign™

WCEPS
Wisconsin Center for
Education Products & Services

Appendix C

Items Flagged for Review and Revision or Removal
Content Analysis of SDSA Items in June 2022

January, 2023

Table 1. Items flagged by reviewers for unmet expectations by grade band

Grades 5			
Item ID (Type)	**Primary Issue (PE)**	**Reason**	**Suggested Resolution**
371 (2938/1581) (Stand alone)	Scoring Assertions omit core part of DCI (3-ESS3-1)	PE states *Make a claim about the merit of a design solution that reduces the impacts of a weather-related hazard. [Clarification Statement: Examples of design solutions to weather-related hazards could include barriers to prevent flooding, wind resistant roofs, and lightning rods.]* In contrast, item focus is water conservation and Scoring Assertion states the item gives evidence that the student "…understands how to evaluate design solutions based on the impact they have on the environment." Reviewers commented that the item would need to incorporate the idea of a weather-related hazard to adequately address the internally coded PE. (1 scoring assertion)	Review and revise
355 (2908/1561) (Stand alone)	Rated Category of Engagement Level 1 (4-PS3-1)	Response requires recall of the relationship between two variables. (1 scoring assertion)	Remove or revise
379 (2950/1465) (Stand alone)	Scoring Assertions weakly address standard + concerns about clarity (5-ESS2-2)	This item was flagged in 2019 for revision due to weak connection of scoring assertions to the standard, intertwined with concern about item clarity. The item asks students to interpret information presented in a diagram and then to select the appropriate graphic representation of the same information. The correct answer requires the student to make an assumption that contrasts with the information presented in the diagram (which shows a change in the flow of fresh water but **no** change in the flow salt water). Weak connection to PE. (2 scoring assertions)	Review and revise
444 (3124/1474) (Stand alone)	Rated Category of Engagement Level 1 (5-ESS2-2)	Requires recall of relative amounts of salt and fresh water on Earth. (1 scoring assertion)	Remove or revise
37 (Stand alone)	Editorial error in Scoring Assertion (5-LS2-1)	Item asks students to demonstrate understanding of flow of energy. There is no reference to matter within item. In the scoring assertion, the word "matter" should be replaced with "energy." (1 scoring assertion)	Review and revise

Grades 8			
Item ID (Type)	**Issue (PE)**	**Reason**	**Suggested Resolution**
378 (2948/1523) (Stand alone)	Scoring Assertions too far from PE (MS-PS4-2)	A majority of panelists found that the scoring assertions reflected the student work, but only a minority agreed they adequately represent the PE. (2 scoring assertions)	Review and revise
375 (2944/1451) (Stand alone)	Rated Category of Engagement Level 1 (MS-PS1-1)	Panelists did not think this was answerable as written. Perhaps could revise prompt to specify "that distinguish the molecular composition…" however this does not resolve absence of phenomenon / DOK 1 (2 scoring assertions)	Remove or revise
395 (2994/1436) (Stand alone)	Rated Category of Engagement Level 1 (MS-LS1-7)	Requires recall of specific input and output molecules for cellular respiration. Reviewers also noted that he assessment boundary for the targeted standard specifies that an assessment of the standard should not include details about the chemical reactions. (3 scoring assertions)	Remove or revise
577 (3514) (Cluster)	Editorial Corrections Needed for Student Interaction (MS-LS4-6)	For Part A, Petal Length, directions and menu dropdown options do not correspond to the question / answer. There is a directive given for students to have a "workaround" to this mismatch. Instead of a workaround, the directions and menu dropdown options should be adjusted so that they correspond to the actual question/answer. (8 scoring assertions)	Review and revise
678 (17790) (Stand alone)	Editorial Corrections Needed for Student Interaction (MS-LS2-3)	Instructions tell students to select "organisms" but students need to select the sun (which is NOT an organism) to get the answer correct. (2 scoring assertions)	Review and revise
682 (17878) (Stand alone)	Inaccurate diagram (MS-ESS1-1)	Image of moon shown for Dec. 16 is incorrect – should be the other half of the moon shown. (1 scoring assertion)	Review and revise
628 (18016) (Stand alone)	Scoring Assertions too far from PE (MS-PS3-4)	A majority of panelists found that the scoring assertions reflected the student work, but only a minority agreed they adequately represent the standard, commenting that the connection to mass was insufficient. (1 scoring assertion)	Review and revise
126 (Stand alone)	Editorial Corrections Needed for Student	Item scoring differentiates between a designation of "neither" (i.e. the evidence provided neither supports nor contradicts a statement) and "contradicts" (i.e. evidence provided contradicts a statement). However, although there is no evidence provided that supports nor contradicts the third statement, answering "neither" is considered incorrect. Panelists recommended switching to two choices and not have	Review and revise

	Interaction (MS-LS1-3)	students need to differentiate between “contradicts” and “neither.” (2 scoring assertions)	
173 (Stand alone)	Editorial Corrections Needed for Graph Label (MS-LS1-6)	Graph label needs correction; First instance of "Fall '01" should say "Fall '00"(1 scoring assertion)	Review and revise
Grade 11			
Item ID (Type)	**Issue (PE)**	**Reason**	**Suggested Resolution**
359 (2915/1477) (Stand alone)	Inaccurate diagram / science (HS-LS3-2)	Item is anchored in inaccurate diagram and confusion about the science ideas overall. (Current diagram shows a common misrepresentation of replicated chromatids, with each chromatid a different color and overlapping in an “x” shape. It is not possible to make sense of the problem, given the misrepresentations. Table 1 shows information about replication errors that occurred in mitosis, not meiosis, and the connection to the overall intent/problem is unclear.)	Remove or revise
564 (3036) (Cluster)	Inaccurate use of terms/units (HS-PS3-3)	Needs corrections for accurate use of kW (to measure power) vs kWh (to measure energy) (9 scoring assertions)	Review and revise
565 (3057) (Stand alone)	Science / Scoring Assertions Issues (HS-ESS2-3)	Item context seems to be drawing on the African Humid Period but the information provided does not allow students to make the inferences needed to arrive at the "correct" answer, and the evidence given is not considered sufficient to explain the actual climate change in the area (multiple factors are thought to have been involved). Panelists were additionally concerned that the emphasis on the "closest point to the sun" could take students toward common misconceptions related to the cause of seasons (i.e. closeness to sun vs tilt of axis). (1 scoring assertion)	Review and revise
476 (17721) (Stand alone)	Editorial Corrections Needed for Student Interaction/Scoring (MS-LS1-6)	Given information provided, it is not possible to determine the extent to which energy and/or matter is a limiting factor, and therefore not possible to infer whether the cause for the increase is due to the increase in available energy or available matter. (1 scoring assertion)	Remove or revise
492 (17800) Stand alone)	Editorial Corrections Needed for Student Interaction /Scoring (MS-LS1-6)	Currently, components of this answer can be correct if a student indicates there is evidence AND if a student indicated there is NOT sufficient evidence. Panelists argued that there are issues with the internal logic of the question if the answer can be "correct" given two opposite conditions (i.e. that it is not reasonable to both HAVE and NOT HAVE enough evidence.) One portion of this item requires students to indicate that individual human action of turning off lights on the beachfront has an impact. This requires an assumption about what that beachfront / lights are like – and	Remove or revise

		would not apply in the context of high-rises, for example, where an individual light would not change the overall beachfront lighting. (1 scoring assertion)	
497 (17814) (Stand alone)	Editorial Corrections Needed for Student Interaction/Scoring (MS-LS2-7)	Currently, the “correct” answer requires assumption, not evidence, and one component of the "correct” answer does not correspond to a science-based assumption – there is no evidence provided nor any reason to assume that creating a wildlife refuge will change the rate of biodiversity outside of the refuge (2 scoring assertions)	Remove or revise
559 (17932) (Cluster)	Science / Scoring Assertions Issues (HS-LS2-5)	Reviewers questioned the science and logic behind several scoring assertions. For example, an increase in biomass should be sufficient evidence for the observation (from Table 1) and it should not be considered incorrect for students to (correctly) state that plants release CO2 through cellular respiration. The standard emphasizes the relationship between photosynthesis and cellular respiration while this item separates them (and penalizes a student for connecting the processes). (9 scoring assertions)	Review and revise
668 (18034) (Cluster)	Overall Content and Scoring Assertions Issues (HS-LS3-2)	Requires students to make assumptions without data and to work outside of standard LS3-2, which does not include speciation by errors in meiosis / polyploidy. Students are not expected to have background experience with polyploidy and therefore the answer to Part D conflicts with the information as currently presented in Table 2. (10 scoring assertions)	Remove or revise
199 (Stand alone)	Editorial Corrections Needed/ Inaccurate Science (HS-PS2-4)	Item states that the mass of an object changes but it should say weight. Replace “mass” with “weight.” Revise sentence for awkward language of “are on is on.” Correct description of Figure 1, which states that it shows a spacecraft at two different points *when it leaves the moon* but actually shows a spacecraft on the moon and after leaving the moon. (1 scoring assertion)	Review and revise
216 (Stand alone)	Editorial Corrections / Graph Corrections (HS-ESS3-6)	Panelists were concerned because item emphasizes use of data but the “correct” graph inaccurately reflects the data in the given table. To resolve, either true up the graph with the data or just use general labels (similar to what is used for distance from stream) which would allow for the general trend without a mismatch of specific data. (1 scoring assertion)	Review and revise

WebbAlign™

WCEPS
Wisconsin Center for
Education Products & Services

Appendix D

Panelists' Notes and Source Of Challenge from the Content Analysis of SDSA Items in June 2022

January, 2023

Grade 5 Batch 48

Table 1 (Grade 5 Batch 48). *Notes by Reviewer*

Notes
Item #1 - 0. none 1. 3D 2. yes 3. no, scoring assertation says "matter" while the question says "energy" 4. no, standard requires both energy and matter while the question says energy and the scoring assertation says matter - 0. None 1. 3D 2. Yes 3. Yes 4. Yes - 0. none 1. yes 2. yes 3. no, the scoring assertation says matter and then ask about energy. 4. no, the question and the actitation do not match fully. Therefore they are both asking or scoring different things. 5. The question in itself is quite confusing. The wording makes it unclear at what the question is asking you to do. - 0. yes i agree 1. 3D 2. yes 3. no scoring and question are not talking about the same things energy vs matter 4. no the question and assertation matches to separate parts of the standard separately and not in alignment. 5. - 0. none 1. 3D 2. yes 3. No, the scoring assertation is suggesting matter whereas the question is talking about energy. 4. no, the standard assertation is referring to matter and energy where as the question and scoring assertation are two different things. - 0. None 1. 3D 2. Yes 3. Yes 4. Yes
Item #2 - 0. none 1. 3D 2. yes 3. yes 4. yes 5. the term "germination" might be too specific for some 3rd graders to know - 0. None 1. 3D 2. Yes 3. Yes 4. Yes - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5. - 0. none 1. 3D 2. yes 3. yes 4. Yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes
Item #3 - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5. - 0. none 1. 3D 2. yes 3. Yes 4. Yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes
Item #4 - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5. - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes
Item #5 - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5. - 0. none 1. 3D 2. Yes 3. Yes 4. Yes 5. Only comment is the ability to click on evidence to toggle it over to answer whereas the middle school did not have that. - 0. None 1. 3D 2. Yes 3. Yes 4. Yes
Item #6 - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes 5. No (5. I believe that this question leads students to believe that all information pertains to male birds. The jump to female birds are green could just mean that male birds that where green where not caught because they are hard to find.) - 0. none 1. 3D 2. yes 3. yes 4. yes 5. The question is a bit long for what is needed. You can eliminate part D and

still get an accurate reading on if the child understands the standard.
- 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5. It was a bit lengthy and the toggle portions tends to lead a student to question their understandings of breeding selections.
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Item #7
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5.
- 0. none 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Item #8
- 0. none 1. 3D 2. yes 3. yes 4. yes 5. An image of the rat storing seeds in cheeks might be useful (though not required)
- 0. None 1. 2D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5.
- 0. none 1. 3D 2. Yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Item #9
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes 5. No (0. I believe that the answer asks the students to make a jump in logic that is not supported by the information given.)
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5.
- 0. none 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Item #10
- 0. none 1. 3D 2. yes 3. yes 4. no, missing finding mates and reproduction components 5. student does not really need the data to answer this question if they have any knowledge at all about camouflage
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. no, this question does not apply to the entire standard. It does not go over breeding or reproduction which is half of the standard.
- 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. No, the scoring assertion does not apply to all of the standard due to it only covering surviving but not finding a mate or reproducing. 5.
- 0. none 1. yes 2. Yes 3. Yes 4. No, the question is only connecting one part of the standard.
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Item #11
- 0. none 1. 3D 2. yes 3. yes 4. no, maybe not broad enough to cover behavior and reproduction?
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes 5. No (5. Tell students they should have a mark in each row or one column may have more then one mark.)
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5.
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Item #12
- 0. none 1. 3D 2. yes (with addition of image mentioned in 5) 3. yes 4. yes 5. since this is such an unusual Australian animal, an image or picture would be useful here
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. No, the scoring assertion does not apply to all of the standard due to it only covering surviving but not finding a mate or reproducing. 5.

- 0. none 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Item #13
- 0. none 1. 3D 2. yes 3. yes 4. no, doesn't really cover finding a mate or reproducing
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes 5. Within this question there seems to be a lot of distractors from the context that is needed. It seems as if there could be a smaller graph or a graph that has more information regarding the exact subject. There is a lot about color but then nothing on the graph about color then the question asks about color. Seems confusing.
- 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5. They are including breeding information in the distractor portion to just to make the question his a majority of the topics in the standard to have it remain applicable.
- 0. none 1. 3D 2. yes 3. yes 4. yes 5. Pulling in extra information in order to meet the needs of the standard.
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Item #14
- 0. none 1. 3D 2. yes 3. yes 4. no, doesn't really cover finding a mate or reproducing
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. Yes I agree 1. 3D 2. yes 3. yes 4.No, the scoring assertion does not apply to all of the standard due to it only covering surviving but not growth, behavior or reproducing.
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Item #15
- 0. none 1. 3D 2. yes 3. yes 4. no, doesn't really cover finding a mate or reproducing 5. question would be strengthened with a description or image of the habitat to show that the larvae are living in the trees (green)
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes 5. No (5. It asks the students to make too many assumption about the time of year and then hiding in green leaves, if it was fall this hypothesis would not be true and it does not say the scientist put out the larva in early summer.)
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5. This question doesnt give any information about the habitat.
- 0. none 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Item #16
- 0. none 1. 3D 2. yes 3. no, on part D as long as students know the survival rate is higher for groups, they can choose the only 2 answer choices that say increase without even reading the rest of the sentence 4. yes 4.
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes 5.No (5. I can not see how question E is supported through the evidence provided.)
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5.
- 0. none 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Item #17
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5.
- 0. none 1. 3d 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Item #18
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5.

- 0. none 1. 3D 2. Yes 3. Yes 4. Yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes
Item #19 - 0. none 1. 3D 2, yes 3. yes 4. yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. Yes I agree 1. 3D 2. yes 3. yes 4. yes 5. - 0- none 1. 3D 2. Yes 3. Yes 4. Yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Table 2 (Grade 5 Batch 48). *Source-of-Challenge Issues by Reviewer*

Sources of Challenge

[NONE]

Grade 5 Batch 49

Table 1 (Grade 5 Batch 49). *Notes by Reviewer*

Notes
Item #1 - 0. none 1. 2D, students are not really asking a question, they are answering one 2. yes 3. yes 4. yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5. - 0. none 1. 3D 2. Yes 3. Yes 4. Yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes
Item #2 - 0. none 1. 3D 2. yes 3. yes 4. yes (maybe a bit of overshooting the standard though... covers more than just that the marshmallow and the air are made of particles, also includes pressure and particle movement) - 0. None 1. 3D 2. Yes 3. Yes 4. Yes - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5. A bit difficult due to amount of cluster items - 0.none 1.3D 2. Yes 3. Yes 4. Yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes
Item #3 - 0. none 1. 2D, no SEP 2. yes 3. yes 4. no, students are not planning or carrying out an investigation, they are simply looking at the data from a scenario - 0. None 1. 3D 2. Yes 3. Yes 4. Yes - 0. none 1. 3D 2. yes 3. yes 4. yes 5. Question seems to be overly simplified. The standard that this aligns with is typically a 4 however I feel this is too simple to be a 4 or even a 3. - 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5. - 0.none 1. 3D 2. Yes 3. Yes 4. Yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes
Item #4 - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5.

- 0. none 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Item #5
- 0. none 1. 2D, missing SEP 2. yes 3. yes 4. no, students are not really planning or even carrying out an investigation, they are just manipulating and analyzing someone else's data
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes 5. Yes Great question real makes students use what they know.
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5. Directions could be cleaned up and a bit more explanatory.
- 0. none 1. 3d 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Item #6
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5.
- 0. none 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Item #7
- 0. none 1. 2D, missing SEP 2. yes 3. yes 4. no, does not as students to define the problem, students are selecting a design
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5.
- 0.none 1. 3d 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Item #8
- 0. none 1. 3D 2. yes 3. yes 4. no, students do not plan or carry out the investigation, they simply analyze the data from the investigation
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes 5. Yes (Great Question)
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5.
- 0. none 1. 3d 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Item #9
- 0. none 1. 3D 2. yes 3. yes 4. no, students are comparing but not creating the solutions
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5.
- 0. none 1. 3d 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Item #10
- 0. none 1. 2D, missing SEP 2. yes 3. yes 4. no, students are not doing the investigation here (see #1) and they are only dealing with heat and light, not sound or electrical current
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. no, Within this question I believe that all things are aligned however there are multiple items in the standard that are not being questioned or even mentioned. I do not feel that this standard is being fully evaluated or tested. The only thing that is mentioned in the question is energy and light. This leaves out sound, heat, electric currents.
- 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5.
- 0.none 1.3d 2. yes 3.yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Item #11
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5. The students are being docked points for performing incorrect trials when the problem set up doesnt give them enough information to have confidence to submit correct trials. If they problem only was "graded" on the correct question then I would say there is no issue. But being docked for incorrect trials without enough information to provide correct trials.
- 0. none 1. 3D 2. Yes 3. Yes 4. Yes 5. a couple notes I am suggesting, is giving more information to the students so they can get the correct answer. The question is pushing for a dok 4 but it is only a dok 3 but taking information away to make it harder.
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Item #12
- 0. none 1. 2D, no SEP 2. yes 3. yes 4. no, doesn't address light, heat, or electric currents
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes 5. No, the final answer hints at movement that could be considered a transfer of energy.
- 0. none 1. 3D 2. yes 3. yes 4. yes 5. This question is very confusing. Adding in the last option as a question is adding in additional confusion and changing the question that it is asking. I do think that they are covering the portion of what they are testing.
- 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes, only sound is being covered as a topic matter in the standard. 5. Both C and D are plausible answers as when energy is transferred out of the system both will come to a stop with no energy in the system. So D being a distractor will throw a student off of the standard being tested.
- 0. none 1. 3d 2. yes 3. yes 4. yes 5. not covering enough of the standard. need to change the distractors in the question as the last option could also be the answer if didn't know the standard was sound.
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Item #13
- 0. none 1. 3D 2. yes 3. yes 4. no, no direct connection to energy levels, student would have to infer that relationship based on prior knowledge
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5.
- 0. none 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Item #14
- 0. none 1. 3D 2. yes 3. yes 4. no, student does not have to design or actually run the test, they are taking first steps to refining with the identification of disadvantages but needs more 5. This question would benefit from the addition of a followup question asking students what their next steps in the process would be
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes 5. No This question would benefit from some additional step that would address the transfer of energy.
- 0. none 1. 3D 2. yes 3. yes 4. yes 5. This question seems to be a little lacking. I feel that this would benefit from additional questions (clustering) or another question in follow up to the ones it is asking.
- 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. No, it is not meeting the full content of the standard 5. This question would need to be taken a step further for them to refine their project or design or reasoning. Would be a great cluster question in order to meet all of the standard needs.
- 0. none 1. 3D 2. yes 3. yes 4. The question needs one more piece in order for it to complete the standard.
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Item #15
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5.
- 0. none 1. 3d 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Item #16 - 0. none 1. 3D 2. yes 3. yes 4. no, doesn't address light, heat, or electric currents - 0. None 1. 3D 2. Yes 3. Yes 4. Yes - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5. - 0. none 1. 3d 2. yes 3. yes 4. yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes
Item #17 - 0. none 1. 3D 2. yes 3. yes 4. no, no direct connection to energy is made, students must infer that stretching the rubber band is adding energy - 0. None 1. 3D 2. Yes 3. Yes 4. Yes - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5. - 0. none 1. 3d 2. yes 3. yes 4. yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes
Item #18 - 0. none 1. 3D 2. yes 3. yes 4. no, students do not actually carryout the investigation, they just begin the planning stages of asking the question 5. may be closer to SEP 1? - 0. None 1. 3D 2. Yes 3. Yes 4. No (This descriptor and question could refer more to about the forces not speed. - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5. - 0. none 1. 3d 2. yes 3. yes 4. yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes
Item #19 - 0. none 1. 3D 2. yes 3. yes 4. no, students didn't plan or carry out the investigation, they simply analu=yzed the data to make a conclusion - 0. None 1. 3D 2. Yes 3. Yes 4. Yes - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5. - 0. none 1. 3d 2. yes 3. yes 4. yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes
Item #20 - 0. none 1. 3D 2. yes 3. yes 4. no, students are not really designing or testing, they are just suggesting a refinement - 0. None 1. 3D 2. Yes 3. Yes 4. Yes - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5. - 0. none 1. 3d 2. yes 3. yes 4. yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes
Item #21 - 0. none 1. 3D 2. yes 3. yes 4. no, students do not have to design or test the device, they just have to suggest a refinement - 0. None 1. 3D 2. Yes 3. Yes 4. Yes - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5. - 0. none 1. 3d 2. yes 3. yes 4. yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Table 2 (Grade 5 Batch 49). *Source-of-Challenge Issues by Reviewer*

Sources of Challenge
[NONE]

Grade 5 Batch 50

Table 1 (Grade 5 Batch 50). *Notes by Reviewer*

Notes
Item #1 - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5. - 0. none 1.3D 2. yes 3. yes 4. yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes
Item #2 - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes 5. No (Represent two regions of a world not place so close . Perhaps a different town in North Dakota that would not panic students when reading and cause them to loss their focus when reading the question.) - 0. none 1. 3D 2. yes 3. yes 4. yes 5. Additional information such as variation in the world needs to be used to insure that we are not using two places from the same place. Some children can be turned off from the names of the places as well when thinking about younger children easier words may ensure the comfortableness of the children. - 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5. Would be better fit into this standard if picked 2 locations further in difference of climates and seasons. If you dont change locations it would be better fit under ESS2-1 - 0. none 1. 3D 2. yes 3. yes 4. yes 5. Note: need to better understand different parts of the world. Need to change it so they are two climates. - 0. None 1. 3D 2. Yes 3. Yes 4. Yes 5. Keep in 2-2 and maybe represent a different region or part of the world as the standard indicates
Item #3 - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5. - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes
Item #4 - 0. none 1. 3D 2. yes 3. yes 4. no, standard asks students to compare to the sun and the sun in not included in the options 5. I am not sure that the graph scale allows the students to differentiate enough between some of the street lights? - 0. None 1. 3D 2. Yes 3. Yes 4. Yes - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5. - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes 5. Part A answer is hard to manipulate
Item #5 - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes - 0. none 1. 3D 2. yes 3. yes 4. yes

- 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5.
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Item #6
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5.
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Item #7
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5.
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Item #8
- 0. none 1. 3D 2. yes 3. yes 4. no, does not address shadows, day and night, or stars (moon is not even part of the standard)
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5. Distractor is not different enough from the correct answer to give a good assessment. A and B are very close in distance and unless and understanding of spatial value, not scientific value can lead to an incorrect answer.
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Item #9
- 0. none 1. 3D 2. yes 3. yes 4. no, does not address shadows, day and night, or stars (moon is not even part of the standard)
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5.
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Item #10
- 0. none (though there could be some connection to 3-ESS3-1 since flooding is a weather related hazard as well) 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5.
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Item #11
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5.
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Item #12
- 0. none 1. 3D 2. yes 3. yes 4. yes

- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5.
- 0. none 1. 3D 2. yes 3. yes 4. yes 5. I would suggest mentioning how long after the solar panels are placed. So essential the pattern could be at what point does the plant life begin to decrease.
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Item #13
- 0. none 1. 3D 2. no, without an image, I am not sure that enough students will know what a "reef" is 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5.
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Item #14
- 0. none 1. 3D 2. yes 3. yes 4. no, students are not generating the solutions, just comparing and evaluating them
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes 5. No (Needs more details in descriptor or limitation for students to make correct answers.)
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5.
- 0. none 1. 3D 2. yes 3. yes 4. yes 5. Data needs more information that needs more. Especially if you have an answer about taller buildings but no data on it.
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Item #15
- 0. none 1. 3D 2. yes 3. yes 4. no, students do not deal with shadows or day and night , only star patterns
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes 5. The coloring and size of the stars that the student is to be determining should be put in another color or changed as it is confusing.
- 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5. Stars to be identified and followed need to be a different color than the distractors in the background.
- 0. none 1. 3D 2. yes 3. yes 4. yes 5. This data does not help the pattern of the story. The color needs to be a yellow.
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Item #16
- 0. none 1. 3D 2. yes 3. yes 4. yes but see note on 5 5. The depth of this task might be more suited to 4th or 5th grade which would make this better connected to 4-ESS3-2 with drying soil being more of an "earth process" and less of a "weather-related hazard"
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5.The clusters are attempting to make it more complex but all it is really doing it making it more difficult.
- 0. none 1. 3D 2. yes 3. yes 4. yes 5. Too much work, trying to make it more complex when they shouldnt.
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes 5. Too complex for student?

Item #17
- 0. none 1. 3D 2. yes 3. yes 4. no, students are not generating solutions, just comparing them
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5.
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Item #18
- 0. none 1. 3D 2. yes 3. yes 4. no, question does not include shadows, or day and night
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes

- 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5. - 0. none 1. 3d 2. yes 2. yes 3. yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes
Item #19 - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5. - 0.none 1. 3d 2. yes 3. yes 4. yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes
Item #20 - 0. none 1. 3D 2. yes 3. yes 4. no, not sure that students are really generating and comparing multiple solutions so much as they are thinking about ways to modify a single solution - 0. None 1. 3D 2. Yes 3. Yes 4. No (This discuses measure of solutions not the comparing after measuring. - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5. - 0.none 1. 3d 2. yes 3. yes 4. yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes 5. What the student is expected to do is a bit confusing
Item #21 - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes 5, No (more info about weather to know that it is not other factors.) - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5. - 0.none 1. 3d 2. yes 3. yes 4. yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Table 2 (Grade 5 Batch 50) *Source-of-Challenge Issues by Reviewer*

Sources of Challenge
[NONE]

Grade 8 Batch 51

Table 1 (Grade 8 Batch 51). *Notes by Reviewer*

Notes
Item #1 - 0. none 1. 3D 2. yes 3. no (see source of challenge) 4. yes - 0. None 1. 3D 2. Yes 3. No (see source of challenge) 4. Yes (Students are unable to answer 126 correctly because the answers do not reflect the evidence in the question. There is no contradictive evidence.) - 0. none 1. 3D 2. yes 3. no (see source of challenge) 4. yes - 0. None 1. 3D 2. yes 3. no (See source of challenge) 4. yes 5.The neither box is vague and doesn't tell the students if there was NO Evidence present to click that box. As an educated adult I was unsure when neither was an applicable answer. - 0.None 1. 3D 2. Yes 3. No (see source of challenge) 4. Yes - 0. None 1. 3D 2. Yes 3. No 4. Yes 5. - 0-none 1.3d 2. yes 3. no (see source of challenge-facts listed are not appropriate for the questions and options available to respond to) 4. yes 5 - O - No disagreement. 1- 3D 2- yes 3-no, if a student selects "neither" then they automatically get no credit. When students select this column, the data don't tell us that students know anything about the evidence. It may need to be worth 4 points, vs 2 points. 4-yes - 0. None 1. 3D 2. yes 3. no (student cannot get correct responses based on structure of question) 4. yes - 0. None. 1. 3D. 2. Yes. 3. No. Neither should not be a choice. 4. Yes. - 0. None 1. 3D 2. Yes 3. No. Scoring assertion is unreasonable based on structure of answer choices noted above. 4. Yes. - 0. none 1. 3D 2. yes 3. no (see source of challenge) 4. yes
Item #2 - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. None 1. 2D 2. Yes 3. Yes 4. Yes - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. yes I agree 1. 3D 2. yes 4. yes 5. - 0. None 1. 3D 2. Yes 3. Yes 4. Yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes - 0. 1. 3d 2.yes 3. yes 4.yes - 1-2D, students are not constructing explanations and designing solutions 2-phenomenon OK 3-yes 4-yes, except the SEP, students aren't really explaining - 0. MS1-6 is the preselected choice however, this standard cites the flow of energy into and out of organisms. This question, even though the correct response involves photosynthesis, is more about the flow of energy among living and nonliving parts of an ecosystem. The response choices reflect this as well - A & C are living, B & D are non-living. After discussion with group, I changed my response for the betterment of the group. 1. 2D (missing SEP) 2. yes 3. yes 4. no - the question assesses the students' ability to read a graph. It does not reflect their understanding of photosynthesis moving energy between organisms. 5. label in graph is wrong - Fall '01 is listed twice - the first instance should read Fall '00 - 0. First I chose MS-LS2-2, but after looking at the answers I can see it is relating to photosynthesis which is standard MS-LS1-6. 1. 3D. 2. Yes. 3. Yes. 4. Yes. - 0. None. 1. 3D. 2. Yes 3. No. A student can get this right without understanding photosynthesis, since it is the only logical answer where carbon dioxide decreases during the summer. 4. No. This question addresses the flow of matter from photosynthesis, but not the flow of energy. - 0. none 1. 3D 2. yes 3. yes 4. yes
Item #3 - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. none 1. 3D 2. Yes 3. Yes 4. Yes - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. Yes I agree 1. 3D 2. yes 3. yes 4. yes 5. - 0.none 1. 3D 2. yes 3. yes 4. yes

- 0. None 1. 2D, looks more like SEP 4 analyzing and interpreting data 2. Yes 3. Yes 4. Yes
- 0. 1.3d 2. yes 3.yes 4. yes 5.
- 1-2D, students are connecting to a pattern of information, not really an SEP of constructing an explanation and designing solutions 2-phenomenon, OK 3-yes 4-yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes

Item #4
- 0.none (though could be connected to 1-4 as well because of the specialized plant structures) 1. 2. 3. 4. 5. calculator is very difficult to manipulate on this question (often covered important info that I needed to see)
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes 5. Toggle on the question with the graph should make it clear that you can apply the same answer to all, or that the answers can be used more than once. It is misleading.
- 0. Yes I agree 1. 3D 2. yes 3. no, it is related to the standard but does not fully address what the standard is asking 4. no, it is related to the standard but does not fully address what the standard is asking 5. Would be better to have a descriptor saying more than an answer can be used more than once in the toggle portion. Also if a student believes that an answer can only be used once, they can be penalized more than once for this misunderstanding of the toggle chart.Too many cluster items for one problem
- 0. None 1. 3D 2. yes 3. No, the scoring do not describe the inferences that can be made from student. 4. No, the scoring did not meet depth or breadth. 5. The question needs to be better broken up as well as there are too many questions.
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. 1.3d 2. maybe geographical bias 3.no-bc if you get the percentages wrong you miss almost half of the points (3 of 8) 4. yes 5. perhaps make the percentage questions only worth 1 point total...showing how to calculate percentage is googleable and a 20 sec fix.
- 1-3D 2-phenomenon = good 3-yes 4-yes 5-Part D of item 577 will be tricky for students because of the similarity of the way choices B & D are worded.
- 0. None 1. 3D 2. yes 3. yes 4. yes 5. Calculator doesn't compute the math - it just enters the equation the student types into the calculator. This will confuse students since it looks like it wants a specific number as a response.
- 0. I originally chose MS-LS4-4, but upon review I can see there is a mathematical aspect which fit MS-LS4-6. 1. 3D 2. Yes. 3. Yes. 4. Yes
- 0. None 1. 3D 2. Yes 3. No. 4. Yes 5. The average petal length question is confusing. The note to select "the same length" for similar mean petal lengths is easy to miss. Making that more obvious or providing additional options like "same as Arizona, etc. would be helpful.
- 0. originally chose LS1-4 but realize the mathematical component and agree with LS4-6 1. 3D 2. yes 3. yes 4. yes 5. calculator with no equal function key

Item #5
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1.3D 2. yes 3. yes 4. yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. yes 1. 3D 2. yes 3. yes 4. yes 5.
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. 1.yes 2.yes 3.yes 4.yes 5. there should be a "does not apply" option because the one that was gone prior to 1980 (paintbrush) doesn't support or not support
- 1-3D 2-phenomenon is good 3-no, see notes in challenge box 4-yes 5-The way the item is scored is a challenge. If a student doesn't get every check correct then it doesn't register that the student knows anything.
- 0. none 1. 3D 2. yes 3. yes 4. yes 5. Is the year for Butterfly Bush a typo? Is it supposed to be 1989 (as it is now) or is it supposed to read 1980?
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes

Item #6
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. yes 3.Yes 4.Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. yes 1. 3D 2. yes 3. yes 4. yes 5.
- 0. None 1. 3D 2. yes 3. yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. 1.3d 2.yes 3yes 4 yes
- 1-2D, I see that students are analyzing data for a pattern, but I'm not convinced that they can interpreting what this data means. Instead they are just aligning to an evolutionary answer they were told about in class. 2-phenomenon is OK 3-no, the distractors are not such that a correct answers tells us that students know about changes in complexity because younger fossils are not always more complex. 4-yes 5-I'm not sure I agree that the "correct answer" is correct. Does getting more ribs really = more advanced?
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes

