West Virginia General Summative Assessment (WVGSA)

2021–2022

Volume 4
Reliability and Validity

West Virginia DEPARTMENT OF EDUCATION

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

1. INTRODUCTION AND OVERVIEW OF RELIABILITY AND VALIDITY EVIDENCE

West Virginia implemented a new assessment program for operational use during the 2017–2018 school year. This new program, named the West Virginia General Summative Assessment (WVGSA), replaced the Smarter Balanced Assessment Consortium (SBAC) in English language arts (ELA) and mathematics, and replaced the West Virginia Educational Standards Test (WESTEST) in science. The WVGSA is delivered as online, adaptive assessments to students in grades 3–8 in ELA and mathematics. For science, the test is administered online for grades 5 and 8 using an adaptive design. Accommodated versions are available for each grade, including braille and a large print Data Entry Interface (DEI) form for ELA, mathematics, and science. Spanish language versions of the mathematics and science tests are also available. Table 1 shows the complete list of tests for spring 2022, which was the fourth year of operational test administration.

Table 1: Test Form and Administration Mode

Subject (language/format)	Administration Mode	Grade
ELA (English/adaptive)	Online	3–8
ELA (English/adaptive-braille)	Online	3–8
ELA (English/fixed-DEI)	Paper	3–8
Mathematics (English/adaptive)	Online	3–8
Mathematics (English/adaptive-braille)	Online	3–8
Mathematics (Spanish/adaptive)	Online	3–8
Mathematics (English/fixed-DEI)	Paper	3–8
Mathematics (English/fixed-braille)*	Paper	3–8
Science (English/adaptive)	Online	5 and 8
Science (Spanish/adaptive)	Online	5 and 8
Science (English/fixed-DEI)	Paper	5 and 8
Science (English/fixed-braille)	Online	5 and 8

**P35 Braille was administered online with a paper supplement.*

Given the intended uses of these tests, both reliability evidence and validity evidence are necessary to support appropriate inferences of student academic achievement from the WVGSA scores. The analyses to support reliability and validity evidence that are reported in this volume were conducted on the basis of students' completed test results, which were obtained through the online administration of English versions of test forms.

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

1. **Reliability.** The reliability of the WVGSA adaptive test forms is estimated using marginal reliability in the item response theory (IRT) framework. The reliability estimates are presented by grade and subject and demographic subgroup. This

discussion also includes the conditional standard error of measurement (CSEM), reliability of performance classifications, and inter-rater reliability (IRR) of ELA writing scores provided by Cambium Assessment, Inc.'s (CAI) Autoscoring Model.

2. **Content Validity.** This section presents evidence showing that test forms were constructed to measure the West Virginia College- and Career-Readiness (WNCCR) Standards with a sufficient number of items targeting each area of the test blueprint.

3. **Internal Structure Validity.** Evidence is provided regarding the internal relationships among the subscale scores to support their use and to justify the IRT measurement model. This type of evidence includes observed and disattenuated Pearson correlations among reporting categories. As explained in detail in Volume 1, Annual Technical Report, for science, the IRT model is a multidimensional model, with an overall dimension representing proficiency in science and nuisance dimensions that account for within-item local dependencies among scoring assertions. In this volume of the technical report, evidence is provided with respect to the presence of item cluster effects. Confirmatory factor analyses (CFAs) have also been performed for the three subjects. Additionally, local item independence, an assumption of unidimensional IRT, was evaluated using the Q_3 statistic in spring 2018 for ELA and mathematics. The CFA and Q_3 statistics were kept as a reference in this document.

4. **Relationship of Test Scores to External Variables.** Evidence of convergent and discriminant validity is provided using observed and disattenuated subscore correlations both within and across subjects. The correlations between the interim and summative assessments, as well as the correlation between SBAC spring 2017 and WVGSA spring 2018 summative assessments in ELA and mathematics are also presented.

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

1.1. RELIABILITY

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

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

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

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

This equation 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 varies over the ability continuum. These heterogeneous errors are a function of a test information function (TIF) that provides different information about test takers depending on their estimated abilities. Often, the TIF is maximized over an important performance cut, such as the proficiency cut score.

Because the TIF indicates the amount of information provided by the test at different points along the ability scale, its inverse indicates the lack of information at different points along the ability scale. This lack of information is the uncertainty, or the measurement error, of the score at various score points. Conventionally, fixed-form tests are maximized near the middle of the score distribution, or near an important classification cut, and have less information at the tails of the score distribution. Refer to Section 3.3, Test Information Curves and Standard Error of Measurement for ELA and Mathematics, for the derivation of heterogeneous errors in the IRT.

1.2. VALIDITY

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

The first source of validity evidence is the relationship between the test content and the intended test construct (refer to Section 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. During these studies, experts review individual items and rate them based on how well they match the test specifications or cognitive skills required for a construct (refer to Volume 2, Test Development for details). Test scores can be used to support an intended validity claim when they contain minimal construct-irrelevant variance. For example, a mathematics item targeting a specific mathematics skill that also requires advanced reading proficiency and vocabulary has a high level of construct-irrelevant variance. Thus, the intended construct of measurement is confounded, which impedes the validity of the test scores.

Statistical analyses, such as factor analysis or multi-dimensional scaling, are also used to evaluate content relevance. The results from factor analysis for the fixed-form spring 2018 WVGSA for ELA and mathematics are presented in Section 5.2, Confirmatory Factor Analysis for Spring 2018 ELA and Mathematics. Factor analysis was not possible for spring 2019 and beyond due to the switch to computer-adaptive testing. Similarly, a linear-on-the-fly (LOFT)/adaptive test design was used for all operational science assessments inspired by the Next Generation Science Standards (NGSS) framework across school years and states, except for Utah in spring 2018 where fixed-form tests were administered. Factor analyses were conducted on Utah's data in 2018 to help provide evidence of the internal structure of WVGSA. Detailed rationales, methods, and results are presented in Section 5.7, Confirmatory Factor Analysis for Spring 2018 Utah Science. Evidence based on test content is a crucial component of validity because construct underrepresentation or irrelevancy can result in unfair advantages or disadvantages to one or more groups of test takers.

Technology-enhanced items should be examined to ensure that no construct-irrelevant variance is introduced. If any aspect of the technology impedes or creates an advantage for a student in their responses to items, this could affect item responses and inferences regarding that student's abilities on the measured construct (refer to Volume 2, Test Development). For ELA and mathematics, the Independent College and Career Readiness (ICCR) item bank uses the technology-enhanced items developed by CAI, and the items are delivered by the same engine used to deliver the SBAC

assessment. Hence, the WVGSA makes use of items that have the same technology-enhanced functionality as those found on other assessments. The same engine is used to deliver the science assessment. Science clusters typically consist of multiple interactions; interactions have the same technology-enhanced functionality as the ELA and mathematics assessments.

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

The third source of validity evidence is based on internal structure: the degree to which the relationships among test items and test components relate to the construct on which the proposed test scores are interpreted. DIF, which determines whether specific items may function differently for subgroups of test takers, is one method for analyzing the internal structure of tests (refer to Volume 1, Annual Technical Report). Other possible analyses to examine internal structure are dimensionality assessment, the goodness-of-fit model to data, and reliability analysis (refer to Section 3, Reliability and Section 5, Evidence on Internal-External Structure for details).

The fourth source of validity evidence is the relationship of test scores to external variables. The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) divides this source of evidence into three parts: (1) convergent and discriminant evidence, (2) test-criterion relationships, and (3) validity generalization. Convergent evidence supports the relationship between the test and other measures intended to assess similar constructs. Conversely, discriminant evidence delineates the test from other measures intended to assess different constructs. A multi-trait multi-method matrix can be used to analyze both convergent and discriminant evidence. Convergent and discriminant validity evidence are discussed in Section 5.4, Convergent and Discriminant Validity.

Additionally, test-criterion relationships indicate how accurately test scores predict criterion performance. The degree of accuracy mainly depends on the test's purpose, such as classification, diagnosis, or selection. Test-criterion evidence is also used to investigate predictions of favoring different groups. Due to construct underrepresentation or construct-irrelevant components, the relation of test scores to a relevant criterion may differ from one group to another. Furthermore, validity generalization is related to whether the evidence is situation-specific or can be generalized across different settings and times. For example, sampling errors or range restrictions may need to be considered to determine whether the conclusions of a test can be assumed for the larger population.

The fifth source of validity evidence should include the intended and unintended consequences of test use in the test-validation process. Determining the validity of the test should depend upon evidence directly related to the test and should not be influenced by external factors. For example, if an employer administers a test to determine hiring rates for different groups of people, an unequal distribution of skills related to the measurement construct does not necessarily imply a lack of validity for the test. However, if the unequal distribution of scores is due to an unintended, confounding aspect of the test, that aspect would interfere with the test's validity. As described in

this volume of the technical report and in Volume 1, Annual Technical Report, test use should align with the test's intended purpose.

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

2. PURPOSE OF THE WEST VIRGINIA GENERAL SUMMATIVE ASSESSMENT

The primary purpose of West Virginia's K–12 assessment system is to yield accurate information on students' achievement of West Virginia's education standards. The WVGSA supports instruction and student learning by measuring growth in student achievement. Assessments can be used as indicators to determine whether students in West Virginia have the knowledge and skills essential for college education and careers.

West Virginia'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 improve their instruction, which in turn has a positive effect on students' learning over time.

The tests are constructed to measure student proficiency on the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014). The test was developed in adherence to the principles of universal design to ensure that all students have access to the test content. Volume 2, Test Development describes the WVGSA standards and test blueprints in more detail. Additional evidence of content validity can also be found in Section 4, Evidence of Content Validity. The WVGSA test scores are useful indicators for understanding individual students' academic achievement of the West Virginia's content standards and evaluating whether students are progressing in their performance over time. Additionally, both individual and aggregated scores can be used to measure test reliability. The reliability of the test scores can be found in Section 3, Reliability.

The WVGSA is a criterion-referenced test designed to measure student performance for English language arts (ELA) and mathematics on the West Virginia College and Career Readiness Standards and for science on the Next Generation Content Standards and Objectives for Science in West Virginia Schools (WV NxGen Science Standards). As a comparison, norm-referenced tests are designed to rank or compare all students with one another. The WVGSA standards and test blueprints are discussed in Volume 2, Test Development.

The scale score and relative strengths and weaknesses at the reporting category (domain) level were provided for each student to indicate student strengths and weaknesses in various content areas of the test relative to other areas and to the district and state. These scores serve as useful feedback that teachers can use to tailor their instruction, provided that they are viewed with the same caution that accompanies using reporting category scores. Thus, to support their practical use across the state, we must examine the reliability coefficients for and the validity of these test scores.

3. RELIABILITY

3.1. RELIABILITY FOR ELA AND MATHEMATICS

The WVGSA for English language arts (ELA) and mathematics are adaptive testing administrations. Because there is no set form in adaptive testing, marginal reliability was computed for the scale scores, considering the varying measurement errors across the ability range. Marginal reliability is a measure of the overall reliability of an assessment based on the average conditional standard error of measurement (CSEM), estimated at different points on the ability scale for all students.

Marginal reliability ($\overline{\rho}$) is defined as

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

where *N* is the number of students; $CSEM_i$ is the CSEM of the theta score for student *i*, and σ^2 is the variance of the theta score. The higher the reliability coefficient, the greater the precision of the test.

Table 2 presents the marginal reliability coefficients for all students. The reliability coefficients for all subjects and grades range from 0.88–0.92. Appendix A: Student Demographics and Reliability Coefficients provides further breakdown, including reliability coefficients for demographic subgroups and reporting categories.

Table 2: Marginal Reliability Coefficients, ELA and Mathematics

Subject	Grade	Reliability	Subject	Grade	Reliability
ELA	3	0.89	Mathematics	3	0.91
	4	0.89		4	0.92
	5	0.90		5	0.90
	6	0.89		6	0.88
	7	0.90		7	0.88
	8	0.91		8	0.89

3.2. RELIABILITY FOR SCIENCE

Classical test theory (CTT)-based reliability indices are not appropriate for science for two reasons. First, in spring 2022, the science test is administered under an adaptive test design. Each student potentially gets a unique set of items, whereas CTT-based reliability indices require the same set of items to 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 in the same way as the marginal reliability defined in Section 3.1, Reliability for ELA

and Mathematics. The marginal reliability in science for the overall sample is reported by grade in Table 3. The overall reliability ranges from 0.87–0.88. The reliability for students who received a complete test (18 items) is about the same as the overall reliability for both grades. Due to the new structure of the science test, Cambium Assessment, Inc. (CAI) has also explored the relationships between reliability and other important factors, such as the effect of nuisance dimension (refer to Volume 1, Annual Technical Report, Section 5.2.1, Model Description). It was found that if the local dependencies among assertions pertaining to the same item are ignored, the marginal reliability increases to approximately 0.90. Ignoring local dependencies can be achieved either by computing the maximum likelihood estimate (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 estimate (MMLE) ability under the one-parameter logistic (1PL) bifactor model (refer to Volume 1, Annual Technical Report, Section 6.2.1, Marginal Likelihood Function).