Item #7
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0.None 1.3D 2.Yes 3. yes 4. yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. yes 1. 3D 2. yes 3. yes 4. yes 5.
- 0. none 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. 1.ok-but sep is wimpy 2yes 3 not necessarily 4 yes 5. the options on the 2nd question need reworked. There are 2 different blanks to fill and they don't necessarily go together. It should be eggs...not chicks for first portion and then the 2nd portion doesn't reflect accurate information based on chart or theory.
- 1-3D 2-phenomenon is good 3-no, (also see SoC box above) student could engage in an argument two different ways, but are only going to get points for selecting one way to look at the outcome. This will not tell us what students know about argumentation from evidence. 4-yes 5-While I get the premise the question is getting at, the population of albatros birds would also immediately decline as chicks would not be born next breading season.
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes

Item #8
- 0. none 1. 3D 2. yes 3. no, graph does not show increase in nutrients that would correspond with the nutrients being released back into the ecosystem by the decay of plankton 4. yes
- 0. None 1. 3D (Also SEP 4- not listed on standard) 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes 5. Scoring for this question seems incorrect. There are three questions but only 2 points. If you miss any of the 3 correct they automatically get only 1/2 instead of 2/2. It should be one point for each.
- 0. yes 1. 3D 2. yes 3. yes 4. yes 5.There is three toggles but there is only two points given. You are either not being counted for a problem or you have to have all 3 correct to receive the 2 points. Not an equitable question.
- 0. None 1. 3D 2. Yes 3. No, The Scoring did not add if it had to have all three correct. 4. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes 5. Because of the zooplankton it makes it 2-3 not 1-6 5. The graphic that comes with this question can be very confusing.
- 0. 1.yes 2.yes 3.no, if it stays written the way it is 4.yes 5.
- 1-3D, Constructing an Explanation is not really as clear as is analyzing and interpreting data for the SEP. 2-phenomenon is OK. 3-yes 4-no, the SEP is aligned better to LS2-1 than to LS1-6 and to the CCC of cause/effect than to energy/matter
- 0. none 1. 2D (missing SEP) 2. no (phenomenon assumes students have prior knowledge of life cycles of marine biome - students may not know that zooplankton eat phytoplankton and that neither of them count as nutrients.) 3. yes 4. no - the question / scoring does not include photosynthesis, it does not require a student to construct an

explanation.
- 0. I originally chose MS-LS2-1 because I felt like resources were more than photosynthesis, but photosynthesis was the basis of all of the resources discussed so MS-LS1-6 is appropriate. 1. 3D 2. Yes 3. Yes 4. Yes
- 0. I disagree with the assigned standard MS-LS1-6. Both the question and scoring assertions relate to energy and matter flow through ecosystems as a whole, not just photosynthesis. 1. 3D 2. Yes 3. Yes 4. No. The scoring assertions provided much better relate to MS-LS2-3 as opposed to MS-LS1-6.
- 0. none 1. 3D 2. yes 3. yes 4. yes

Item #9
- 0. none 1. 3D 2. yes 3. yes 4. no, students don't really obtain or evaluate the information
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. yes 1. 3D 2. yes 3. yes 4. yes 5.
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. Yes 3. Yes 4. Yes
- 0. 1.yes 2. yes 3.yes 4.yes 5. the articles do not say DNA so the "altered DNA" means the students need to infer which isn't the SEP.
- 1-2D, not SEP strong 2-phenomenon is ok. 3-no, see SoC box above 4-yes 5-Scoring is a source of concern as one incorrect check box indicates the student doesn't know anything. 4 checks = 1 point doesn't allow for data about much to be gathered about what students are thinking.
- 0. none 1. 2D (missing CCC) 2. yes 3. no (a student has to be correct on all four items in order to earn the single point. A student may be able to respond correctly to some and miss others. 4. no (the standard is obtaining, evaluating, and communicating information - the scoring assertions is synthesizing information)
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. no - the item only treats information about the processes and does not address the effect of technological changes

Item #10
- 0. none 1. 3D 2. yes 3. no, students could potentially complete the last column of the table without looking at the model 4. no, students are not developing the model, simply using it (and not even really needing it for part of the answers)
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. yes 1. 3D 2. yes 3. yes 4. yes 5. I labeled the last problem as B which gave me a right point on the first, wrong on last. I changed the last problem to an E and then it changed my first problem to wrong and my last one to correct.
- 0. None 1. 3D 2. yes 3. no, check button are correct 4. Yes 5. Check to see the correct answer in in corresponding answer.
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. 1.yes 2.no necessarily...if I put the outside membrane (cell wall) as the bodyguard to let water in and out...but if I put it's name as cell membrane I only get 1/2 points...this doesn't seem reflective of my knowledge...just forgot a googleable term. 4.yes 5. if I put the outside membrane (cell wall) as the bodyguard to let water in and out...but if I put it's name as cell membrane I only get 1/2 points...this doesn't seem reflective of my knowledge...just forgot a googleable term.
- 1-3D 2-phenomenon is ok, but it is not needed to answer the item questions. That is sad because the phenomenon is not driving the student responses. 3-no, (also see SoC above) It is not telling us that a student know what the cell membrane is or what it does if they get correct scoring points for selecting D-cell wall. 4-yes 5-The way the item allows for cell wall to be a correct answer can misguide about what students know.
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. Yes 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes

Item #11
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0.None 1. 3D 2. Yes 3. Yes 4. Yes

- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. I see this as standard MS-LS2-1 as it is directly talking about effects of resource availability organisms in a populations. 1.3D; Should be Sep 4 Should be CCC Cause/effect 2. yes 3. yes 4. yes 5.
- 0. None 1. 3D because of the change in standard that it should be cause/effect 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. No 4. Yes 5. The data is incomplete. The student will be unable to draw a conclusion without data before the introduction of the carp and the data afterwards.
- 0. 1.3d 2.yes 3.yes 4.yes
- 1-3D 2-phenomenon is ok. 3-no, see SoC notes above 4-yes 5-The carp shares features of it's diet with all the native fish on the chart. Why couldn't any of them be selected for the fish population that decreases due to similar diet?
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. I felt MS-LS2-1 was correct because the source of food would be resource availability, but I can see that MS-LS2-4 is describing a change affecting the population. 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes

Item #12
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. yes 1. 3D 2. yes 3. yes 4. yes 5.There is two toggles but there is only one point given. You are either not being counted for a problem or you have to have both correct to receive the 1 point. Not an equitable question.
- 0. None 1. 3D 2. Yes 3. No need to have more than one possible point 4. Yes 5. Need a corrected amount of points.
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. I think it's more ls 2.1 because it's about resources, it hardly talked about changes (ls2.4) 1. 3d 2.yes 3.yes 4.yes
- 1-3D 2-phenomenon is ok. 3-yes 4-yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes

Item #13
- 0. none 1. 3D 2. yes 3. yes 4. yes (but seems keyed more to cause and effect than stability and change?)
- 0. None 1. 3D 2. Yes (3. There is not any evidence to see how resource are impacted. To me the negative percentages did not necessarily mean resources it could be nonliving things such as pollination by wind.) 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. yes 1. 3D 2. yes 3. yes 4. yes 5.
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. I think it's more ls2.2 due to different locations =different ecosystems and the ccc is patterns...looking at data, shows patterns, not ls2.4 1.3d 2.yes 3.yes 4.yes
- 1-2D (lacks SEP) 2-phenomenon is weak, its just some data. 3-no, From the information given, how do we know Kudzo is the problem? Trumpet Honeysuckle and Virginia Creeper could be affected by other things too, like deer populations. 4-no, the item doesn't really ask students to do anything with patterns in the data. 5-The way the distractors are worded is concerning. What is considered "much more harmful"? B could be considered correct.
- 0. none 1. 3D 2. yes 3. no (student could also choose B based on the data given and would also support the assertion made.) 4. yes 5. The data given shows the impact is greater to the honeysuckle (the negative percentages are greater implying the impact was greater). Although it says "much more", define much. This makes option B correct but a student is marked wrong if B is selected.
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes 5. Descriptor B - key word Much more could be missed
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes 5. The answer option that kudzu is "much more harmful" to one plant than the other is vague and should be re-worded. Students may interpret the difference in percent change between the two plants as significant to mark this option.
- 0. none 1. 3D 2. yes 3. yes 4. yes

Item #14
- 0. none 1. 3D 2. yes 3. yes 4. yes 5. the term "suture" might be confusing to students, maybe add a diagram showing what a suture is?
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. yes 1. 3D 2. yes 3. yes 4. yes 5.
- 0. None 1. 3D 2. yes 3. yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. 1.yes 3d 2.yes 3.no 4.yes 5.there are 2 possible correct answers. the "over time organisms developed more complex sutures" as well as shells at the bottom could survive with lower oxygen levels, which must have been true too, bc they survived long enough to make fossils.
- 1-1D - students aren't really asked to do anything with patterns, or with the SEPs, it is just recall earth science information that younger fossils are generally more advanced. 2-phenomenon is ok. 3-yes, 4-no, see reasons on #1
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. No - I believe students could infer choice C from the information given. 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes

Item #15
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. yes 1. 3D 2. yes 3. yes 4. yes 5.
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. 1.3d 2 yes...landlocked students do not know estuaries 3.yes 4. yes 5. terminology is wrong in a distractor...oxygen isn't "dissolved" it could be more or less saturated, or, has more or less oxygen...but definitely not dissolved.
- 1-2D, lacks SEP 2-phenomenon is ok. 3-yes 4-yes, but doesn't really address argumentation from evidence
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. I originally thought it was MS-LS2-1 because it was discussing resource availability effect on populations. MS-LS2-4 is talking about changes to ecosystems that affect populations (a subtle difference). 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes

Item #16
- 0. none 1. 3D 2. yes 3. yes 4. no, does not address matter, only energy 5. instructions ask student to organize the "organisms" in the drop down menus but the first answer has to be "sun"
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. yes 1. 3D 2. yes 3. yes 4. yes 5. Using the word organisms misleads the student to not select the sun which is a non living part of the cycle.You need the sun to get the problem correct. The word organisms make you want to chose all living items.
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes 5. Need to remove organism from the directions.
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. 1.yes 2.yes 3.not necessarily 4.yes 5. The question asks for "organisms" in flow of energy. one of the correct answers should have been "sun" and the sun isn't an organism.
- 1-3D 2-phenomenon is ok. 3-yes 4-yes 5-I like how there are multiple pathways available to getting a correct answer.
- 0. none 1. 3D 2. yes 3. yes 4. no - other than the sun - where are the nonliving parts of this model? The question asks about flow of energy, but where is cycling of matter?
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. No. LS2-3 includes the cycling of matter AND energy. 5. The prompt for this question states says "Using the information from Table 1, click the blank boxes and select organisms in order to complete the diagram showing the flow of energy to the boa constrictor." However, the correct first answer is

"sun," which is not an organism. A more appropriate prompt would be to "select the item," NOT the organism.
- 0. none 1. 3D 2. yes 3. yes 4. yes

Item #17
- 0. none 1. 3D 2. yes 3. yes 4. yes 5. Isn't the measurement actually called the Shannon-Weiner index?
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0.yes 1. 3D 2. yes 3. yes 4. yes 5.
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0 I think it should be ls2.1 1.yes 2.yes-although "abundance" may be a word not known by many 8th graders 3. no-the graph shows that partial removal of the plant had LESS diversity of species than NO removal of plants. So FULL removal that created LOTS of diversity...so there is some contradictive evidence...so it should have an answer of cannot be determined...but it said the correct answer is "negative impact" 4. 5. see note a
- 1-2D 2-phenomenon is ok. 3-yes 4-yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 2D - Missing SEP 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. no - assertion does not address effects on specific populations; only diversity

Item #18
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. yes 1. 3D 2. yes 3. yes 4. yes 5.
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. 1.3d 2.yes 3.yes 4.yes
- 1-2D, the item has students make predictions, but not construct explanations or design solutions. 2-phenomenon is OK. 3-yes. 4-no, see comments to #1.
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. I originally chose MS-LS2-4 because we were using a change to see effect on a population, but I can see that MS-LS2-2 addresses multiple ecosystems. 1. 3D 2. Yes 3. Yes 4. Yes
- 0. I initially picked LS2-4, but after re-reading think LS2-2 is a better choice. 1. 3D 2. Yes 3. Yes 4. Yes
- 0. I first considered LS2-4 but agree that LS2-2 relates better considering the emphasis on interactions across multiple ecosystems 1. 3D 2. yes 3. yes 4. yes

Item #19
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. yes 1. 3D 2. yes 3. yes 4. yes 5.
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. 1.3d 2.yes-snout may be a bias against ELL students 3.yes 4. yes 5. the word snout isn't a common word and could impinge ELL students
- 1-3D 2-phenomena is OK. 3-yes. 4-yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes

Item #20
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. No (3. It would help students understand that the second box is advantage if it was put before it and the the third box reflected disadvantage. I would say by not doing that there is a bias for test takes who understand how a test is put together.) 4. Yes

- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. yes 1. 3D 2. yes 3. yes 4. yes 5.There is three toggles but there is only 1 point given. You are either not being counted for 1-2 problem or you have to have all 3 correct to receive the 1 points1. Not an equitable question.
- 0.none 1. 3D 2. yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. 1.3d 2.yes 3.no...almost any answer is correct, using the table word for word 4.yes 5.if pesticides are chosen, the right option at the end states "does not harm other organisms." Yet, it does. It states "plants must be removed to create space."
- 1-2D, students are not engaging with the CCC. 2-phenomenon is OK 3-no, students can enter something logical (does not change the environment) and still get the point incorrect. This doesn't produce data about what a student knows. See also notes in SoC above. 4-no, missing a scoring assertion to the DCI and to the CCC. Scoring is a problem. This item should score more like item 671 about tomato genes. 5-The choice, "is less expensive" as an answer is problematic bc the table just says it is "not expensive" and doesn't tell anything about the cost of the other control measures.
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. No - I think expense is something farmers would consider 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes 5. great that the item allows for individual solutions and reasoning

Item #21
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. N0 (0. I like to see the mention of photosynthesis in the question so student know what they are addressing.)
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. yes 1. 3D 2. yes 3. yes 4. yes 5.There is three toggles but there is only 1 point given. You are either not being counted for a problem or you have to have all 3 correct to receive the 1 point. Not an equitable question.
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 1-3D 2-phenomenon is OK. 3-no, having just 1 score point for this item does not allow for telling what kinds of gaps in student understanding exist. 4-no, similar to #3, having just one score point does not allow for telling what piece (SEP, DCI, or CCC) a student is not connecting with correctly. 5-Scoring is a problem. This item should score more like MS-LS item 671 about tomato genes.
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. none 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. I initially chose LS 1-5 which specifically addresses the cause and effect of environmental factors influence on the growth of organisms. I changed to LS 1-6 to emphasize the transfer of matter by photosynthesis. 1. 3D 2. yes 3. yes 4. Yes
- 0. 1.3d 2.yes 3.yes 4.yes

Item #22
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. Yes 1. 3D 2. Yes 3. Yes 4. Yes 5.
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes 5. Students could easily be confused by the order of answers. They might want to change one answer to a given statement to help the student better figure out the order.
- 0. 1. 3d 2.yes 3.yes 4yes
- 1-3D 2-phenomenon is good. 3-YES 4-no, the scoring assertations are all related to the DCI & CCC, none to the SEP
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- o. None 1. 3D 2. Yes 3. No – I think students could have multiple correct answers. 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Item #23
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. Yes 1. 3D 2. Yes 3. Yes 4. Yes 5.There is 9 toggles but there is only 3 points given. You are either not being counted for a problem or you have to have all 3-9 correct to receive the 3 points. Not an equitable question.
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. 1. 2d…doesn't develop and use a model to describe WHY 2.yes 3.yes 4. Yes 5.The picture doesn't accurately display what they say should be the answers. Perhaps 3 different pictures (or more) are needed to provide the evidence for the question
- 1-2D, lacks connections to CCC. 2-phenomenon, ugh! "Scientists made a strain of bread mold that produces proteins that glow green under certain light." Tell the students how the scientists did this (I'm assuming it was with jellyfish genes.) The stimulus leads to asking questions that interfere with train of thought towards answering the item. There needs to be a diagram that accompanies the first paragraph. There is a lot of background information students need to know and bring to this questions before correct answers can be selected. 3-no, there is no evidence in student choice of answers that they can "improve" the model. 4-no, no connection to CCC 5-There is not enough information in this stimulus for a student to understand what is going on with hyphae and spores unless they have specifically studied it in class.
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. No – not explaining why as standard states
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. No – assertion inference is limited to describing bread mold reproduction and does not explain reason for differing results in offspring

Item #24
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. Yes 1. 3D 2. Yes 3. Yes 4. Yes 5.
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. 1.yes 2.yes 3.yes 4.yes 5. This is the best question I've seen. It lists what criteria are important for the grower. Therefore, NO student bias…examples-low income will never do things that cost.
- 1-2D, this item isn't as much about evaluating competing design solutions as it is about selecting the answer that meets the criteria, so it lacks the SEP connection. 2-phenomenon is ok. 3-no, rubric does not indicate evidence about supporting a claim because it is about matching to a criteria list. 4-no, lacks connections to SEP
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Table 2 (Grade 8 Batch 51). *Source-of-Challenge Issues by Reviewer*

Sources of Challenge
Item #1 - Students are unable to correctly answer this question as written. “Neither” and “Contradicted” could be correct answers for two of the answer keys. - Students are not able to answer correctly how it is written. “Neither” option could also be the same for “contradicted” option when choosing the correct answer. - Students are not able to answer questions accurately. “Neither” could be the correct answer for all questions. - Students not able to answer correctly. Contradicted is confusing as it was not contradicted it was none existent. - The way the item is scored is a source of challenge. See #3 below in notes. - Contradicting evidence and neither are not distinguishable for MS student. There are statements that have zero evidence given but when student marks neither, the question is marked zero points. - 3. Choices should be Supported by the Evidence or Not Supported by the Evidence. - The difference between “contradictory evidence” and “neither” is very subjective. Both columns should be replaced with one column, “no evidence.” - The table is confusing regarding answer choices. Appropriate responses would be supported and no evidence.

Grade 8 Batch 52

Table 1 (Grade 8 Batch 52). *Notes by Reviewer*

Notes
Item #1 - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. Yes I agree with the standard given 1. 3D 2. yes 3. yes 4.yes 5. - 0. None 1. 3D 2. yes 3. Yes 4. yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes - 0. 1.3d 2.yes 3.yes 4no 5.The wording above question is poor. If students do not know the independent dependent variable. They will not show knowing forces...just scientific terms or scientific method. - 1-1D, can be answered correctly with only knowledge about experimental design, no DCI or CCC needed to answer this item. 2-phenomenon is ok. 3-no, rubric says it connects to the DCI, but the item really only connects to students knowing parts of experimental design correctly. No DCI knowledge is needed to select correct answers, thus the score points don't tell us what students know about forces, only about experimental design. 4-no, related to comments on #3 5-I challenge that this item has any real connections to DCI and CCC - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes - 0. none 1. 3D 2. yes 3. yes 4. yes

Item #2
- 0. none 1. 3D 2. yes 3. yes 4. yes 5. scoring seems like it should be split between knowing the relationship between sound and amplitude for 1 pt and then giving a separate point for making the correct calculation
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5.
- 0. none 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0 1.yes 2.yes 3.yes 4. no...it doesn't show they know amplitude is energy in wave
- 1-3D 2-phenomenon is OK. 3-no, with only 1 score point, the item won't report that students may conceptually get the concept but may not get the math. 4-no, scoring assertation does not disaggregate student knowledge about CCC (patterns) vs (DCI) content vs (SEP) mathematical thinking. 5-SoC - with only 1 score point, the item won't report that students may conceptually get the concept but may not get the math
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. none 1. 3D 2. Yes 3. Yes 4. No - Does not address energy of wave, only loudness.
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes

Item #3
- 0. none 1. 3D 2. yes 3. yes 4. no, students do not have to construct the graphical displays
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5.
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0 1.yes 2.yes 3.yes 4.yes
- 1-2D, though it will be hard for students to analyze and interpret data without a formula and there is really no connection to the CCC. 2-phenomena is ok, although Joules would not be the choice of units in a physical science class except when doing chemistry topics. 3-no, the scoring assertation only gets data that a student selects the right answer but doesn't help with what the student knows about the standard. 4-no, the scoring does not represent the depth and breadth of the standard due to the item being a single-select MC answer. 5-Students are going to struggle with getting the correct answer without a formula for KE
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. No: the standard asks students to construct and analyze. In this question, they only construct. 5. The units on the graph are difficult to interpret. A smaller scale or larger graphs would aid students in thinking mathematically about the question.
- 0. none 1. 3D 2. yes 3. yes 4. yes

Item #4
- 0. none 1. 3D 2. yes 3. yes 4. no, students do not have to develop the model 5. The word "calculate" in the instructions could be a distractor since it might lead students to think that they have to actually make a calculation to find the answer?
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5. When same numbers input to provide evidence of conservation the problem was marked as incorrect.
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. 1.yes 2.yes 3.no-students can know that mass is conserved and still get it wrong bc they put the wrong total number of atoms...maybe they think it's unbalanced because it's counted wrong 4.yes 5.see note above
- 1-1D, content knowledge is all that is required to answer this item. 2-phenomenon is not engaging at all the way it is written. 3-no, the scoring rationale does not describe what a student may do to answer this question correctly. 4-no, the scoring assertions do not address the depth and breadth of the standards because the item doesn't require that to answer it correctly. 5-This is a DOK1 item because a student only needs to know the law of conservation of mass to answer the question, no actual calculations are necessary to complete this item.

- 0. none 1. 2D (missing SEP) 2. yes 3. yes 4. no (student does not have to develop a model. The model is given to them and the question is asking more about mathematical equations than developing a model.)
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. No. A student could completely lack an understanding of conservation of mass, but get this question correct by adding the masses of the atoms present using a periodic table. 4. Yes
- 0. none 1. 2D - no SEP engagement necessary 2. yes 3. yes 4. no - the item does not require student to engage in developing a model

Item #5
- 0. none 1. 3D 2. yes 3. yes 4. no, students do not have to develop a model
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5.
- 0. none 1. 3D 2. yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. 1.no 2d 2.yes 3.no-kids will just click...no describing needed, as well as it doesn't provide evidence or prove the standard 4.no see above 5. see above plus challenge notes of: unless they say look at the dye to determine the motion of the particles of the water, it doesn't cover the standard, nor is it describing the model
- 1-2D, lacks CCC connections 2-phenomenon is weak 3-no, there isn't enough in the item to elicit information about student knowledge of thermal and molecular motion 4- no, there isn't enough in the item to elicit information about student knowledge of cause and effect. A single MC item doesn't tell much about what students know. 5-SoC - The item needs more to illicit knowledge about how a student is applying CCC knowledge about cause and effect.
- 0. This question does partially align with PS3-4 which discusses change in kinetic energy based on temperature of sample. 1. 3D 2. yes 3. yes 4. no - does not address the state of matter of a substance based on particle motion, and temperature.
- 0. None 1. 3D 2. Yes 3. Yes 4. No - state in standard, not addressed in question
- 0. This does not align to the standard listed for this question. The movement of dye through water at varying temperatures is assessing student knowledge of diffusion, not of particle motion of a pure substance and state changes as stated in PS1-4. 1. 3D 2. Yes 3. Yes 4. No. While this question does accurately assess whether students understand that particles in water are moving quickly at higher temperatures, it fails to address the concept of state change and makes things confusing by asking about a pure substance, but a solution consisting of a dye and water as solvent. 5. This question should be reworked, perhaps by using the visual selected in the multiple choice as something for students to view, and then using dropdown boxes to have students construct an explanation that the dye molecules disperse more quickly in the warmer solutions due to the higher kinetic energy of the water molecules.
- 0. none 1. 3D 2. yes 3. yes 4. no - the assertion does not address change in state

Item #6
- 0. none 1. 3D 2. yes 3. yes 4. no, students do not have to develop a model
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5. Really like this question!!
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes' 4. Yes
- 0. 1.3d 2.yes 3.yes 4.yes
- 1-3D, but the item only functions correctly if the model is illustrated with accuracy. I am not convinced it is completely accurate. 2-phenomena is good, but the illustration is distracting to what is happening with the light waves. 3-3-no, see comments on#5 below. 4-no, see comments on #5 below. 5-There needs to be more to this question to elicit information about why a student chose D because otherwise if they chose any incorrect answer the data doesn't tell us what students misconceptions are. Without direct experience such as this investigation, a student could select B yet the data wouldn't tell us they knew any more than a student who selects A or C. Conversely, a student could select D by einy-miney between B & D and we still wouldn't recognize any gaps in instruction. 5-SoC - I don't think the image is totally correct, the way it is drawn, there should also be light rays from the lamp to the fish tank.
- 0. none 1. 2D (missing CCC) 2. yes 3. yes 4. no (does not address absorbed or transmitted - also does not address

the CCC structure)
- 0. none 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes

Item #7
- 0. none 1. 3D 2. yes 3. yes 4. no, students do not have to develop a model
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5.
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. 1. yes 2.yes 3. no-see note 4.yes 5.Whoa...this needs to be redone. If students cannot follow the reading (which is poorly worded with the chosen illustration) they will get it wrong, even if they understand the concept. Perhaps use 3 pictures, instead of trying to describe the three different outcomes. Or do both. It took me 3 times rereading and relooking at the illustration to comprehend what the outcome of each step of the experiment was. You don't need to "color" the material, just show where the light lands...which screens.
- 1-3D 2-phenomenon is ok. 3-no, Anytime 3 student responses are required for 1 score point, it limits the amount of information that can be ascertained from an item score. It would be more useful to score like LS item671. 4-yes 5-Anytime 3 student responses are required for 1 score point, it limits the amount of information that can be ascertained from an item score. It would be more useful to score like LS item671.
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. none 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes

Item #8
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. Yes I agree, but in the stimulus of the problem there is no negative impact stated, only in the drop down toggles for answers. 1. 3D 2. Yes 3. Yes 4. Yes 5. Distractors in this problem lead the problem to be difficult and not deepening the complexity which I feel was the intent. Also, in the stimulus of the problem there is no negative impact stated, only in the drop down toggles for answers.
- 0. none 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0.no it doesn't address the dci 1. no it doesn't address the dci 2 yes-bias-African peoples seem to love or hate mining for personal reasons 3.no-see notes 4.no-see notes 5...really just toss this question. the information is inaccurate. First -melting materials infer heat...not chemicals. Second. using chemicals has to produce emeralds has to have some type of pollution. As well as the "expensive equipment" that is used also had pollution when it was manufactured. the health risks are not listed, but students are told "chemicals" have significant health risks (chromium oxide can cause gene mutations) by omitting facts, the answer is incorrect
- 1-2D, the item does not address the natural resources the synthetic materials come from. 2-although the phenomenon is cool, the item is a poor attempt at addressing student thinking about the topic. It funnels students to one correct choice which may not be supported by global data and lacks information that connects synthetic materials from natural resources. 3-no, the item & rubric only test if students are thinking from a limited world view to get a "right" answer (which may only be corrected based on the limited choices provided.) 4-yes, albeit I don't like this item for reasons in the comments already mentioned. 5-This item contains bias. The answer to this question is subjective. Just because synthetic has less pollution and fewer health risks in the Zambian area doesn't mean it has a positive effect on society. Synthetic emeralds are not likely to be made in Zambia, thus removing jobs from the area causing more poverty in Zambia.
- 0. none 1. 2D (missing CCC) 2. yes 3. yes 4. yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. This question does not align to PS1-3. The standard in question asks students to understand how natural resources are used to create synthetic materials, like food, drugs, or fuels like ethanol and the associated impacts. 1. 3D 2. Yes 3. No. This does not show students' ability to analyze information about natural and synthetic

materials. There are multiple factors to consider and only one thing is considered "correct." 4. No. The standard clearly asks students to evaluate how natural resources are used to create synthetic materials. This question instead asks students to compare natural and synthetic materials but does not explain where the natural materials used to make the synthetic emeralds come from or to ask students to synthesize that information. 5. This question should be reworked with more information about synthetic emeralds and the costs and benefits associated with creating synthetic materials from natural materials.
- 0. none 1. 3D 2. yes 3. no - assertion does not relate well to item. The item does not consider resources needed to make synthetic emerald so negative impacts cannot be considered. The item is incomplete. 4. no -see #3

Item #9
- 0. none 1. 3D 2. yes 3. yes 4. no, students do not have to develop a model
- 0. None 1. 3D 2. Yes 3. Yes 4. No (0. I can not see how the model and data show the difference between reflection and transmission.)
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5.
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. 1.yes 2.yes-like the example of noise pollution 3.no-see note 4.yes 5.all, according to data and info given show noise energy will be reflected (not accounting for any absorption -which is weird?)...in essence causing less transmission. Emissions have nothing to do with evidence given.
- 1-3D 2-phenomenon is ok. 3-yes, with one SoC, students might guess at the first answer because the diagram doesn't support any of the choices given. 4-yes. 5-SoC - why isn't "No effect" a choice? According to the diagram, sounds waves from the vehicle aren't hitting the upper part of the wall anyway, so making the barrier taller shouldn't have an effect on how much sound gets to the house. 5-SoC - why isn't "No effect" a choice? According to the diagram, sounds waves from the vehicle aren't hitting the upper part of the wall anyway, so making the barrier taller shouldn't have an effect on how much sound gets to the house.
- 0. none 1. 3D 2. yes 3. no (the item does not provide enough information for the student to be able to respond correctly. The data does not give the student information on absorption, reflection, and transmission of sound waves through the two substances. Therefore the assertions are not supported.) 4. no (The question does not address absorption.) 5. The question does not provide a student with enough information to respond. The model shows arrows of reflection and transmission but does not show if each substance absorbs, reflects, or transmits the sound waves. Allowing a student to run actual experiment to see how the thickness of a material changes the reflection and the transmission of sound waves would then give the student the information they need in order to respond to the questions being asked.
- 0. none 1. 3D 2. yes 3. no - increasing reflection would also decrease transmission 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. no - the item provides no data about barrier height. 4. no - there is no treatment of absorption of sound by the materials

Item #10
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes 5. No (5. Include the names of both coolers in the second box of criteria to prevent students from think about just the cooler they interact with often. Something like the evaporating and insulating cooler.)
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5.
- 0. none 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. 1.yes 2.yes. 3.yes 4.yes 5.why wouldn't you use a constant to test if it is working...like temperature of the vegetables and not both temp of veggies and how much ice is left? It doesn't make sense.
- 1-3D 2-phenomena is good, the way the question is written about the phenomena is bad. 3-no, because I don't think the question is written to elicit good understanding of planning and carrying out experiments, it cannot tell us accurately what students know and understand. 4-no, ditto #3 5- oh, I have so many problems with this item! Since it is supposed to reflect a scientific experiment, "the temperature of the vegetables after 1 hour" should be the dependent variable being measured for both the evaporative and the insulating cooler. Otherwise, to get the "right" answer implies we are doing bad science by comparing cool, dry vegetables to ice. SoC - the item is written to set

students up to perform science in an incomparable way between the two scenarios = bad science
- 0. none 1. 3D 2. yes 3. yes 4. no (The devices were already designed, and constructed for the student. The student doesn't test the devices either, but simply starts to create a plan to test.)
- 0. none 1. 3D 2. yes 3. yes 4. no- question is not same depth as standard
- 0. None 1. 3D 2. Yes 3. Yes 4. No. The standard ask students to design, construct, and test.
- 0. none 1. 3D 2. yes 3. yes 4. yes

Item #11
- 0. none 1. 3D 2. yes 3. yes 4. no, students do not construct or test the box
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5.
- 0.none 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. 1.yes 2.yes. 3.no 4.yes 5.if you pick remove some food, it would be correct, due to less thermal energy needing to be overcome.
- 1-2D, lacks student use of CCC 2-phenomenon is ok 3-yes, although the item needs more to it to elicit student knowledge of WHY they would add aluminum foil. 4-yes
- 0. none 1. 3D 2. yes 3. yes 4. no (The device is already designed and constructed. The student is not asked to test to find the correct response.)
- 0. none 1. 3D 2. yes 3. yes 4. no-depth of standard is 4, question does not match this depth
- 0. None 1. 3D 2. Yes 3. Yes 4. No. The standard ask students to design, construct, and test
- 0. weak relationship- item does not provide opportunity for design, construct or test 1. 3D 2. yes 3. yes 4. no - does not allow for design or testing

Item #12
- 0. none 1. 3D 2. yes 3. no, diagram does not provide enough information for students to know that there is something blocking the signal or that the distance between towers is long or short 4. no, students could possibly come up with the opposite answer given the missing information from the diagram
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. Yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5. One distractor to the problem in the secondary toggle is a plausible answer that is also stated in the data table above which could leave to a incorrect answer from the given information present to use.
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. 1.yes 2.yes. 3.yes 4.yes 5.
- 1-2D (response doesn't differentiate that students know the content of analog vs digital) 2-phenomenon is ok. 3-no, completing the statement does not provide evidence for supporting a claim. 4-no, the scoring doesn't support that students know the DCI content. 4-The item needs just a little bit more in the student response to gather data that students know A&C are analog signals.
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. none 2. 3D 3. no - unclear from text that light signals are also converted to radio waves 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes

Item #13
- 0. none 1. 3D 2. yes 3. yes 4. no, not sure that the connection to mass is strong enough to meet the standard
- 0. None 1. 3D 2. Yes 3. No (3. I think it is a poorly constructed answers. A student could change variables and it would work to get answers that would address the question.) 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes 5. I could see that this question has some relation to MS-PS3-3. The only way to measure is by measuring the temperature.
- 0. Yes I agree but I had originally thought it was a PS3-3 due to not a strong mass connection. 1. 3D 2. Yes 3. Yes 4. No, the mass connection isn't strong enough. 5.
- 0. none 1. 3d 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0.no. it doesn't include mass in the variables. 1.yes 2.yes. 3.no 4.no...due to not enough mass information 5.could

have more than 2 answers. type of coffee could make a difference because milk, sugar, creamer all have different specific heats. also, if you only put 1 checkmark, you get the answer wrong. Could you make it that the responder HAS to check 2 boxes before they move on.
- 1-2D, the CCC is not connected to in the thinking required by the student 2-phenomenon is ok 3-no, an item with 1 point for needing two correct choices does not allow for different ways of thinking about the problem. 4-yes 5-In the stimulus, "A student runs an experiment to determine which cup design would keep coffee hot for the longest period of time," leaves the item open to interpretation. If I assume the two cups are made of the same material and I change the shape of the cup, I get the item wrong, even if I also select "temp outside of cup". However, the shape is related to design, so should be a viable option.
- 0. The student is testing a device for its ability to minimize thermal energy transfer based on the temperature of the coffee inside the cup. 1. 3D 2. yes 3. yes - the assertion simply states energy transfer, in this case thermal energy. 4. yes
- 0. I thought this would fit MS-PS3-3, but the correct standard discusses change in kinetic energy 1. 3D 2. yes 3. no - shape of cup would affect how much is open to air at top and amount of energy transferred out 4. no - mass is included in standard, but missing from question.
- 0. I disagree with PS3-4. Students are using temperature as a way to measure thermal energy transfer, using coffee cups as a device. 1. 3D 2. Yes 3. Yes 4. No. The standard asks students to design construct, and test a device.
- 0. none 1. 3D 2. yes 3. yes 4. no - the assertion does not correlate well as there is only evaluation and no planning involved.

Item #14
- 0. none 1. 3D 2. yes 3. yes 4. no, students are designing the system but not really testing their variable choices
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. Yes I agree but I had originally thought that it was PS3-4 (planning and investigation) due to no design and testing and constructing. 1. 3D 2. Yes 3. Yes 4. Yes 5.This problem is attempting to be complex and deepen the level of knowledge needed to complete it. But all it is difficult and leading to pathways for incorrect submission. You get initial toggle incorrect and you will then input the wrong independent and dependent variables according to the question filler you choose.
- 0. none 1. 3D 2. Yes 3. Yes 4. Yes 5. I think one beneficial item would be to eliminate one question slots as well as one variable slot as well so students do not have to chose between 4-5 choices
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. 1.yes 2.yes 3.yes. 4.yes 5. the data table isn't needed. none of the info in it is required...it distracts.
- 1-3D 2-phenomenon is ok, although not written in an engaging way 3-no, anytime a rubric has one scoring point for multiple thinking on the student part, it doesn't allow to know what the gaps in student understanding are. (SoC - How would one measure the heat generated by the engine without measuring heat transferred to the surroundings? Those seem to be inherently the same thing.) 4-yes 5-SoC - How would one measure the heat generated by the engine without measuring heat transferred to the surroundings? Those seem to be inherently the same thing.
- 0. The student is being asked to plan an investigation. After discussion, I have changed to the predetermined standard, however, the action the student is being asked to do is to choose variables for an investigation. The student doesn't even need to know the material or information about the devices given to be able to answer this question. 1. 3D 2. yes 3. yes (The student is being asked to pick the variables being tested in an investigation). 4. no (Does not require a student to design, construct, or test a device in any way possible.)
- 0. none 1. 3D 2. yes 3. yes 4. no - standard depth is a 4, question is a 3
- 0. I originally was inclined towards PS3-4, but because this question meets SEP 6 and focuses on heat, I agree with PS3-3 1. 3D 2. Yes 3. Yes 4. No. Students are asked to design and construct a device.
- 0. none 1. 3D 2. yes 3. yes 4. no- assertion does not address design, construct and test a device

Item #15
- 0. none 1. 3D 2. yes 3. yes 4. yes 5. the scoring assertion packs a lot into a single point on this one, given the complexity of this question and the various correct answers, maybe this should be worth more points?
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes 5 No (5. I found the answers system confusing.)
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. yes I agree 1. 3D 2. Yes 3. Yes 4. Yes 5. Problem needs to state that the two people in the problem are using

the same sled. That is not made clear in the directions.
- 0. none 1. 3D 2. yes 3. yes 4. yes 5. I believe this question should have more than one point especially with the difficult and the time needed to answer question.
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. 1.yes 2.no 3.yes 4.yes 5.student may know information (dci) but not vocabulary of scientific method...controlled (in fact in another question you used "constant" instead of controlled), independent, dependent.
- 1-2D, does not engage student understanding of CCC 2-phenomenon is ok 3-yes, but only if students answer based on mass or force. However, students aren't going to know what standard the item is aligned to. 4-no, the score does not reflect CCC application. 5-SoC - although the item is written to align to the standard, a student could actually design a legitimate controlled experiment that wouldn't earn them score points.
- 0. none 1. 3D 2. yes 3. yes 4. no (Does not address mass, not considered a variable.)
- 0. none 1. 3D 2. yes 3. yes 4. no - question addresses forces or mass, standard states forces and mass
- 0. None 1. 3D 2. Yes 3. Yes 4. No. This question does not address the impact of changing mass on an object's motion.
- 0. none 1. 3D 2. yes 3. yes 4. yes

Item #16
- 0. none 1. 3D 2. yes, though the phenomenon seems very intricate and complex and might loose students along the way? 3. yes 4. yes (exceedingly so!) 5. this question overshoots the intended standard and could benefit from shortening... maybe leave out part D
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes 5. Yes (0. The complication of the question seem to exceed the standard.)
- 0. none 1. 3D 2. yes 3. yes 4. yes 5. I do agree with all of the parts of the standard being met, however, I also think that the standard had ben met by Part c. This question is quite large and honestly too overwhelming after part c. The complexity of the questions is not going up it is just adding confusion to this problem. Can this be different questions instead of one large question.
- 0. Yes I agree 1. 3D 2. Yes 3. yes 4. yes 5.This problem is beating the standard to death when it is covered in full complexity at the end of Part C. Its an attempt to create more complexity when really it is creating more difficulty and confusion. It is too long for the level to which the standard is written. Its a poor attempt to stretch a level 2 standard into a level 3 question.
- 0. none 1. 3D 2. Yes 3. Yes 4. Yes, however you overshot the standard by adding too much complexity. 5. That is too many questions just for one page. A student is going to freak out doing this question.
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0.no 1.n/a 2.n/a 3.n/a 4.n/a 5 this is high school content...not middle school HS ps4.1 hsps4.2 hsps4.5
- 1-3D 2-phenomenon is ok, not sure why the cup of water is even mentioned. 3-yes, but it probably better aligns to an HS standard. 4-yes, but it probably better aligns to a HS standard. 5-SoC - this is a lot more here than I would expect for a DOK2 standard. It is also aligned to the HS-PS4-1 standard more than to the MS-PS4-1.
- 0. This is beyond the MS PS4-1 standard. This should be aligned to the HS-PS4-1. MS standard only refers to amplitude. HS standard refers to frequency, wavelength, and speed. 1. 3D 2. no (Information student is asked to digest and provide is not grade level appropriate.) 3. no (The assertions state evidence of mathematical representation of frequency - outside the standard.) 4. no (The scoring requires student to meet expectations beyond the MS standard.) 5. This is beyond the MS PS4-1 standard. This should be aligned to the HS-PS4-1. MS standard only refers to amplitude. HS standard refers to frequency, wavelength, and speed.
- 0. none 1. standard only state how amplitude is related to energy; question includes frequency and wavelength. 2. yes 3. yes 4. no - question is level 3, standard is level 2
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes 5. Part E is confusing in that it asks students to restate the relationship between a change in volume and amplitude in multiple ways. It should be more clear to students that the energy of the wave is being assessed, not its frequency or amplitude.
- 0. none 1. 3D 2. yes 3. yes 4. yes 5. The cluster addresses the standard in first two sections. Repeated questions increase difficulty and may confuse a student. Phenomenon could be tied in more clearly and earlier than section E to avoid some of the confusion.

Table 2 (Grade 8 Batch 52). *Source-of-Challenge Issues by Reviewer*

Sources of Challenge
Item #16 - SoC - this is a lot more here than I would expect for a DOK2 standard. It is also aligned to the HS-PS4-1 standard more than to the MS-PS4-1.

Grade 8 Batch 53

Table 1 (Grade 8 Batch 53). *Notes by Reviewer*

Notes
Item #1 - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. Yes I agree with the intended standard 1. 3D 2. Yes 3. Yes 4. Yes 5. - 0. none 1. 3D 2. yes 3. Yes 4. yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes - 0. 1.yes 2.yes 3.yes 4.yes 5. - 1-2D, doesn't address CCC(cause/effect) 2-phenomenon is ok 3-no, I don't see the connection between the claim and support to limited resources. It presumes students know there is limited amount of water available to humans on earth. 4-no, the item needs more to align to all aspects of the standards. 5-SoC - I don't see how the item asks students to support how the consumption of water impacts Earth's systems. - 0. none 1. 3D 2. yes 3. yes 4. no (Does not address impact on Earth's systems) - 0. Standard states design a method for monitoring human impact, question interprets data to evaluate human impact. 1. 2D - no SEP 6 (constructing explanation present, designing solutions absent) 2. yes 3. yes 4. no - question does not design a method for monitoring and minimizing human impact - 0. None 1. 3D 2. Yes 3. Yes 4. Yes - 0. none 1. 3D 2. yes 3. yes 4. yes
Item #2 - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. Yes I agree with the intended standard 1. 3D 2. Yes 3. Yes 4. Yes 5. Instead of saying formation of large swamps I would rephrase that statement to say presence of large swamps as the question is not asked to be labeled in a sequence. It may also lead a bubble understanding student to think the formation should be in the second model not all in the first model. - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes - 0. 1.yes 2.yes 3.yes 4.yes 5. - 1-3D 2-phenomenon is good. 3-yes 4-yes - 0. none 1. 3D 2. yes 3. no (The first assertion students will choose large swamps form is inaccurate. Students who have no knowledge of how coal forms, will choose this only if they recognize that the large swamp areas in figure 1 overlap the coal mining areas in figure 2. The standard does not require students to know how all of Earth's minerals are formed. 4. yes (although it does focus only on one of three focuses (mineral, energy, and groundwater) but that is reasonable. - 0. none 1. 2D - missing SEP6 (constructing explanations present, no designing solutions) 2. yes 3. no - students may interpret swamps as shallow seas, so shallow seas recede might be chosen 4. no- only addresses mineral resources; standard states mineral, energy, and groundwater resources

- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes

Item #3
- 0. none 1. 3D 2. yes 3. yes 4. no, students did not develop the model that they used to form their description, also doesn't tightly connect the rotation idea to the answer key
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes 5. No (5. I felt the answering structure was confusing.)
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. Yes I agree with the intended standard 1. 3D 2. Yes 3. Yes 4. Yes 5. The question is vague in the model descriptors for students to make inferences from. I know the intent of this is a show that students know how air masses move. But I feel its to vague to be equitable for all students to analyze the given models.
- 0. none 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. 1.yes 2.yes 3.yes 4.yes 5.
- 1-3D 2-phenomenon is ok 3-yes 4-yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. none 1. 3D 2. yes 3. yes 4. no - standard DOK is level 2; I believe this question is level 3
- 0. None 1. 3D 2. Yes 3. Yes 4. No. The question does not address changes in oceanic circulation
- 0. none 1. 3D 2. yes 3. yes 4. yes

Item #4
- 0. none 1. 3D 2. yes 3. no, the scoring seems to be missing some possible combinations on cause and effect relationships with putting cause and effect in correct order 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. Yes I agree with the intended standard 1. 3D 2. Yes 3. Yes 4. Yes 5. Long and repeats concepts on a similar level. Length of problem does not deepened complexity.
- 0. none 1. 3D 2. yes 3. yes 4. yes 5. I felt that some of the parts were just being recovered again.
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes 5. Very in depth question
- 0. 1.Yes 2.yes 3.no-see notes 4.yes 5. If a student gets one wrong in the sequencing portion (part c), they could lose up to 3 of the 11 points. It is likely that once one is sequenced wrong, the others that follow will be wrong, just due to the wrong order. The evidence citing is misleading too. If I pick vegetation + energy and reflection option because there is a relationship. I get it wrong. the only "right" answer is the one for the "reason" (this is for ALL evidence questions) which doesn't support the ccc of consequences action...the test only wants the effect. I'm don't think it accurately shows student knowledge.
- 1-3D 2-phenomena is ok 3-yes, although some of the scoring criteria are not flexible for other lines of thinking 4-yes 5-Part C, step 3, if a student selects one incorrect answer they lose 2 points. This should not work this way. Also, cutting down trees may be necessary to build more roads which causes an increase in air and water temps (that should work too, but does not score as "right")
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. No. Parts C and E are in need of serious attention. In part C, it is logical for a student to select "cut down trees" followed by "build roads," and it is unclear that "stream ecosystems are disturbed" is the logical endpoint, as an increase in temperature is also a disruption to earth's systems. 4. Yes 5. Part C should be reworked as explained above, either by simplifying the answer options or by making it clear that the intended end is disruption of stream ecosystems. Part E should be made more clear that students are evaluating how the evidence applies to Part D, NOT whether the statements are supported the provided evidence.
- 0. none 1. 3D 2. yes 3. yes 4. yes

Item #5
- 0. none 1. 3D 2. yes 3. yes 4. no, students are only looking at the lunar phases portion of the standard (no eclipses or seasons present)
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes 5. No (5, I did not feel that there was enough evidence to answer the question correctly.)
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. Yes I agree with the intended standard, but it only covers the moon phases, not circular pattern, or seasons. 1. 3D 2. Yes 3. Yes 4. Yes 5. Poor attempt to cover the full standard.

- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. 1.yes 2.yes 3.yes 4.yes 5.
- 1-3D 2-phenomenon is ok 3-yes 4-no, the item doesn't elicit information about patterns, although it is infered
- 0. none 1. 3D 2. yes 3. yes 4. no (Does not address eclipses or seasons).
- 0. none 1. 3D 2. yes 3. yes 4. no - eclipses and seasons are in standard, not addressed in question.
- 0. None 1. 3D 2. Yes 3. Yes 4. No. Question does not address eclipses or seasons.
- 0. none 1. 3D 2. yes 3. yes 4. yes

Item #6
- 0. none 1. 3D 2. yes 3. yes 4. no, students are only dealing with the eclipse portion of the standard and not the moon phases or seasons portions)
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. Yes I agree with the intended standard 1. 3D 2. Yes 3. Yes 4. Yes 5.
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. 1.yes 2.yes 3.yes 4.yes 5.
- 1- 3D 2-phenomenon is good 3-yes 4-yes
- 0. none 1. 3D 2. yes 3. yes 4. no (Does not address, lunar phases or seasons)
- 0. none 1. 3D 2. yes 3. yes 4. no - lunar phases and seasons are included in the standard but are not addressed in the question
- 0. None 1. 3D 2. Yes 3. Yes 4. No. The question does not address seasons or lunar phases.
- 0. none 1. 3D 2. yes 3. yes 4. yes

Item #7
- 0. none 1. 3D 2. yes 3. yes 4. no, students are not developing the model that they are using to describe the phenomenon
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. Yes I agree with the intended standard 1. 3D 2. Yes 3. Yes 4. Yes 5. Dotted lines to show a greater difference in lengths for students to apply the idea of the greater the gravity the greater the speed.
- 0. none 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. 1.yes 2.yes 3.yes 4.yes 5.great question with diagrams
- 1-3D 2-phenomenon is good! 3-no, selecting the correct answer does not indicate that students know the strength of the force of gravity has anything to do with speed of the moons. 4-yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. none 1. 3D 2. yes 3. yes 4. no - standard states motions within galaxies and solar system; question addresses only solar sysetem.
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes

Item #8
- 0. I see the standards match here but it may be a bit out of line since the standard says "within galaxies and the solar system" but the question is between galaxies in the universe 1. 3D 2. yes 3. yes 4. no, see my comment above.., the standard says "within galaxies", not between galaxies, students are also not developing the model that they are using for their description
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes 5. No (5. I felt there was not enough information to answer this.)
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. Yes I agree with the intended standard 1. 3D 2. Yes 3. Yes 4. Yes 5.
- 0. none 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. 1.yes 2.yes 3.yes 4.yes 5. I don't like that mass isn't included and speed is. The two factors that affect gravity are mass and distance from (or to) an object. NOT speed.
- 1-2D, not really applying the CCC of systems, although it is inferred 2- phenomenon is good 3-no, the assertion really only indicates students know that things closest to the black hole orbit faster than those far away. In order to

make a relationship to gravity, mass of the objects (BH, sun, CasA, crabN) would also be included. 4-no, students are not describing the role of gravity, they are making predictions based on distance
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. none 1. 3D 2. yes 3. yes 4. no - standard states motions within galaxies and solar system; question addresses only galaxies.
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes

Item #9
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. Yes I agree with the intended standard 1. 3D 2. Yes 3. Yes 4. Yes 5. You have 6 toggles for 2 points. Not equitable. I know in chart one toggle will influence the secondary one. But that concept would lead the problem to be worth 3 points not 2.
- 0. none 1. 3d 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. 1.yes 2.yes 3.yes 4.yes 5. I don't like that mass isn't included and speed is. The two factors that affect gravity are mass and distance from (or to) an object. NOT speed.
- 1-3D 2-phenomenon is ok. 3-no, I'm not convinced that a student selecting the correct arrows is related to knowing how gravity affects speed as it can also be determined simply by knowing the distance from the sun. 4-yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. none 1. 3D 2. yes 3. yes 4. no standard states motions within galazies and the solar system; question only addresses solar system.
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes

Item #10
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 2D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. Yes I agree with the intended standard 1. 3D 2. Yes 3. Yes 4. Yes 5.
- 0. none 1. 3D 2. Yes 3. Yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. 1.yes 2.yes 3.yes 4.yes 5.
- 1- 3D 2-phenomenon is good 3-yes 4-yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. I chose the correct standard, however I think it could also fit MS-ESS3-1 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes

Item #11
- 0. none 1. 3D 2. yes 3. yes 4. no, students did not develop the model that they used to describe, and they are only addressing the eclipse portion of the standard rather than the lunar phases and seasons
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. Yes I agree with the intended standard 1. 3D 2. Yes 3. Yes 4. Yes 5.
- 0. none 1. 3d 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. 1.yes 2.yes 3.yes 4.no-see notes 5. the answer rationale states it is a total eclipse. However, the information in the question talks about the timeframe and rust color (which isn't ALWAYS seen with a total eclipse-especially that long of timeframe) might indicate that it's a total eclipse but that's not the standard to indicate...it's to use a model to show patterns of events.
- 1-2D, there needs to be an extension that relates to student thinking in order to know if students recognize why this phenomenon is a pattern because otherwise just reading the prompt can give them the answer. 2-phenomenon is good. 3-no, students can answer this question just given the information in the stimulus not needing to describe

or use patterns. 4-no, same reason as #3.
- 0. none 1. 3D 2. yes 3. no - the assertion is that the student will be able to recognize that the description is referring to a lunar eclipse but the information in the question tells the student they can see the red rusty moon for 3.5 hours. This would not be a description of a total lunar eclipse as the assertion states. 4. no - does not address phases and seasons.
- 0. none 1. 3D 2. yes 3. yes 4. no - standard states cyclic patterns of lunar phases, eclipses, and seasons; question does not address seasons
- 0. This question is poorly aligned with the standard. While a red moon can occur around a total eclipse, it is not clear that solar eclipses are being assessed in the question as currently written. 1. 3D 2. Yes 3. No. The diagram depicted does not make it clear that light is passing through the Earth's atmosphere, only showing shadow and light. Because of this, it would be difficult for students to choose the correct image. 4. No. The color of the moon is not relevant to how a total solar eclipse occurs. 5. This question should be reworked so that the depiction of the Earth in the diagram shows light passing through it, or removed because the color of the moon is not relevant to a students' understanding of a total solar eclipse. 4.
- 0. none 1. 3D 2. yes 3. yes 4. yes

Item #12
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. Yes I agree with the intended standard 1. 3D 2. Yes 3. Yes 4. Yes 5.
- 0. none 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. 1.yes 2.no-see notes 3.yes 4.yes 5. students with language barriers may not understand "tick" therefore they will get the question wrong based on nonscience information
- 1-2D, lacks DCI application, although the item is written around DCI content, a student only needs to elicit math concepts to answer the questions 2-phenomenon is ok. 3-yes 4-no, lacks rubric alignment to DCI
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. none 1. 3D 2. yes 3. yes 4. yes 5. Calculator will not show answer for division, but will count calculation as correct.
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes

Item #13
- 0. none 1. 3D 2. yes 3. yes 4. no, students are not developing the model that they use to describe, and question only addresses the seasons portion of the question rather than lunar phases and eclipses
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. Yes I agree with the intended standard 1. 3D 2. Yes 3. Yes 4. Yes 5.
- 0. none 1. 3D 2. Yes 3. Yes 4. Yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes
- 0. 1.yes 2.yes 3.yes 4.yes 5. great question because the distractors each mean something...the sun is either: higher/lower, bigger/smaller, same size and level or moon vs sun...so you can tell what the student isn't understanding.
- 1-3D 2-phenomenon is good 3-yes 4-yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. none 1. 3D 2. yes 3. yes 4. no - standard states dydlic patterns of lunar phases, eclipses of the sun and moon, and seasons; question only addresses seasons.
- 0. None 1. 3D 2. Yes 3. Yes 4. No. Question does not include lunar phases or eclipses. 5. The distractor question that includes the moon in winter is inappropriate for students at this level.
- 0. none 1. 3D 2. yes 3. yes 4. yes

Item #14
- 0. none 1. 3D 2. yes 3. no, students seem to be able to get the entire question correct as long as they understand placement of a new moon 4. no, only addresses the phases of the moon portion of the standard 5. the quarter moon diagram is backwards and the dates do not reflect the correct timing for the moon cycle
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes 5. No (5. Image does not agree with the actual patterns of the moon.)

- 0. none 1. 3D 2. yes 3. yes 4. yes - 0. Yes I agree with the intended standard 1. 3D 2. Yes 3. Yes 4. Yes 5. - 0. none 1. 3D 2. Yes 3. Yes 4. Yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes 5. There are some problems with the graphics. The graphic for Dec. 16th should be flipped. The visible moon should be on the right not the left. The dates of the phases are incorrect - 0. 1.yes 2.yes 3.yes 4.yes 5. - 1-3D, but because of the way the diagram choices are ordered, a student really only needs to look for the diagram of the new moon first and doesn't need to analyze beyond making that choice. 2-phenomenon is good. 3-no, because of the way the diagram choices are ordered, a student really only needs to look for the diagram of the new moon first and doesn't need to analyze beyond making that choice. 4-yes - 0. none 1. 3D 2. yes 3. yes 4. yes 5. The image of the moon for Dec. 16th is wrong. The correct answer shows the moon in the 1st quarter position, so the image should show the right half of the moon with light and left side of the moon dark. - 0. none 1. 3D 2. yes 3. no - view of moon does not match Dec 16 position shown in correct response. If moon were in this position, right half would be lit as viewed from the earth. 4. no - standard includes lunar phases, eclipses, and seasons; question only addresses lunar phases. - 0. None 1. 3D 2. Yes 3. Yes 4. No. Does not address seasons. - 0. none 1. 3D 2. yes 3. yes 4. yes
Item #15 - 0. standard 2-1 and 2-4 are similar in nature though 2-4 seems to allude more directly to water where glaciers may be taught than to the more generic "Earth's materials" from 2-1 1. 3D 2. yes 3. yes 4. no, does not address gravity at all (even as a distractor) - 0. None 1. 3D 2. Yes 3. Yes 4. Yes - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. Yes I agree with the intended standard 1. 3D 2. Yes 3. Yes 4. Yes 5. - 0. none 1. 3D 2. Yes 3. Yes 4. Yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes - 0. 1.yes 2.yes 3.yes 4.yes 5. - 1-3D 2-phenomenon is good 3-yes 4-yes - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. I chose the correct standard; however the question does not really concern cycling of Earth's materials. 1. 3D 2. yes 3. yes 4. no- question addresses energy driving a process, but not cycling of materials as standard states. - 0. None 1. 3D 2. Yes 3. Yes 4. Yes - 0. none 1. 3D 2. yes 3. yes 4. yes

Table 2 (Grade 8 Batch 53). *Source-of-Challenge Issues by Reviewer*

Sources of Challenge
[NONE]

Grade 11 Batch 54

Table 1 (Grade 11 Batch 54). *Notes by Reviewer*

Notes
Item #1 - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. 1.yes 2.yes 3.yes 4.yes 5. - 1-3D 2-phenonomen is good 3-yes 4-yes - 0. The stem does not mention any factors about the two different parts of the park that would explain the difference in biodiversity and populations. The correct response specifically states carrying capacity. 1. 3D 2. yes 3. no - The information given does not allow the student to distinguish factors that affect the biodiversity. Given the data, the student would choose location A as having a higher carrying capacity simply due to the larger number of plants found there. 4. no - The question does not provide for the student or have the student identify specific factors that would affect the carrying capacity. - 0. none 1. 3D 2. yes 3. no - I think number of species present should be used instead of species richness 4. yes - 0. None 1. 3D 2. Yes 3. Yes 4. No. Species richness is a rote memorization that not all students should be expected to know. Having one question rely on students' mastery of this definition that is not addressed in the standard 5. Species richness should be included in this question as a provided definition.
Item #2 - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. 1.yes 2.yes 3.yes 4.yes 5. - 1-3D 2-phenomenon is ok 3-yes 4-yes - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes
Item #3 - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. 1.yes 2.yes 3.yes 4.no 5.students do not need to ask the inheritance question because the pedigree SHOWS it, if the pedigree is used correctly. - 1-2D, missing CCC application 2-phenomenon is good 3-yes 4-yes to DCI and SEP - 0. none 1. 3D 2. yes 3. yes 4. no - does not address DNA and chromosomes in coding the instructions for characteristic traits. - 0. none 2. 3D 3. yes 4. yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes
Item #4 - 0. none 1. 3D 2. yes 3. yes 4. no, the scoring assertion only address that light and carbon dioxide are involved, it does not address the process of converting to chemical energy - 0. 1.yes 2.yes 3.yes 4.yes 5. - 1-2D, doesn't quite get to the content of the DCI about transforming light energy into stored chemical energy. 2-phenomenon is ok 3-yes, except if a student shows only partial understanding the data isn't going to indicate what the gap is. Item analysis would be needed to reveal that. I don't think that level of analysis is available on reports. Thus, this should probably be a 2pt item. 4-no, doesn't get to the DCI of how photosynthesis transforms light energy into stored chemical energy, but it is on the very lower edge of it. - 0. none 1. 3D 2. yes 3. yes 4. no - does not address stored chemical energy. - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes
Item #5 - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. 1.yes 2.yes 3.yes 4.yes 5. - 1-3D, although weak as to how the student in engaging in argument & using cause/effect 2-phenomenon is good 3-yes 4-no, without more for the student to do other than pick 1 right answer it is hard to tell if the student is engaging in SEP application

- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Item #6
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. 1.yes 2.yes 3.yes 4.yes 5.
- 1-2D, because a correct answer may not have to do with the structure/function of the gene. 2-phenomenon is not complete if ProteinC is important 3-no, due to distractor A being a plausible answer 4-no, due to structure/function may not be necessary to reasonably answer
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Item #7
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. 1.yes 2.yes 3.yes 4.yes 5.
- 1-2D, the structure/function CCC is lacking for engagement. 2-no, it's a typical cell analogy project poorly executed. 3-if the purpose it to simply illustrate hierarchal organization then it might be ok 4-no, there should be a relationship to structures with functions, not just a list from smallest to largest 5- An item should NOT be written where both columns are out of order. If the intent is to analogize to levels of organization then the levels should be listed in order.
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. none 2. 3D 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Item #8
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. 1.yes 2.yes 3.yes 4.yes 5.
- 1-3D 2-phenomenon is ok 3-yes, but the question isn't written to clearly align with any HS-LS standard 4- no because the question isn't written to clearly align with any HS-LS standard
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Item #9
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. 1.yes 2.yes 3.no 4.yes 5. if a student picks "a change in the shape of the amino acid protein -which could be accurate- and put it in the middle box instead of a change in the sequence, both points would be lost, even though it may be correct.
- 1-3D 2-phenomenon is good 3-yes 4-yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. I don't see how this protein carries out essential functions of life 1. 3D 2. yes 3. no - I think students may select a change in the RNA sequence as the starting point 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Item #10
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. 1.yes 2.yes 3.yes 4.yes 5.
- 1-3D 2-phenomenon is good, although the graph doesn't totally support the claim because the trend in ratio of male:females is getting better from subadults to juveniles. 3-yes, but see comments on #2. 4-yes, but see comments on #2.
- 0. none 1. 3D 2. yes 3. yes 4. yes 5. The graph is confusing. We have ages that overlap. We have data given that is not assigned to juvenile, subadult, or adult (look at the data bars on the far right side of the graph). Students are not told when turtles reproduce (adult, subadult, and/or juvenile).
- 0. I thought this was also appropriate to HA-LS3-3 1. 3D 2. yes 3. yes 4. no- standard states increases in number of species and emergence of new species; question only addresses extinction of a species

- 0. I first thought LS2-2 was a better fit, but I now agree with LS4-5 1. 3D 2. Yes 3. Yes 4. No. Does not address increase or emergence of new species.

Item #11
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. 1.yes 2.yes 3.yes 4.yes 5.
- 1-3D 2-phenomenon is good 3-yes 4-yes 5-I like how there are 4 score points because the data will then show where misconceptions occur. The item would be made awesome if there were one more column that allowed for students to select choices of why they think each factor decreases. (Facilitate making student thinking visible!)
- 0. none 1. 3D 2. yes 3. yes 4. no - does not address all the inputs and outputs, anaerobic, nor the bonds broken and formed in the process.
- 0. none 1. 3D 2. yes 3. no - I think an assumption is made that less food is consumed at night and less oxygen is consumed, but is not evident from the data presented 4. yes
- 0. I originally chose LS1-3, thinking of the temperature change as a homeostatic mechanism, but I can understand why LS1-7 was chosen. 1. 3D 2. Yes 3. Yes 4. Does not include anaerobic respiration and does not include all inputs and outputs of cellular respiration.

Item #12
- 0. none 1. 3D 2. yes 3. yes 4. yes 5. calculator does not have an equal (=) function key
- 0. 1.yes 2.yes 3.yes 4.yes 5.
- 1-3D 2-phenomenon is good 3-yes 4-yes
- 0. none 1. 3D 2. yes 3. yes 4. no - does not address cycling of matter
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. No. Does not include flow of matter 5. This question is a weak DOK 2 an exercise in applying a vocabulary-level concept, the rule of 10.

Item #13
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. 1.yes 2.yes 3.no 4.yes 5. students could say another type of system, like nervous system because they use calcium too, yet would get the question wrong.
- 1-3D 2- phenomenon is good, analogy is unnecessary 3-yes 4-yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Item #14
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. 1.yes 2.yes 3.no 4.yes 5. there are 2 correct answers. one is "the available H, O, and C from glucose" but the other is.." the available energy" because when they break the glucose apart to get the H, O, C, it releases that energy. sugar=potential energy
- 1-3D 2-phenomenon is good 3-yes, but it is a stretch to think that students make the connection to "because the bacteria rearrange the molecules in glucose to make PIA." This could only be inferred because the other distractors are unreasonable. 4-yes.
- 0. none 1. 3D 2. yes 3. yes 4. yes 5. Students may leave the scientist's claim as is since the biofilm does increase and will trust that a "scientist" will know the reason why. Removing this option would remove the chance of a student getting it wrong for the wrong reason.
- 0. none 1. 2D - SEP6 - constructed explanation but no solution designed 2. yes 3. yes 4. yes
- 0. None 1. 3D. 2. Yes 3. Yes 4. Yes

Item #15
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. 1.yes 2.no, frond has a geographical bias. 3.yes 4.yes 5.
- 1-3D 2-phenomenon is good 3-yes, but very weak to presume that students are "using this information to make predictions about the frequency in a population in the presence of selection pressure in the environment," because they are probably just using mathematical reasoning as there is only 1 correct answer based on the data. This MC-form does not allow for making student thinking about the reason WHY visible. 4-no, for same reason as #3. The item only tells us a student can mathematically reason through the choices.
- 0. none 1. 3D 2. yes 3. yes 4. yes

- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Item #16
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. 1.yes 2.yes 3.yes 4.yes 5.
- 1-3D 2-phenomenon is good 3-yes 4-yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. none 1. 3D 2. yes 3. no - I think tail structure is as valid as paw structure 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Item #17
- 0. none 1. 3D 2. yes 3.yes 4. yes
- 0. 1.yes 2.yes 3.yes 4.yes 5.
- 1-1D, only mathematical reasoning (SEP4) is required to answer this question. 2-phenomenon is ok 3- yes, but it is a stretch to assume that students "understands that environmental conditions can lead to natural selection against certain forms of a species," because the item can be answered by just applying mathematical reasoning from table 1. No engagement of thought about the DCI or CCC is required. The SEP is probably more of 4 than it is of 7. 4-no, see comments on #3.
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. none 1. 3D 2. yes 3. yes 4. no-standard discusses emergence of new species and extinction of species; question only addresses increases in number of individuals in a species
- 0. This question does not align to LS4-5, which assesses students knowledge of whether species will go extinct, form or the increase in members of a species. This question relates to the traits within a single species. 1. 3D 2. Yes 3. No. The scoring assertions refer to "forms" of a species. This is related to a specific trait, not a form. 4. Yes

Item #18
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. 1.yes 2.yes 3.yes 4.yes 5. I like how each one is labeled individually, so one answer doesn't affect the rest of the answers.
- 1-3D 2-phenomenon is good 3-yes 4-yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Item #19
- 0. none 1. 3D 2. yes 3. yes 4. no, the scoring assertion does not reflect the breadth and depth of the corresponding standard. The assertion shows protein relatedness but does not include multiple lines of evidence in support of common ancestry and biological evolution.
- 0. 1.yes 2.yes 3.yes 4.yes 5.
- 1-2D because it is a stretch to say this item is directly related to either 4-1 common ancestry and biological evolution or to 1-1 (structure of DNA determines the structure of proteins which carry out the essential functions) 2-no phenomenon is present to introduce the table of information 3-no, "What is the amino acid sequence?" is a reasonable question no matter which protein is selected for the answer. This item score does not reveal anything about student thinking by the way it is scored. 4-no, for reasons noted on #3.
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. I originally thought LS1-1 was better aligned, but I agree with LS 4-1 1. 3D 2. Yes 3. Yes 4. Yes

Item #20
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. 1.yes 2.yes 3.yes 4.yes 5.
- 1-3D 2-phenomenon is ok, although the diagram is confusing having all the steps connected to each other. Each step should be separated by a little space. 3-no, the blastema wasn't present in step 1,2,3 so wouldn't it becoming present in step 4 be evidence of differentiation occurring? 4-yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Item #21
- 0. none 1. 3D 2. yes 3. yes 4. no - scoring assertion includes only evaluation and does not include design or refinement of the solution
- 0. 1.yes 2.yes 3.yes 4.yes 5.poorly written, needs major revamp
- 1-2D, does not engage the CCC 2-phenomenon is ok 3- no, there is not enough known/given about hatchlings and there are too many clicks necessary to "get correct" for one score point. Even after reading "the answer key" I had trouble getting the score point. There is no way of telling what a student was thinking by clicking the choices that they have to in order to earn 1 point. 4-no, for reasons listed in #3
- 0. none 1. 3D 2. yes 3. no - There is not enough information given for students to accurately predict the responses indicated. For example, for the light covers, the student can select all three OR not enough information and get the answer correct. If that is acceptable, then there isn't enough information given to have a "right" answer. 4. yes 5. A student has to get click all correct boxes for all three scenarios in order to get the point. There is not enough information given for the student to be able to know the "right" answers for each scenario.
- 0. Also fits with HS-LS4-6 1. 3D 2. yes 3. no - choice A could also be correct given the distractor 4. no - not designing, only evaluate and refine; Standard is DOK level 4, question is level 2
- 0. None 1. 3D 2. Yes 3. Yes 4. No. Students do not design or refine.