Table 3: Marginal Reliability Coefficients, Science

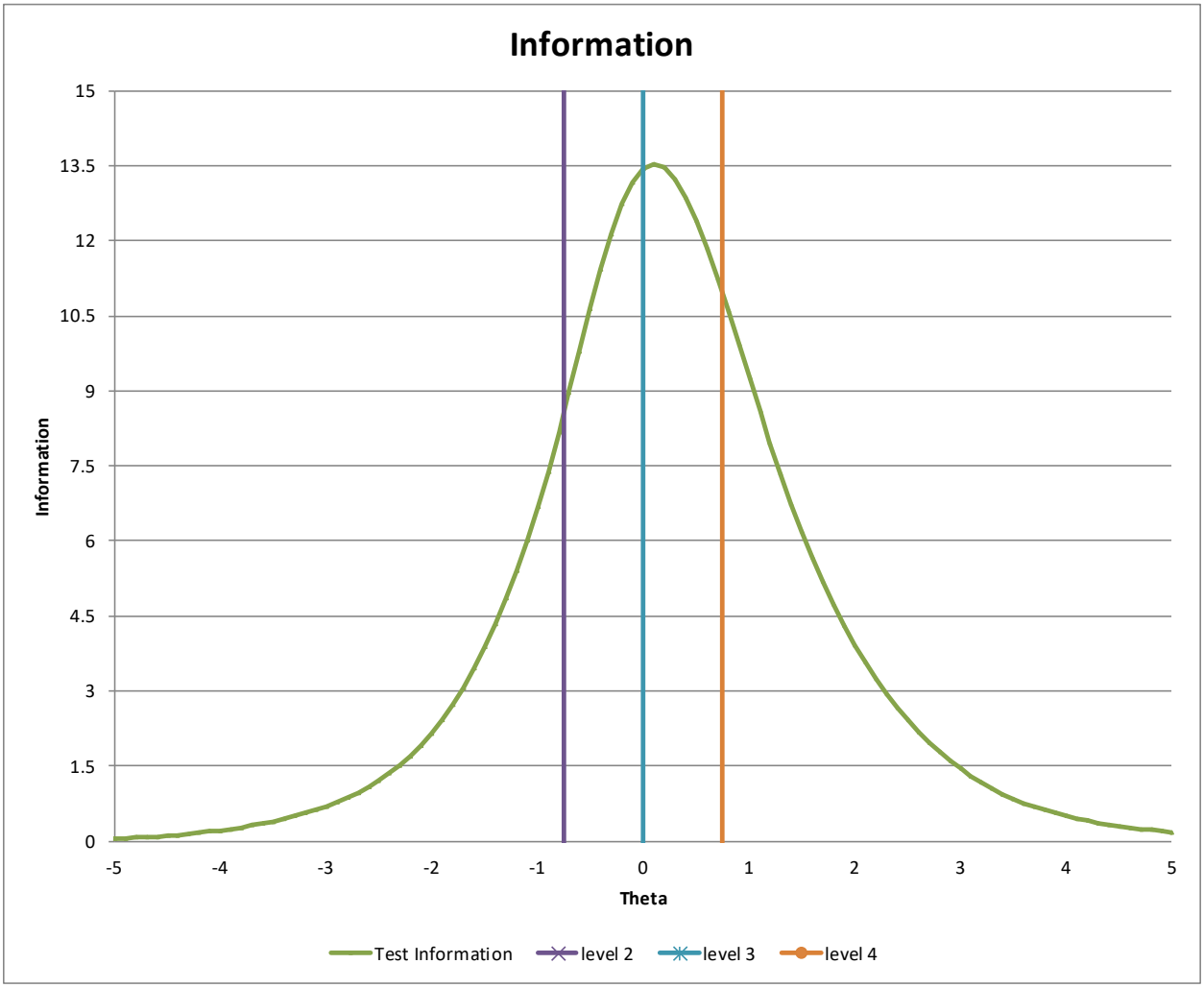
Grade	Sample Size	Reliability
5	17,698	0.88
8	18,694	0.87

3.3. Test Information Curves and Standard Error of Measurement for ELA and Mathematics

Within the IRT framework, measurement error varies across the range of ability as a result of the test, providing varied information across the range of ability as displayed by the test information function (TIF). The TIF describes the amount of information provided by the test at each score point along the ability continuum. The inverse of the TIF is characterized as the conditional measurement error at each score point. For instance, if the measurement error is large, less information is being provided by the assessment at the specific ability level.

Figure 1 displays a sample TIF with three vertical lines indicating the performance cuts. The graphic shows that this test information is maximized in the middle of the score distribution, meaning it provides the most precise scores in this range. The curve is lower at the tails, indicating that the test provides less information about test takers at the tails relative to the center.

Figure 1: Sample Test Information Function



Computing these TIFs is useful to evaluate where the test is maximally informative. In IRT, the TIF is based on the estimates of the item parameters in the test, and the formula used for the WVGSA is calculated as

$$TIF(\theta_s) = \sum_{i=1}^{N_{GPCM}} D^2 a_i^2 \left(\frac{\sum_{j=1}^{m_i} j^2 Exp\left(\sum_{k=1}^{j} Da_i(\theta_s - b_{ik})\right)}{1 + \sum_{j=1}^{m_i} Exp\left(\sum_{k=1}^{j} Da_i(\theta_s - b_{ik})\right)} - \left(\frac{\sum_{j=1}^{m_i} jExp\left(\sum_{k=1}^{j} Da_i(\theta_s - b_{ik})\right)}{1 + \sum_{j=1}^{m_i} Exp\left(\sum_{k=1}^{j} Da_i(\theta_s - b_{ik})\right)} \right)^2 \right) + \sum_{i=1}^{N_{3PL}} D^2 a_i^2 \left(\frac{Q_i}{P_i} \left[\frac{P_i - c_i}{1 - c_i} \right]^2 \right),$$

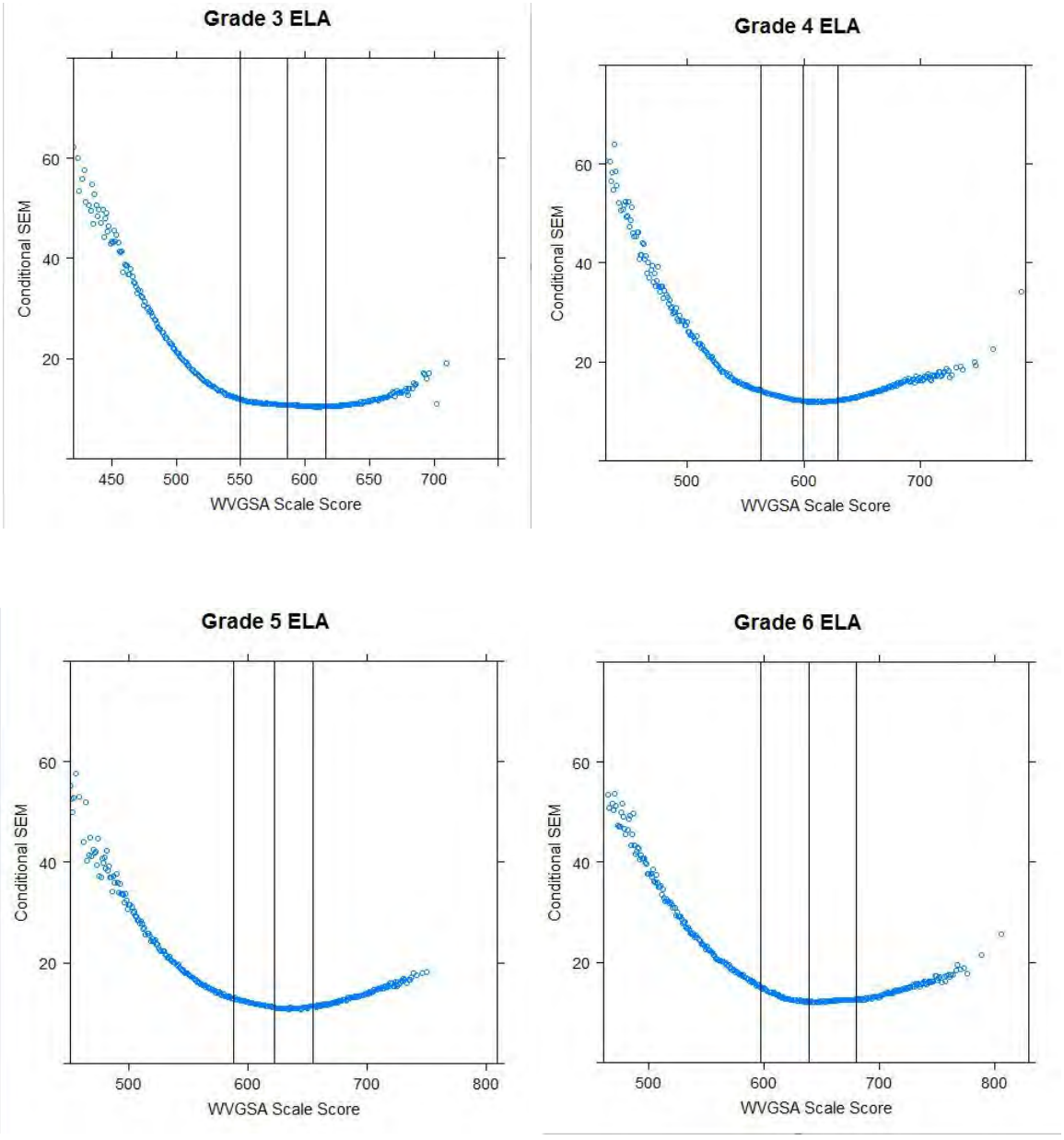
where N_{GPCM} is the number of items scored using the generalized partial credit model (GPCM) items; N_{3PL} is the number of items scored using the three-parameter logistic (3PL) or two-parameter logistic (2PL) model; i indicates item i ($i \in \{1,2,\ldots,N\}$); m_i is the maximum possible score of the item; s indicates student s; and θ_s is the ability of student s.

The standard error of measurement (SEM) for estimated student ability (theta score) is the square root of the reciprocal of the TIF as follows:

$$se(\theta) = \frac{1}{\sqrt{TIF(\theta_i)}}.$$

It is typically more useful to consider the inverse of the TIF rather than the TIF itself, as the SEMs are more useful for score interpretation. For this reason, CSEM curves are presented in Figure 2 and Figure 3 for ELA and mathematics, respectively, instead of the TIFs. The plots presented in this section are based on the scaled scores reported in spring 2022. Vertical lines represent the three achievement level cut scores.

Figure 2: Conditional Standard Error of Measurement, ELA



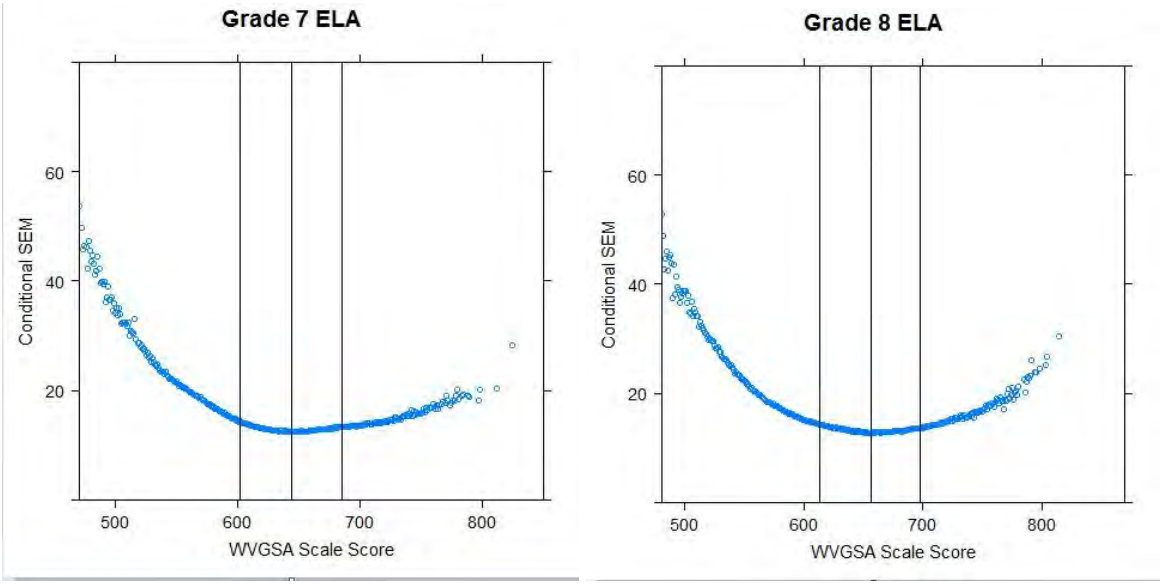
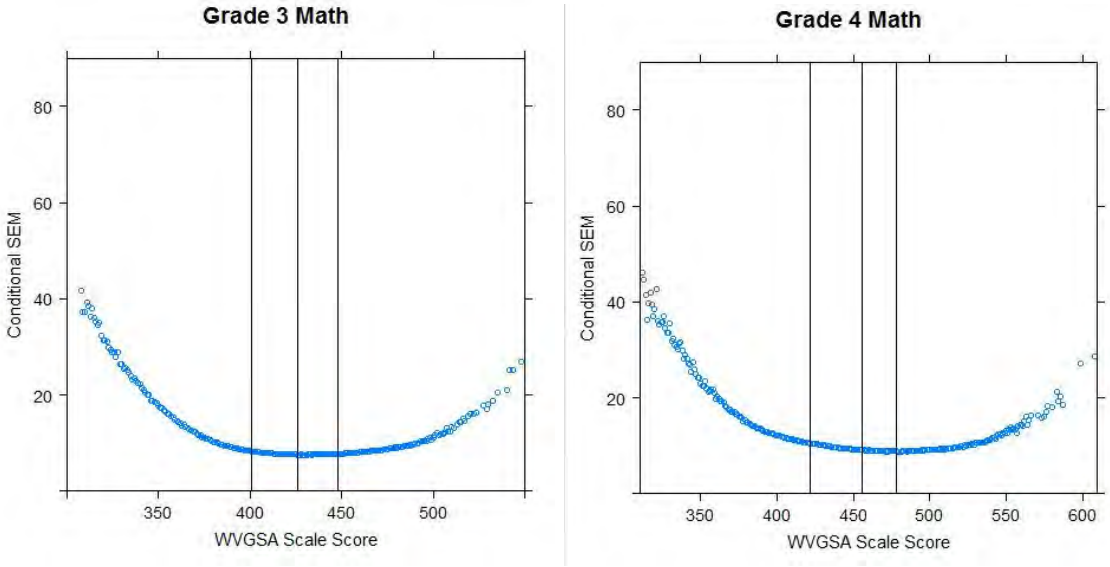
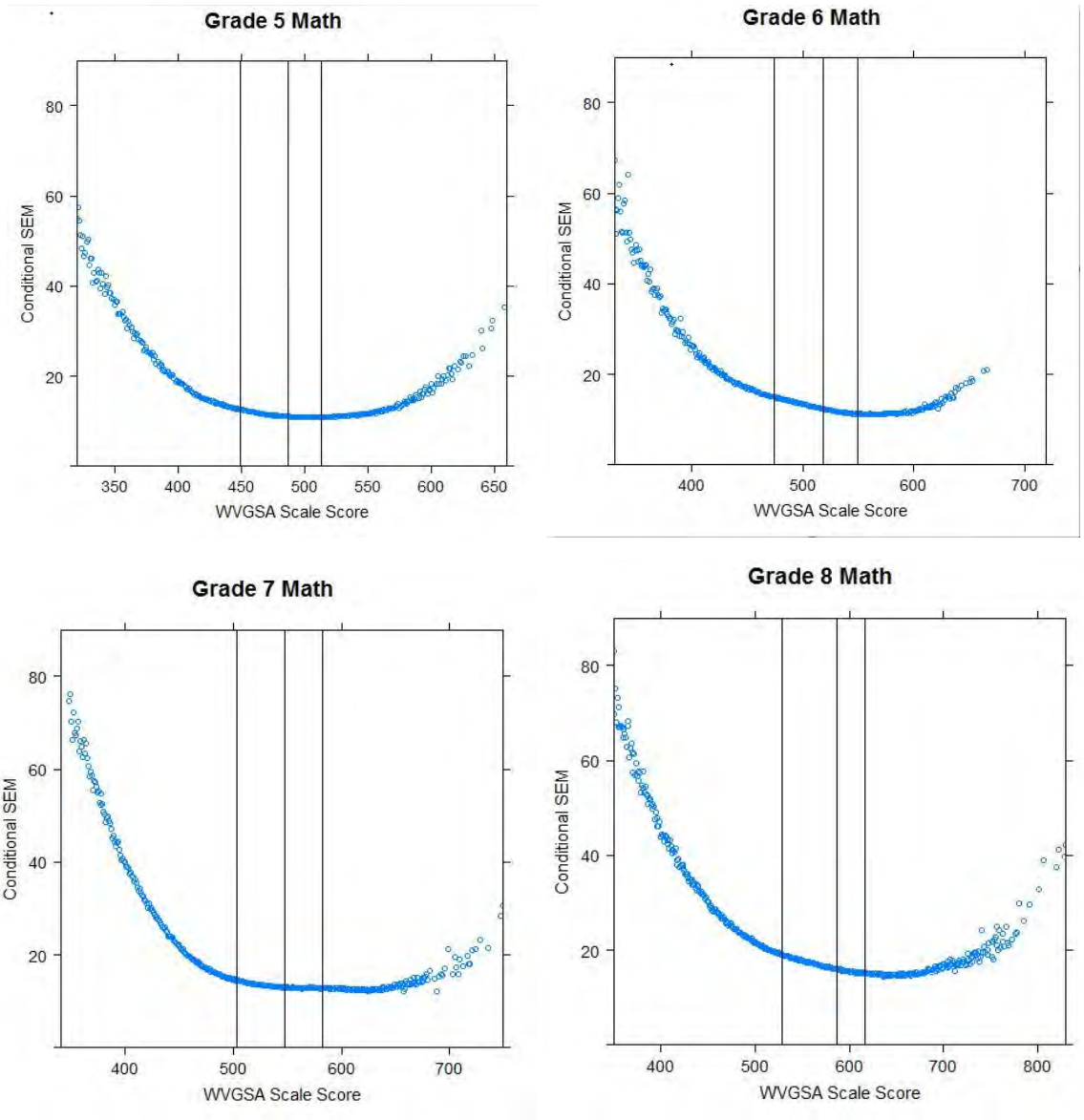


Figure 3: Conditional Standard Error of Measurement, Mathematics



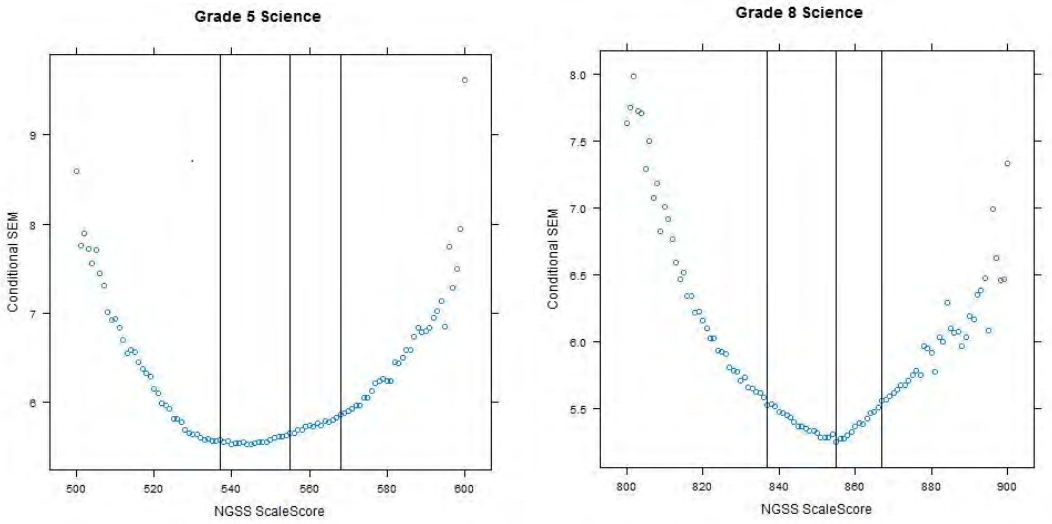


The CSEM curves follow the typical expected trends with the smallest values observed near the middle of the score scale. Desirably, the lowest SEMs are observed at the proficiency cut (the middle vertical line between *Partially Meets Standard* and *Meets Standard* score ranges) for most tests. Appendix B: Conditional Standard Error of Measurement includes the CSEM at each scale score point and corresponding achievement levels.

3.4. STANDARD ERROR OF MEASUREMENT FOR SCIENCE

The computation method of conditional standard error for science has been described in Volume 1, Annual Technical Report, Section 6.2, Marginal Maximum Likelihood Estimation for Science. Figure 4 presents the CSEM curves for science. The lowest standard errors are observed near the proficiency cut for both grades, which is a desirable test property.

Figure 4: Conditional Standard Error of Measurement, Science



3.5. RELIABILITY OF ACHIEVEMENT CLASSIFICATION

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

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

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

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

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

where c_{Ll} and c_{Ul} denote the score corresponding to the lower and upper limits of the performance level l, respectively.

CA and CC for all students and subgroups by achievement level are shown side by side for comparison in Appendix C: Classification Accuracy and Consistency Indices by Subgroups.

3.5.1. Classification Accuracy

Using p_{jl}, the expected number of students at level l, based on students from observed level *k*, can be expressed as

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

where pl_j is the *j*th student's performance level. The values of E_{Akl} are the elements used to populate the matrix $\boldsymbol{E_A}$, an $L \times L$ matrix of conditionally expected numbers of students to score within each performance level based on their true scores. 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 performance level k.

The CA for the pth cut (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:

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

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 4 through Table 6 provide the overall CA and the CA for the individual cuts for ELA, mathematics, and science. The overall CA of the test ranges from 78% to roughly 81% for ELA, from 80%–82% for mathematics, and roughly 76%–78% for science. The individual cut accuracy rates are high across all grades, forms, and subjects, with the minimum value being 90.61% for grade 8 science. These cut accuracy rates denotes that more than 90% of the time we can accurately differentiate students between adjacent achievement levels in the spring 2022 WVGSA.

Table 4: Classification Accuracy Index, ELA

Grade	Overall Accuracy (%)	Cut Accuracy (%)		
		Cut 1	Cut 2	Cut 3
3	80.4	92.38	93.06	94.93
4	78.33	92.37	92.18	93.7
5	79.73	92.59	92.57	94.54
6	79.25	92.58	91.65	95.01
7	80.03	93.02	91.94	95.05
8	80.66	92.98	92.77	94.9

Table 5: Classification Accuracy Index, Mathematics

Grade	Overall Accuracy (%)	Cut Accuracy (%)		
		Cut 1	Cut 2	Cut 3
3	80.67	93.77	92.71	94.17
4	81.06	93.66	92.49	94.87
5	80.4	92.02	93.09	95.24
6	81.94	91.47	93.81	96.62
7	82.33	92.25	93.66	96.38
8	82.48	92.31	93.97	96.06

Table 6: Classification Accuracy Index, Science

Grade	Overall Accuracy (%)	Cut Accuracy (%)		
		Level 2 Cut	Level 3 Cut	Level 4 Cut
5	76.97	91.02	90.64	95.24
8	78.15	90.61	91.56	95.90

3.5.2. Classification Consistency

Similar to CA, assuming the test is administered twice independently to the same group of students, 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 performance level l in test one, and p_{jk} is the probability of the true score at performance level k in test two for the jth student. The classification consistency index for the cuts (CCC) and overall CC were estimated in a way similar to CAC and CA.

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

and

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

Table 7 through Table 9 provide the overall CC and CC for the individual cuts for ELA, mathematics, and science. The overall CC of the test ranged from 70%–73% for ELA, from 73%–76% for mathematics, and roughly 68%–70% for science. The individual cut consistency rates were high across all grades, forms, and subjects, with the minimum value being 86.79% for grade 8 science. In all achievement levels, CA was slightly higher than CC. CC rates can be lower than CA; the consistency is based on two tests with measurement errors, but the accuracy is based on one test with a measurement error and the true score. The accuracy and consistency rates for each achievement level were higher for the levels with smaller SEM.

Table 7: Classification Consistency Index, ELA

Grade	Overall Consistency (%)	Cut Consistency (%)		
		Cut 1	Cut 2	Cut 3
3	72.51	89.09	90.22	92.9
4	69.87	89.13	88.92	91.1
5	71.68	89.5	89.43	92.32
6	70.85	89.45	88.22	92.96
7	71.95	90.07	88.62	93.03
8	72.88	90.05	89.78	92.84

Table 8: Classification Consistency Index, Mathematics

Grade	Overall Consistency (%)	Cut Consistency (%)		
		Cut 1	Cut 2	Cut 3
3	73.08	91.2	89.72	91.84
4	73.68	91.05	89.44	92.72
5	72.83	88.76	90.27	93.26
6	74.59	87.91	91.18	95.18
7	75.21	89	91.03	94.86
8	75.66	89.03	91.46	94.34

Table 9: Classification Consistency Index, Science

Grade	Overall Consistency (%)	Cut Consistency (%)		
		Level 2 Cut	Level 3 Cut	Level 4 Cut
5	68.42	87.46	86.95	93.26
8	69.86	86.79	88.20	94.21

3.6. PRECISION AT CUT SCORES

Table 10 through Table 12 present the mean CSEM at each achievement level by grade and subject. These tables also include achievement-level cut scores and the associated CSEM.