Item #22
- 0. none 1. 3D 2. yes 3. yes 4. no - does not include solution design
- 0. 1.yes 2.yes 3.yes 4.yes 5.
- 1-2D - this item needs a bit of extension to engage the SEP of constructing explanations 2-phenomenon is ok 3-no, for reasons in#1 4-no, for reasons in #1
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. also fits standard HS-LS4-6 1. 3D 2. yes 3. yes 4. no - not designing solution/Standard is DOK level 4, question is level 2
- 0. None 1. 3D 2. Yes 3. Yes 4. No. Students do not design or refine.

Item #23
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. 1.yes 2.yes 3.yes 4.yes 5.
- 1-3D 2-phenomenon is good 3-no, if only one line of evidence is necessary to get the correct answer (male sex chromosomes) then there is no point to have the other lines of evidence present. This is deceiving because all the lines of evidence (except habitat) relate to DNA, and even the beaver shares similar sex chromosomes. Thus, this defeats the point of "supported by multiple lines of evidence" in the standard. 4-yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. none 1. 3D 2. yes 3. yes 4. no - standard state "supported by multiple line of empirical evidence"; question only uses one evidence.
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Item #24
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. 1.yes 2.yes 3.yes 4.yes 5.great question
- 1-3D 2-phenomenon is good 3-yes 4-yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Item #25
- 0. none 1. 3D 2. yes 3. yes 4. no - item does not address design or refinement of solution; evaluation and identification of potential issues only
- 0. 1.yes 2.yes 3.yes 4.yes 5.
- 1-3D 2-phenomenon is good 3-no, not reasonable answer key given the facts of the situation 4-no, given the SoC
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. none 1. 3D 2. yes 3. yes 4. no- standard includes designing and refining solution, question is only evaluating solution; question is DOK level 4, question is level 2.
- 0. None 1. 3D 2. Yes 3. No. "Biodiversity decreasing faster in areas outside the refuge" assumes that by creating a protected area, trawling will increase in other areas, which is not necessarily true. A better option would be

"biodiversity will continue to decline in areas outside the refuge." 4. No. Students do not design or refine their solution.
Item #26 - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. 1.yes 2.yes 3.yes 4.yes 5. - 1-1D, only mathematical reasoning is needed to answer this item. 2-phenomenon is good. 3-no, this item can be answered with mathematical reasoning, not application of the DCI 4-no, this is an SEP of 4, not 7 - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. I thought it would fit better with HS-LS4-5, but because the number of species is increased and not individuals within a species it fits with this standard. 1. 3D 2. yes 3. yes 4. no-standard state under stable conditions numbers and types of organisms maintain consistent numbers; question has no stable conditions in it. - 0. LS2-6 does not apply to this question. While the prompt of the question represent the process of biological succession, disturbance is not present. In this question, students are using a mathematical representation (a graph) to explain how the age of a forest and the height of the trees affect biodiversity. 1. 3D 2. Yes 3. No. The scoring assertion 1 refers to stability in an ecosystem. This ecosystem is not stable, it is changing over time. 4. No, The scoring assertions do not support the standard.
Item #27 - 0. none 1. 3D 2. yes 4. yes 5. no - scoring assertion does not make an inference regarding the emergence or extinction of species - 0. 1.yes 2.yes 3.yes 4.yes 5. - 1-1D, only mathematical reasoning is needed to answer this item. 2-phenomenon is good. More engagement with questions about the phenomenon is needed to make it a 3D item (or bundle) 3-no, only mathematical reasoning is needed to answer this item. The rubric makes assertions about the DCI, SEP, and CCC that are not necessary to answer the item. 4-no, this is an SEP of4, not 7. - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. noone 1. 3D 2. yes 3. yes 4. no - standard includes emergence of new species and extinction of species; question only discusses increases in number of individuals of a species. - 0. I disagree with LS4-5. LS4-5 asks about the process of speciation and the potential for the increase in a population over time. This question is asking about how environmental conditions may affect carrying capacity. 1. 3D 2. Yes 3. No. The population would actually be increasing in all situations, since the population growth rate is positive in all conditions listed in table 2. See #5. 4. Yes. 5. Table 2 in this question needs to seriously reworked. There is a POSITIVE growth RATE for all three conditions. Growth rate (r), defined as r = births-deaths/total population, is only positive if a population is increasing. Further, a population growth rate of 2 suggests a doubling of the population each year, which is highly unlikely. The answer options should be changed to ask about population growth rate instead of population AND the table should be updated to provide more realistic values, perhaps as percentages.
Item #28 - 0. none 1. 3D 2. yes 3. yes 4. no - uses only one line of evidence; multiple lines of evidence are not considered - 0. 1.yes 2.yes 3.yes 4.yes 5. - 1-2D, only looking at a pattern in DNA sequence (CCC) is needed to answer this item. all the rest of the item is irrelevant information. 2-not sure what the phenomenon is because it is not told to the students what kind of organism "C. unicinctus and several other related species" are. 3-yes 4-yes - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. none 1. 2D - SEP8 - did not obtain information 2. yes 3. yes 4. no - standard states multiple lines of evidence; question uses only DNA evidence - 0. None 1. 3D 2. Yes 3. Yes 4. Yes
Item #29 - 0. none 1. 3D 2. yes 3. yes 4. no - evidence is simply identified and not evaluated - 0. 1.yes 2.yes 3.yes 4.yes 5. - 1-3D 2-phenomenon is good. 3-yes 4-yes - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. none 1. 3D 2. yes 3. yes 4. no - standard includes increase of individuals in a species and extinction of a

species; question only addresses emergence of a new species - 0. None 1. 3D 2. Yes 3. Yes 4. No. Only addresses emergence of new species
Item #30 - 0. none 1. 3D 2. yes 3. yes 4. yes 5. calculator has no equal (=) function key - 0. 1.yes 2.yes 3.yes 4.yes 5. - 1-2D, no engagement of CCC to answer the item 2-phenomenon is ok 3-yes 4-yes, minus the CCC is not evaluated - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. none 1. 3D 2. yes 3. yes 4. yes 5. Cannot work calculator - 0. None 1. 3D 2. Yes 3. Yes 4. No. Does not address cycling of matter 5. This question is a rote use of the rule of 10 and a routine calculation.
Item #31 - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. 1.yes 2.yes 3.yes 4.yes 5. - 1-2D, no SEP model is actually needed to answer this question, it is based on applying rote memory about respiration. 2-phenomenon is ok 3-yes 4-yes, minus the SEP is not evaluated - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. None 1. 3D 2. Yes 3. Yes 4. No. Does not include aerobic respiration
Item #32 - 0. none 1. 3D 2. yes 3. no - simply calculating the mass of sea turtles in the first response does not support the inference stated in the scoring assertion 4. no - simply calculating the mass of sea turtles in the first response does not support the inference stated in the scoring assertion 5. calculator does not have an equal (=) function key - 0. 1.yes 2.yes 3.yes 4.yes 5. - 1-3D 2-phenomenon is ok 3-yes 4-yes - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. none 1. 3D 2. yes 3. yes 4. yes 5. Cannot work calculator - 0. None 1. 3D 2. Yes 3. Yes 4. Yes
Item #33 - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. 1.yes 2.yes 3.yes 4.yes 5. I feel that the first step should include geosphere because it's getting nutrients in the soil. - 1-3D 2-phenomenon is good 3-yes 4-yes - 0. none 1. 3D 2. yes 3. no - the first step ignores the requirement of water for photosynthesis - thus requiring the student to either mark hydrosphere. 4. yes - 0. none 1. 3D 2. yes 3. yes 4. yes 5. Step 3 in question does not mention carbon dioxide in atmosphere - 0. None 1. 3D 2. Yes 3. Yes 4. Yes
Item #34 - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. 1.yes 2.yes 3.yes 4.yes 5. I don't like that if a student makes a mistake in the sequencing, it could affect multiple points, instead of really just being 1 off. - 1-3D 2-phenomenon is ok 3-yes 4-yes - 0. none 1. 3D 2. yes (hectare should be defined) 3. yes 4. yes - 0. none 1. 3D 2. yes 3. yes 4. no - Standard is DOK level 2, but question is level 3. - 0. None 1. 3D 2. Yes 3. No. Part C is difficult to complete. Plants do use cellular respiration, but that step is considered incorrect. 4. Yes
Item #35 - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. 1.yes 2.yes 3.yes 4.yes 5. - 1-3D 2-phenomenon is good. 3-yes 4-yes - 0. none 1. 3D 2. yes 3. no - the assertions do not match the assumed knowledge the student has. 4. no - goes beyond the standard. 5. The information given and what a student is asked to infer is beyond what is reasonable. - 0. none 1. 3D 2. yes 3. yes 4. no - Standard is DOK level 2, but question is level 3.

- 0. This question is not aligned to the standard LS3-2 1. 3D 2. Yes 3. No. Part C's claim that all are divisible by 7 is vague and could be chosen for the wrong reasons. 4. No. None of the assertions listed here actually relate back to LS3-2. Students answering these questions correctly has no bearing on their mastery of this standard. 5. This question should be removed.
Item #36 - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. 1.yes 2.yes 3.yes 4.yes 5. - 1-3D 2-ok 3-yes 4-yes - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes
Item #37 - 0. none 1. 3D 2. yes 3. yes 4. no - the scoring assertions do not describe the inferences that can be made from a student's successful interaction with the item. The item only asks the student to make an observation without supporting and revising an explanation of the factors that affect the population as the standard states. - 0. 1.yes 2.yes 3.yes 4.yes 5. - 1-1D, only mathematical reasoning is needed to answer the questions 2-phenomenon is good 3-no, the item does not tell us that students know about individuals vs populations 4-no, only mathematical reasoning is necessary (SEP is only dimension measured) - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. None 1. 3D 2. Yes 3. Yes 4. No. Does not address biodiversity.

Table 2 (Grade 11 Batch 54). *Source-of-Challenge Issues by Reviewer*

Sources of Challenge
Item #6 - SoC - because the stimulus tells us nothing about the production of ProteinC in the two cows, it is reasonable that A is also a correct answer. - Students are given the fact "Protein C breaks down Factor V." Based on this information, it is reasonable that a student may choose option A as their response and be justified.
Item #10 - The respondent may look at the subadult and juvenile data and conclude that the "current" trend is that things are improving. The answer selections don't allow for a line of reasoning that the data refutes the claim because the current trend shows improvement and the respondent could interpret if that trend continues the claim is not supported. The respondent may get the "right" answer for the wrong reason by following the line of reasoning of the item writer and forsaking their own conceptual u - SoC - the graph doesn't totally support the claim because the trend in ratio of male:females is getting better from subadults to juveniles. - No reasoning that supports refutation of claim would allow a student to get this question right without understanding the supporting evidence.
Item #25 - SoC - The answer "Biodiversity in areas outside of the refuge would decrease faster" is not reasonable for the question because this is already happening. Creating refuges is not going to change the biodiversity in the areas outside of the refuges. It requires an assumption that is not given.

- Choosing that biodiversity outside the refuge asks the student to make an assumption that is not justified based on the question.
Item #35 - The information given and what a student is asked to infer is beyond the standard. It requires students to make assumptions that data is not present. - Speciation by errors in meiosis is not addressed by LS3-2. Students answering these questions correctly has no bearing on their understanding of LS3-2.

Grade 11 Batch 55

Table 1 (Grade 11 Batch 55). *Notes by Reviewer*

Notes
Item #1 *- 0. none 1. 3D 2. yes 3. yes 4. yes 5. This scenario is excellent but needs tweaking. Energy (kWh) and power (kW) is consistently mislabeled and misused. For example the energy usage should be 30 kWhr for 24 hr.* [THIS COMMENT BELONGS WITH ITEM #4] - 0. 1.yes 2.yes 3.yes 4.yes 5. - 1-1D, really only related to knowing reactants & products 2-ok 3-yes 4-no, not accessing SEP or CCC knowledge - 0. none 1. 2D (missing SEP) 2. yes 3. yes 4. yes - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes
Item #2 - 0. none 1. 3D 2. yes 3. yes 4. yes 5. The mass of the astronauts is constant. It is the weight that changes. There is also an error in first sentence - "are on is on" - 0. 1.yes 2.yes 3.yes 4.yes 5. - 1-2D, this is rote knowledge about gravitational forces 2-phenomenon is ok 3-no, there is nothing in the model that gets students to a factor of 4 except rote knowledge they may have learned in class. 4-no, due to notes above - 0. none 1. 3D 2. yes 3. yes 4. yes 5. Mass does not change. This should say weight. - 0. none 1. 3D 2. yes 3. no - factor of 4 was acceptable, however factor of 1/4 was not 4. yes 5. Introduction to question state that mass is changed, not weight. Calculator did not work. - 0. None 1. 3D 2. Yes 3. Yes 4. Yes 5. The question references the MASS, saying it is different in space. The correct word to use here would be WEIGHT.
Item #3 - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. 1.yes 2.yes 3.yes 4.yes 5. - 1-1D, utilizing only mathematical reasoning to answer the question. 2-phenomenon is ok 3-no, because there is only one graph that goes up from reactants to products therefore selecting the correct choice does not show that students know anything about activation energy. 4-no, this is purely a mathematical reasoning item - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. none 1.3D 2. yes 3. yes 4. yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes
Item #4 - 0. none 1. 3D 2. yes 3. yes 4. yes 5. great opportunity for students to engage at a complex and in-depth process - 0. 1.yes 2.yes 3.yes 4.yes 5. - 1-3D 2-phenomenon is good 3-yes, cluster that has students run trials is excellent 4-yes - 0. none 1. 3D 2. yes 3. no - assertions are based on information the student does not have 4. yes 5. there are errors in labels. The student is given an option that is not testable. A required response for additional testing is the

same as an option for initial testing. A student would not select their original focus as an addition point of testing.
- 0. none 1. 3D 2. yes 3. no - Power is given as kwh not kw which would prevent students from calculating correct answer. Energy should be given for each house as 30 kwh per day. 4. no - Standard is DOK level 4 which cannot be assessed with this type of test; question is level 3.
- 0. None 1. 3D 2. Yes 3. No. The scoring of part D if a student selects "longest-lasting is unclear" as there are no variables to test that return information regarding how long the panel lasts. 4. Yes 5. This question needs serious revision. In Part C, power is given as kilowatt-hours (kWh) which is a measurement of energy use. The correct unit should be kilowatts both in the list and the table. In part E, the description of the house using "30kw of power in 24 hours" is incorrect. To get the calculation correct, we should assume that 30 KILOWATT-HOURS (kwh) are used by the house per 24 hours. This needs to be changed as well.

Item #5
- 0. none 1. 3D 2. yes 3. yes 4. yes 5. Engaging stimulus
- 0. 1.yes 2.yes 3.yes 4.yes 5.
- 1-3D 2-phenomenon is good 3-yes 4-yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. none 1. 3D 2. yes 3. yes 4. no - standard is DOK level 2, but I feel that question is level 3.
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Item #6
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. 1.yes 2.yes 3.yes 4.yes 5.
- 1-3D 2-phenomenon is good 3-yes 4-yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. none 1. 3D 2. yes 3. no - I disagree that sodium would form stronger bonds than potassium based on their placement on the periodic table and their electronegativity difference. 4. no - Standard is DOK level 2 and I believe that the question is level 3.
- 0. None 1. 3D 2. Yes 3. No. A student who gets part D right does not understand electronegativity trends. 4. Yes 5. For part D, potassium should form a stronger bond because it is less electronegative, meaning the electronegativity difference and thus bond strength should be higher between potassium and oxygen.

Item #7
- 0. none 1. 3D 2. yes 3. yes 4. yes 5. clearly understood simulation
- 0. 1.yes 2.yes 3.yes 4.yes 5.
- 1-3D 2-phenomenon is good 3-yes 4-yes
- 0. none 1. 3D 2. yes 3. yes 4. yes 5. Part D is not scored so why is it included?
- 0. I believe this question also addresses HS-PS1-6 since the reaction is at equilibrium. 1. 3D 2. yes 3. no - there is no score result for part D of the question. 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes 5. The change in reaction rate when modifying temperature is much different than when modifying hydrogen or carbon dioxide.

Item #8
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. 1.yes 2.yes 3.yes 4.yes 5.I have an issue with saying a different energy in the molecule changes the chance of reaction...because heat doesn't necessarily change the energy "in a molecule", to me "in a molecule" means the chemical energy. Therefore students will get 0/2 on that portion of the test.
- 1-3D 2-phenomenon is good 3-yes 4-yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Item #9
- 0. none 1. 3D 2. yes 3. yes 4. yes
- 0. 1.yes 2.yes 3.yes 4.yes 5.
- 1-3D (although the SEP on this item is #4, not #3 as in the standard) 2-phenomenon is good 3-yes 4-no, see notes on #1
- 0. none 1. 2D (missing SEP) 2. yes 3. yes 4. no - does not address plan and carrying out an investigation - students are just analyzing data

- 0. none 1. 2D - SEP3 missing - did not plan and carry out investigation 2. yes 3. yes 4. no - Standard is DOK level 4, question is level 2. Also question does not have a closed system or uniform energy distribution as stated in the standard. - 0. None 1. 3D 2. Yes 3. Yes 4. No. Students do not plan and carry out.
Item #10 - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. 1.yes 2.yes 3.yes 4.yes 5. - 1-3D 2-phenomenon is ok 3-yes 4-yes - 0. none 1. 3D 2. yes 3. yes 4. no - does not address revising the explanation. - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes
Item #11 - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. 1.yes 2.yes 3.yes 4.yes 5. - 1-3D 2-phenomenon is good 3-yes 4-yes - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes
Item #12 - 0. none 1. 2D, SEP missing - student is not actually engaged in communication which is difficult in the assessment context 2. yes 3. yes 4. no 5. great practical application item - 0. 1.yes 2.yes 3.yes 4.yes 5. - 1-3D (although the SEP is probably #2 in the item and not #8 as in the standard) 2-phenomenon is good 3-yes 4-yes, with consideration of note on #1 - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes
Item #13 - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. 1.yes 2.yes 3.yes 4.yes 5. - 1-3D 2-phenomenon is good 3-yes 4-yes - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Table 2 (Grade 11 Batch 55) *Source-of-Challenge Issues by Reviewer*

Sources of Challenge
Item #4 - This question has so many issues, that it needs to go back to the drawing board and be reworked. The concept is good, but the current form is not acceptable. See specifics in note #5.

Grade 11 Batch 56

Table 1 (Grade 11 Batch 56). *Notes by Reviewer*

Notes
Item #1 - 0. none 1. 3D 2. yes 3. yes 4. no - the assertion does not correlate as the modification of relationships between humans and Earth systems due to human activity is not addressed - 0. 1.yes 2.yes 3.yes 4.yes 5.The correct answer graph is VERY distorted mathematically. It should be fixed. - 1-2D, is really SEP of 4 (not 5) and only applies analyzing and interpreting data not really applying CCC-system-thinking. 2-phenomenon is OK, but the answer choices don't reflect the data in the diagram or table. 3-no, based on the data given, the blue line for well output could just as easily be vertical as 16->15 is not as drastic change as 25->5 is on the given answer. 4-no, only mathematical reasoning is really needed to answer this question. 5-SoC - the item doesn't address the standard of "how those relationships are being modified due to human activity." - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. none 1. 3D 2. yes 3. yes 4. no - I do not see any relationship between Earth systems. - 0. None 1. 3D 2. Yes 3. Yes 4. Yes
Item #2 - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. 1.yes 2.yes 3.yes 4.yes 5.good analogy - 1-1D because students can answer this using rote memory of thermal convection concepts learned in middle school. 2-phenomenon = yeck! 3-no, students are not developing the model, only labeling parts. They are also not using evidence to build the model. 4-no not aligned to the SEP4 or to CCC application of stability/change or technology. 5-SoC - this item is written to the SEP2, not SEP4. Thus, this model is not based on evidence, students can answer it from rote memory from information learned in a MS Earth Science class. That makes is more like a MS item for MS-ESS2-1 - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes
Item #3 - 0. none 1. 3D 2. yes 3. no - the assertion cannot be inferred from answering this question 4. no - the assertion cannot be inferred from answering this question (see source of challenge) 5. the science of this question is flawed - 0. meh 1.no 2.no 3.no 4.no 5. Science is incorrect, incomplete and distorted. See https://www.youtube.com/watch?v=b3TRUDKpoAs - 1-3D 2-phenomenon is ok, although students might wonder [How do we know this "fact"?] 3-yes 4-yes 5-SoC - the item lays out a "fact" 9,000 years ago, Earth’s closest point to the sun during its orbit occurred during summer in the Northern Hemisphere. Today, Earth’s closest point to the sun during its orbit occurs during winter in the Northern Hemisphere) without any information about how we know this "fact" - 0. none 1. 3D 2. yes 3. yes 4. yes 5. The idea that the tilt of the earth would affect the energy flow is acceptable, but that is not what is presented to student. The information is presented that the change in flow is because of the change of season due to distance from the sun. - 0. none 1. 3D 2. yes 3. no - energy amount would be negligible when earth's closest point changes. Is affected by tilt of the earth. 4. yes - 0. None 1. 3D 2. Yes 3. No. Earth's distance from the sun does not cause seasonal changes in climate. The tilt of the earth does. 4. No. Earth's distance from the sun does not cause seasonal changes in climate. The tilt of the earth does.
Item #4 - 0. none 1. 3D 2. yes 3. yes 4. yes 5. good cluster item that is engaging - 0. 1.yes 2.yes 3.yes 4.yes 5. - 1-3D 2-phenomenon is good 3-yes 4-yes - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes

Item #5 - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. 1.yes 2.yes 3.yes 4.yes 5. the wording in the paragraph's first two sentences makes it sound like there is algae growing with the corn, which is confusing. Please tweak. - 1-3D 2-phenomenon is good 3-yes, but see SoC above 4-yes 5-SoC - in the partC answer choice of ALGAE as "this crop creates cleaner energy" because according to the table, both algae and corn reduce emissions, thus making them both cleaner. The student could get this item wrong for using the right reasoning. - 0. none 1. 3D 2. yes 3. yes 4. yes - 0. I think this question could also align with HS-ESS3-2 1. 3D 2. yes 3. yes 4. yes - 0. None 1. 3D 2. Yes 3. Yes 4. Yes 5. Part 2 could be easy to a student to miss if there is no warning that it is left uncomplete in the student application.

Table 2 (Grade 11 Batch 56) *Source-of-Challenge Issues by Reviewer*

Sources of Challenge
Item #3 - This is one insignificant factor in a system of multiple subsystems affecting climate. - Science is incorrect, incomplete and distorted. - The premise of distance from sun dictating seasons is scientifically inaccurate. The tilt of the Earth is what causes seasons. - This question is based on false information and should be removed.
Item #5 - SoC - in the partC answer choice of ALGAE as "this crop creates cleaner energy" because according to the table, both algae and corn reduce emissions, thus making them both cleaner. The student could get this item wrong for using the right reasoning.

WebbAlign™

WCEPS
Wisconsin Center for
Education Products & Services

Appendix E

DOK – Category of Engagement Definitions for Science

January, 2023

WebbAlign®

DOK: Categories of Cognitive Engagement for Science

This tool supports educators, educational content developers, assessment writers, and other stakeholders in interpreting, evaluating, operationalizing, and communicating about shared goals related to the types of complex cognitive engagement expected within current science standards, including NGSS and other Framework-influenced standards. This tool can be used to differentiate between and among the different types of complexity of cognitive engagement required by learning expectations along with corresponding questions, prompts, and tasks used within curriculum, instruction, and assessments.

*The four broad DOK Categories of Cognitive Engagement for science are described in this document. These categories **do not** represent a progression or sequence in terms of learning. Students may engage directly with a higher complexity task and later incorporate tasks of lower complexity–that all together contribute to an overall learning goal. Verbs should not be relied upon to determine task complexity; complexity is dependent on the way(s) in which students are required to interact with or engage with science ideas, practices, and concepts.*

Importantly, this tool differentiates the complexity of cognitive engagement from difficulty, from cognitive load, and from sophistication of thinking as well as from other important but distinct factors and considerations, including the dimensionality of the NGSS and other Framework-influenced standards. This is consistent with the NGSS, which includes three-dimensional performance expectations requiring cognitive engagement at DOK Categories 2, 3, and 4. The standards expect "deeper understanding of content," "application of content," "putting…knowledge to use," greater depth and rigor," "conceptual understanding," "engage[ment] in scientific investigations and argumentation," etc. These expectations for complexity of cognitive engagement apply across grades with "increasing sophistication of student thinking" developing as students move through the grade bands (Appendices A, C, E; The Framework).

Acknowledgements:

These definitions were developed with the thoughtful input of many educators and other stakeholders, including Kevin J. B. Anderson, Aneesha Badrinarayan, Greg Bartley, Roy Beven, Samantha Bunting, Joseph Durand, Brian Gong, Hannah Graham, Cora James, Thomas E. Keller, Peter McClaren, April McCrae, Emily Miller, Brett Moulding, Stephen Pruitt, Karen Whisler, and Ted Willard.

WebbAlign®

Using DOK to Interpret the Complexity of Cognitive Engagement Represented within the NGSS PEs:

A Framework for K-12 Science Education and the resulting NGSS both emphasize a conceptual shift in science standards, related to the **complexity** of student engagement with science ideas, concepts, and practices (NGSS Appendix A, Conceptual Shift #4). As one of the central conceptual shifts specified in the standards, attention must be given to determine if and in what ways different types of student cognitive engagement (i.e. cognitive complexity) are being interpreted in the expectations, in curriculum / learning opportunities, and in assessments (of all types). Use of Webb's DOK – Categories of Engagement helps educators interpret, communicate about, and evaluate the **complexity** of cognitive engagement required by learning expectations, along with the corresponding questions, tasks, and prompts used in curriculum and assessment.

Use of DOK helps all stakeholders to work purposefully to attain our existing goals of an aligned system. As a reflective lens, DOK is used to foster intentionality in teachers' and in content writers' practices, to help ensure that the complexity of expected learning outcomes are clearly understood, that (formative/summative/etc.) assessments provide opportunities to make reasonable inferences about attainment of the intended learning outcomes, and that appropriate educational opportunities are provided to allow students to engage at the level(s) of complexity intended. The critical role of alignment in the success of Framework-influenced science standards, including but not limited to the NGSS, was called out in the very first chapter of A Framework for K-12 Science Education:

> *"The committee recognizes that the framework and subsequent standards will not lead to improvements in K-12 science education unless the other components of the system – curriculum, instruction, PD, and assessment – also change so they are aligned with the framework's vision."* (NRC, 2012)

In other words, in order to achieve the shift in the complexity of student engagement with science–an explicit goal of the standards–students must be provided with learning opportunities that are as cognitively complex as what students are expected to know and do as stated in the corresponding standards. Similarly, what is elicited from students on assessments must be as cognitively complex as what students are expected to know and do as stated in the corresponding standards.

The Framework and NGSS documentation specify that DOK Category 1 type expectations are not intended as summative assessment targets. Because "[p]erformance expectations are the assessable statements of what students should know and be able to do" and are intended to "to make clear the intent of the assessments" (p. 1, NGSS Release and p. 2, Appendix A, April 2013) it can be inferred that no PE should be considered to expect only DOK Category 1 type work. Although DOK Category 1 expectations are not intended as summative assessment targets, they *are* expected to be necessary and included in curriculum and instruction. One example given is that although "[n]o part of the NGSS specifies the student outcome of defining a gene – it is…implicit that in order to demonstrate proficiency on MS-LS3-1, students will have to be introduced to the concept of a gene through curriculum and instruction" (NGSS, Appendix B, p. 6).

Individual PEs are used by some and/or in some cases as curriculum and assessment targets. Bundles of PEs are used by some and/or in some cases as curriculum and assessment targets. When bundled, dimensions may be shuffled and regrouped, affecting the complexity of the expectation(s) and corresponding curriculum and assessment tasks. No matter the approach, meeting the goals of the NGSS to effect a conceptual shift in science standards, related to the **complexity** of student engagement with science concepts and scientific thinking (NGSS Appendix A, Conceptual Shift #4) means that it is necessary to differentiate between and among the different types of student cognitive engagement (i.e. cognitive complexity) explicit in the standards. Use of DOK – Categories of Cognitive Engagement allows stakeholders in all parts of the system to identify and name the referents for complexity, adding clarity to the interpretation and operationalization of the standards, and informing instructional, curricular, and assessment choices and design.

Science – DOK – Category 1

Category 1 is defined by the recall of information, such as a discrete fact, definition, or term, as well as performance of a clearly defined process, scripted series of steps, or set procedure (e.g. use a balance, read information from a Periodic Table, follow a protocol). Category 1 tasks may require a rote response or use of a well-known formula. Finding a particular point on a graph or otherwise directly reading information from graphs, charts, diagrams, or maps is considered Category 1 work. In the context of multidimensional science standards, Category 1 tasks are, by definition, unidimensional—for example, requiring recall of a particular disciplinary core idea or widely accepted "fact." Category 1 expectations and tasks, by definition, do not require students to engage in sense-making and do not require knowledge-in-use. If working with NGSS or other Framework-based standards, it is important to note that while performance of Category 1 tasks are expected as a part of curriculum and instruction (NGSS Appendix B), an explicit goal of Framework-based standards is to promote a shift away from Category 1 tasks as ultimate learning expectations and, correspondingly, as summative assessment targets. Students will, however, engage in Category 1 tasks in the classroom in the context of broader work to make sense of a phenomenon. Across all grades, for example, students are expected to properly use measurement tools, recognize specific structures or relationships, recall appropriate safety protocols, and learn relevant terminology. Students may be expected to develop fluency with Category 1 expectations. Although not complex, Category 1 expectations can be difficult, and may require time and effort to learn.

Importantly, Category 1 expectations do **not** necessarily need to be mastered before engaging in more complex expectations. For example, it is possible to plan and conduct an investigation to provide evidence that feedback mechanisms maintain homeostasis (HS-LS1-3) without first memorizing vocabulary terms for the structures involved in the feedback system and without first determining the atomic composition of molecules involved in the feedback system. In fact, engaging in complex tasks can promote, motivate, and facilitate mastery of DOK 1 learning expectations because they are encountered in a relevant and meaningful context.

Some examples that represent (but do not constitute all of) Category 1 expectations and tasks:

- Recall or recognize a fact, term, relationship, structure, or property.
- Reproduce in words or diagrams a typical or routinely used representation or model of a scientific concept or relationship, such as labeling a diagram of a life cycle or labeling a diagram of the water cycle with the correct terms.
- Provide or recognize a standard scientific representation for common phenomena or relationships, such as reading directly from or adding arrows to a food web diagram.
- Perform a (grade-level-appropriate) routine procedure, such as measuring length or completing a basic Punnett square.

Science – DOK – Category 2

Category 2 expectations and tasks require knowledge-in-use rather than in isolation of purpose or context. In general, Category 2 tasks require application of underlying conceptual understanding and therefore engagement in mental processing beyond recalling or reproducing a response. In other words, Category 2 tasks require students to interact with and make use of science ideas and concepts. Students may need to make some decisions about how to approach a question or problem, including applying knowledge and making connections between and among related ideas and concepts. Category 2 tasks require students to use observations, data, and/or other information to make sense of a phenomenon. Sense-making within Category 2 involves fairly straightforward or routine relationships or interactions between and among ideas and concepts. Using one's own observations to make original comparisons or to draw connections between and among science ideas and concepts are tasks that are typically Category 2. Tasks that require purposeful interpreting, organizing, and displaying of data in tables, graphs, and charts are also considered Category 2. Students may represent ideas mathematically or use routine mathematical and statistical concepts and processes to represent relationships between variables. At Category 2, students use evidence in the context of tasks such as explaining relationships in terms of observations or science concepts. A task requiring a rationale equivalent to an explanation grounded in conceptual understanding would be Category 2.

Some examples that represent (but do not constitute all of) Category 2 expectations and tasks:

- Specify and explain in one's own words the relationship between ideas, concepts, properties, or variables; draw meaning from observing, describing, and/or comparing patterns.
- Differentiate between and among ideas that are considered scientific fact, reasoned hypothesis, and speculation.
- Engage in sense-making related to the relationships between and among ideas and concepts in the context of a fairly routine phenomenon or problem, given data and conditions.
- Organize and represent data to show basic patterns or relationships relevant to making sense of a phenomenon.
- Interpret data to make sense of concrete relationships or to inform an explanation or design solution relevant to a phenomenon.
- Interpret or explain phenomena in terms of science ideas and concepts.
- Develop a fairly basic model that demonstrates underlying conceptual understanding and/or use a model that is a common representation of a phenomenon or concept to solve a problem, make sense of a relationship, etc.
- Apply conceptual understanding of disciplinary ideas to identify limitations of models.
- Make predictions for cause-and-effect relationships that are fairly direct but that require some consideration of the factors that influence outcomes.