Table 10: Achievement Levels and Associated CSEM, ELA

Grade	Achievement Level	Mean CSEM	Cut Score (Scale Score)	CSEM at Cut Score
3	1	17.63	-	
	2	11.22	550	11.97
	3	10.64	586	10.85
	4	10.97	616	10.5
4	1	19.69	-	
	2	13.05	563	14.42
	3	12.06	599	12.18
	4	13.52	629	12.32
5	1	17.71	-	
	2	12.06	588	12.94
	3	11.12	622	11.25
	4	12.44	655	11.37

Grade	Achievement Level	Mean CSEM	Cut Score (Scale Score)	CSEM at Cut Score
6	1	20.91	-	
	2	13.15	597	15.15
	3	12.36	639	12.2
	4	13.46	680	12.72
7	1	19.69	-	
	2	13.16	602	14.46
	3	12.86	644	12.63
	4	14.04	685	13.52
8	1	18.69	-	
	2	13.46	613	14.42
	3	13.17	656	12.77
	4	15.07	698	13.76

Table 11: Achievement Levels and Associated CSEM, Mathematics

Grade	Achievement Level	Mean CSEM	Cut Score (Scale Score)	CSEM at Cut Score
3	1	13.29	-	
	2	7.87	401	8.3
	3	7.67	426	7.61
	4	8.55	448	7.81
4	1	16.05	-	
	2	9.74	422	10.41
	3	8.96	456	9.1
	4	9.42	478	8.89
5	1	19.71	-	
	2	11.74	449	12.72
	3	10.98	487	11.11
	4	11.71	513	10.94
6	1	22.99	-	
	2	13.72	474	15.03
	3	11.81	518	12.4
	4	11.48	550	11.38

Grade	Achievement Level	Mean CSEM	Cut Score (Scale Score)	CSEM at Cut Score
7	1	25.15	-	
	2	13.64	503	14.59
	3	12.94	548	12.96
	4	12.86	583	12.89
8	1	30.51	-	
	2	17.45	529	18.97
	3	15.55	587	16.07
	4	15.39	617	15.19

Table 12: Achievement Levels and Associated CSEM, Science

Grade	Achievement Level	Mean CSEM	Cut Score (Scale Score)	CSEM at Cut Score
5	1	5.93	-	-
	2	5.56	537	5.59
	3	5.73	555	5.66
	4	6.13	568	5.86
8	1	5.88	-	-
	2	5.39	837	5.53
	3	5.35	855	5.25
	4	5.74	867	5.56

3.7. WRITING PROMPTS INTER-RATER RELIABILITY

The 2021–2022 writing responses were scored using a combination of Cambium Assessment, Inc.'s (CAI) Autoscore engine and human scoring. In this section, we describe the engine, how engine scores are combined with human scores, the performance of the engine on a held-out validation sample, and the performance of the engine during live scoring.

3.7.1. Automated Scoring Engine

CAI's automated scoring engine, Autoscore, uses a statistical process to evaluate writing prompts. Autoscore evaluates student essays against the same rubric used by human raters, and uses a statistical process to analyze each essay and assign a score for each of the three traits. Autoscore's training/calibration process creates prompt-specific scoring models used for scoring responses for each prompt.

As noted above, Autoscore analyzes response characteristics and human-provided scores and predicts what a human rater would do. The response characteristics are collected using features,

which are then used to predict scores. Autoscore uses features associated with writing quality and response meaning. Writing quality features include measures of syntax, grammatical/mechanical correctness, spelling correctness, text complexity, paragraphing quality, and sentence variation and quality. Measures of response meaning include the use of latent semantic analysis (LSA) and deep learning methods which consider not just the pattern of word frequencies in a response, but also order of words in the response. LSA ignores word order but identifies key topics associated with the sets of words in a response. Deep learning methods use word order and sets of localized word patterns that are related to scores humans have assigned. Finally, in Autoscore, two models are built in parallel and the outputs of these models are optimally combined to predict the response score. This approach allows for a more stable score estimate, similar to the use of two or more human raters.

Aside from rubric-based score, Autoscore can generate condition codes—that is, conditions indicating that the response provided by the student is considered invalid and therefore incorrect. The machine-generated condition codes are as follows:

- NO_RESPONSE: No non-blank characters are detected in the response.
- NOT_ENOUGH_DATA: Student response has less than the minimum number of words configured in the rubric (currently set to 11 words).
- PROMPT_COPY_MATCH: Student response is copied from the passage or item prompt (currently flagged when a 70% match is found, but this parameter is configurable).
- DUPLICATE_TEXT: Student response is repeated text copied over and over (currently flagged when a 43% match is found, but this parameter is configurable).
- OUT_OF_VOCAB: Student response is comprised mostly of words that do not overlap with those in the training set vocabulary (currently set to 50%).
- NONSPECIFIC: Essay-scoring engine predicts the assignment of a condition code. Even after training the system, there can be responses that do not fall into any of the pre-set categories. For those responses, the system will generate a condition code of NONSPECIFIC.

Additionally, Autoscore produces a confidence index for a response, indicating how confident the engine is that its score is correct. This index is on a percentile scale and is computed in a two-stage process. In the first stage, for each item, a confidence level is estimated on each trait using the held-out validation sample; this level can be interpreted as the probability that a trait score is accurately produced by the engine and is influenced by whether a response has a borderline score or has unusual characteristics. An overall item confidence level can be interpreted as an average of the confidence levels of each trait. Then, a sample of approximately 5,000 responses gathered from an operational test administration and unseen by the engine is scored by Autoscore, and percentile tables are computed based on the overall confidence level.

3.7.2. Handscoring Data Used to Train the Engine

CAI uses approximately 2,000 responses to train and validate Autoscore performance. These responses are divided into three samples: train, ensemble, and held-out validation. The training sample is used to train competing models and to pick the best performing model. The ensemble sample is used to estimate parameters of a categorical logistic regression (one-vs-rest) using as inputs the probabilities from a model comprised of LSA features and writing features and the logits

from a deep learning model. Once the ensembling model parameters are estimated, the held-out validation data are scored and the performance of the engine is examined on these data. The engine is trained on the best available score (the final, resolved score) coming out of the handscoring process described in the following paragraphs.

The 2,000 responses were selected using stratified random sampling and scored by two human raters. Essay responses to the grades 3–7 writing prompts were sent to Measurement Incorporated (MI) and responses for grade 8 were sent to Data Recognition Corporation (DRC) for human scoring. Using anchor papers selected by content experts and finalized rubrics (Table 13), human raters were trained to score writing responses at a rangefinding meeting. Raters revisited anchor papers and rubrics to refamiliarize themselves with scoring, including a range of sample responses and scores.

Raters were assigned to groups. Training the raters occurred as the leader of each group read student responses out loud to raters; the raters independently referred back to the anchors and rubrics and shared what they thought the score for the particular response should be. If the decision among raters was unanimous, there was a brief discussion and they moved to the next response. If the decision was not unanimous, the raters had a discussion referring to the anchors and rubrics to reach a consensus.

Two trained raters scored each writing item response. Where scores from reader 1 and reader 2 were not in exact agreement, the response was sent for resolution scoring by a team leader or scoring director. The final item score was based on the resolution score, when present, or on the initial read.

Table 13: Writing Rubrics

Trait	Rubric	Score Points
Conventions	The response demonstrates an adequate command of basic conventions. The response may include the following: • Some minor errors in usage but no patterns of errors • Adequate use of punctuation, capitalization, sentence formation, and spelling	0,1,2
Evidence & Elaboration	The response provides thorough and convincing support, citing evidence for the controlling idea or main idea that includes the effective use of sources, facts, and details. The response includes most of the following: • Smoothly integrated, thorough, and relevant evidence, including precise references to sources • Effective use of a variety of elaborative techniques (including but not limited to definitions, quotations, and examples), demonstrating an understanding of the topic and text • Clear and effective expression of ideas, using precise language • Academic and domain-specific vocabulary clearly appropriate for the audience and purpose • Varied sentence structure, demonstrating language facility	1,2,3,4

Trait	Rubric	Score Points
Purpose, Focus, & Organization	The response is fully sustained and consistently focused within the purpose, audience, and task; and it has a clear controlling idea and effective organizational structure creating coherence and completeness. The response includes most of the following: • Strongly maintained controlling idea with little or no loosely related material • Skillful use of a variety of transitional strategies to clarify the relationships between and among ideas • Logical progression of ideas from beginning to end with a satisfying introduction and conclusion • Appropriate style and objective tone established and maintained	1,2,3,4

3.7.3. Engine Evaluation Methods

Statistics used to examine human-human agreement and Autoscore-human agreement were percent exact agreement and quadratic weighted kappa (QWK). Percent exact agreement is the total number of responses in which scores from both raters are equal, divided by the number of responses that were scored twice. In addition to the percentage agreement rates, the QWK values were computed for the training sample and the validation sample for the writing prompts.

Cohen's kappa (Cohen, 1968) is an index of inter-rater agreement that accounts for the agreement that could be expected due to chance. This statistic can be computed as

$$K = \frac{P_o - P_c}{1 - P_c},$$

where P_o is the proportion of observed agreement, and P_c indicates the proportion of agreement by chance. Cohen's kappa treats all disagreement values with equal weights. QWK coefficients (Cohen, 1968), however, allow unequal weights, which can be used as a measure of validity. QWK coefficients were calculated using the formula below:

$$K_w = \frac{P'_o - P'_c}{1 - P'_c},$$

where

$$P'_o = \frac{\sum w_{ij} p_{oij}}{w_{max}},$$

$$P'_c = \frac{\sum w_{ij} p_{cij}}{w_{max}},$$

where p_{oij} is the proportion of the judgments observed in the *ij*th cell, p_{cij} is the proportion in the *ij*th cell expected by chance, and w_{ij} is the disagreement weight. QWK ranges from 0 to 1, where values of 0 indicate no agreement and values of 1 indicate perfect agreement.

The standardized mean difference (SMD) was used to compare the mean scores assigned by Autoscore relative to the final resolved score. The SMD calculated from these values examines the

mean differences in standard deviation units, which are then interpretable across items and traits. For this calculation, the human mean is subtracted from the Autoscore mean and divided by the square root of the average of the two variances. A positive SMD value indicates that Autoscore assigned a higher mean score than the human rater.

3.7.4. Engine Performance on the Held-Out Evaluation Sample

Autoscore-human agreement was on average higher than human-human agreement for exact agreement and QWK on the held-out validation sample (Table 14). The agreement metrics were computed between the two human raters (H1-H2) and between the final, resolved score and Autoscore (HS-AS). Because Autoscore is trained on and evaluated against a more reliable score (the final, resolved score), the agreement between the final, resolved score and Autoscore should be higher than that of two human raters. On average, Autoscore assigned slightly higher scores in Conventions compared to the final, resolved score and slightly lower scores in Elaboration and Organization, as evidenced by the SMD values.

Table 14: Average, Standard Deviation, Minimum, and Maximum Agreements of Autoscore with Human Raters on the Held-Out Validation Sample

Statistic	Trait	Exact Agreement		Quadratic Weighted Kappa		SMD
		H1-H2	HS-AS	H1-H2	HS-AS	HS-AS
Average	Conventions	70%	78%	0.59	0.70	0.08
	Elaboration	70%	77%	0.63	0.67	-0.03
	Organization	67%	74%	0.63	0.68	-0.03
Standard Deviation	Conventions	6%	3%	0.05	0.04	0.05
	Elaboration	6%	3%	0.07	0.07	0.09
	Organization	4%	3%	0.07	0.06	0.07
Minimum	Conventions	60%	73%	0.48	0.60	-0.05
	Elaboration	61%	73%	0.55	0.60	-0.18
	Organization	61%	70%	0.56	0.59	-0.16
Maximum	Conventions	78%	83%	0.66	0.79	0.14
	Elaboration	78%	83%	0.79	0.80	0.16
	Organization	75%	80%	0.78	0.79	0.10

Using Williamson, Xi, & Breyer (2012) recommendations, we expect almost all item traits will be such that the HS-AS QWK is no more than .1 less than the H1-H2 QWK. Although not an industry recommendation, we also expect that the HS-AS exact agreement is no more than 5.25% less than the H1-H2 exact agreement rate. Looking at the individual item and trait performance (Table 15), HS-AS agreements were the same or higher than H1-H2 exact agreements for 35 of the 36 item traits. No items had HS-AS exact agreement more than 5.25% lower than the H1-H2 exact agreement. HS-AS QWK agreements were the same or higher than H1-H2 QWK agreements for 32 of the 36 item traits. No items had HS-AS QWK more than .1 lower than the H1-H2 QWK.