Science – DOK – Category 3

Well-designed Category 3 tasks are likely to promote productive struggle as students may need to grapple with the context and information provided to figure out how to even begin to make sense of a phenomenon or problem. The complexity does not result only from the fact that there could be multiple approaches and solutions to a problem (also a possibility for both Category 1 and 2) but because the task requires more demanding, thorough, and abstract reasoning grounded in evidence. Category 3 tasks require planning with consideration of purpose and constraints. Students must use robust evidence to make original arguments. Tasks that require students to provide an evidence-based rationale for a novel solution or engage in scientific argumentation that involves heavy reasoning grounded in appropriate evidence are Category 3. An authentic science or engineering problem that has more than one possible solution and requires students to justify the response with appropriate evidence would most likely be a Category 3. Work may require application of ideas across diverse concepts, contexts, and disciplines. Category 3 expectations and tasks typically involve the use of science and engineering practices to solve non-routine problems. Conceptual understanding of science ideas and concepts may be applied to hypothetical contexts or used to support design solutions, claims, and arguments. Category 3 tasks include a scope of work that can be completed in a discrete period of time (i.e. “in one sitting”).

Some examples that represent, but do not constitute all of Category 3 expectations and tasks:

- Identify appropriate research questions and design brief investigations to help make sense of a phenomenon or science/engineering problem.
- Engage in abstract sense-making related to a complex and non-routine phenomenon or problem, given data and conditions, to develop hypotheses, logical conclusions, or original scientific arguments grounded in evidence.
- Develop and/or use a model (likely novel to the student) to describe a complex, non-routine phenomenon or concept.
- Conduct critical analyses of models, requiring the synthesis of disciplinary ideas.
- Form robust and defensible conclusions about non-routine problems or phenomena based on experimental data.
- Evaluate the bias, credibility, or accuracy of a scientific claim expressed in a text.
- Critically analyze causes for different conclusions based on scientific investigations of or reports about the same phenomenon.
- Evaluate alternative design solutions to an engineering problem.
- Propose revisions for aspects of experimental design grounded in evaluative review.
- Define authentic constraints and incorporate considerations for these constraints into problem-solving work.
- Analyze data to inform revisions to a proposed process or system.
- Develop a mathematical or computational simulation of a phenomenon.

Science – DOK – Category 4

Category 4 demands are at least as complex as those of Category 3, but a main factor that distinguishes the two categories is the need to perform activities over days or weeks (Category 4) rather than in one sitting (Category 3). The extended time that accompanies this type of task allows for more extensive planning and consideration of potentially intricate contingencies (dependent and interacting pieces) within and across systems. Category 4 tasks likely require thinking about implications of choices across time and require sustained metacognitive awareness. Category 4 science tasks parallel the types of extended iterative and non-linear engagement involved in authentic science inquiry and engineering design processes. Broad and abstract thinking is likely required to synthesize diverse ideas, concepts, contexts, and disciplines.

Note that an extended time period is not a distinguishing factor if the required work is only repetitive and does not require applying significant higher-order thinking. For example, if a student is expected to measure the water temperature from a river each day for a month and then construct a graph, this would be considered to fit within Category 2. However, if the student is engaged not only in the data collection and representation but in all aspects of planning and carrying out an authentic scientific investigation or design solution, then the overall task would be Category 4. While some science standards expect students to engage at Category 4, on-demand assessment instruments are inappropriate tools for judging student proficiency as relates to the full scope of Category 4 expectations; these are most appropriate for classroom assessment.

The scope of a Category 4 task requires demonstration of multiple Category 1, 2, and 3 expectations in the service of the larger goal. Note that educators may choose to design Category 4 tasks that promote, motivate, and facilitate Category 1, 2, and 3 work. These Category 4 tasks may be grounded in PE bundles or other groupings of learning goals. Phenomenon-based learning, problem-based learning, and the 5E Model, are some examples of common pedagogical strategies that may be used to support this approach.

Some examples that represent, but do not constitute all of, Category 4 expectations and tasks:

- Plan and carry out an authentic scientific investigation that will yield appropriate data that could be used as evidence to answer scientific questions related to real-world problems.
- Plan, test, and revise a design solution for a real-world problem.
- Analyze the results of multiple studies on a particular science topic or design solution to form an original conclusion about the subject.
- Use trials of a scientific investigation or design solution to evaluate strengths and weaknesses of an experimental design and develop a revised and more optimized approach.
- Conduct broad-scope, systems-level analyses of non-routine problems.

Bibliography and References:

Badrinarayan, A, Christopherson, S., Gong, B., McCrae, A. (2018) *Developing a Common Language to Understand Content Complexity for Alignment Studies of the NGSS.* Presented at the National Conference on Student Assessment.

Badrinarayan, A., Christopherson, S., Davis-Becker, S., Everson, H., and Forte, E. (2019). *Representing Cognitive Complexity in Test Design and Evaluation*. Presented at ATP Innovations in Testing.

Fulmer, G.W., Tanas, J., Weiss, K.A. (2018) The challenges of alignment for the Next Generation Science Standards. *J Res Sci Teach*. 55: 1076– 1100. https://doi.org/10.1002/tea.21481

National Research Council. (2000). *How People Learn: Brain, Mind, Experience, and School: Expanded Edition*. Washington, DC: The National Academies Press. https://doi.org/10.17226/9853.

National Research Council. (2005). *How Students Learn: History, Mathematics, and Science in the Classroom. Committee on How People Learn, A Targeted Report for Teachers*, M.S. Donovan and J.D. Bransford, Editors. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

National Research Council. (2012). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. Washington, DC: The National Academies Press. https://doi.org/10.17226/13165.

National Research Council. (2014). *Developing Assessments for the Next Generation Science Standards*. Washington, DC: The National Academies Press. https://doi.org/10.17226/18409.

NGSS Lead States. (2013). *Next Generation Science Standards: For States, By State*s. Washington, DC: The National Academies Press.

NGSS Lead States. (2013). *NGSS Release: How to Read the Next Generation Science Standards (NGSS)*. Washington, DC: The National Academies Press.

Schneider, M.C, Huff, K.L., Egan, K.L, Gaines, M.L., and Ferrara, S. (2013). *Relationships Among Item Cognitive Complexity, Contextual Demands, and Item Difficulty: Implications for Achievement-Level Descriptors*. Educational Assessment, 18:99-121

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education*. Council of Chief State School Officers and National Institute for Mathematics Education Research Monograph No. 6. Madison, WI: University of Wisconsin Center for Education Research.

Webb, N. L. (1999). *Alignment of science and mathematics standards and assessments in four states*. Council of Chief State School Officers and National Institute for Science Education Research Monograph No. 18. Madison, WI: University of Wisconsin Center for Education Research.

Webb, N. L. (2002). *Alignment study in language arts, mathematics, science, and social studies of state standards and assessments for four states*. A study of the State Collaborative on Assessment & Student Standards (SCASS) Technical Issues in Large-Scale Assessment (TILSA). Washington, D.C.: Council of Chief State School Officers.

Webb's DOK - Category 1 as relates to NGSS and Framework-based standards: FAQ	
Do DOK Category 1 tasks require students to interweave multiple dimensions of the standards to make sense of phenomena?	No. A DOK Category 1 task is not multidimensional and does not involve sense-making. Students do not need to engage multiple dimensions to reproduce or recall a response. DOK Category 1 tasks typically separate content from practice. If a practice is at all involved, it is scripted or rote; a reproduction of a response. Although all Performance Expectations are written to incorporate all three dimensions of the standards, related curriculum and assessment tasks must be analyzed to determine if they are one-, two-, or three-dimensional. The dimensionality of a task is related to but different than the complexity of a task and must be evaluated separately. Dimensionality is best evaluated with reference to the foundation boxes for PEs and the corresponding Appendices of the NGSS. A DOK Category 1 task may relate to content that falls outside of the scope of the NGSS, and, therefore, it is possible for a DOK Category 1 task to be "zero-dimensional."
Do DOK Category 1 tasks require sense-making?	No. A DOK Category 1 problem cannot be reasoned through or "figured out;" the answer (or defined procedure or steps to find the answer) must be known. By definition, a DOK Category 1 task is one that is **not** completed via sense-making but instead by a rote or reproduced response. For many DOK Category 1 tasks, "either you know it or you don't."
Do DOK Category 1 tasks require knowledge-in-use?	No. A DOK Category 1 task typically involves knowledge in isolation and, by definition, does **not** require meaningful application, conceptualization, or integration of content, ideas, practices, or concepts. Any application of content required by a DOK Category 1 task would be nominal to the point of insignificant. For example, an assessment task may present a novel phenomenon but (typically unintentionally) ask students to provide a rote response. For example, a student may be given a diagram of a novel food web and asked to determine the ultimate source of energy for a particular animal's food. Because the energy in animals' food was once energy from the sun–no matter which animal and what type of food–the answer to this type of question is typically the same, no matter the context. This overall issue applies in all cases where a student response does not actually require using the information provided. Use of DOK can help educators and other content developers differentiate between complexity, difficulty, cognitive load, dimensionality, etc – to ensure that questions, prompts, and tasks are indeed providing students with opportunities that allow engagement with the intended categories of complexity.
Do NGSS and other Framework-based standards expect students to engage with DOK Category 1 tasks?	Yes. Some engagement with DOK Category 1 work is expected in the classroom and is understood as a necessary contributing component of the goals of the Performance Expectations. One example given in the NGSS documentation is that although "[n]o part of the NGSS specifies the student outcome of defining a gene – it is…implicit that in order to demonstrate proficiency on MS-LS3-1, students will have to be introduced to the concept of a gene through curriculum and instruction" (NGSS, Appendix B, p. 6). It is important to note that an explicit goal of Framework-based standards is to promote a shift away from DOK Category 1 tasks as ultimate learning goals or summative assessment targets.

Webb's DOK - Category 2 as relates to NGSS and Framework-based standards: FAQ	
Do DOK Category 2 tasks require students to interweave multiple dimensions of the standards to make sense of phenomena?	Maybe. A DOK Category 2 task could, for example, require students to use a model to characterize a phenomenon in terms of patterns or interpret graphical displays to make sense of cause-and-effect relationships as relates to a phenomenon. DOK Category 2 cognitive engagement does not, however, guarantee that a task requires students to engage science or engineering practices in the context of disciplinary core ideas. Because crosscutting concepts (CCCs) "unit[e] core ideas throughout the fields of science and engineering" (NGSS, Appendix G, p. 1) DOK Category 2 tasks are likely to *relate* to one or more crosscutting concepts even if the tasks do not necessarily require a student to explicitly invoke a CCC. Although all Performance Expectations are written to incorporate all three dimensions of the standards, related curriculum and assessment tasks must be analyzed to determine if they are one-, two-, or three-dimensional. The dimensionality of a task is related to but different than the complexity of a task and must be evaluated separately. Dimensionality is best evaluated with reference to the foundation boxes for PEs and the corresponding Appendices of the NGSS. It is possible for DOK Category 2 tasks to interweave ideas, concepts, and/or practices that fall outside of the scope of NGSS.
Do DOK Category 2 tasks require sense-making?	Yes. DOK Category 2 tasks require students to connect science ideas and make sense of relationships and interactions between and among science ideas. Sense-making within DOK Category 2 draws on conceptual understanding of science ideas.
Do DOK Category 2 tasks require knowledge-in-use?	Yes. By definition, DOK Category 2 tasks involves purposeful application, conceptualization, or integration of content, idea(s), practice(s), and/or concept(s) within context. At DOK Category 2, knowledge is put to use in the context of tasks that involve underlying conceptual understanding. Some DOK Category 2 tasks may require students to consider relationships between and among or to apply ideas from one concept, context, or discipline to another.
Do NGSS and other Framework-based standards expect students to engage with DOK Category 2 tasks?	Yes. The conceptual understanding emphasized by DOK Category 2 expectations and tasks are a central focus of the goals outlined in the Framework and NGSS documentation. For example, Appendix A conceptual shift number four states that "[t]he NGSS focus on deeper understanding of content as well as application of content" (NGSS, Appendix A, p. 4). Appendix C also underscores this key shift, noting that "the NGSS focus [is] on understanding rather than memorization" (NGSS, Appendix C, p. 6). This, in turn, reflects the Framework committee's intent to "give time for students to…achieve depth of understanding of the core ideas" (A Framework, p. 11). The language of the Framework and the NGSS, viewed through the lens of DOK – Categories of Engagement, specify a shift away from DOK 1 expectations as the ultimate learning outcomes and a strong emphasis on DOK 2 expectations as learning outcomes, instead.

Webb's DOK - Category 3 as relates to NGSS and Framework-based standards: FAQ	
Do DOK Category 3 tasks require students to interweave multiple dimensions of the standards to make sense of phenomena?	Maybe. DOK Category 3 tasks likely require students to engage in science and/or engineering practices in the context of disciplinary core ideas. Because crosscutting concepts (CCCs) "unit[e] core ideas throughout the fields of science and engineering" (NGSS, Appendix G, p. 1) DOK Category 3 tasks are likely to also involve one or more crosscutting concepts. Although all Performance Expectations are written to incorporate all three dimensions of the standards, related curriculum and assessment tasks must be analyzed to determine if they are one-, two-, or three-dimensional. The dimensionality of a task is related to but different than the complexity of a task and must be evaluated separately. Dimensionality is best evaluated with reference to the foundation boxes for PEs and the corresponding Appendices of the NGSS. It is possible for DOK Category 3 tasks to interweave ideas, concepts, and/or practices that fall outside of the scope of NGSS. At DOK Category 3, the scope of work can be completed in a discrete amount of time (i.e., "in one sitting").
Do DOK Category 3 tasks require sense-making?	Yes. DOK Category 3 tasks require students to engage deeply in sense-making, involving abstract, analytical, hypothetical, non-routine, and innovative thinking. Sense-making at DOK Category 3 involves crafting reasoned arguments and novel solutions based on evidence.
Do DOK Category 3 tasks require knowledge-in-use?	Yes. By definition, DOK Category 3 tasks involve purposeful application, conceptualization, and/or integration of content, idea(s), practice(s), and/or concept(s) within contexts that may be non-routine. At DOK Category 3, knowledge is put to use in the context of tasks that involve deep reasoning and development of novel solutions grounded in critical, evaluative, analytical, argumentative, hypothetical, etc. thinking. DOK Category 3 tasks require broad and abstract thinking in order to synthesize diverse ideas, concepts, contexts, and disciplines.
Do NGSS and other Framework-based standards expect students to engage with DOK Category 3 tasks?	Yes. For example, DOK Category 3 expectations and tasks are reflected in the Framework committee's intent to "give time for students to engage in scientific…argumentation" (A Framework, p. 11) and support the goal of supporting students as they "discove[r] new knowledge, solv[e] challenging problems, and generat[e] innovations" including addressing "problems not previously encountered" (NGSS, Appendix C, p. 1-2; 5). The language of the Framework and the NGSS, viewed through the lens of DOK – Categories of Engagement, specify an intent for inclusion of DOK 3 expectations.

Webb's DOK - Category 4 as relates to NGSS and Framework-based standards: FAQ	
Do DOK Category 4 tasks require students to interweave multiple dimensions of the standards to make sense of phenomena?	Most likely, yes. Because of the scope of DOK Category 4 tasks, they are almost certainly three-dimensional. At DOK Category 4, the scope of work requires sustained and extended engagement, over days or weeks (or more) rather than in one sitting (DOK Category 3). DOK Category 4 tasks involve authentic and extended engagement with science practices, ideas, and concepts. Although all Performance Expectations are written to incorporate all three dimensions of the standards, related curriculum and assessment tasks must be analyzed to determine if they are one-, two-, or three-dimensional. The dimensionality of a task is related to but different than the complexity of a task and must be evaluated separately. Dimensionality is best evaluated with reference to the foundation boxes for PEs and the corresponding Appendices of the NGSS. It is still possible for DOK Category 4 tasks to interweave ideas, concepts, and/or practices that fall outside of the scope of NGSS.
Do DOK Category 4 tasks require sense-making?	Yes. DOK Category 4 tasks require students to engage deeply in extended and iterative sense-making, involving abstract, analytical, hypothetical, non-routine, and innovative thinking. Sense-making at DOK Category 4 involves extended and iterative thinking related to crafting reasoned arguments and novel solutions based on research and evidence.
Do DOK Category 4 tasks require knowledge-in-use?	Yes. Inherent to DOK Category 4 tasks is the purposeful application, conceptualization, and/or integration of content, idea(s), practice(s), and/or concept(s) within contexts that may be non-routine. At DOK Category 4, knowledge is put to use in the context of extended and iterative tasks that involve deep reasoning and development of novel solutions grounded in critical, evaluative, analytical, argumentative, hypothetical, etc. thinking. DOK Category 4 tasks require broad and abstract thinking in order to synthesize diverse ideas, concepts, contexts, and disciplines.
Do NGSS and other Framework-based standards expect students to engage with DOK Category 4 tasks?	Yes. DOK Category 4 expectations and tasks correspond to the "expectation…that students generate and interpret evidence and develop explanations of the natural world through sustained investigations" and that students "carry out empirical investigations in order to develop or evaluate knowledge claims" (A Framework, p. 255; 252)

WebbAlign™

WCEPS
Wisconsin Center for
Education Products & Services

Appendix F

Agenda, Coding Instructions, and other Materials Provided to Panelists; Panelist Responses to Study Evaluation and Demographic Forms

January, 2023

SDSA Alignment Institute Summary Agenda

June 21-23, 2022

Lodging:
Holiday Inn, 110 Stanley Rd, Fort Pierre, SD 57532

Meeting:
Casey Tibbs / Mattie Goff Newcombe Conference Center
210 Verendrye Drive, Ft. Pierre, SD 57532

Tuesday, June 21, 2022	
Starting at 6:30 am	Breakfast buffet provided at Holiday Inn (on your own) – Shuttle available to meeting site (or carpool with others)
8:30 am –10:00 am	Introductions and orientation; alignment processes – practice and calibration (at Conference Center)
10:00 am – 10:15 am	Break
10:15 am – 11:45 am	Standards calibration and consensus – Grade 6-8; log in to WATv2 Group ID 322; log in to Content Rater; calibration on item-level analysis (Batch 51)
11:45 am – 12:30 pm - Lunch (catered, on-site)	
12:30 pm – 2:45pm	Code and adjudicate Batch 51
2:45 pm – 3:00 pm	Break
3:00 pm – 4:00 pm	Complete coding for Batch 52
4:00 pm – 4:30 pm	OPTIONAL: Debrief / Q&A with State Officials
Dinner on your own	

WebbAlign™

WCEPS
WISCONSIN CENTER for EDUCATION PRODUCTS & SERVICES

Wednesday, June 22, 2022		
Starting at 6:30 am	Breakfast buffet provided at Holiday Inn (on your own) – Shuttle available to meeting site (or carpool with others)	
8:30 am –11:45 am	Adjudicate Batch 52; Code and adjudicate Batch 53 (breaks as needed).	
11:45 am – 12:30 pm - Lunch (catered, on-site)		
IN TWO PANELS:	**Grade 5 Panel (Group ID 321)**	**Grade 5 Panel (Group ID 323)**
12:30 pm – 1:30 pm	Grades 3-5 Standards: Calibration and Consensus	Grades 9-12 Standards: Calibration and Consensus
1:30 pm – 4:00 pm	Code and adjudicate Batch 48; Start Batch 49 (breaks as needed)	Code and adjudicate Batch 54 (breaks as needed)
4:00 pm – 4:30 pm	OPTIONAL: Debrief / Q&A with State Officials	
6:00 pm	OPTIONAL: Group Dinner at Drifters	
Thursday, June 23, 2022		
Starting at 6:30 am	Breakfast buffet provided at Holiday Inn (on your own) – Shuttle available to meeting site (or carpool with others)	
IN TWO PANELS:	**Grade 5 Panel**	**Grade 11 Panel**
8:30 am –11:45 am	Adjudicate Batch 49; Code and adjudicate Batch 50 (breaks as needed)	Code and adjudicate Batch 55 and Batch 56 (breaks as needed)
11:45 am – 12:30 pm - Lunch (Catered, on-site)		
12:30 – 4:30 pm	Final adjudications; complete debriefing comments; wrap up; complete evaluation form	
DEPART – THANK YOU!		

South Dakota Science Assessment (SDSA) Alignment Analysis: Information and Instructions for Coding

The SDSA includes both **stand-alone items** and **item clusters**. Stand-alone items may be multi-part, but you will consider the entire item as a single unit of analysis. Item clusters are made up of multiple interactions. You will analyze each of the component interactions and will report on the overall cluster as the unit of analysis.

Each item (stand-alone OR cluster) is intended to target a single standard.

To conduct this analysis, you will need: 1) access to items (online via Content Rater); 2) access to the data entry system (online via WATv2); 3) definitions of the Categories of Engagement - DOK for Science (print copy provided); 4) South Dakota Science Content Standards (print copy provided + online).

Use the following information to guide your item-level evaluation:

Relationship of the item/cluster to a South Dakota Science Content Standard:
Does a student's correct response allow for a reasonable inference about the student's proficiency as relates to the expectations within a standard?

- Work through the item or item cluster (all parts) as if you are the student.
- Determine what the item is measuring.
- On your own, identify a corresponding standard.
- For an **item cluster**, successful completion of the task should require students to engage with the specific three dimensions identified in the corresponding standard. For a **stand-alone item**, successful completion of the task should require students to engage with at least two of the specific three dimensions identified in the corresponding standard.
- After you complete your independent selection of standard, your panel facilitator will share the internally assigned standard (within the CAI metadata). If it is the same as the standard you chose, select this standard from the drop-down menu in the WATv2.
- If the internally assigned standard is different from what you chose, decide if you think a student's correct response to the item/cluster would allow for a reasonable inference about the student's proficiency as relates to this internally assigned standard.
- If anyone thinks an item **does not** address (or does not adequately address) the internally assigned standard, alert the group leader. Your group leader will then facilitate a discussion of individual responses.
- After discussion, if you agree with the internally assigned standard, select that standard from the drop-down menu in the WATv2. If you do not agree with that standard, instead select and record the standard you prefer.
- If a panel majority agrees with the internally assigned standard, you will continue with the content analysis of the item (see steps below).
- If a panel majority does **NOT** agree with the internally assigned standard, enter a comment in the notes box to explain why and move on to the next item/cluster.
- If your initial standard assignment was different from the internally coded standard, enter a comment in the notes box to record this change of mind.
- Any comments related to this evaluation step should be recorded in the Notes Box and identified as step **0**.

Notes Box 1. Dimensionality: Does a correct response require the student to engage with one, two, or three of the dimensions specified within the standard?

- An **item cluster** will be coded to a standard only if successful completion of the task requires students to engage with the specific three dimensions identified in the corresponding standard.
- For a **stand-alone item**, note if the item is three-dimensional (write "3D") or two-dimensional (write "2D"). If 2D, state which dimension is missing. Comments should be identified as step **1**. For example, you might record "1. 2D. No CCC."
- Any items that one or more panelists consider to be 1D only will be discussed and flagged if needed.

Notes Box 2. Phenomenon: Does the stimulus meet the test development criteria provided for a "phenomenon"?

Overall, does the stimulus presented meet the following criteria:

- The phenomenon is based on a specific real-world scenario and focused enough to require students' application of a SEP in the context of a DCI and CCC as intended by the standard in order to make sense of the phenomenon.
- Is grade appropriate context and complexity
- Is presented in way(s) that all students can access and comprehend based on information provided (including text, graphics, data, images, animations, etc.)
- Is free of cultural bias, insensitivity, or depiction of unsafe situation
- Is puzzling and/or intriguing for students to engage in; focused on real-world observations that students can connect with or have direct experience with
- Record **yes** or **no** in the notes box as step **2.** If no, explain why.

Complexity of Engagement (DOK): What category of engagement is required for successful completion of an item or interaction?

- Use the printed definitions for Categories of Engagement (DOK) for this part of the analysis. Think about the degree of processing of concepts and skills along with the other factors discussed in our initial training that influence the complexity of an expectation or task.
- You will conduct a content analysis of the task to make an inference about the cognitive complexity required for successful completion of the item or item cluster.
- For each item or item cluster, assign a DOK Category 1, 2, or 3. (Note that DOK Category 4 expectations require extended time and are not fully assessed in an on-demand setting.) You will analyze the complexity of each subcomponent of a stand-alone item or item cluster and record the highest Category of Engagement in the text box marked "Depth."

Notes Box 3. Relationship of Scoring Assertions to Item/Item Cluster: Do the scoring assertions reflect the inferences that can be made from successful student interactions with the item/cluster?

Look at the scoring assertions and compare each assertion against the actual student interaction. Overall, do the scoring assertions reflect the inferences that can be made from successful student interactions? Record your response as step **3**.

- **YES.** Overall, the scoring assertions describe the inferences that can be made from a student's successful interaction with the item/item cluster. You may think that one or more of the assertions may slightly misstate, overstate, or understate the inferences that can be made but a large majority (~75%) of the scoring assertions should describe a direct inference that can be made from the student's correct response.
- **NO.** Overall, the scoring assertions do not describe the inferences that can be made from a student's successful interaction with the item/item cluster. If no, record why.

If any scoring assertion is considered to be completely unreasonable (i.e. not at all something that could be inferred based on the student's response), mark the assertion as a Source of Challenge.

Notes Box 4. Relationship of Scoring Assertions to the corresponding standard Do the scoring assertions reflect the expectations in the corresponding standard?

Consider the scoring assertions holistically. Do the assertions, in aggregate, represent the depth and breadth of the corresponding standard (including its multidimensionality)? Record your response as step **4**.

YES. Overall, the scoring assertions represent the depth and breadth of the corresponding standard, including its multidimensionality.

NO. Overall, the scoring assertions do not represent the depth and breadth of the corresponding standard. If no, record why.

Notes Box 5. Miscellaneous qualitative feedback

As you analyze the items/item clusters you may leave qualitative and descriptive feedback in the Notes text box as step **5**. Leave notes only as time allows.

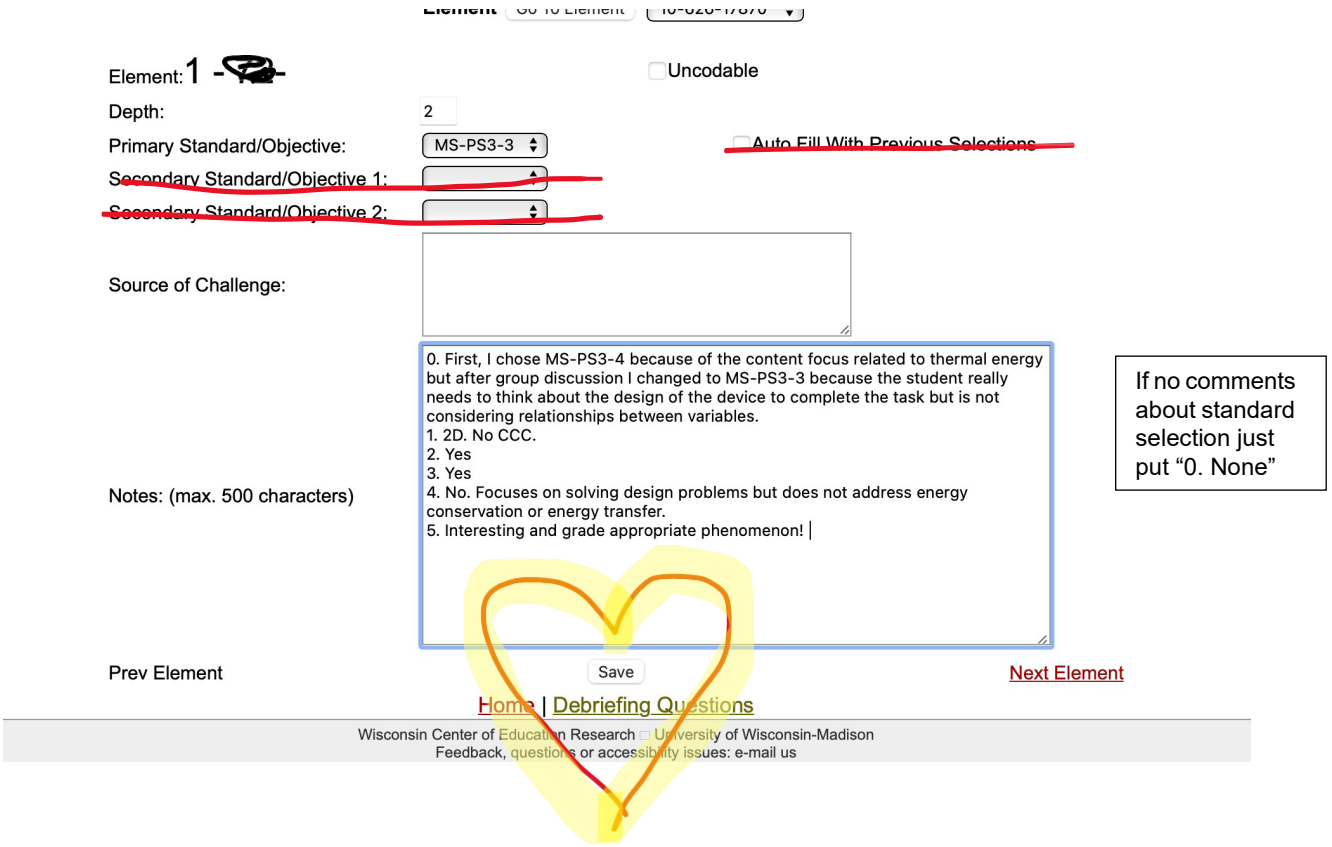
Source of Challenge

A Source of Challenge is a technical issue with the item that could cause a student to get a right answer for the wrong reason or a wrong answer for the wrong reason. If you identify any technical issue with an item, make note in the Source of Challenge text box. Any Source of Challenge is critical to record.

Debriefing Notes

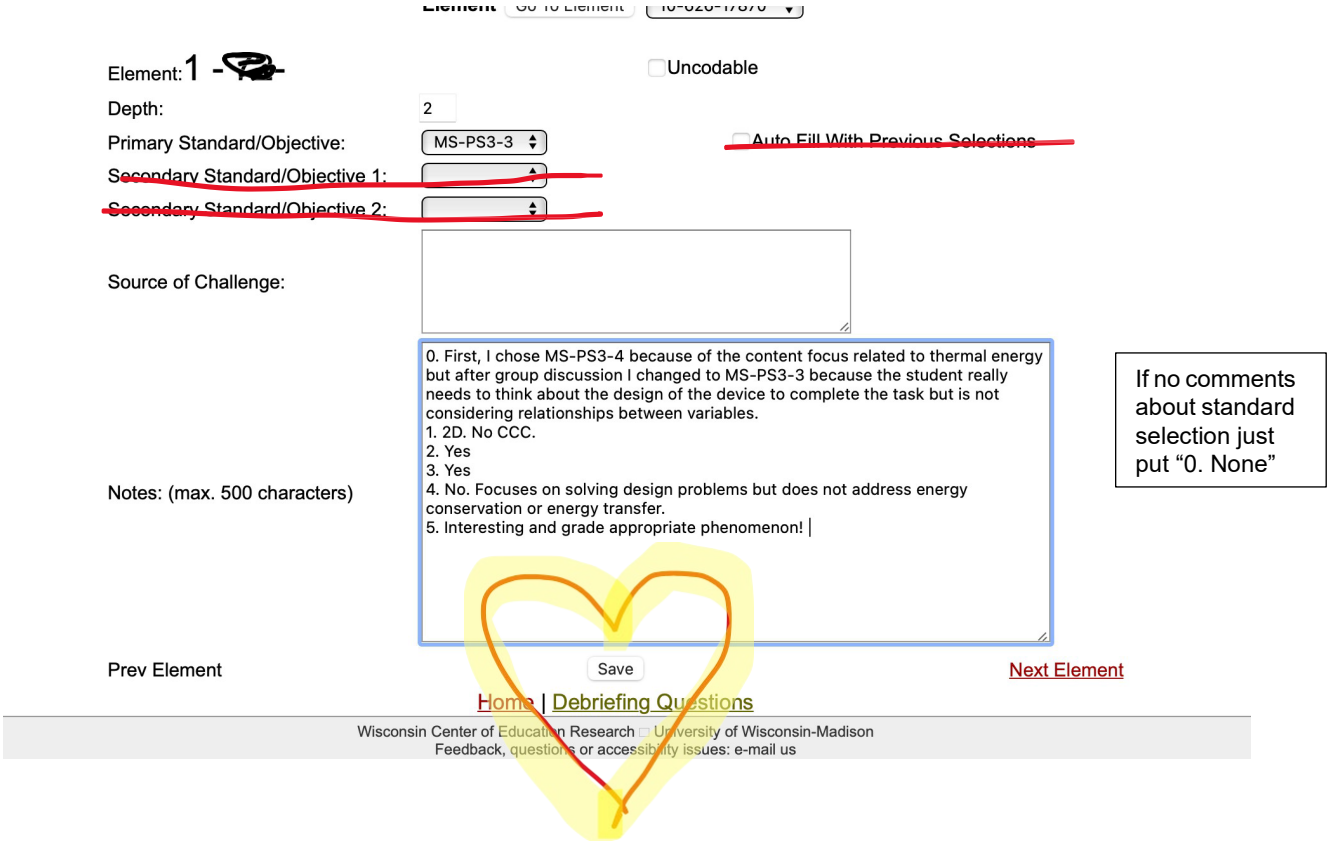
- At the end of each item batch, you will have the opportunity to leave observations, feedback, and comments about broader topics and themes related to the assessment items overall. For example, to what extent did you see South Dakota culture reflected in the assessment items?
- You do NOT need to enter any information that you have already included in the item-level inputs. This is a space to capture any qualitative input that you were not able to communicate in the item-level coding.

EXAMPLE data entry in WATv2



If no comments about standard selection just put “0. None”

EXAMPLE data entry in WATv2



If no comments about standard selection just put “0. None”

How to read the South Dakota Science Standards

Each of the three-dimensions from *A Framework for K-12 Science Education* can be referenced in every standard. This information can be used to interpret a deeper meaning for each of the three dimensions. **Below is a legend** to decode the components involved within each standard.

SEP = Science and Engineering Practices (Chapter 3: Page 41 of *Framework*)

1. Asking Questions and Defining Problems
2. Developing and Using Models
3. Planning and Carrying Out Investigations
4. Analyzing and Interpreting Data
5. Using Mathematics and Computational Thinking
6. Constructing Explanations and Designing Solutions
7. Engaging in Argument from Evidence
8. Obtaining, Evaluating, and Communicating Information

The reader will notice engineering is integrated through inclusion as a Disciplinary Core Idea, Crosscutting Concept or by use of a Science and Engineering Practice. All standards with an emphasis on engineering are marked by an asterisk (*). For more information on Engineering see the Framework page 201and Appendix C of the South Dakota Science Standards.

DCI: Disciplinary Core Idea (Chapter 5: Page 103 of *Framework*)

These are listed as written in *A Framework for K-12 Science Education*. For example PS1 stands for Physical Science Core Idea 1: Matter and Its Interactions. You will notice that next to the standard it will read, for example, PS1.A. In this case, the coding is referring to Physical Science Core Idea 1: Matter and Its Interactions, Component Idea A: Structure and Properties of Matter.

PS = Physical Science
LS = Life Science
ESS = Earth and Space Science
ETS = Engineering, Technology and Applications of Sciences

CCC = Crosscutting Concept (Chapter 4: Page 83 of *Framework*)

Patterns = Patterns
Cause/Effect = Cause and Effect
Scale/Prop. = Scale, Proportion, and Quantity
Systems = Systems and System Models
Energy/Matter = Energy and Matter
Structure/Function = Structure and Function
Stability/Change = Stability and Change

The Framework specifies two core ideas that relate science, technology, society and the environment: the interdependence of science, engineering and technology, and the influence of science, engineering and technology on society and the natural world. These two core ideas may accompany or replace crosscutting concepts related to standards that include engineering. In this instance, we refer to them as connection statements because they are not true crosscutting concepts. When this occurs, we use the following legend.

Technology = Connections to Engineering, Technology, and Applications of Science

Panelist Responses:
SDSA and SDSAA Item-Level Alignment Study Evaluation

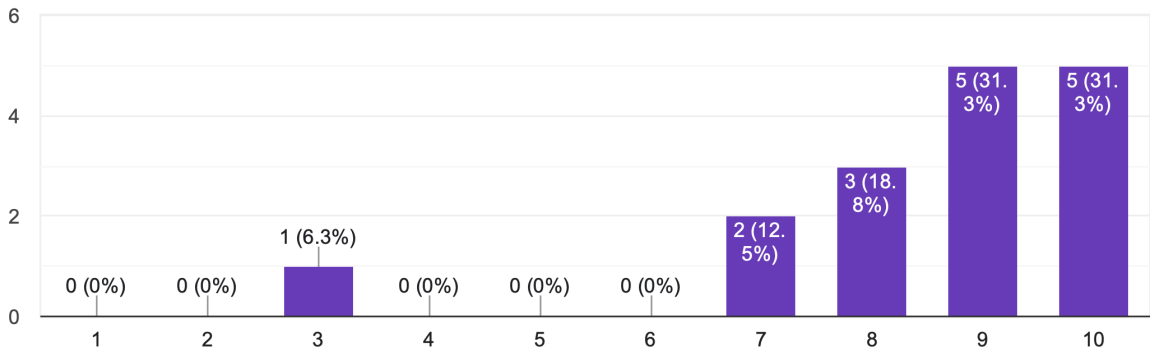
June 2022

- *Linear scale questions required a selection; open response questions were optional.*
- *For linear scale questions, panelists responded on a scale of "Not at all" (0) to "Completely" (10).*

1.

To what extent did the orientation provided in the launch meeting and information provided in the event webspace help prepare you for the alignment work?

16 responses



2. What aspect(s) of the launch meeting and/or materials was (were) most helpful?
   - Honestly all of the information at the meeting was helpful as I was very unsure at what I was going to be doing. One of the materials stating how we were going to be filling out the forms online was very confusing.
   - DOK review/calibration; coding instructions and sample half-sheet
   - I really liked the DCI/ Standard worksheet
   - The standards, coding directions, DOK descriptors sheet, and collages
   - All of it, it was nice to preview the materials in advance so that I knew what I was looking at when asked.

- Talking through everything in person
- going through the standards and understanding the DOK
- The working together on the first item
- Absolutely everything. Honestly, the information that was sent out and the meeting to kick things off were very helpful.
- Explain the coding.
- It was a good background. Some practice problems distinguishing between complexity and difficulty would have been helpful.
- How to coordinate the two websites.
- Printed copies were helpful for reference.
- having the standards and dok listed on one document
- Reviewing the standards was helpful.
- Review and reminder of DOK Classifications.

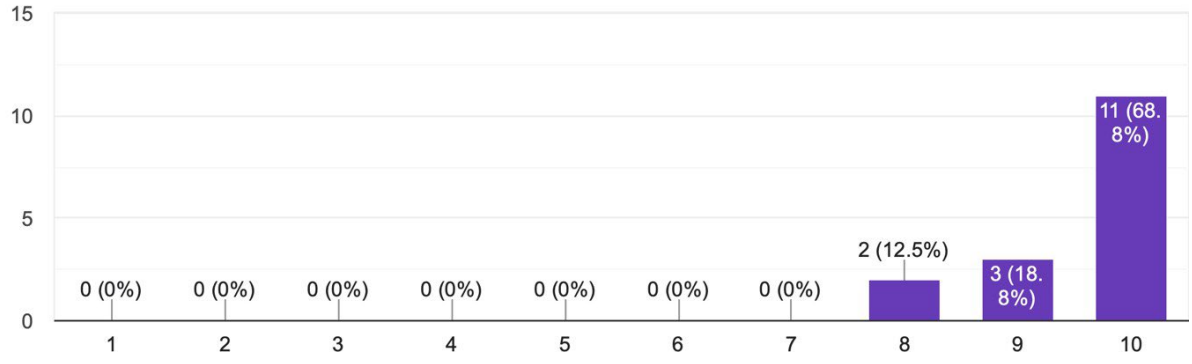
3. What additional information or materials would have been helpful?
Better clarification of Source of Challenge; rearranging order of coding instructions to match watv2 order

- I think a more direct approach on what exactly we would be doing
- Putting the coding directions in the same order as the portal. It was confusing to keep the coding directions and my information in order
- None, we had all we needed!
- More video samples of what we would be doing
- Maybe do two together
- Nothing, I think a person just has to do it and ask questions to understand completely.
- More in depth about whether to code 2D or 3D.
- None
- none, that I can think of
- When working on the SDSAA, having a document that shows the progression of standards K-12 would be helpful when trying to see which grade level the material aligns to best.

4.

To what extent did you feel supported and included in your panel's group work, including adjudication meetings?

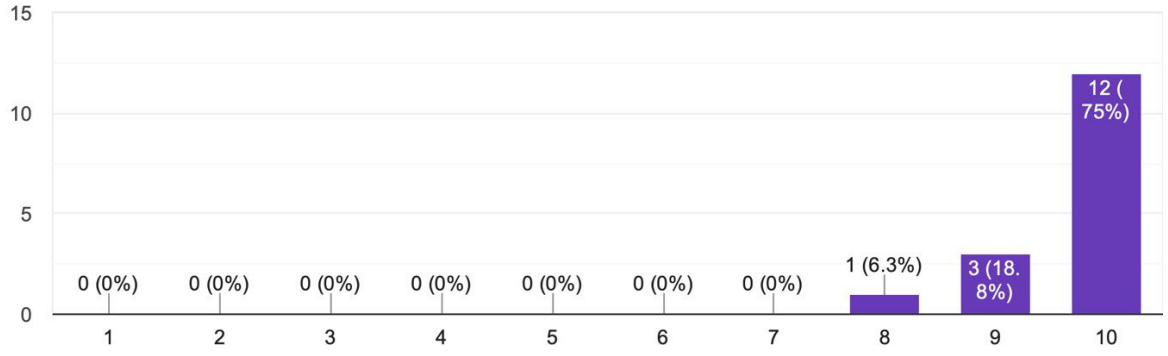
16 responses



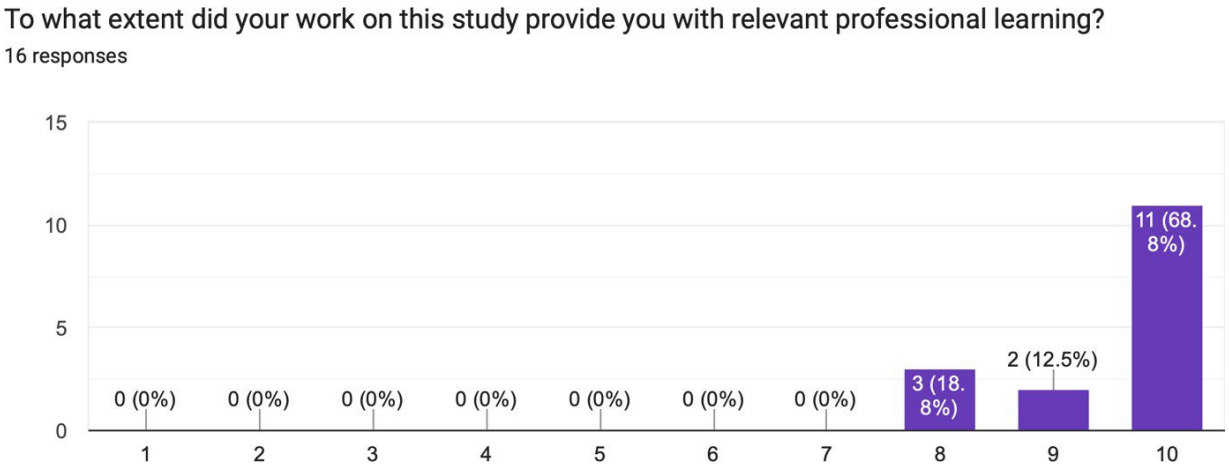
5.

How well do you think the in-person format for the alignment study was managed?

16 responses



6.



7. The South Dakota Department of Education is interested in hearing more about what you feel you gained from your experiences during this week's meetings. For example, did you learn more about the South Dakota Science Standards and/or Core Content Connectors? Did you learn more about alignment, DOK, other? Did you learn things that you will take back to the classroom (and if so, what/how?)? Opportunity for dialogue / exchanging ideas with colleagues? Etc. - Please share some specific examples about what you are taking away from this work!

- I did learn more about the standards. I also got to gain a better understanding of the test and how they chose/what the questions are. I was able to gain a better DOK understanding as it is something that I used to struggle with. I liked being able to exchange my thoughts and concepts to the people around me and make sense of item I was unsure of or felt strongly about.
- The collaborative effort and exchange of ideas help deepen understanding of DOK and alignment of the assessment items to the standards.
- I learned a lot about connecting the test to the DOK. I also learned what kind of questions they present as a test.
- I learned more about South Dakota Science Standards in more than just the area that I teach in and I had never worked with Core Content Connectors much prior to this workshop. I learned a lot more about alignment in testing. I was able to better practice DOK alignments and descriptors. I got a chance to see a wide variety of problems for

each one of the DOK levels to better enhance my own assessments in my classroom. I know I will take the things I learned about the verbs in the standards vs how the verbs are used in a test form and apply those to my classroom and my assessment preparations. I would like to participate in something like this again, as it makes me as an educator feel like I am helping advocate for my students, my profession and what the state DOE stands for.

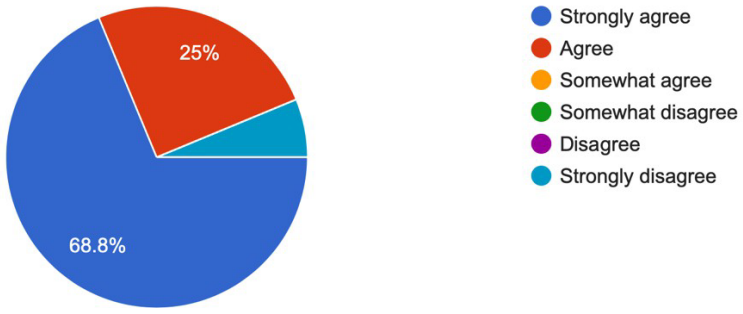
- I definitely learned more about the CCC's as per SDSAA...this project helps me check myself as a teacher of science in South Dakota! Thank you for the opportunity!
- I am more knowledgeable on what I need to focus on and how as an educator to better prepare my students.
- I learned a lot because I don't specifically work with the standards with my students. I learned what would be beneficial to work with my students.
- I learned more about the DOK and will take many things back into the classroom. It was refreshing to know that what I am teaching in my classroom is what the standards are tested on.
- I am very guilty of not being thrilled about the assessments we offer our students every spring and I am certain that I have said, on more than one occasion, "Who writes this stuff?" It was very interesting to see this side of the assessment. I still struggle with the amount of time we put each of our cognitive children through to assess them and quite honestly, so many of the questions are above their level of understanding. I know that we "have" to do them, I just wish there was a better way of adjusting the eval to the student. Thanks so much for listening to our concerns this week and for allowing us the opportunity to take part in this experience.
- I gained a deeper understanding of the SD standards I learned about the Core Content Connectors. I will take how to better construct test questions to my classroom. I gained a better understanding of DOK. I loved the exchange with my peers from around the state. I am a 5th grade teacher so talking with my colleges will help me to better prepare my students for middle school. I am sorry if this is not easy to read it is typing in in reverse, very odd.
- It was helpful to spend time in the standards and learn about the core content connectors. I also enjoyed networking with other teachers.
- I did not know about Core Content Connectors before this meeting.
- Gaining better understanding the standards/connectors and alignment process will help how some classroom concepts are presented and evaluated. Discussions with people from different backgrounds and perspectives was helpful to see the bigger picture.
- This was extremely important; more teachers should have the opportunity. As, being able to review information from the test and having days to be immersed in talking

through standards, what they mean and how they could be tested with other educators give a new perspective. Plus, please tell Drifters, they are amazing, in both food and friendliness and deserve more money :)

- I learned about Core Content Connectors, which I had never been exposed to prior. I will use this in my role in my district as our SPED department pushes to integrate RISE students into general ed science classrooms. The opportunity for dialogue with other educators from other school districts is always beneficial. I always gain a deeper understanding of the standards each and every time I have an opportunity to work with other teachers in this type of format.
- Every time I do one of these studies, I deepen my understanding of how state standards can assist or mislead instructional approaches. So, the exposure to all different standards from all different states and the intricacies of each is always very enlightening.

8.

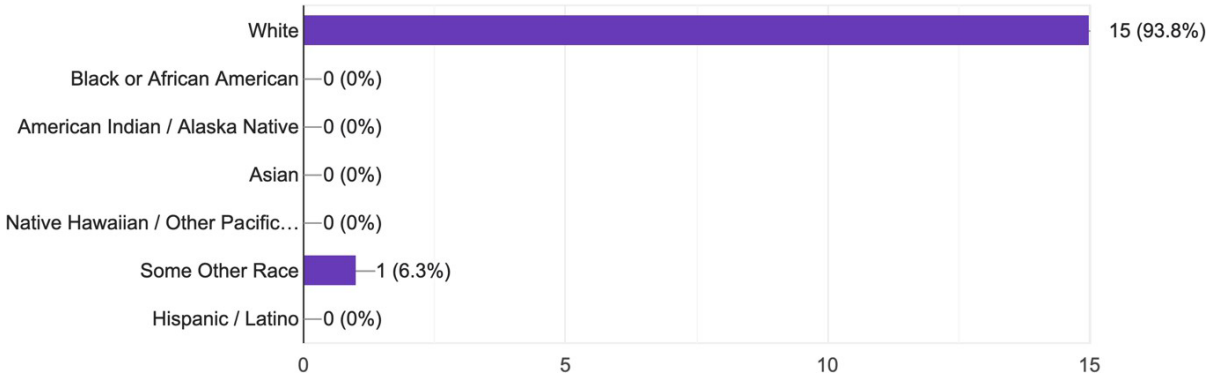
The South Dakota Department of Education includes the following question in all PD evaluations. Please answer this question as relates to the infor... statement? The information presented was useful."
16 responses



Demographic Self-Reporting

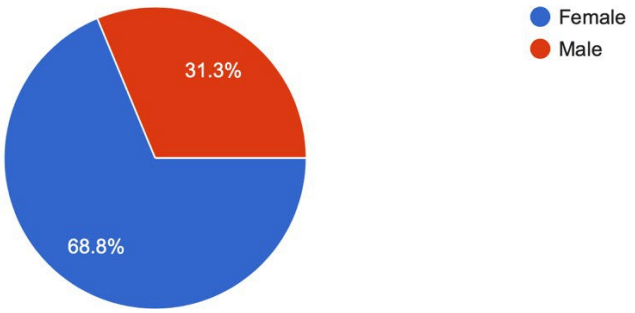
Race/Ethnicity

16 responses



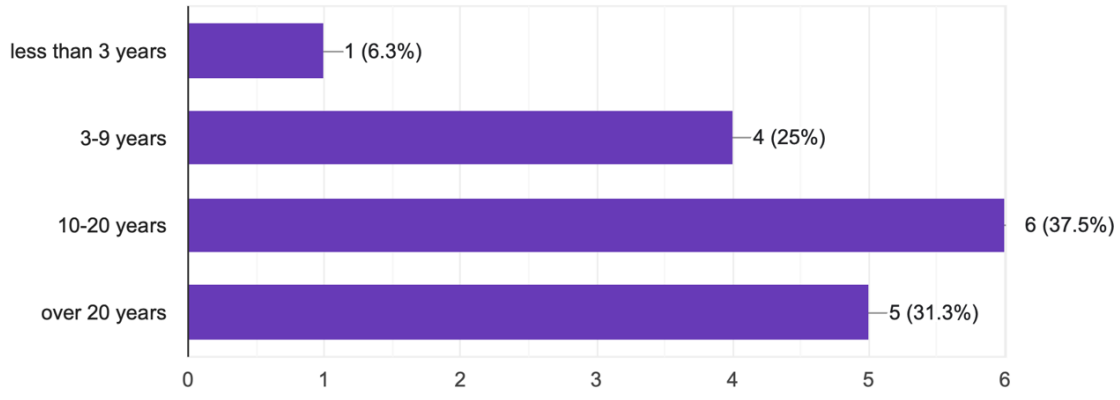
Gender

16 responses



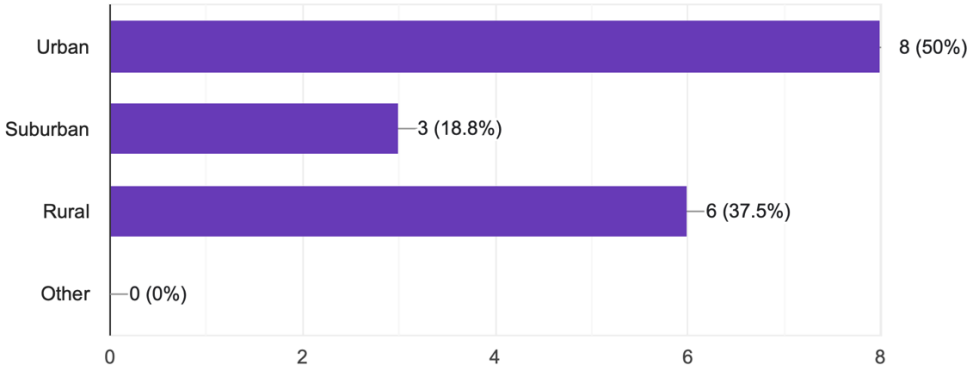
Years of teaching experience

16 responses



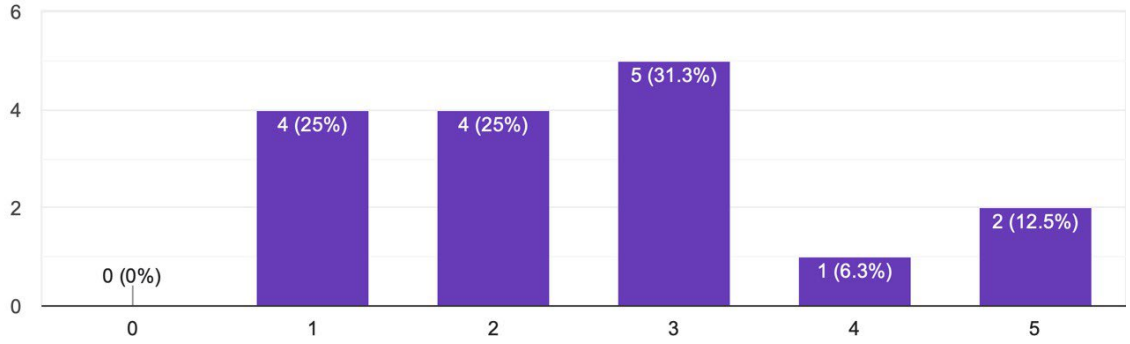
Geographic region(s) in which you have experience teaching

16 responses



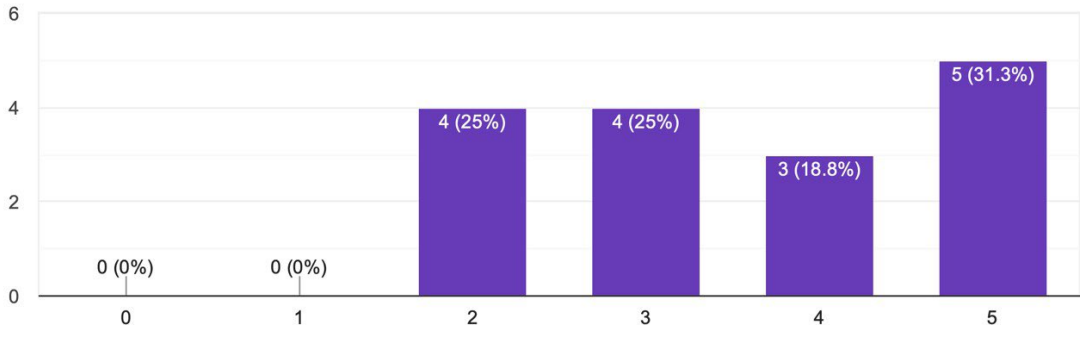
Please report your experience working with English learners / multilingual learners.

16 responses



Please report your experience working with Special Education.

16 responses



Appendix 4-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).

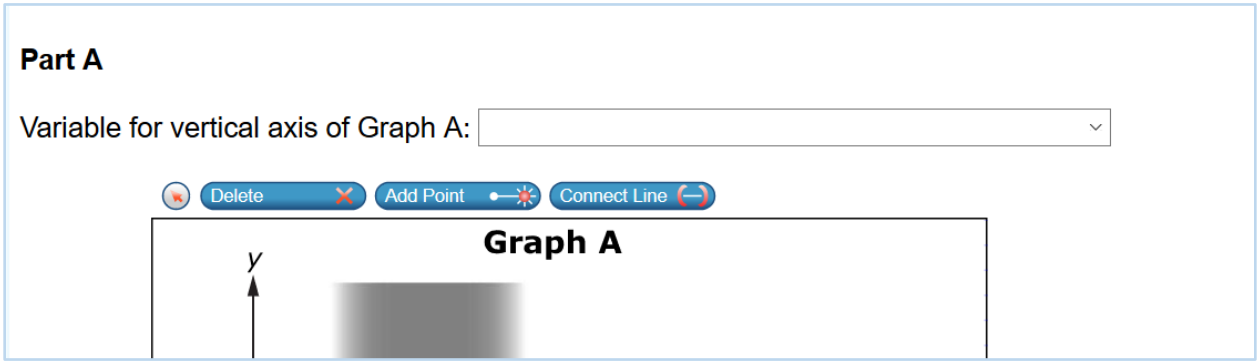
Finally, there were issues associated with the way in which drop-down boxes were handled by JAWS. Some students were not familiar with the term "combo boxes," which was used to describe these boxes, and many students were confused by the ways in which JAWS handled the response options for these boxes. 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.

Appendix 4-F

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.

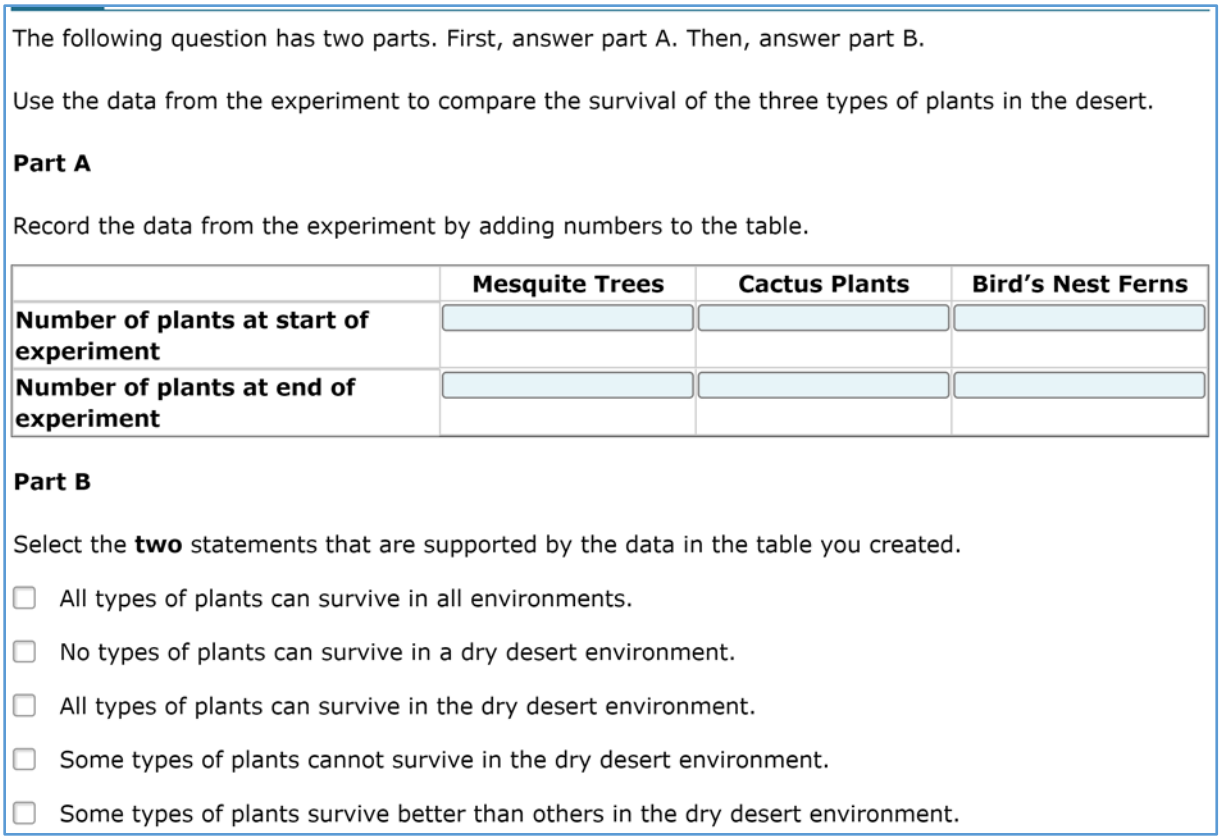
Ⓑ The Grand Canyon region was once underwater.

Ⓒ 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:

Part B

From the list of additional experiments, select the evidence that would support your answer in part A.

- ☐ Scientists grow a sample of wild-type *Mycobacterium tuberculosis* in the lab. Over time, some of the bacteria show resistance to rifampin.
- ☐ Scientists plate a colony of wild-type *Mycobacterium tuberculosis* and a colony of *Escherichia coli* in one petri dish. Some of the new colonies show resistance to rifampin.
- ☐ Scientists plate a colony of wild-type *Mycobacterium tuberculosis* and a colony of mutant *Mycobacterium tuberculosis* in one petri dish. Some of the new colonies show resistance to rifampin.
- ☐ Scientists create additional *Mycobacterium tuberculosis* mutants by creating substitution mutations in the DNA that codes for amino acids 36-67. Many of the mutants are resistant to rifampin.

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:

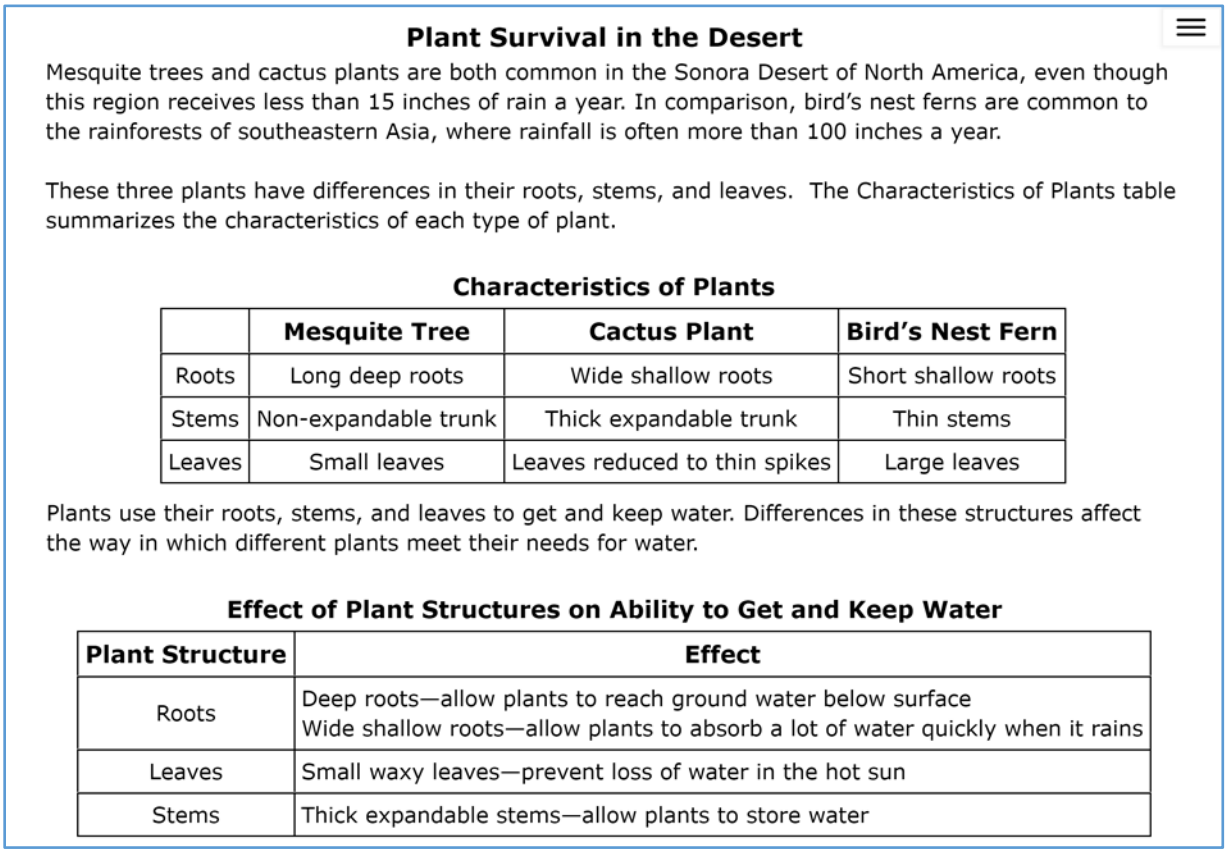
- Organize or summarize data to highlight trends and patterns and/or determine relationships between the traits of an organism and survival in its environment.
- Understand and generate simple bar graphs or tables that document patterns, trends, or relationships between traits of an organism and its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

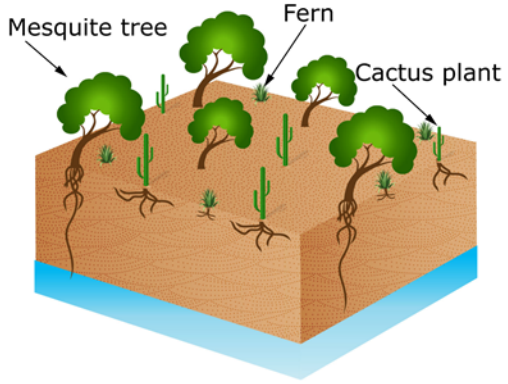
Plants use their roots, stems, and leaves to get and keep water. Differences in these structures affect the way in which different plants meet their needs for water.