Table 15: Item Trait-Level Exact and QWK Agreement of Autoscore with Human Raters on the Held-Out Validation Sample

Grade	Item ID	Trait	Number of Responses	Exact Agreement			Quadratic Weighted Kappa		
				H1-H2	HS-AS	Diff	H1-H2	HS-AS	Diff
3	31041	Conventions	281	62%	78%	16%	0.48	0.69	0.20
		Elaboration	281	61%	74%	13%	0.65	0.75	0.10
		Organization	281	61%	73%	12%	0.66	0.73	0.06
	31042	Conventions	277	70%	81%	11%	0.63	0.79	0.15
		Elaboration	277	62%	74%	12%	0.64	0.74	0.10
		Organization	277	61%	70%	8%	0.64	0.72	0.07
4	31043	Conventions	288	61%	74%	13%	0.53	0.73	0.21
		Elaboration	288	78%	78%	1%	0.62	0.61	-0.01
		Organization	288	69%	75%	6%	0.56	0.60	0.04
	31044	Conventions	296	68%	74%	6%	0.62	0.68	0.06
		Elaboration	296	71%	77%	6%	0.55	0.60	0.05
		Organization	296	68%	76%	8%	0.56	0.60	0.04
5	31045	Conventions	294	75%	79%	4%	0.64	0.71	0.06
		Elaboration	294	63%	74%	11%	0.55	0.64	0.09
		Organization	294	70%	74%	4%	0.62	0.59	-0.03
	31046	Conventions	293	71%	76%	6%	0.59	0.68	0.09
		Elaboration	293	70%	78%	8%	0.58	0.62	0.04
		Organization	293	71%	71%	0%	0.67	0.66	-0.01
6	31047	Conventions	281	70%	78%	8%	0.59	0.67	0.08
		Elaboration	281	66%	74%	8%	0.61	0.61	0.00
		Organization	281	63%	70%	7%	0.62	0.66	0.04
	31048	Conventions	298	60%	73%	13%	0.55	0.60	0.05
		Elaboration	298	74%	73%	0%	0.57	0.61	0.04
		Organization	298	65%	72%	7%	0.60	0.68	0.09
7	31049	Conventions	304	75%	83%	7%	0.56	0.69	0.13
		Elaboration	304	70%	81%	11%	0.64	0.74	0.10
		Organization	304	64%	78%	14%	0.59	0.70	0.10
	31050	Conventions	283	71%	79%	8%	0.55	0.71	0.16
		Elaboration	283	74%	83%	8%	0.61	0.64	0.03
		Organization	283	68%	80%	12%	0.58	0.63	0.06
8	31051	Conventions	379	77%	78%	1%	0.66	0.69	0.04
		Elaboration	379	77%	79%	3%	0.79	0.80	0.01
		Organization	379	75%	76%	1%	0.78	0.79	0.01
	31052	Conventions	365	78%	82%	4%	0.64	0.71	0.07
		Elaboration	365	69%	75%	5%	0.73	0.72	-0.01
		Organization	365	68%	74%	5%	0.73	0.74	0.01

**Essays that were given a condition code by Autoscore or human readers were excluded.*

Item traits with HS-AS SMD magnitudes lower than .15 indicate similar distributions produced by the final, resolved score and Autoscore (Williamson, Xi, & Breyer, 2012). Thirty-three item traits

met this criterion and three item traits (31045 Elaboration and 31050 Elaboration and Organization) failed this criterion (Table 16). The SMD values above .15 are underlined in the table.

Table 16: Item Trait-Level Autoscore and Human Rater Mean Scores and SMDs on the Held-Out Validation Sample

Grade	Item ID	Trait	Number of Responses	Human		Autoscore		SMD
				Mean	SD	Mean	SD	
3	31041	Conventions	281	1.53	0.67	1.59	0.65	0.09
		Elaboration	281	1.96	0.75	1.96	0.70	0.01
		Organization	281	2.04	0.75	2.05	0.71	0.01
	31042	Conventions	277	1.45	0.71	1.55	0.67	0.14
		Elaboration	277	1.99	0.77	1.99	0.68	0.00
		Organization	277	2.01	0.77	1.94	0.71	-0.09
4	31043	Conventions	288	1.27	0.71	1.23	0.70	-0.05
		Elaboration	288	1.34	0.51	1.35	0.54	0.01
		Organization	288	1.60	0.60	1.54	0.55	-0.10
	31044	Conventions	296	1.26	0.67	1.32	0.64	0.09
		Elaboration	296	1.39	0.54	1.37	0.52	-0.03
		Organization	296	1.56	0.58	1.53	0.55	-0.05
5	31045	Conventions	294	1.52	0.63	1.60	0.60	0.13
		Elaboration	294	1.69	0.64	1.79	0.59	0.16
		Organization	294	1.94	0.61	1.93	0.51	-0.02
	31046	Conventions	293	1.48	0.61	1.53	0.60	0.07
		Elaboration	293	1.63	0.60	1.55	0.50	-0.148
		Organization	293	1.85	0.66	1.91	0.63	0.10
6	31047	Conventions	281	1.57	0.60	1.60	0.58	0.07
		Elaboration	281	1.69	0.63	1.63	0.57	-0.11
		Organization	281	1.90	0.69	1.89	0.66	-0.02
	31048	Conventions	298	1.47	0.69	1.54	0.60	0.12
		Elaboration	298	1.48	0.65	1.44	0.61	-0.06
		Organization	298	1.68	0.70	1.63	0.67	-0.08
7	31049	Conventions	304	1.61	0.55	1.66	0.52	0.09
		Elaboration	304	1.60	0.62	1.56	0.58	-0.07
		Organization	304	1.78	0.65	1.79	0.56	0.02
	31050	Conventions	283	1.49	0.60	1.53	0.59	0.07
		Elaboration	283	1.83	0.51	1.74	0.47	-0.18
		Organization	283	1.91	0.55	1.83	0.49	-0.16
8	31051	Conventions	379	1.55	0.60	1.60	0.62	0.07
		Elaboration	379	2.02	0.76	2.01	0.73	-0.02
		Organization	379	2.07	0.80	2.09	0.74	0.02
	31052	Conventions	365	1.61	0.58	1.64	0.58	0.04
		Elaboration	365	2.08	0.75	2.12	0.68	0.05

Grade	Item ID	Trait	Number of Responses	Human		Autoscore		SMD
				Mean	SD	Mean	SD	
		Organization	365	2.20	0.76	2.24	0.72	0.06

**Essays that were given a condition code by Autoscore or human readers were excluded.*

3.7.5. Engine Condition Codes, Confidence, and Routing to Handscoring

During live testing, CAI uses a hybrid, human/machine-scoring approach whereby low confidence responses or other unusual responses are flagged for human scoring. Responses that received a confidence percentile value lower than 15 and any responses that receive a condition code of NONSPECIFIC, OUT OF VOCAB, or DUPLICATE TEXT were routed for human verification. Because the confidence percentile is based on a sample, there will be variation across items in the actual percentage of responses receiving a "low confidence" score.

Human verification was conducted by the following process:

- If the first verification reader assigned scores in a trait that were the same as the machine-assigned scores, the machine-assigned scores were accepted as the final trait scores.
- If the first verification reader did not assign the same trait score as the machine-assigned score, the essay was sent to the second verification reader, who then assigned a score in the trait of disagreement. If the second reader's trait score matched with either machine or the first reader's score, the matching score was accepted as the final score.
- If the second verification reader's trait score did not match the machine or first reader's scores, the essay was sent to the scoring supervisor to assign the final score in that trait.
- If a backreader's score was available, their score was accepted to be the final score regardless of all others' scores assigned.

In addition to the essays sent for human verification due to the low confidence flag or condition codes, the first 500 essays that did not receive a NO RESPONSE, NOT ENOUGH DATA, or PROMPT COPY MATCH code were routed for human scoring. The purpose of handscoring the first 500 essays is to ensure that the human scoring and the engine scoring are performing as expected, recognizing the inherent complexities in the dynamics of live scoring. While the first 500 essays cannot be thought to be representative of the tested population, they should be reasonably indicative of the performance of the essay-scoring system for responses encountered after the first 500.

Finally, a random sample of 5% of responses not in the first 500, routed condition codes, or low confidence scores was drawn and sent for verification. The purpose of this sample was to provide agreement data across the test administration.

Table 17 presents the number and percentage of responses routed for human verification, overall and by the condition for routing. As expected, 500 responses were routed as part of the First 500 routing condition, and these percentages were 5%–6% of the tested population for each item. The percentage of responses routed for condition codes ranged between 0.1% and 1.2%. The percentage of responses routed due to low confidence ranged from 11% to 17%, with most (N=9)

items routed at rates between 13% and 17%. The percentage routed due to random 5% was 5% for almost all items, with the grade 3 item 31042 being the exception. The total percentage of responses routed ranged from 21% to 27%, with most items (N=9) having 23% to 27% routed for handscoring.

Table 17: Number and Percentage of Responses Routed for Human Verification, by Routing Condition

Grade	Item ID	Total Tested	First 500		Condition Code		Low Confidence		Random 5%		Total Routed	
			%	N	%	N	%	N	%	N	%	N
3	31041	8862	6%	500	1.1%	92	15%	1214	5%	404	25%	2210
	31042	8712	6%	500	1.2%	99	11%	936	4%	369	22%	1904
4	31043	8734	6%	500	0.8%	62	15%	1210	5%	409	25%	2181
	31044	8622	6%	500	0.5%	42	17%	1384	5%	423	27%	2349
5	31045	8783	6%	500	0.3%	28	14%	1131	5%	422	24%	2081
	31046	8944	6%	500	0.4%	32	15%	1276	5%	412	25%	2220
6	31047	8868	6%	500	0.1%	12	15%	1270	5%	431	25%	2213
	31048	8861	6%	500	0.2%	18	11%	917	5%	459	21%	1894
7	31049	9214	5%	500	0.3%	22	12%	1044	5%	472	22%	2038
	31050	9109	5%	500	0.2%	13	16%	1378	5%	401	25%	2292
8	31051	9361	5%	500	0.2%	17	13%	1142	5%	482	23%	2141
	31052	9403	5%	500	0.2%	16	14%	1219	5%	485	24%	2220

**Data do not include responses receiving the NO RESPONSE, NOT ENOUGH DATA, or PROMPT COPY MATCH condition codes. Percentages of condition code, low confidence, and random 5% are computed on the total tested minus the first 500 count to examine that routing worked as intended. The total routed percentages are computed using the total tested.*

3.7.6. Engine Performance on the First 500 Sample

The performance of the scoring on the First 500 sample can be examined using the handscoring agreements of the held-out validation sample as a benchmark. Currently, there are no standards in the industry for examining live scoring, in part because handscoring is a dynamic and complex process and because the processes used during handscoring when obtaining the training and validation data may not match those used during live scoring. We should expect, however, that Autoscore agreement with the human rater on the First 500 sample approximates that agreement observed in the held-out validation sample.

Table 18 presents the average results for exact agreement and QWK for the benchmarks across items, for the human rater relative to Autoscore (H1-AS), and for the human rater and Autoscore relative to the final, resolved score (H1-Final and AS-Final, respectively). Recall that the final, resolved score is produced from a process involving Autoscore and human scoring. The aggregated results show that Autoscore agreements with the human score are similar to the benchmark for each trait, and that the agreements of the human rater and of Autoscore and the final resolved score are similar as well. Because the final, resolved score is adjudicated using Autoscore and the human rater, we expect these agreements to be substantially higher than that of the benchmark.

Table 18: Average, Standard Deviation, Minimum, and Maximum Agreements of Autoscore with Human Raters on the First 500 Sample

Statistic	Trait	Exact Agreement				Quadratic Weighted Kappa			
		Bench-mark	H1-AS	H1-Final	AS-Final	Bench-mark	H1-AS	H1-Final	AS-Final
Average	Conventions	70%	70%	86%	84%	0.59	0.64	0.83	0.81
	Elaboration	70%	73%	86%	86%	0.63	0.60	0.80	0.78
	Organization	67%	70%	84%	86%	0.63	0.62	0.80	0.79
Standard Deviation	Conventions	6%	7%	4%	5%	0.05	0.05	0.05	0.03
	Elaboration	6%	5%	4%	3%	0.07	0.09	0.05	0.06
	Organization	4%	5%	3%	5%	0.07	0.10	0.05	0.08
Minimum	Conventions	60%	59%	80%	75%	0.48	0.59	0.73	0.74
	Elaboration	61%	63%	78%	82%	0.55	0.49	0.70	0.70
	Organization	61%	59%	80%	76%	0.56	0.48	0.74	0.64
Maximum	Conventions	78%	80%	91%	92%	0.66	0.73	0.91	0.86
	Elaboration	78%	83%	92%	91%	0.79	0.79	0.89	0.89
	Organization	75%	76%	89%	91%	0.78	0.79	0.87	0.90

CAI used thresholds of 7.5% and .15 for exact agreement and QWK, respectively, to identify item and trait agreements on the First 500 sample that lie below these thresholds as a way to monitor the scoring process. We use larger thresholds for monitoring the operational data (compared to the thresholds for monitoring the held-out validation data) because of the complexities surrounding live scoring situations, because the engine scores were compared to a less reliable score (i.e., a non-expert score), and because the scoring occurs early in the window when raters are still cementing their understanding and application of the rubric. However, we do note that the very early raters in the First 500 sample tend to be the expert raters with typical raters being used later in the sample. Thus, the First 500 sample is a blend of expert raters and typical raters.

Table 19 presents the exact agreement and QWK of Autoscore with the human rater (H1-AS), the human score and the final resolved score (H1-Final), and Autoscore and the final resolved score (AS-Final) at the item and trait level. One item trait had an H1-AS exact agreement rate more than 7.5% below the benchmark (underlined in the table), and no item traits had H1-AS QWK agreement rates more than .15 below the benchmark.