Effect of Plant Structures on Ability to Get and Keep Water

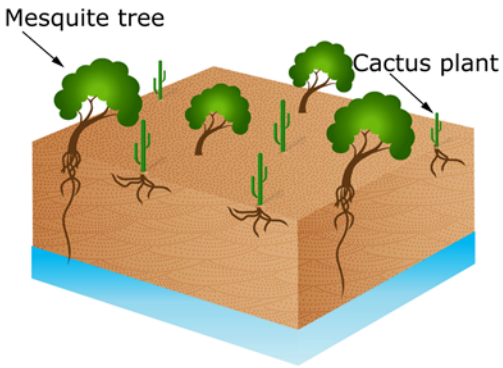
Plant Structure	Effect
Roots	Deep roots—allow plants to reach ground water below surface Wide shallow roots—allow plants to absorb a lot of water quickly when it rains
Leaves	Small waxy leaves—prevent loss of water in the hot sun
Stems	Thick expandable stems—allow plants to store water

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



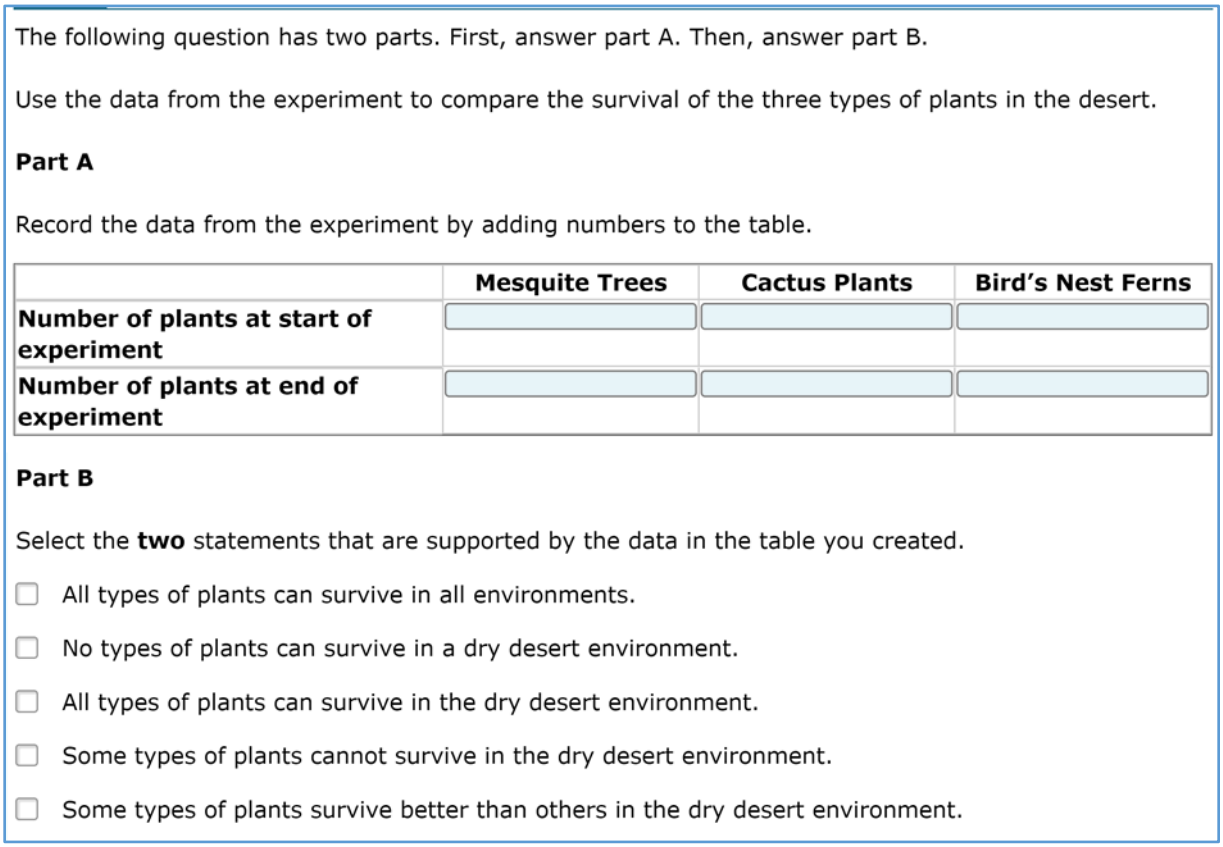
In the questions that follow you will construct an argument for why some plants survive better in the desert than others.

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.

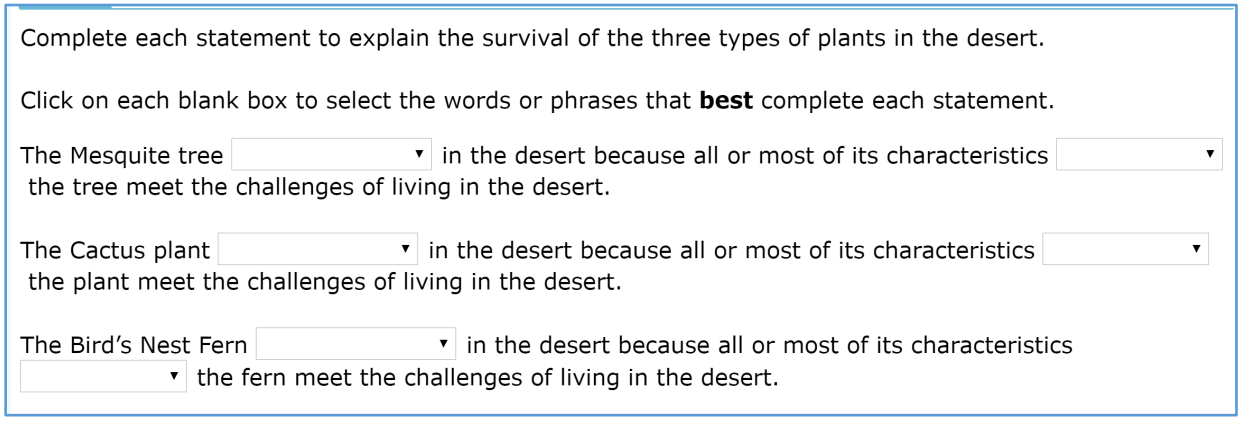
COMPREHENSION

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

A German pyramid candle is a decoration whose parts only move when the candles are lit. The parts that move are driven by a fan that sits on the top of the pyramid. As the fan turns, other parts of the pyramid turn. The animation shows an example of a German pyramid candle. Click the small gray arrow to begin the animation.

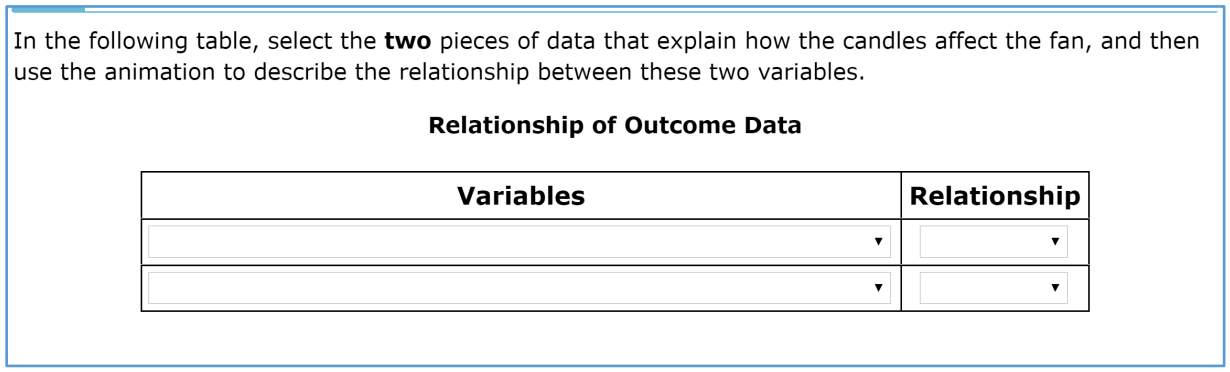
Use the following questions to determine how energy is transferred from the candles to the fan blades.

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



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

Use the table below to correctly order the statements based on what you have observed. Use the numbers 1 through 4 to order your statements, 1 being the first step and 4 being the last step. Use the "-" sign to indicate that the statement is not a part of the process you observed.

Step	Statement
▾	Air moves upward past the fan blades
▾	Light from candles transfers energy to the air
▾	Air gets hotter
▾	Moving air transfers energy to the fan blades
▾	Air transfers heat energy to the fan blades
▾	Heat from candles transfers energy to the air
▾	Light energy carries the air upwards past the fan blade

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

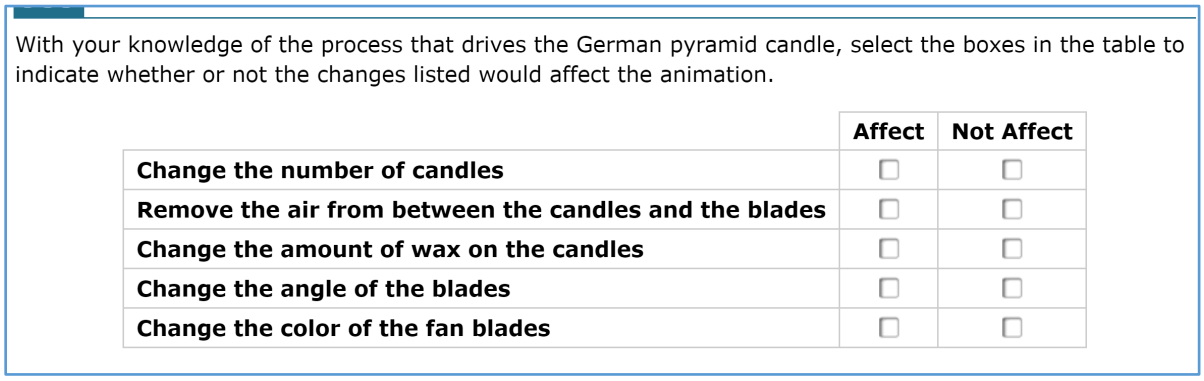
Students had the same issues with lack of content knowledge as they did with Item 1.

> 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

The Grand Canyon is a huge canyon located in Arizona. The canyon has been formed by the Colorado River. The river has cut down into the ground, exposing rock layers that were deposited millions of years ago. The picture shows part of the Grand Canyon.

Portion of Grand Canyon

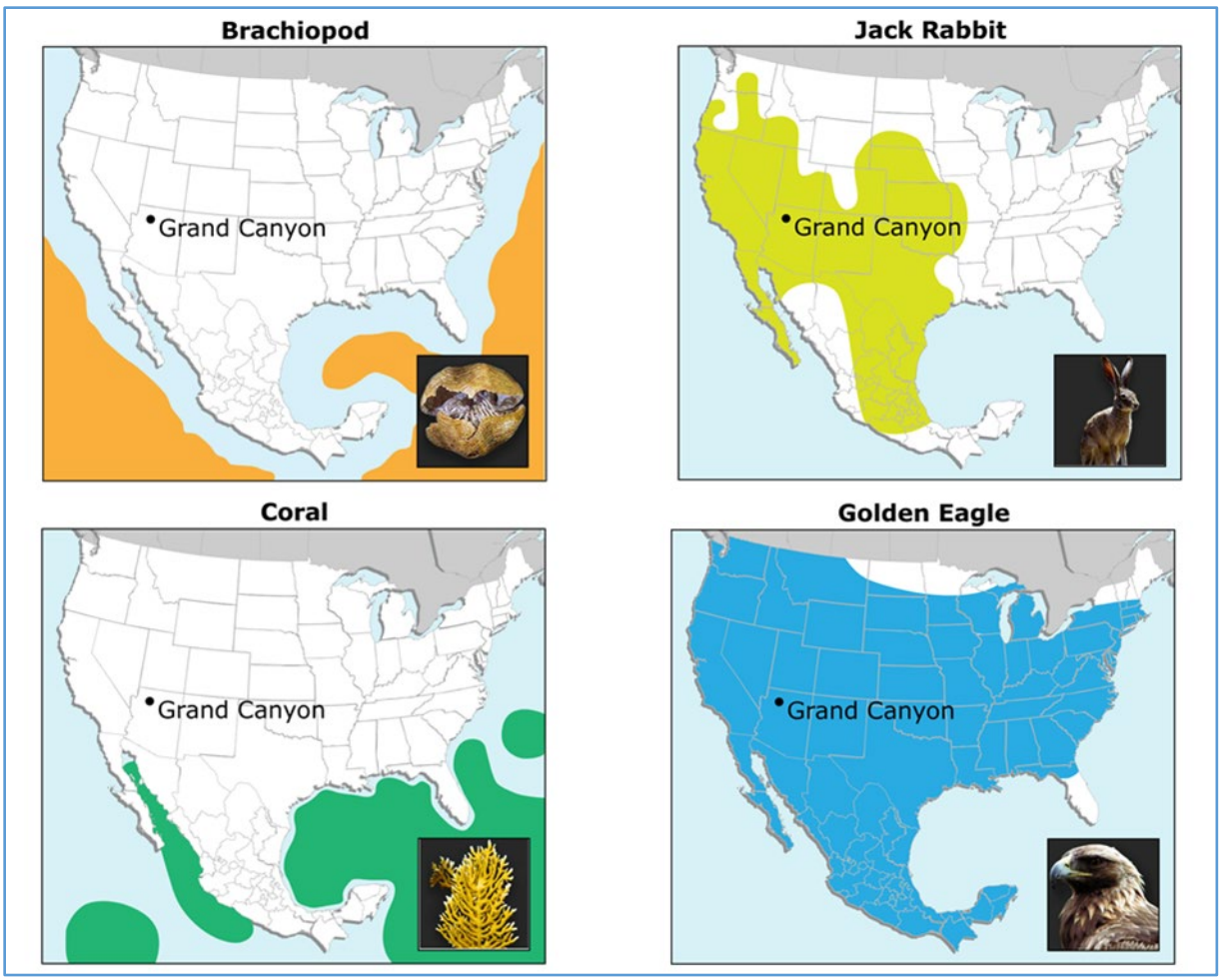
One of these rock layers is called the Redwall Limestone. The Redwall Limestone contains many different fossils, including corals, clams, octopi, and fish.

In the questions that follow, you will study six animals in order to lean about what Arizona was like when the Redwall Limetone was deposited millions of years ago.

The pictures show the animals and maps of where they are found. The colored regions show where the animals live.

Bighorn Sheep

Octopus



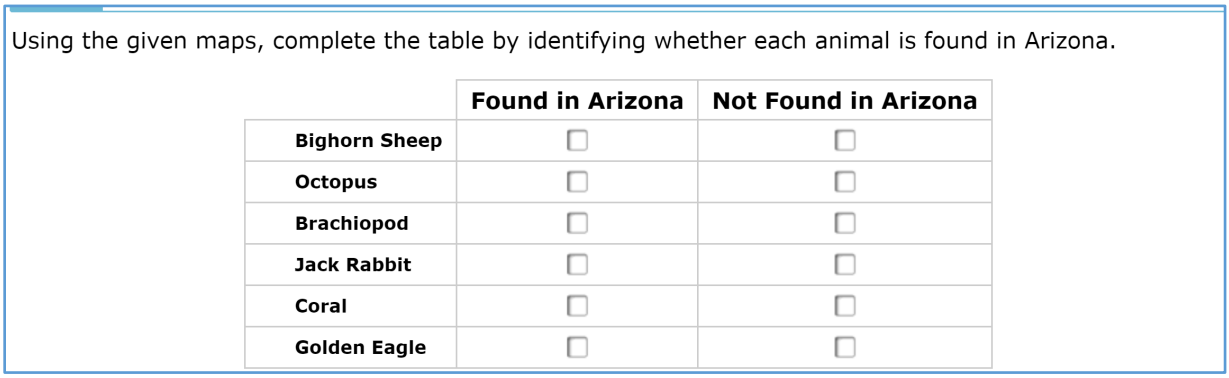
Despite some incorrect responses, nearly all the students seemed comfortable navigating through the maps to decide where the animals are found and filling out the tables in Items 1 and 2. One student did not make any use of the maps.

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.

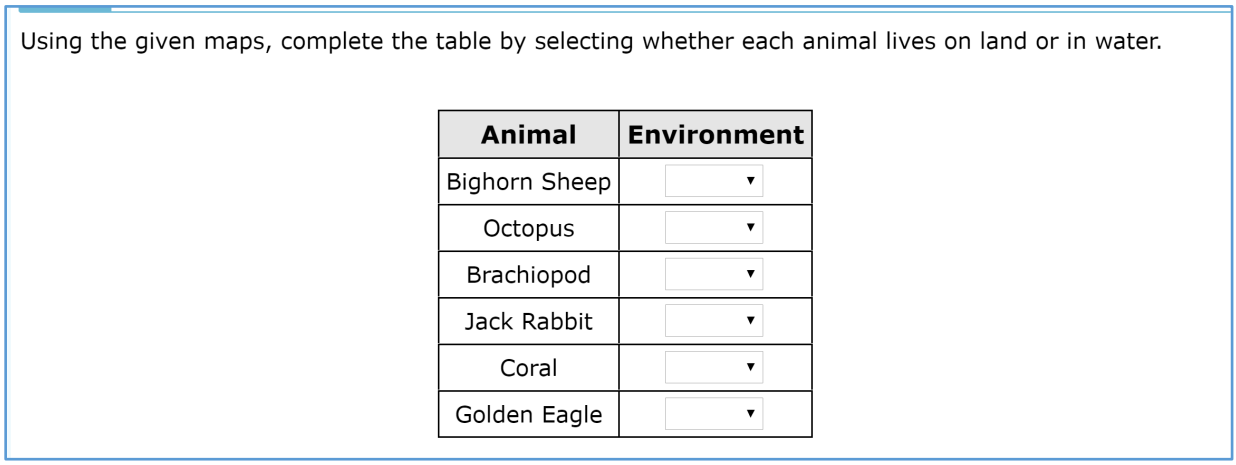
For example, one student

- 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."

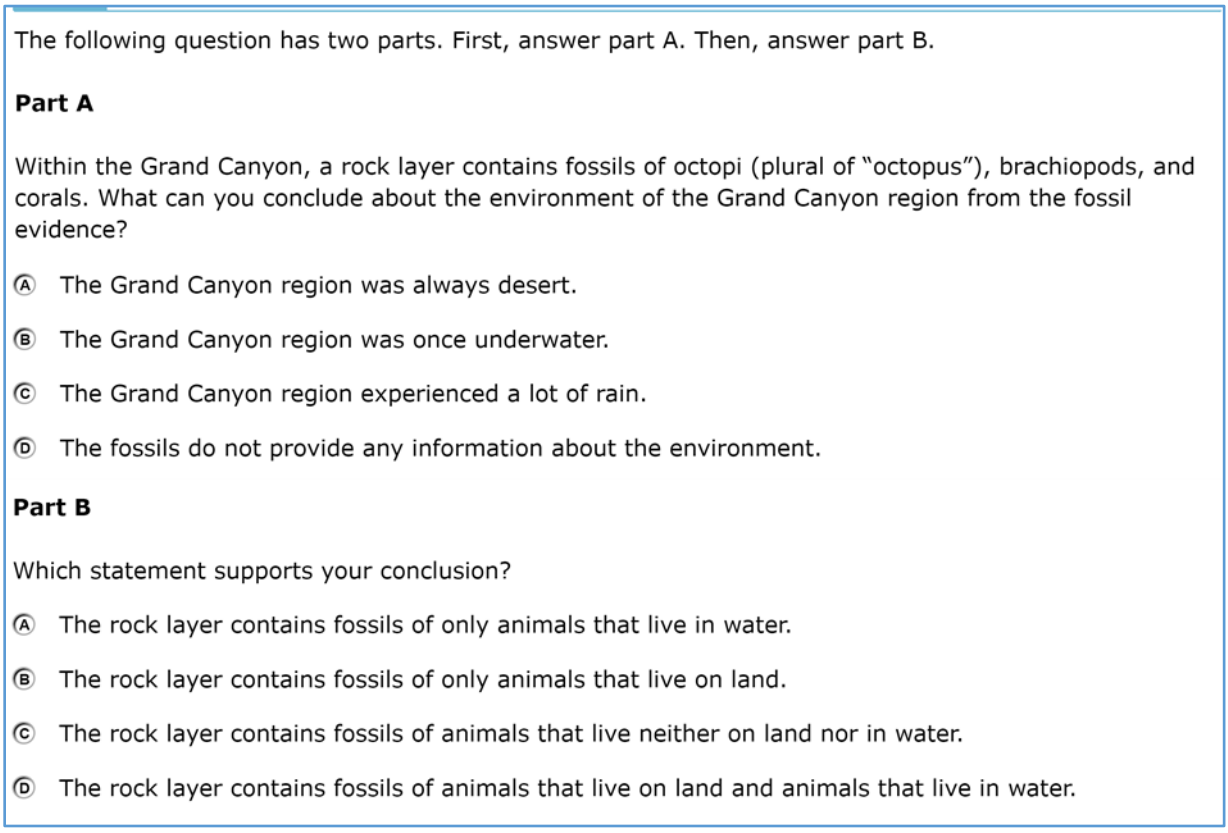
Students who did not earn credit for this item mis-located either the brachiopod or the coral; one student also mis-located the golden eagle. These students also relied on background knowledge for their responses. For example, one student explained his choices as follows:

- 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.

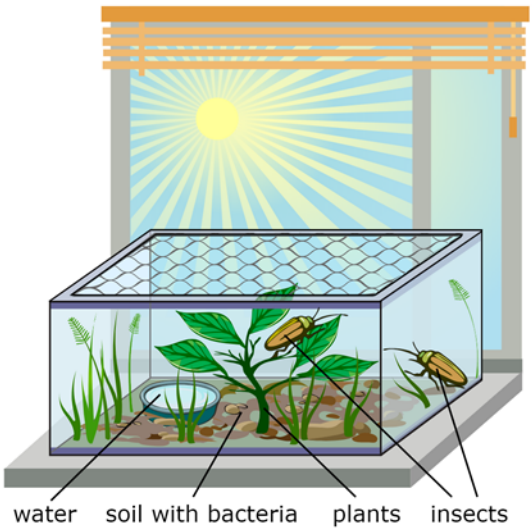
- Manipulate the components of a model to demonstrate properties, processes, and/or events that result in the movement of matter among plants, animals, decomposers, and the environment including the relationships of organisms and/or the cycle(s) of matter and/or energy.
- Articulate, describe, illustrate, 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.

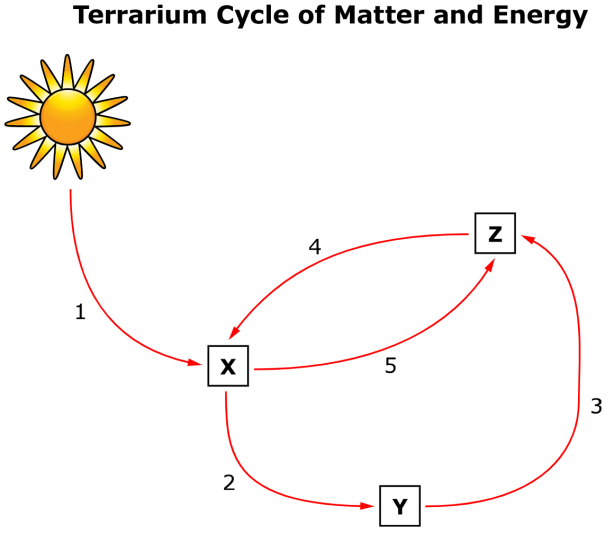
> For example, one student selected option 1, and when asked how she knew, the student said, "if insects don't have food or water they'll die, and I know that just from background knowledge." The student selected option 3 because, "plants need nutrients from the soil, or they will die too… I just used 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

The Terrarium Cycle of Matter and Energy diagram shows an incomplete model of the terrarium environment.



The following question has four parts. First, answer part A. Next, answer part B. Then, answer part C. Finally, answer part D.

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

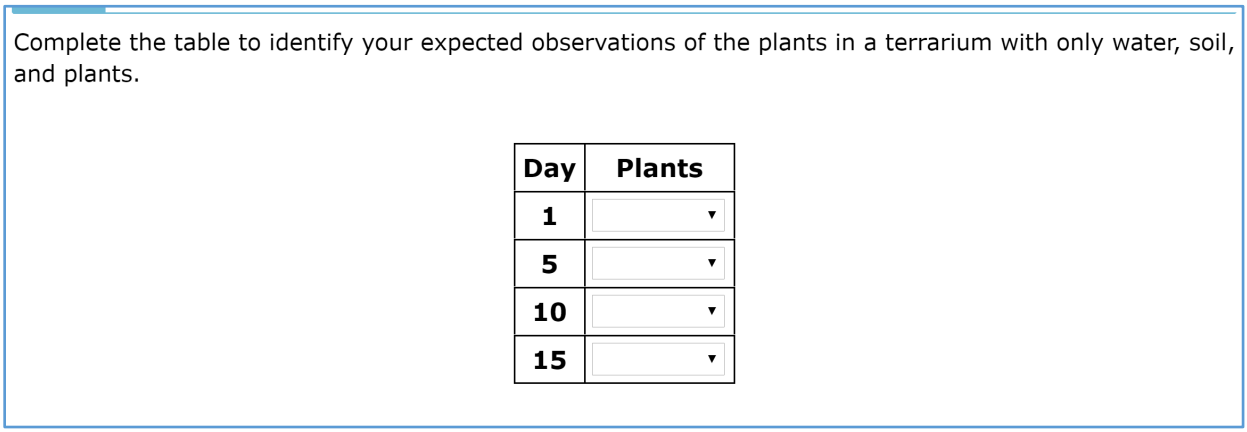
Several students simply lacked the content knowledge to answer this question.

> 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:

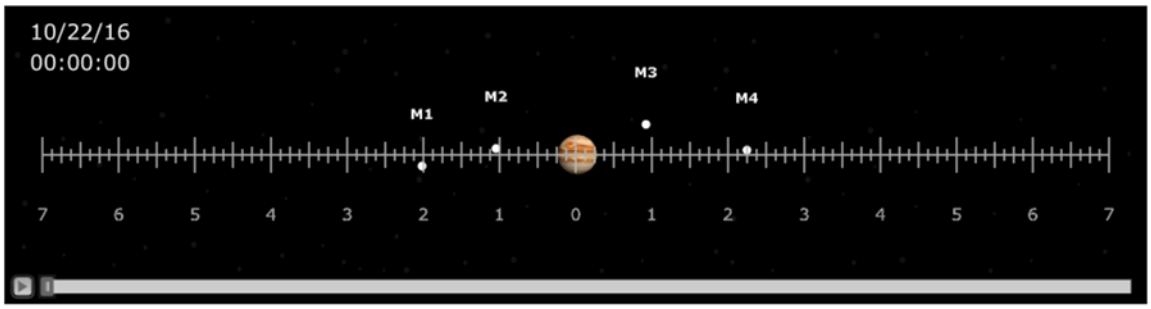
- Make simple calculations using given data to estimate the properties (e.g., mass, surface temperature, diameter) and locations of different solar system objects relative to a given reference point/object (Item 1).
- Calculate or estimate or identify properties of objects or relationships among objects 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

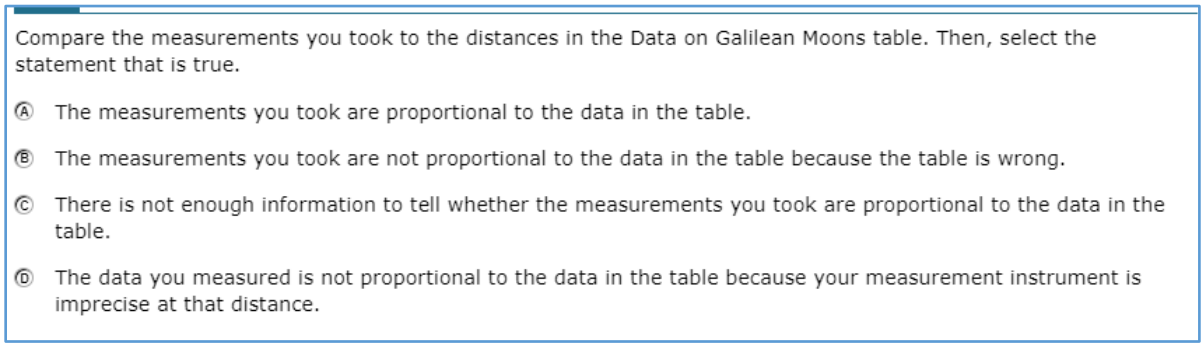
Nearly all the students reasoned their way to a solution using the stimulus materials as intended.

> 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

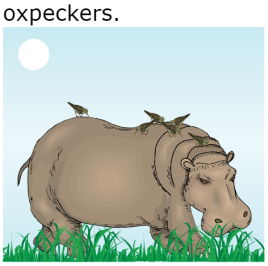
The stimulus for the Hippos cluster is shown in Figure 21.

Figure 21. Stimulus: Hippos

In Africa, a variety of organisms coexist with others in distinct ecosystems. For example, hippopotamuses spend time in both aquatic and savannah ecosystems.

When found in aquatic environments, hippopotamuses are often surrounded by carp.

When found in a savannah environment, hippopotamuses are often surrounded by birds called oxpeckers.

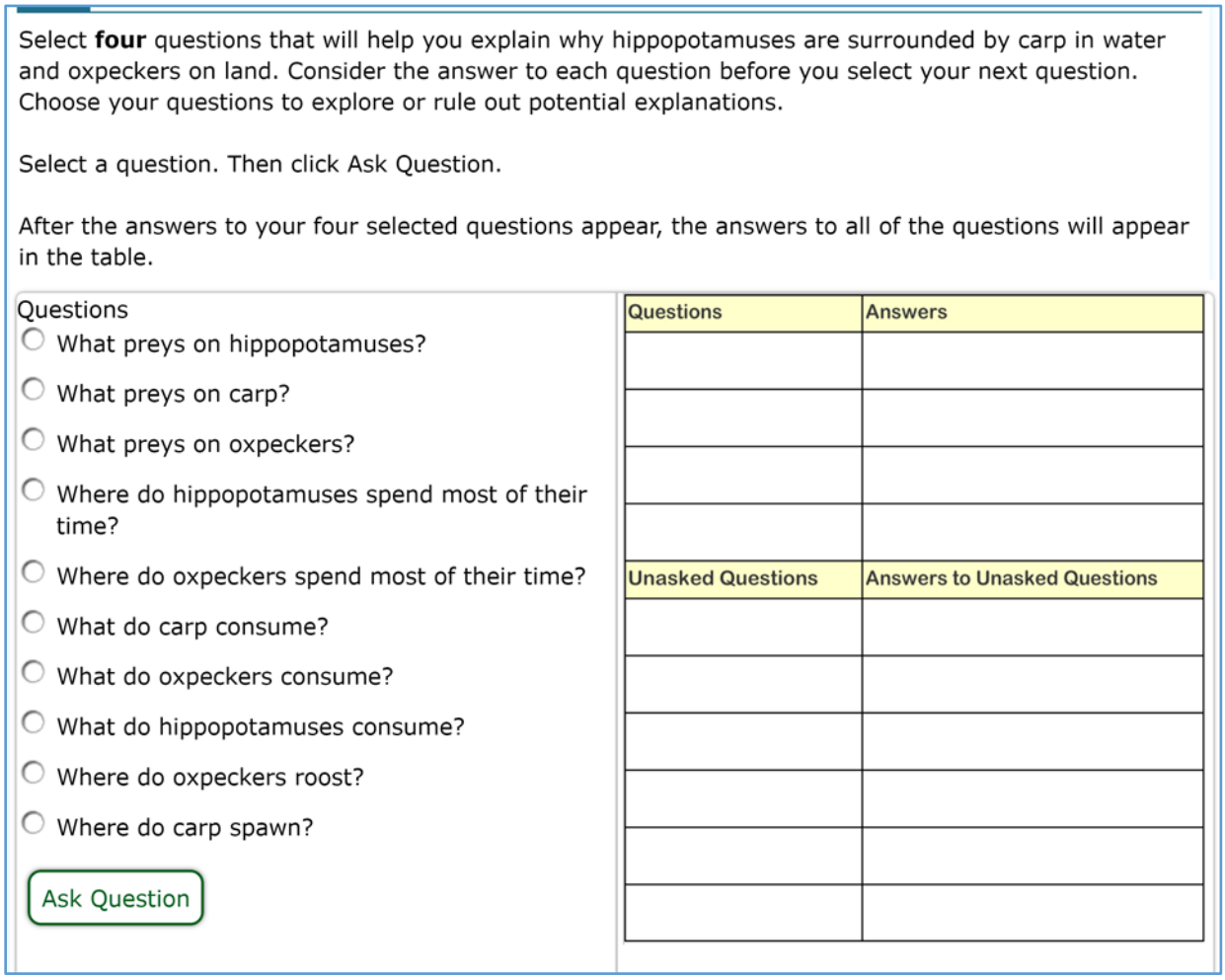


Details by Item

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

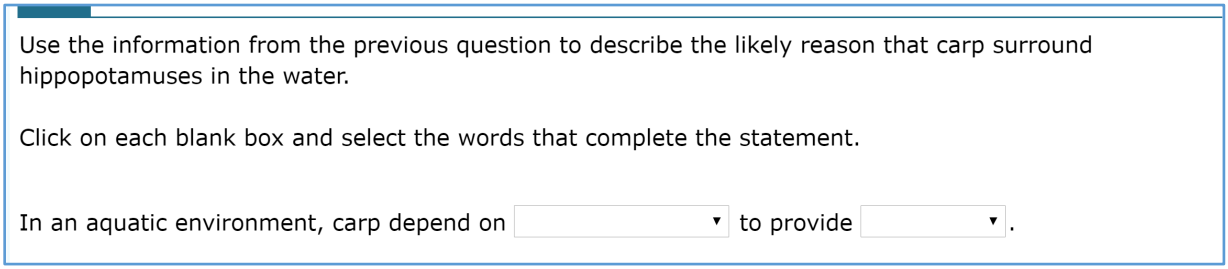
As evidenced from their reasoning in the think-alouds, students understood that they were to choose questions they thought would be helpful to explain the relationships between hippos and oxpeckers or carp, although, as can be seen from the score distribution, they did not necessarily know what those 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."

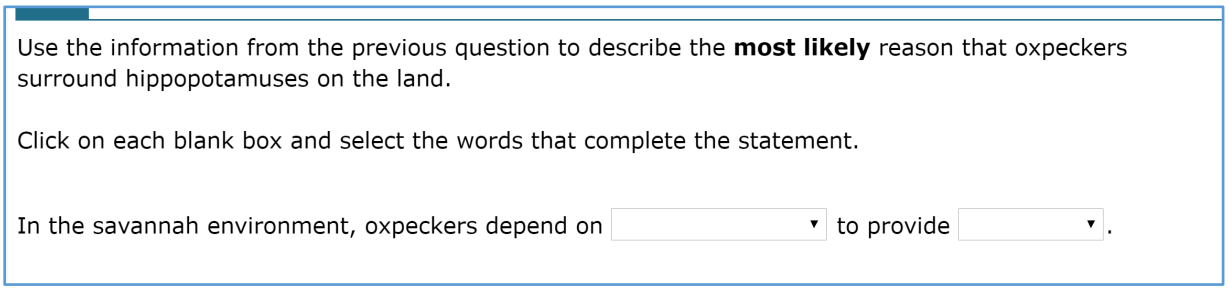
Those who did not respond correctly simply made wrong inferences from the data—some of which were wrong but plausible.

> 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.

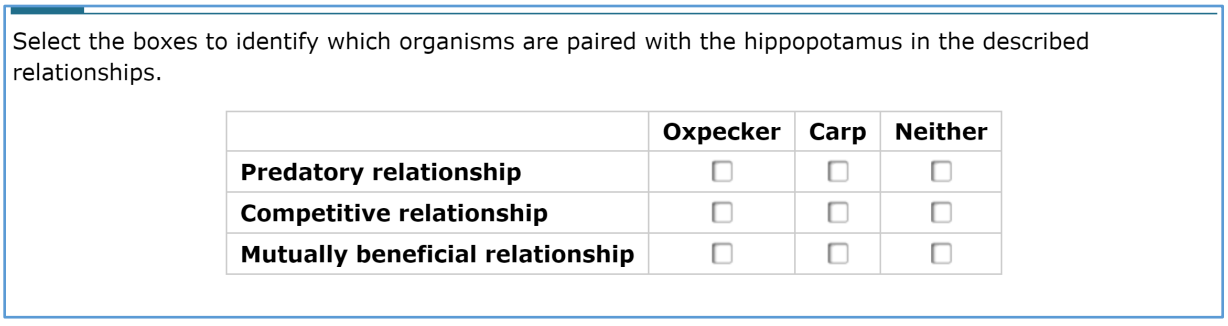
Reasoning

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

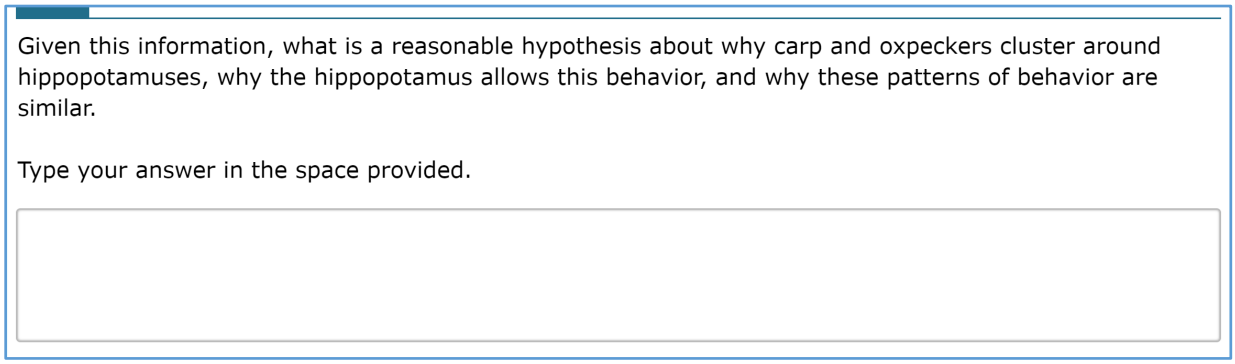
Although students generally understood the concept of transactional relationship among animals, some lacked prior knowledge of the terms used in the item.

> 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

Morning Fog in a Valley

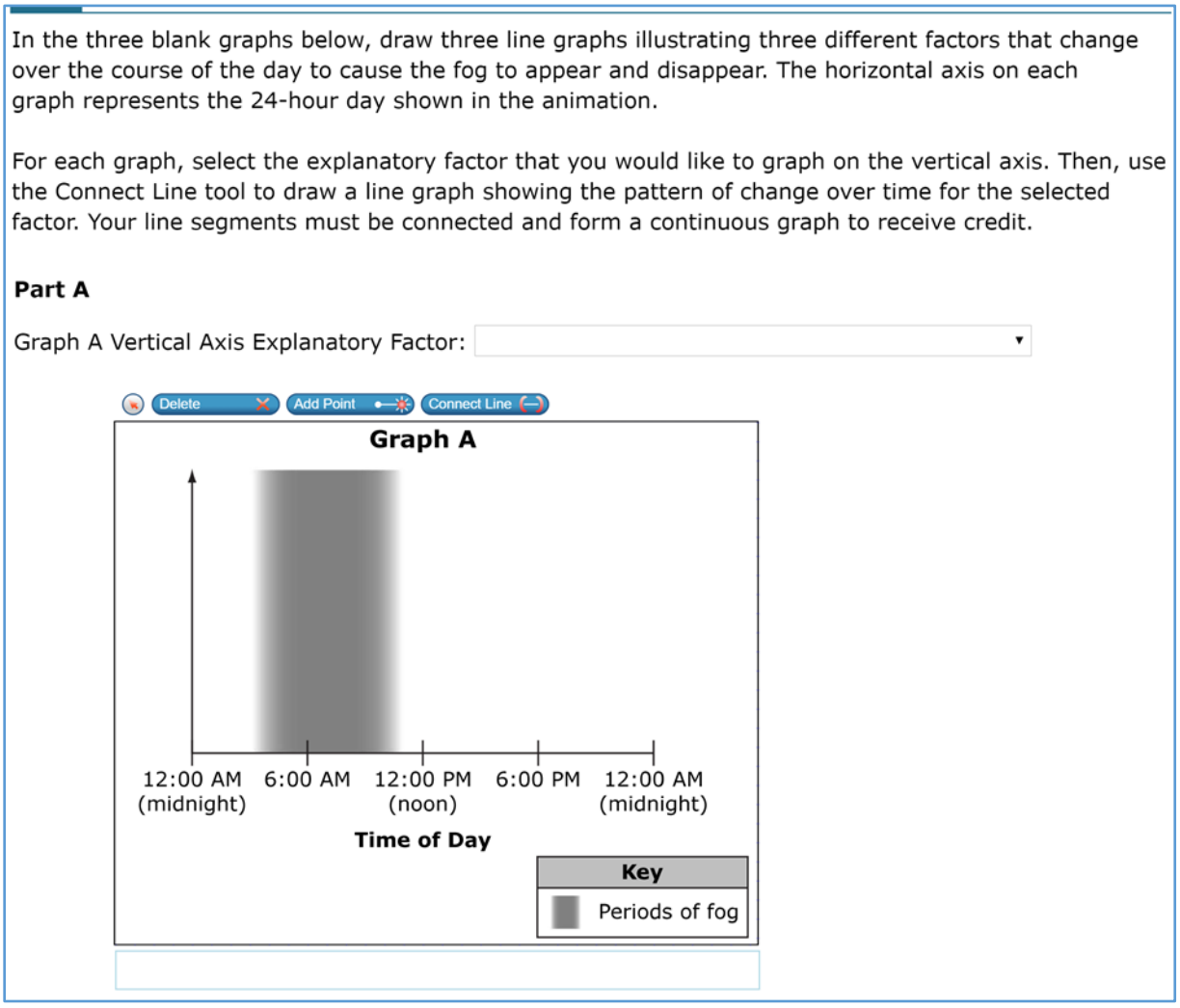
Fog appears and disappears over the course of the morning in the Willamette Valley in Oregon. The animation shows the appearance and disappearance of fog in the valley during a 24-hour day. The sun rises at 6 AM and later sets at 6 PM.

Details by Item

Item 1

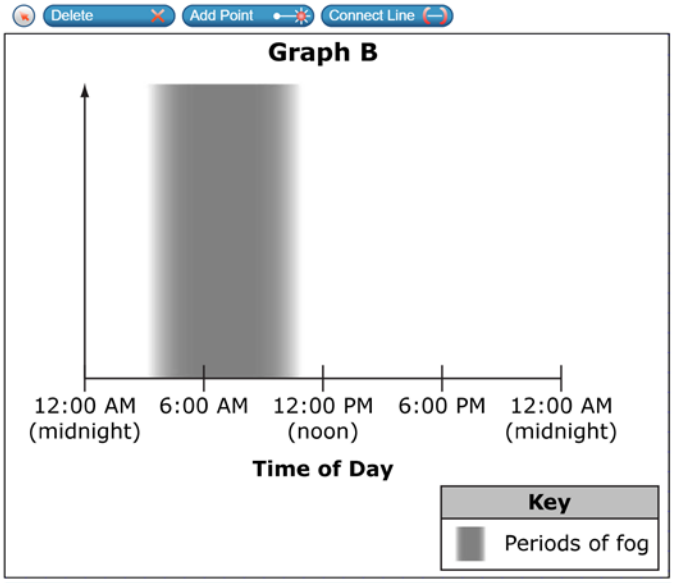
Item 1 of the Morning Fog cluster is shown in Figure 28.

Figure 28. Item 1: Morning Fog



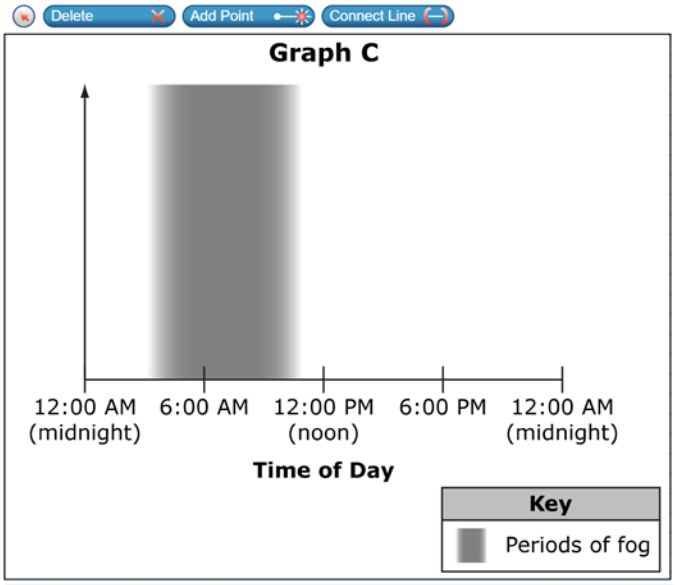
Part B

Graph B Vertical Axis Explanatory Factor:



Part C

Graph C Vertical Axis Explanatory Factor:



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

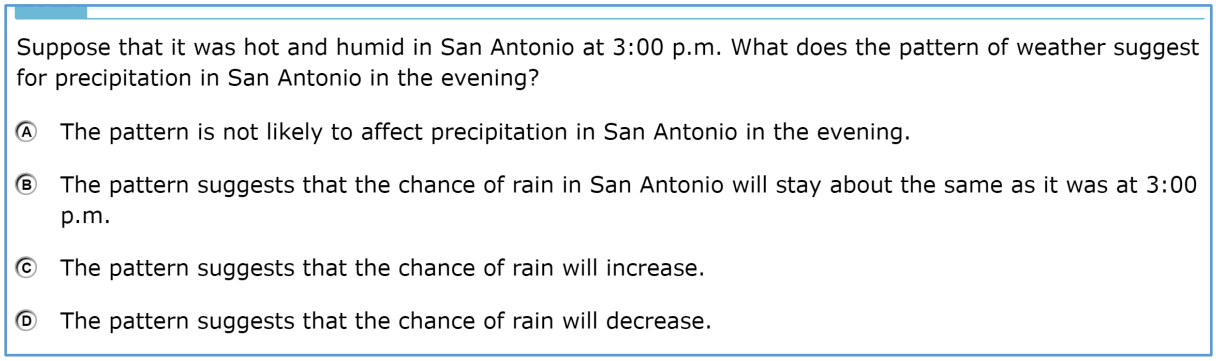
Given their performance on Part A, students had little to work with in Part B, even if they understood what they were supposed to do.

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.

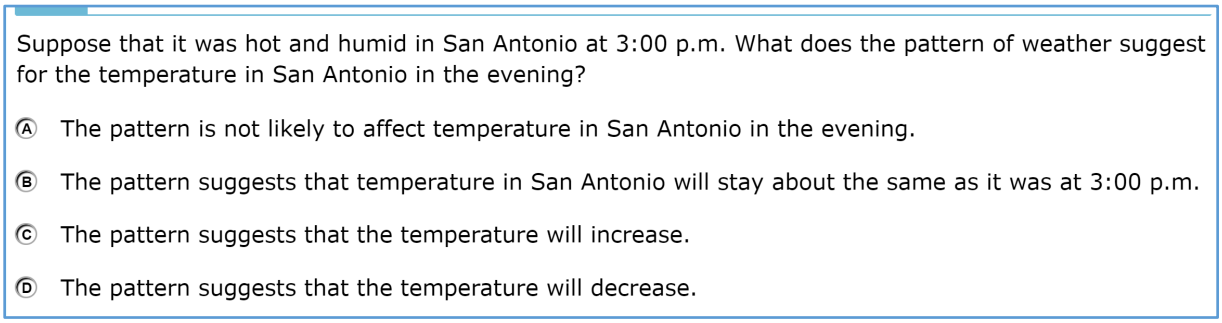
Some students understood "pattern of weather" as referring to the hour-by-hour weather report shown in the stimulus, and it's not clear that any of the students realized that the question pertained to a different location than the weather report (or Item 1).

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.

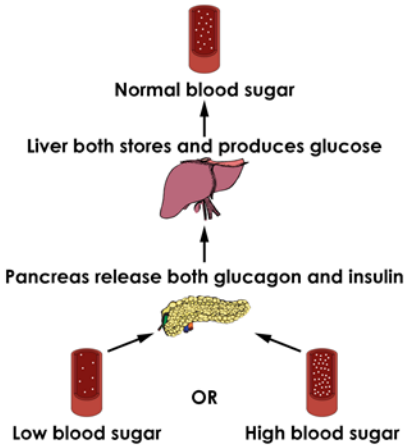
Stimulus

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.



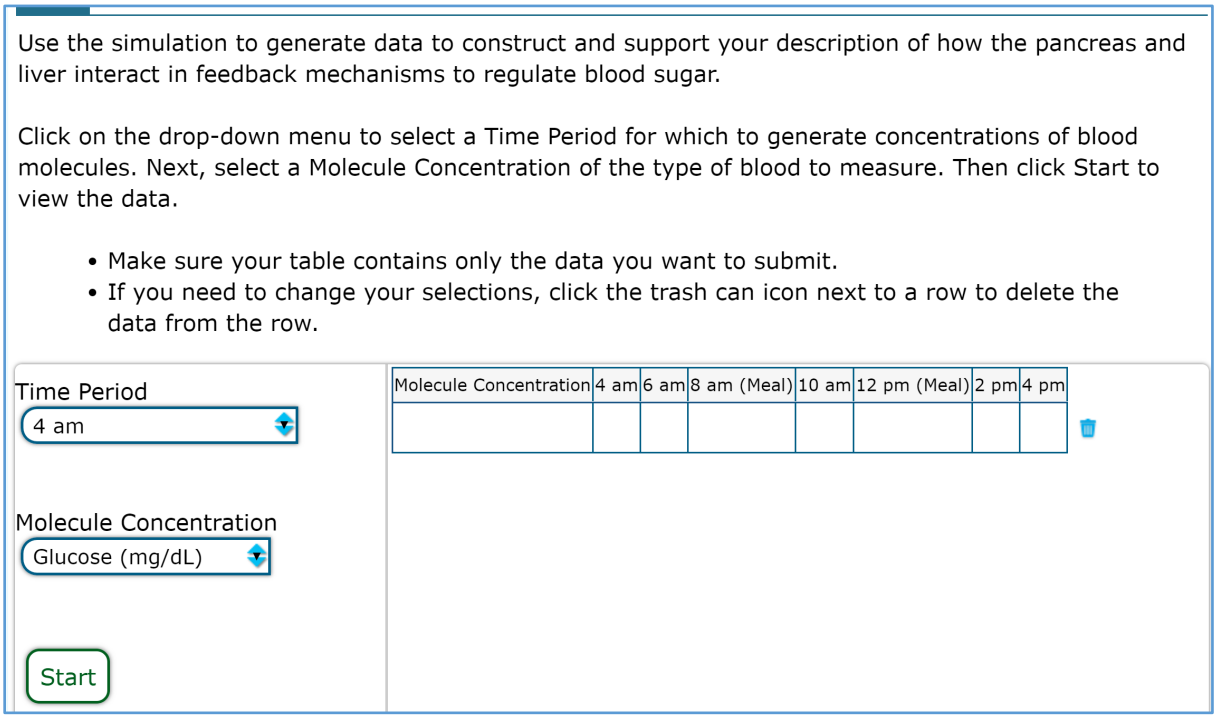
In the questions that follow, investigate and describe how the molecules produced and stored by the pancreas and liver interact in feedback mechanisms to regulate blood sugar.

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.

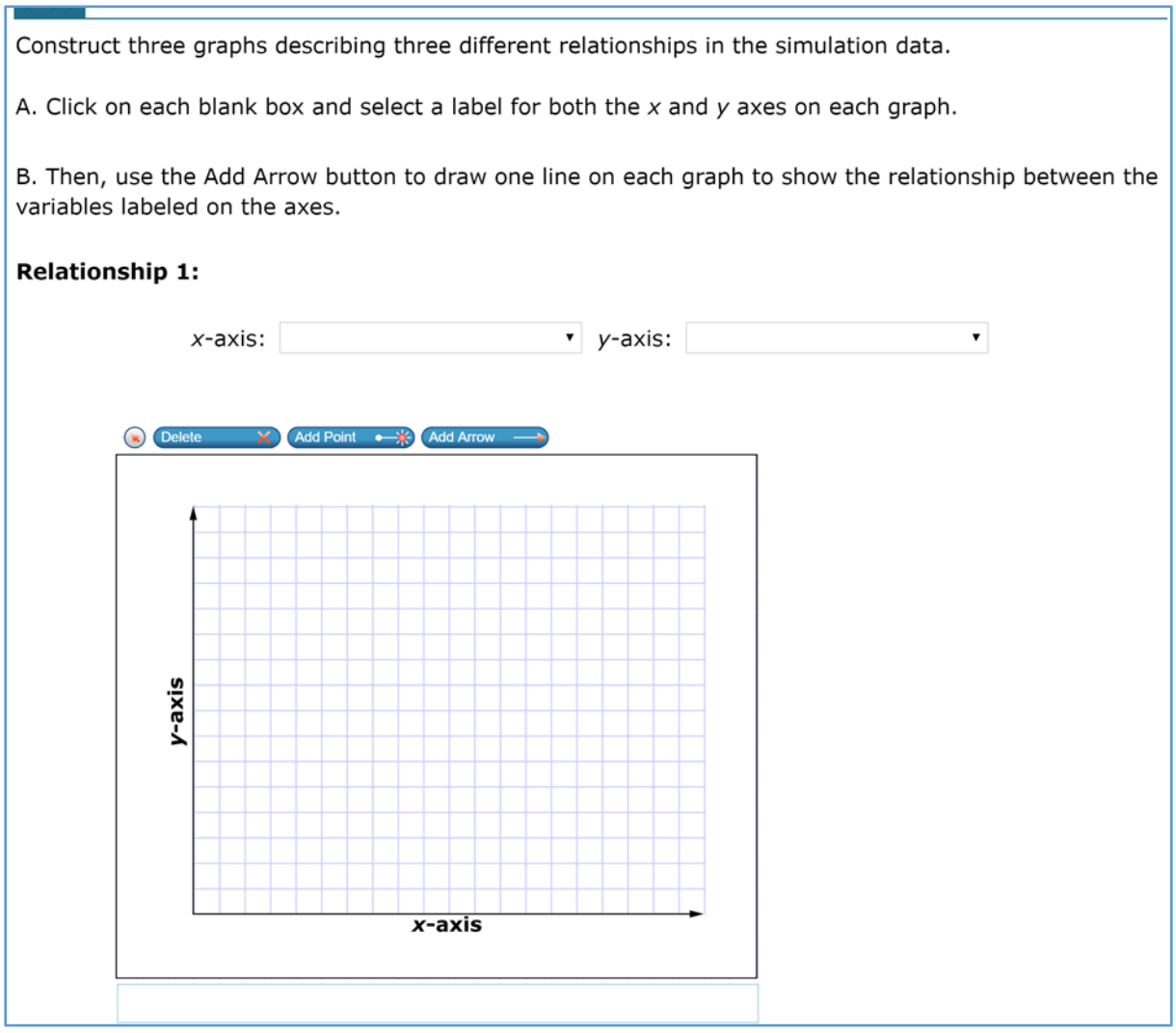
REASONING

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.

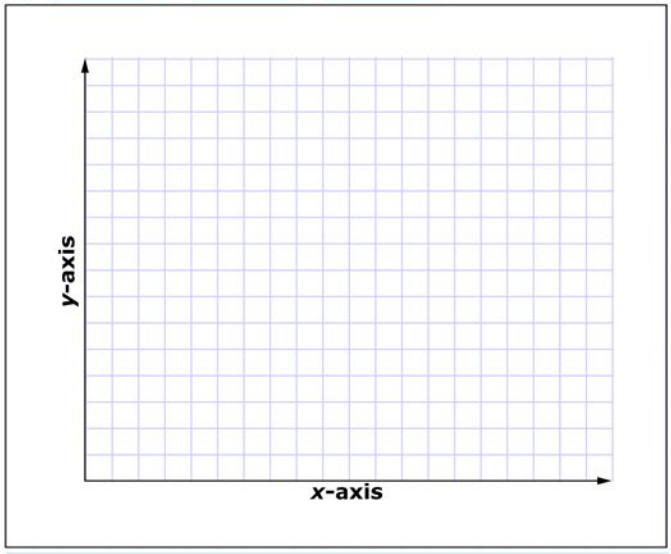
Figure 35. Item 2: Blood Sugar Regulation



Relationship 2:

x-axis: ▾ *y*-axis: ▾

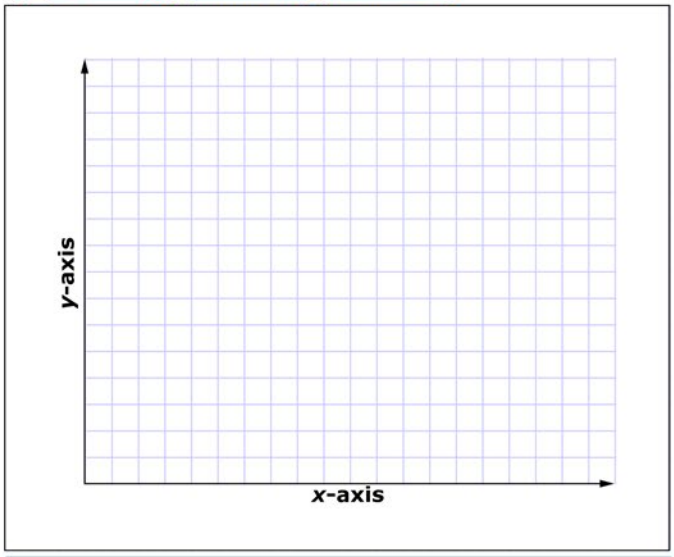
Delete Add Point Add Arrow



Relationship 3:

x-axis ▾ *y*-axis ▾

Delete Add Point Add Arrow



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

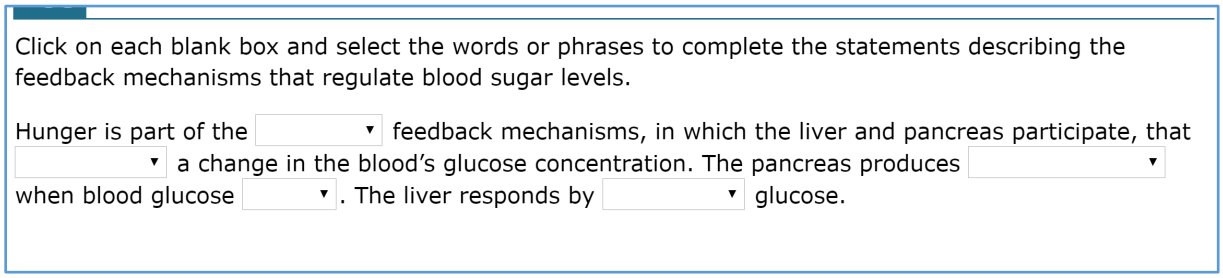
The following is an example of how one student reasoned through the construction of one of the graphs.

> 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.

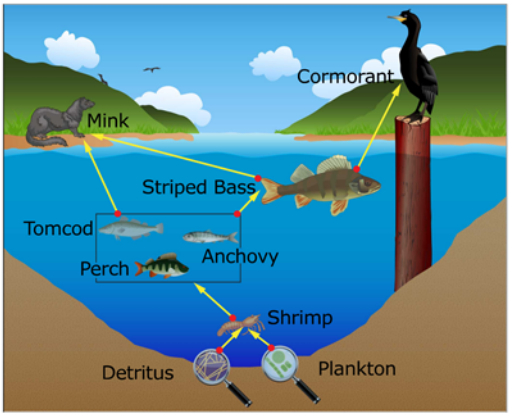
Figure 40. Stimulus: Tomcods

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.

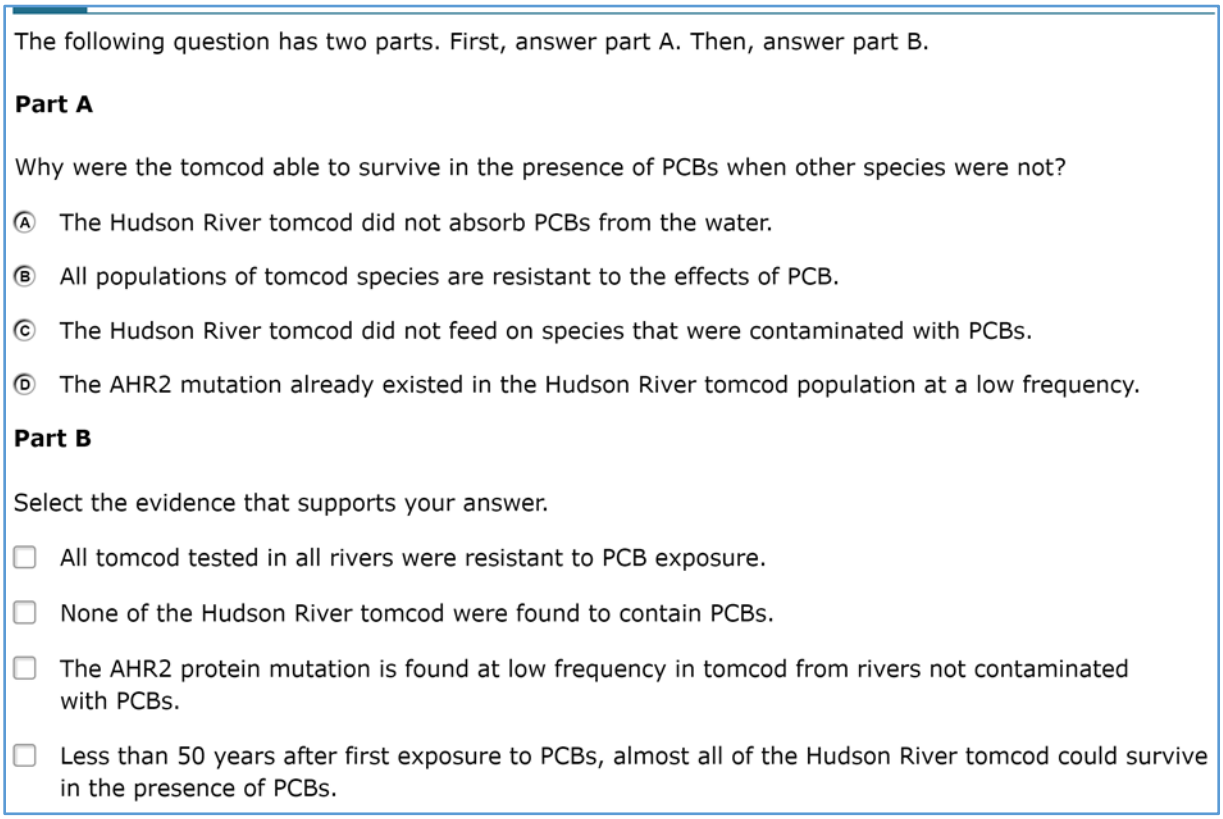
> For example, one student read out loud Hypothesis 1 and 2 in the introduction. She said, "So there's a higher percentage in the Hudson River than in rivers not contaminated," and selected Supports Hypothesis 1 for line 1 "because it's talking about how this one is saying that it's from the water and 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

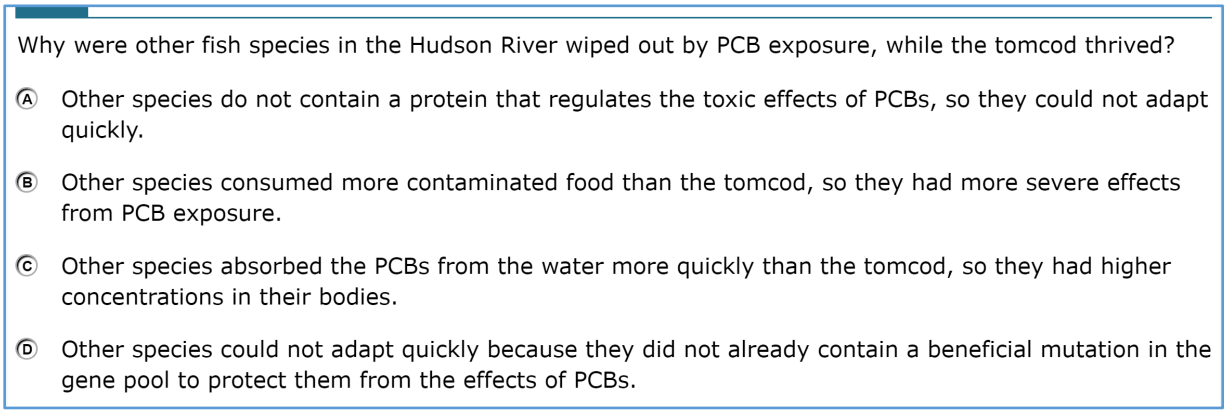
The following is an example of the reasoning of one of the students who correctly identified option D as the reason why Tomcod survived in Part A,

> 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.

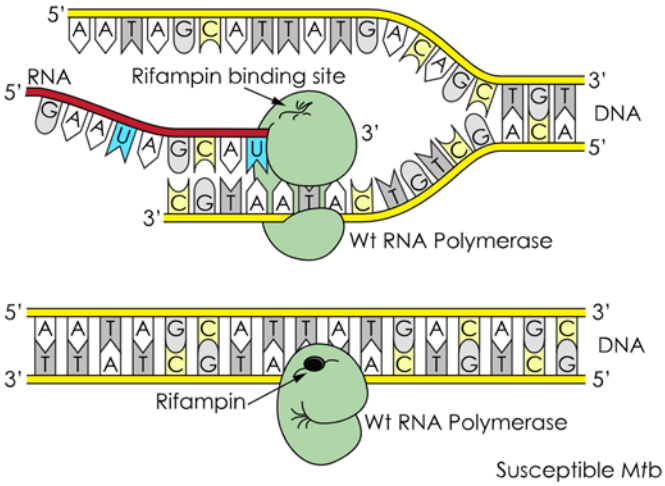
Stimulus

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

If the rifampin cannot bind to the RNA polymerase protein in *Mycobacterium tuberculosis*, this leads to antibiotic resistance. Mutations in the rifampin binding site can block binding of the antibiotic. Based on the information provided, determine which mutants are likely to be resistant to rifampin by this mechanism.

Click on each blank box to select the correct words or phrases.

Resistance of Mutant *Mycobacterium Tuberculosis* Strains

Strain	Resistance	Explanation
Mutant 1	▾	▾ ▾ of rifampin
Mutant 2	▾	▾ ▾ of rifampin
Mutant 3	▾	▾ ▾ of rifampin

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.

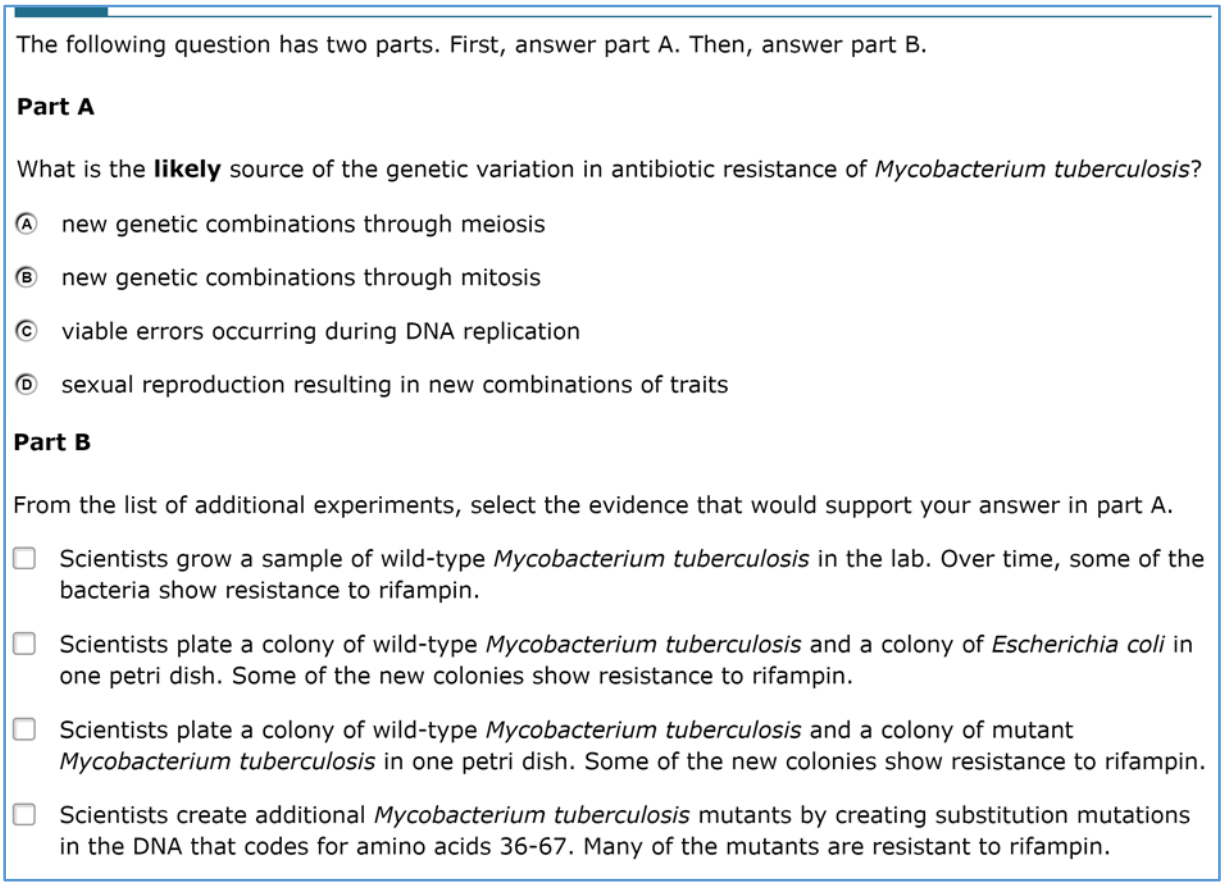
Although only one student had the correct responses for all three of the mutant strains, several used the stimulus materials in the intended manner to reason through the problem.

> 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