Table 19: Item Trait-Level Agreement of Autoscore with Human Raters on the First 500 Sample

Grade	Item ID	Trait	Number of Responses	Exact Agreement				Quadratic Weighted Kappa			
				Bench-mark	H1-AS	H1-Final	AS-Final	Bench-mark	H1-AS	H1-Final	AS-Final
3	31041	Conventions	493	62%	59%	83%	75%	0.48	0.63	0.84	0.78
		Elaboration	486	61%	63%	78%	84%	0.65	0.54	0.70	0.79
		Organization	486	61%	76%	84%	91%	0.66	0.69	0.81	0.87
	31042	Conventions	490	70%	<u>62%</u>	80%	82%	0.63	0.59	0.79	0.79
		Elaboration	484	62%	74%	86%	89%	0.64	0.67	0.82	0.85
		Organization	484	61%	72%	84%	88%	0.64	0.69	0.83	0.85

Grade	Item ID	Trait	Number of Responses	Exact Agreement				Quadratic Weighted Kappa			
				Bench-mark	H1-AS	H1-Final	AS-Final	Bench-mark	H1-AS	H1-Final	AS-Final
4	31043	Conventions	500	61%	66%	83%	82%	0.53	0.61	0.82	0.81
		Elaboration	496	78%	77%	88%	90%	0.62	0.49	0.73	0.76
		Organization	496	69%	72%	83%	90%	0.56	0.58	0.74	0.81
	31044	Conventions	499	68%	64%	82%	82%	0.62	0.59	0.80	0.80
		Elaboration	499	71%	83%	92%	91%	0.55	0.62	0.82	0.79
		Organization	499	68%	74%	83%	90%	0.56	0.57	0.74	0.81
5	31045	Conventions	500	75%	73%	90%	82%	0.64	0.70	0.91	0.80
		Elaboration	500	63%	69%	85%	83%	0.55	0.56	0.78	0.74
		Organization	500	70%	67%	81%	86%	0.62	0.56	0.75	0.78
	31046	Conventions	498	71%	75%	89%	87%	0.59	0.73	0.88	0.85
		Elaboration	497	70%	76%	89%	88%	0.58	0.57	0.80	0.76
		Organization	497	71%	73%	88%	85%	0.67	0.63	0.82	0.79
6	31047	Conventions	499	70%	68%	89%	79%	0.59	0.62	0.88	0.74
		Elaboration	498	66%	72%	85%	88%	0.61	0.58	0.77	0.77
		Organization	498	63%	67%	83%	84%	0.62	0.59	0.80	0.78
	31048	Conventions	498	60%	68%	83%	84%	0.55	0.61	0.82	0.78
		Elaboration	498	74%	70%	87%	82%	0.57	0.61	0.84	0.76
		Organization	498	65%	59%	82%	76%	0.60	0.57	0.83	0.73
7	31049	Conventions	500	75%	75%	88%	87%	0.56	0.67	0.86	0.79
		Elaboration	499	70%	70%	88%	82%	0.64	0.55	0.82	0.70
		Organization	499	64%	67%	89%	78%	0.59	0.54	0.83	0.64
	31050	Conventions	500	71%	77%	91%	86%	0.55	0.69	0.87	0.82
		Elaboration	500	74%	68%	81%	85%	0.61	0.52	0.78	0.71
		Organization	500	68%	64%	80%	82%	0.58	0.48	0.76	0.67
8	31051	Conventions	498	77%	80%	89%	91%	0.66	0.64	0.80	0.85
		Elaboration	496	77%	75%	88%	87%	0.79	0.79	0.89	0.89
		Organization	496	75%	74%	86%	88%	0.78	0.78	0.86	0.90
	31052	Conventions	500	78%	78%	86%	92%	0.64	0.59	0.73	0.86
		Elaboration	500	69%	74%	86%	86%	0.73	0.72	0.84	0.84
		Organization	500	68%	76%	87%	89%	0.73	0.79	0.87	0.90

**Essays that were given a condition code by Autoscore or human readers were excluded.*

The human score, Autoscore, and final, resolved score means and standard deviations can also be compared. Again, there are no industry standards around how best to monitor the distributional characteristics of these sources. Table 20 presents the average SMDs across items and traits for H1-AS and for Final-H1 and Final-AS. A positive value for the H1-AS SMD indicates that Autoscore produced a higher mean score than H1. A positive value for the H1-Final SMD indicates that the final, resolved mean score was higher than the H1 mean score. A positive value for the AS-Final SMD indicates that the final, resolved mean score was higher than the Autoscore mean score.

On average, Autoscore assigned higher scores in Conventions relative to the human read and lower scores in Elaboration and Organization (Table 20). This trend continues when Autoscore is

compared to the final resolved score, although the magnitude of the differences is slightly lower. The H1 showed almost identical mean scores in Conventions with the final, resolved score and slightly higher means in Elaboration and Organization. Note that there are large standard deviations in the H1-AS SMD values (.23 - .25), and large minimum and maximum values as well. Thus, Autoscore and H1 standardized mean scores differed somewhat substantially for some items and traits.

Table 20: Average, Standard Deviation, Minimum, and Maximum SMD of Autoscore with Human Raters on the First 500 Sample

Statistic	Trait	SMD		
		H1-AS	H1-Final	AS-Final
Average	Conventions	0.12	0.01	-0.11
	Elaboration	-0.12	-0.05	0.08
	Organization	-0.16	-0.06	0.11
Standard Deviation	Conventions	0.23	0.12	0.13
	Elaboration	0.25	0.13	0.13
	Organization	0.23	0.11	0.14
Minimum	Conventions	-0.27	-0.24	-0.29
	Elaboration	-0.39	-0.19	-0.16
	Organization	-0.46	-0.22	-0.10
Maximum	Conventions	0.40	0.16	0.06
	Elaboration	0.41	0.25	0.30
	Organization	0.12	0.14	0.34

Table 21 presents the H1, AS, and Final resolved score means and standard deviations for each item and trait, as well as the SMD values. The H1-AS SMD values that exceed .225 in magnitude are underlined in the table. Again, we use a larger threshold for the operational data than for the held-out validation in consideration of the complexities inherent to live scoring. Using this threshold, 17 item traits were identified and the SMD values are underlined in the table. The SMD values of the H1-Final and AS-Final show less variation.

Table 21: Item Trait-Level Autoscore and Human Rater Means and Standard Deviations on the First 500 Sample

Grade	Item ID	Trait	H1		AS		Final		SMD		
			Mean	SD	Mean	SD	Mean	SD	H1-AS	H1-Final	AS-Final
3	31041	Conventions	1.14	0.70	1.43	0.79	1.22	0.75	0.40	0.12	-0.27
		Elaboration	1.44	0.62	1.70	0.65	1.60	0.63	0.41	0.25	-0.16
		Organization	1.70	0.65	1.77	0.64	1.79	0.65	0.11	0.14	0.03
	31042	Conventions	1.26	0.66	1.41	0.75	1.36	0.70	0.22	0.14	-0.08
		Elaboration	1.61	0.63	1.72	0.62	1.68	0.63	0.18	0.12	-0.06
		Organization	1.74	0.71	1.68	0.63	1.73	0.67	-0.10	-0.01	0.09
4	31043	Conventions	1.24	0.69	1.18	0.67	1.19	0.69	-0.10	-0.08	0.02
		Elaboration	1.31	0.49	1.22	0.48	1.28	0.49	-0.18	-0.06	0.12
		Organization	1.59	0.62	1.38	0.50	1.47	0.55	-0.38	-0.22	0.16

Grade	Item ID	Trait	H1		AS		Final		SMD		
			Mean	SD	Mean	SD	Mean	SD	H1-AS	H1-Final	AS-Final
5	31044	Conventions	1.16	0.69	1.26	0.66	1.16	0.68	0.15	-0.01	-0.16
		Elaboration	1.36	0.49	1.24	0.44	1.30	0.47	-0.26	-0.13	0.13
		Organization	1.52	0.58	1.32	0.49	1.40	0.53	-0.37	-0.22	0.15
	31045	Conventions	1.35	0.73	1.56	0.64	1.39	0.72	0.31	0.05	-0.25
		Elaboration	1.61	0.62	1.67	0.61	1.63	0.57	0.10	0.04	-0.06
		Organization	1.78	0.68	1.86	0.54	1.80	0.57	0.12	0.03	-0.10
	31046	Conventions	1.37	0.72	1.38	0.65	1.33	0.69	0.01	-0.05	-0.07
		Elaboration	1.56	0.54	1.41	0.49	1.48	0.52	-0.28	-0.15	0.13
		Organization	1.74	0.61	1.72	0.63	1.73	0.59	-0.02	-0.02	0.01
6	31047	Conventions	1.39	0.72	1.64	0.61	1.45	0.69	0.37	0.08	-0.29
		Elaboration	1.69	0.62	1.57	0.52	1.63	0.54	-0.22	-0.11	0.11
		Organization	1.89	0.67	1.78	0.58	1.83	0.62	-0.17	-0.08	0.09
	31048	Conventions	1.33	0.74	1.57	0.60	1.45	0.66	0.36	0.16	-0.20
		Elaboration	1.60	0.69	1.40	0.59	1.52	0.64	-0.31	-0.12	0.20
		Organization	1.87	0.74	1.61	0.66	1.79	0.71	-0.36	-0.10	0.26
7	31049	Conventions	1.50	0.65	1.66	0.58	1.54	0.63	0.25	0.06	-0.19
		Elaboration	1.67	0.65	1.43	0.52	1.60	0.57	-0.39	-0.11	0.30
		Organization	1.90	0.63	1.64	0.54	1.83	0.56	-0.44	-0.12	0.34
	31050	Conventions	1.59	0.58	1.53	0.64	1.56	0.60	-0.11	-0.05	0.06
		Elaboration	1.91	0.73	1.67	0.47	1.78	0.60	-0.39	-0.19	0.21
		Organization	2.01	0.76	1.72	0.46	1.89	0.61	-0.46	-0.17	0.32
8	31051	Conventions	1.76	0.52	1.68	0.57	1.69	0.55	-0.14	-0.12	0.02
		Elaboration	2.08	0.79	1.96	0.75	2.03	0.74	-0.15	-0.07	0.09
		Organization	2.07	0.75	2.10	0.82	2.09	0.76	0.03	0.02	-0.01
	31052	Conventions	1.81	0.47	1.66	0.58	1.68	0.54	-0.27	-0.24	0.04
		Elaboration	1.99	0.72	2.00	0.70	1.97	0.65	0.02	-0.02	-0.04
		Organization	2.02	0.75	2.08	0.76	2.05	0.72	0.07	0.04	-0.04

**Essays that were given a condition code by Autoscore or human readers were excluded.*

3.7.7. Engine Performance on the Random 5% Sample

As with the First 500 sample, the performance of the scoring on the Random 5% can be examined using the handscoring agreements of the held-out validation sample as a benchmark. Recall that the 5% sample is taken throughout the test administration window and from responses not routed due to low confidence or condition code (and occurring after the First 500 responses). Note that this sample is not necessarily representative of the full sample, as the low confidence responses are not included. Finally, note that this sample is scored primarily by typical raters, as opposed to a mix of expert and typical raters used in the First 500 sample.

Table 22 presents the average results for exact agreement and QWK for the benchmarks, for H1-AS, H1-Final, and AS-Final. The aggregated results show that Autoscore agreements with the H1

rater are similar to the benchmark for each trait, and that the H1-Final and AS-Final agreements are similar as well. The Elaboration trait has slightly lower AS-H1 agreements relative to the benchmark, and the AS-Final has slightly lower agreements relative to the H1-Final.

Table 22: Average, Standard Deviation, Minimum, and Maximum Agreements of Autoscore with Human Raters on the Random 5% Sample

Statistic	Trait	Exact Agreement				Quadratic Weighted Kappa			
		Bench-mark	H1-AS	H1-Final	AS-Final	Bench-mark	H1-AS	H1-Final	AS-Final
Average	Conventions	70%	70%	86%	84%	0.59	0.65	0.83	0.81
	Elaboration	70%	67%	84%	82%	0.63	0.56	0.79	0.74
	Organization	67%	67%	83%	83%	0.63	0.61	0.80	0.78
Standard Deviation	Conventions	6%	8%	5%	5%	0.05	0.05	0.05	0.04
	Elaboration	6%	6%	5%	5%	0.07	0.06	0.04	0.08
	Organization	4%	8%	5%	7%	0.07	0.07	0.05	0.10
Minimum	Conventions	60%	54%	76%	72%	0.48	0.57	0.73	0.74
	Elaboration	61%	59%	74%	75%	0.55	0.48	0.75	0.62
	Organization	61%	55%	73%	71%	0.56	0.47	0.73	0.60
Maximum	Conventions	78%	82%	90%	92%	0.66	0.70	0.91	0.88
	Elaboration	78%	75%	90%	89%	0.79	0.65	0.87	0.86
	Organization	75%	77%	91%	90%	0.78	0.68	0.90	0.91

Table 23 presents the exact agreement and QWK at the item and trait level. There were 10 item traits with H1-AS exact agreement rates more than 7.5% below the benchmark and no item traits with H1-AS QWK agreement rates more than .15 below the benchmark.

Table 23: Item Trait-Level Agreement of Autoscore with Human Raters on the Random 5% Sample

Grade	Item ID	Trait	Number of Responses		Exact Agreement				Quadratic Weighted Kappa			
			H1	H2	Bench-mark	H1-AS	H1-Final	AS-Final	Bench-mark	H1-AS	H1-Final	AS-Final
3	31041	Conventions	403	185	62%	54%	80%	72%	0.48	0.57	0.82	0.75
		Elaboration	396	118	61%	70%	84%	84%	0.65	0.59	0.76	0.78
		Organization	396	97	61%	75%	86%	88%	0.66	0.66	0.78	0.85
	31042	Conventions	368	154	70%	58%	76%	81%	0.63	0.60	0.77	0.82
		Elaboration	360	113	62%	69%	83%	86%	0.64	0.63	0.78	0.81
		Organization	360	119	61%	67%	80%	86%	0.64	0.62	0.78	0.85
4	31043	Conventions	384	140	61%	64%	81%	82%	0.53	0.66	0.82	0.84
		Elaboration	380	107	78%	72%	88%	84%	0.62	0.52	0.81	0.71
		Organization	380	89	69%	77%	89%	87%	0.56	0.66	0.85	0.81
	31044	Conventions	423	125	68%	70%	85%	85%	0.62	0.68	0.83	0.84
		Elaboration	421	110	71%	74%	86%	88%	0.55	0.54	0.76	0.76
		Organization	421	111	68%	74%	84%	90%	0.56	0.59	0.77	0.81
5	31045	Conventions	409	108	75%	74%	90%	82%	0.64	0.70	0.91	0.80

Grade	Item ID	Trait	Number of Responses		Exact Agreement				Quadratic Weighted Kappa			
			H1	H2	Bench-mark	H1-AS	H1-Final	AS-Final	Bench-mark	H1-AS	H1-Final	AS-Final
6	31046	Elaboration	407	154	63%	62%	86%	75%	0.55	0.48	0.79	0.66
		Organization	407	112	70%	72%	87%	85%	0.62	0.60	0.82	0.75
		Conventions	412	124	71%	70%	89%	80%	0.59	0.70	0.91	0.79
		Elaboration	410	101	70%	75%	86%	89%	0.58	0.55	0.75	0.78
		Organization	410	102	71%	75%	87%	88%	0.67	0.68	0.83	0.84
	31047	Conventions	431	111	70%	74%	86%	87%	0.59	0.68	0.85	0.84
		Elaboration	430	112	66%	74%	89%	84%	0.61	0.65	0.87	0.77
		Organization	430	140	63%	67%	85%	83%	0.62	0.67	0.86	0.80
	31048	Conventions	459	151	60%	67%	86%	80%	0.55	0.63	0.87	0.74
		Elaboration	458	157	74%	66%	90%	75%	0.57	0.51	0.85	0.62
		Organization	458	170	65%	63%	91%	72%	0.60	0.58	0.90	0.66
7	31049	Conventions	472	93	75%	80%	89%	91%	0.56	0.67	0.85	0.82
		Elaboration	471	192	70%	59%	82%	77%	0.64	0.52	0.78	0.67
		Organization	471	214	64%	55%	81%	71%	0.59	0.48	0.79	0.62
	31050	Conventions	401	100	71%	75%	89%	86%	0.55	0.67	0.83	0.81
		Elaboration	400	154	74%	62%	83%	77%	0.61	0.50	0.81	0.63
		Organization	400	171	68%	57%	81%	75%	0.58	0.47	0.79	0.60
8	31051	Conventions	477	111	77%	77%	90%	86%	0.66	0.57	0.78	0.78
		Elaboration	476	185	77%	61%	78%	82%	0.79	0.64	0.77	0.82
		Organization	476	183	75%	61%	77%	84%	0.78	0.64	0.75	0.85
	31052	Conventions	483	86	78%	82%	90%	92%	0.64	0.64	0.73	0.88
		Elaboration	482	189	69%	61%	74%	87%	0.73	0.63	0.75	0.86
		Organization	482	182	68%	62%	73%	89%	0.73	0.64	0.73	0.91

**Essays that were given a condition code by Autoscore or human readers were excluded.*

On average, Autoscore assigned higher scores in Conventions relative to the human read and lower scores in Elaboration and Organization (Table 24). This trend continues when Autoscore is compared to the final resolved score although the magnitude of the differences is slightly lower. The human reader showed almost identical mean scores in Conventions with the final, resolved score and slightly higher means in Elaboration and Organization. Note that there are large standard deviations in the H1-SMD values (.28–.32), and large minimum and maximum values as well.

Table 24: Average, Standard Deviation, Minimum, and Maximum SMD of Autoscore with Human Raters on the Random 5% Sample

Statistic	Trait	SMD		
		H1-AS	H1-Final	AS-Final
Average	Conventions	0.14	0.04	-0.10
	Elaboration	-0.20	-0.10	0.11
	Organization	-0.22	-0.09	0.14

Statistic	Trait	SMD		
		H1-AS	H1-Final	AS-Final
Standard Deviation	Conventions	0.28	0.11	0.18
	Elaboration	0.32	0.13	0.22
	Organization	0.30	0.14	0.20
Minimum	Conventions	-0.33	-0.16	-0.33
	Elaboration	-0.63	-0.31	-0.27
	Organization	-0.70	-0.30	-0.17
Maximum	Conventions	0.51	0.17	0.22
	Elaboration	0.34	0.14	0.41
	Organization	0.25	0.12	0.49

Table 25 presents the score means and standard deviations at the item and trait level. The H1-AS SMD values that exceed .225 in magnitude are underlined in the table. Using this threshold, 27 item traits were identified and are underlined in the table.

Table 25: Item Trait-Level Autoscore and Human Rater Means and Standard Deviations on the Random 5% Sample

Grade	Item ID	Trait	H1		AS		Final		SMD		
			Mean	SD	Mean	SD	Mean	SD	H1-AS	H1-Final	AS-Final
3	31041	Conventions	1.14	0.75	1.53	0.78	1.27	0.78	0.51	0.17	-0.33
		Elaboration	1.57	0.60	1.78	0.61	1.66	0.60	0.34	0.14	-0.20
		Organization	1.72	0.60	1.83	0.63	1.79	0.60	0.17	0.12	-0.06
	31042	Conventions	1.19	0.73	1.43	0.79	1.30	0.77	0.31	0.14	-0.16
		Elaboration	1.71	0.68	1.79	0.66	1.74	0.62	0.11	0.03	-0.08
		Organization	1.79	0.70	1.76	0.66	1.79	0.66	-0.05	0.00	0.05
4	31043	Conventions	1.24	0.74	1.28	0.76	1.24	0.73	0.06	0.01	-0.06
		Elaboration	1.47	0.58	1.33	0.54	1.43	0.56	-0.24	-0.06	0.18
		Organization	1.60	0.60	1.50	0.58	1.58	0.59	-0.18	-0.04	0.14
	31044	Conventions	1.13	0.73	1.30	0.67	1.20	0.70	0.24	0.10	-0.14
		Elaboration	1.44	0.56	1.31	0.49	1.38	0.51	-0.25	-0.13	0.13
		Organization	1.53	0.62	1.42	0.51	1.50	0.56	-0.19	-0.06	0.14
5	31045	Conventions	1.43	0.76	1.63	0.63	1.47	0.73	0.29	0.05	-0.24
		Elaboration	1.57	0.60	1.76	0.62	1.60	0.59	0.32	0.06	-0.27
		Organization	1.73	0.61	1.87	0.56	1.78	0.57	0.25	0.09	-0.17
	31046	Conventions	1.27	0.77	1.52	0.64	1.34	0.74	0.36	0.10	-0.26
		Elaboration	1.55	0.58	1.55	0.50	1.53	0.51	0.00	-0.04	-0.04
		Organization	1.78	0.64	1.82	0.61	1.78	0.60	0.07	0.01	-0.06
6	31047	Conventions	1.49	0.71	1.63	0.58	1.54	0.68	0.22	0.07	-0.15
		Elaboration	1.76	0.69	1.61	0.57	1.70	0.63	-0.24	-0.09	0.16
		Organization	1.96	0.76	1.79	0.66	1.91	0.71	-0.25	-0.08	0.18
	31048	Conventions	1.34	0.74	1.54	0.63	1.40	0.72	0.30	0.09	-0.21

Grade	Item ID	Trait	H1		AS		Final		SMD		
			Mean	SD	Mean	SD	Mean	SD	H1-AS	H1-Final	AS-Final
		Elaboration	1.59	0.65	1.36	0.54	1.55	0.63	-0.39	-0.07	0.32
		Organization	1.77	0.70	1.52	0.62	1.72	0.69	-0.37	-0.08	0.30
7	31049	Conventions	1.55	0.62	1.70	0.51	1.61	0.58	0.26	0.10	-0.16
		Elaboration	1.81	0.76	1.40	0.51	1.64	0.65	-0.63	-0.24	0.41
		Organization	2.03	0.72	1.59	0.53	1.87	0.63	-0.70	-0.23	0.49
	31050	Conventions	1.67	0.55	1.50	0.67	1.60	0.59	-0.27	-0.12	0.16
		Elaboration	2.04	0.79	1.65	0.48	1.88	0.67	-0.60	-0.23	0.39
		Organization	2.12	0.78	1.70	0.49	1.96	0.65	-0.66	-0.23	0.46
8	31051	Conventions	1.81	0.47	1.63	0.61	1.76	0.52	-0.33	-0.11	0.22
		Elaboration	2.13	0.71	1.86	0.73	2.00	0.70	-0.38	-0.20	0.19
		Organization	2.19	0.66	1.91	0.77	2.03	0.71	-0.39	-0.23	0.16
	31052	Conventions	1.83	0.44	1.68	0.62	1.75	0.56	-0.28	-0.16	0.12
		Elaboration	2.27	0.69	1.98	0.73	2.05	0.71	-0.40	-0.31	0.09
		Organization	2.32	0.64	2.05	0.77	2.11	0.75	-0.38	-0.30	0.07

**Essays that were given a condition code by Autoscore or human readers were excluded.*

3.7.8. Summary

This set of results shows that Autoscore models well the responses and scores from the same sample on which it was trained, as indicated by the performance on the held-out validation sample. Autoscore shows adequate agreement relative to the human score and the final, resolved score on the First 500 sample, a sample in which both expert and typical raters are used. Autoscore also shows adequate agreement relative to the human score and the final, resolved score on the Random 5% sample, a sample in which primarily typical raters are used and which is restricted to response with higher confidence values. Autoscore shows mean score differences relative to the human scores in both samples (with more differences in the Random 5% sample) and slightly higher differences relative to the final, resolved score compared to the human scores and the final, resolved score. With the current scoring design, it is not possible to disentangle the source of the difference; in future years, CAI recommends the use of a vetted validity sample to better determine whether the source of the differences is due primarily to changes in how human raters are scoring or in how Autoscore is scoring new responses.

4. EVIDENCE OF CONTENT VALIDITY

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

4.1. CONTENT STANDARDS

The WVGSA was aligned to the English language arts (ELA), mathematics, and science standards adopted by West Virginia in April 2017. The standards are available for review at the following URL: https://wvde.us/assessment/scaled-score-information/wvgsa-in-grades-3-8/. Blueprints were developed to ensure that the test and the items were aligned to the prioritized standards that they were intended to measure. A complete description of the blueprint and test construction process can be found in Volume 2, Test Development.

Table 26 through Table 28 present the number of items in the operational item pool measuring each reporting category by grade for ELA, mathematics, and science, respectively.

Table 26: Number of Items for Each Reporting Category, ELA

Grade	Reporting Category	Number of Items
3	Informational Text	194
	Literary Text	159
	Writing and Language	105
4	Informational Text	193
	Literary Text	132
	Writing and Language	119
5	Informational Text	174
	Literary Text	155
	Writing and Language	106
6	Informational Text	242
	Literary Text	172
	Writing and Language	101
7	Informational Text	199
	Literary Text	149
	Writing and Language	91
8	Informational Text	179
	Literary Text	110
	Writing and Language	93

Table 27: Number of Items for Each Reporting Category, Mathematics

Grade	Reporting Category	Number of Items
3	Measurement, Data, and Geometry	194
	Numbers and Operations in Base Ten & Fractions	273
	Operations and Algebraic Thinking	186

Grade	Reporting Category	Number of Items
4	Measurement, Data, and Geometry	176
	Numbers and Operations in Base Ten & Fractions	391
	Operations and Algebraic Thinking	124
5	Measurement, Data, and Geometry	152
	Numbers and Operations in Base Ten & Fractions	317
	Operations and Algebraic Thinking	92
6	Expressions and Equations	200
	Geometry & Statistics and Probability	140
	Ratios and Proportional Relationships & Number System	335
7	Expressions and Equations	91
	Geometry	100
	Ratios and Proportional Relationships & Number System	179
	Statistics and Probability	128
8	Expressions and Equations & Number System	220
	Functions	117
	Geometry & Statistics and Probability	229

Table 28: Number of Items for Each Reporting Category, Science

Grade	Reporting Category	Cluster	Stand-Alone
5	Earth and Space Science	35	38
	Life Science	43	40
	Physical Science	41	42
8	Earth and Space Science	39	34
	Life Science	59	44
	Physical Science	48	38

4.2. INDEPENDENT ALIGNMENT STUDY

While it is critically important to develop and strictly enforce an item development process that works to ensure alignment of test items to content standards, it is also important to independently verify the alignment of test items to content standards. The WebbAlign team of the not-for-profit Wisconsin Center for Education Products and Services (WCEPS) conducted alignment studies in 2019 for ELA, mathematics, and science. Refer to Volume 7, Special Studies of this technical report for more information.

5. EVIDENCE ON INTERNAL-EXTERNAL STRUCTURE

This section explores the internal structure of the assessment using the scores provided at the reporting category level. The relationship of the subscores is just one indicator of the test dimensionality.

In English language arts (ELA), there are three reporting categories per grade: Reading Standards for Informational/Nonfiction Text, Reading Standards for Literature/Fiction, and Writing and Language Standards. In mathematics, reporting categories differ in each grade (refer to Table 29 and Table 30).

Scale scores and relative strengths and weaknesses based on each reporting category were provided to students. Evidence is needed to verify that scale scores and relative strengths and weaknesses for each reporting category provide both different and useful information for student achievement.

It may not be reasonable to expect that the reporting category scores are completely orthogonal—this would suggest that there are no relationships among reporting category scores and would make justification of a unidimensional item response theory (IRT) model difficult, though reporting these separate scores could then easily be justified. On the contrary, if the reporting categories were perfectly correlated, a unidimensional model could be justified, but reporting separate scores could not.

One pathway to explore the internal structure of the test is via a second-order factor model, assuming a general mathematics construct (first factor) with reporting categories (second factor), and that the items load onto the reporting category they intend to measure. If the first-order factors are highly correlated and the model fits data well for the second-order model, this provides evidence of unidimensionality and reporting subscores. A second-order factor model was fit to the spring 2018 fixed forms for ELA and mathematics.

The science assessment is modeled with the Rasch testlet model (Wang & Wilson, 2005). Unlike the models for ELA and mathematics, the IRT model for science is a high-dimensional model, incorporating a nuisance dimension for each item cluster and an overall dimension representing the overall proficiency in science. 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.

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

5.1. CORRELATIONS AMONG REPORTING CATEGORY SCORES

Table 29 and Table 30 present the observed and disattenuated correlation matrix of the reporting category scale scores for ELA and mathematics, respectively. Values in the lower triangle are observed correlations, and values in the upper triangle are disattenuated. Diagonals (highlighted in gray) are the reliability coefficient. In ELA, the observed correlations ranged from 0.52–0.65,

and disattenuated correlations ranged from 0.72–0.87. In mathematics, the observed correlations among the reporting categories ranged from 0.59–0.78 and the disattenuated correlations ranged from 0.87–0.99.

Table 31 presents the observed and disattenuated correlation matrix of the reporting category scale scores for science. The observed correlations ranged from 0.68–0.69 and the disattenuated correlations ranged from 0.98–1.00.

In some instances, these correlations were lower than one might expect. However, as previously noted, the correlations were subject to a large amount of measurement error at the strand level, given the limited number of items from which the scores were derived. Consequently, over-interpretation of these correlations, as either high or low, should be made cautiously.

Table 29: Correlations among Reporting Categories, ELA

Grade	Reporting Category	Mean # of Items per Student	IT	LT	WL
3	Reading Informational Text (IT)	13.3	0.69*	0.72	0.73
	Reading Literary Text (LT)	16.4	0.52	0.75*	0.75
	Writing and Language (WL)	9.4	0.55	0.59	0.83*
4	Reading Informational Text (IT)	13.2	0.7*	0.8	0.73
	Reading Literary Text (LT)	16.1	0.58	0.75*	0.84
	Writing and Language (WL)	9.2	0.55	0.65	0.8*
5	Reading Informational Text (IT)	13.0	0.71*	0.79	0.76
	Reading Literary Text (LT)	16.2	0.58	0.75*	0.75
	Writing and Language (WL)	9.1	0.58	0.59	0.82*
6	Reading Informational Text (IT)	16.5	0.71*	0.78	0.77
	Reading Literary Text (LT)	13.2	0.55	0.7*	0.77
	Writing and Language (WL)	9.0	0.59	0.58	0.82*
7	Reading Informational Text (IT)	16.5	0.74*	0.87	0.82
	Reading Literary Text (LT)	13.1	0.62	0.69*	0.82
	Writing and Language (WL)	9.1	0.64	0.62	0.82*
8	Reading Informational Text (IT)	16.5	0.73*	0.81	0.83
	Reading Literary Text (LT)	13.3	0.59	0.73*	0.77
	Writing and Language (WL)	9.0	0.65	0.6	0.84*

**Diagonal value represents the reliability coefficient of the reporting category. Observed correlations are below the diagonal, and disattenuated are above.*

Table 30: Correlations among Reporting Categories, Mathematics

Grade	Reporting Category	Mean # of Items per Student	MDG	NBTF	OAT	
3	Measurement, Data, and Geometry (MDG)	10.0	0.73*	0.98	0.92	-
	Numbers and Operations in Base Ten & Fractions (NBTF)	14.0	0.76	0.83*	0.94	-
	Operations and Algebraic Thinking (OAT)	10.0	0.71	0.77	0.81*	-
4	Measurement, Data, and Geometry (MDG)	9.0	0.74*	0.95	0.87	-
	Numbers and Operations in Base Ten & Fractions (NBTF)	16.0	0.75	0.85*	0.96	-
	Operations and Algebraic Thinking (OAT)	8.9	0.66	0.78	0.77*	-
5	Measurement, Data, and Geometry (MDG)	10.0	0.73*	0.94	0.92	-
	Numbers and Operations in Base Ten & Fractions (NBTF)	16.0	0.73	0.82*	0.92	-
	Operations and Algebraic Thinking (OAT)	8.0	0.68	0.72	0.75*	-
			EE	**GSP**	**RPNS**	
6	Expressions and Equations (EE)	11.9	0.75*	0.87	0.99	-
	Geometry & Statistics and Probability (GSP)	8.0	0.59	0.61*	0.88	-
	Ratios and Proportional Relationships & Number System (RPNS)	14.1	0.76	0.61	0.79*	-
			EE	**G**	**RPNS**	**SP**
7	Expressions and Equations (EE)	8.5	0.69*	0.88	0.92	0.91
	Geometry (G)	8.0	0.61	0.7*	0.89	0.88
	Ratios and Proportional Relationships & Number System (RPNS)	8.9	0.68	0.66	0.79*	0.93
	Statistics and Probability (SP)	8.6	0.63	0.61	0.69	0.69*
			EENS	**F**	**GSP**	
8	Expressions and Equations & Number System (EENS)	12.0	0.78*	0.94	0.99	-
	Functions (F)	8.0	0.67	0.65*	0.95	-
	Geometry & Statistics and Probability (GSP)	14.0	0.77	0.67	0.77*	-

**Diagonal value represents the reliability coefficient of the reporting category. Observed correlations are below the diagonal, and disattenuated are above.*

Table 31: Correlations among Reporting Categories, Science

Grade	Reporting Category	ESS	LS	PS
5	ESS	0.69*	0.99	0.98
	LS	0.69	0.70*	0.98
	PS	0.68	0.68	0.69*
8	ESS	0.69*	1.00	1.00
	LS	0.69	0.69*	1.00
	PS	0.69	0.69	0.68*

ESS: Earth and Space Science; LS: Life Science; PS: Physical Science

**Diagonal value represents the reliability coefficient of the reporting category. Observed correlations are below the diagonal, and disattenuated are above the diagonal.*

5.2. CONFIRMATORY FACTOR ANALYSIS FOR SPRING 2018 ELA AND MATHEMATICS

In the 2018–2019 school year, the WVGSA was administered as an adaptive test. Unlike the fixed-form tests administered in the 2017–2018 school year, the number of students who took each item was not always sufficient for conducting confirmatory factor analysis (CFA). Due to this restriction, the internal structural validity evidence supported by 2017–2018 WVGSA student data is summarized in this section. The 2017–2018 WVGSA and 2018–2019 WVGSA were constructed based on the same content standards and similar test blueprints. The internal structure of the two assessments is expected to be equivalent, with some degree of variability in model coefficients.

The WVGSA had test items designed to measure different standards and higher-level reporting categories. Test scores were reported as an overall performance measure. Additionally, scores on the various reporting categories were also provided as indices of strand-specific performance. The strand scores were reported in a fashion that aligned with the theoretical structure of the test derived from the test blueprint.

The results in this section are intended to provide evidence that the methods for reporting WVGSA strand scores align with the test's underlying structure and evidence for the appropriateness of the selected IRT models. This section is based on a second-order CFA, in which the first-order factors load onto a common underlying factor. The first-order factors represent the dimensions of the test blueprint, and items load onto the factors that they are intended to measure. The underlying structure of the ELA and mathematics tests was generally common across all grades, which is useful for comparing the results of our analyses across grades.

Although the test consisted of items targeting different standards, all items within a grade and subject were calibrated concurrently using the various IRT models described in this technical report. This implies the pivotal IRT assumption of local independence (Lord, 1980). Formally stated, this assumption posits that the probability of the outcome on item *i* depends only on the student's ability and the item's characteristics. Beyond that, the score of item *i* is independent of the outcome of all other items. From this assumption, the joint density (i.e., the likelihood) is

viewed as the product of the individual densities. Thus, the maximum likelihood estimation of person and item parameters in traditional IRT is derived based on this theory.

The measurement model and the score reporting method assume a single underlying factor, with separate factors representing each reporting category. Consequently, it is important to collect validity evidence on the internal structure of the assessment to determine the rationality of conducting concurrent calibrations and using these scoring and reporting methods.

5.2.1. Factor Analytic Methods

A series of CFAs were conducted using the statistical program Mplus [version 7.31] (Muthén & Muthén, 2012) for each grade and subject assessment. Mplus is commonly used for collecting validity evidence on the internal structure of assessments. Weighted least squares means and variance adjusted (WLSMV) was employed as the estimation method because it is less sensitive to the sample size and model and is shown to perform well with categorical variables (Muthén, du Toit, & Spisic, 1997).

As previously stated, the method of reporting scores for West Virginia implied separate factors connected by a single underlying factor for each reporting category. This model is subsequently referred to as the implied model. In factor analytic terms, this suggests that test items load onto separate first-order factors, with the first-order factors connected to a single underlying second-order factor. The use of CFA in this section establishes some validity evidence for the degree to which the implied model is reasonable.

A chi-square difference test is often applied to assess model fit. However, this test is sensitive to sample size, almost always rejecting the null hypothesis when the sample size is large. Therefore, instead of conducting a chi-square difference test, other goodness-of-fit indices were used to evaluate the implied model for the WVGSA.

If the internal structure of the test was strictly unidimensional, then the overall person ability measure, theta (θ), would be the single underlying common factor, and the correlation matrix among test items would suggest no discernable pattern among factors. Therefore, there would be no empirical or logical basis to report scores for the separate performance categories. In factor analytic terms, a strictly unidimensional test structure implies a single-order factor model in which all test items load onto a single underlying factor. The development below expands the first-order model to a generalized second-order parameterization to show the relationship between the models.

The factor analysis models are based on the matrix $\boldsymbol{S}$ of tetrachoric and polychoric sample correlations among the item scores (Olsson, 1979), and the matrix $\boldsymbol{W}$ of asymptotic covariances among these sample correlations (Jöreskog, 1994) is employed as a weight matrix in a weighted least squares estimation approach (Browne, 1984; Muthén, 1984) to minimize the fit function:

$$F_{WLS} = \text{vech}(\boldsymbol{S} - \widehat{\boldsymbol{\Sigma}})' \boldsymbol{W^{-1}} \text{vech}(\boldsymbol{S} - \widehat{\boldsymbol{\Sigma}}).$$

In the preceding equation, $\widehat{\boldsymbol{\Sigma}}$ is the implied correlation matrix, given the estimated factor model, and the function vech vectorizes a symmetric matrix. That is, the vech stacks each column of the matrix to form a vector. Note that the WLSMV approach (Muthén, du Toit, & Spisic, 1997) employs a weight matrix of asymptotic variances (i.e., the diagonal of the weight matrix) instead of the full asymptotic covariances.

We posit a first-order factor analysis in which all test items load onto a single underlying common factor as the base model. The first-order model can be mathematically represented as

$$\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Lambda\Phi\Lambda}' + \boldsymbol{\Theta},$$

where $\boldsymbol{\Lambda}$ is the matrix of item factor loadings (with $\boldsymbol{\Lambda}'$ representing its transpose), and $\boldsymbol{\Theta}$ is the uniqueness or measurement error. The matrix $\boldsymbol{\Phi}$ is the correlation among the separate factors. For the base model, items are thought to load onto a single underlying factor only. Hence $\boldsymbol{\Lambda}$ is a *p x 1* vector, where *p* is the number of test items and $\boldsymbol{\Phi}$ is a scalar equal to 1. Therefore, it is possible to drop the matrix $\boldsymbol{\Phi}$ from the general notation. However, this notation is retained to facilitate comparisons to the implied model more easily, such that it can subsequently be viewed as a special case of the second-order factor analysis.

For the implied model, we posit a second-order factor analysis in which test items are coerced to load onto the reporting categories they are designed to target, and all reporting categories share a common underlying factor. The second-order factor analysis can be mathematically represented as:

$$\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Lambda}(\boldsymbol{\Gamma\Phi\Gamma}' + \boldsymbol{\Psi})\boldsymbol{\Lambda}' + \boldsymbol{\Theta},$$

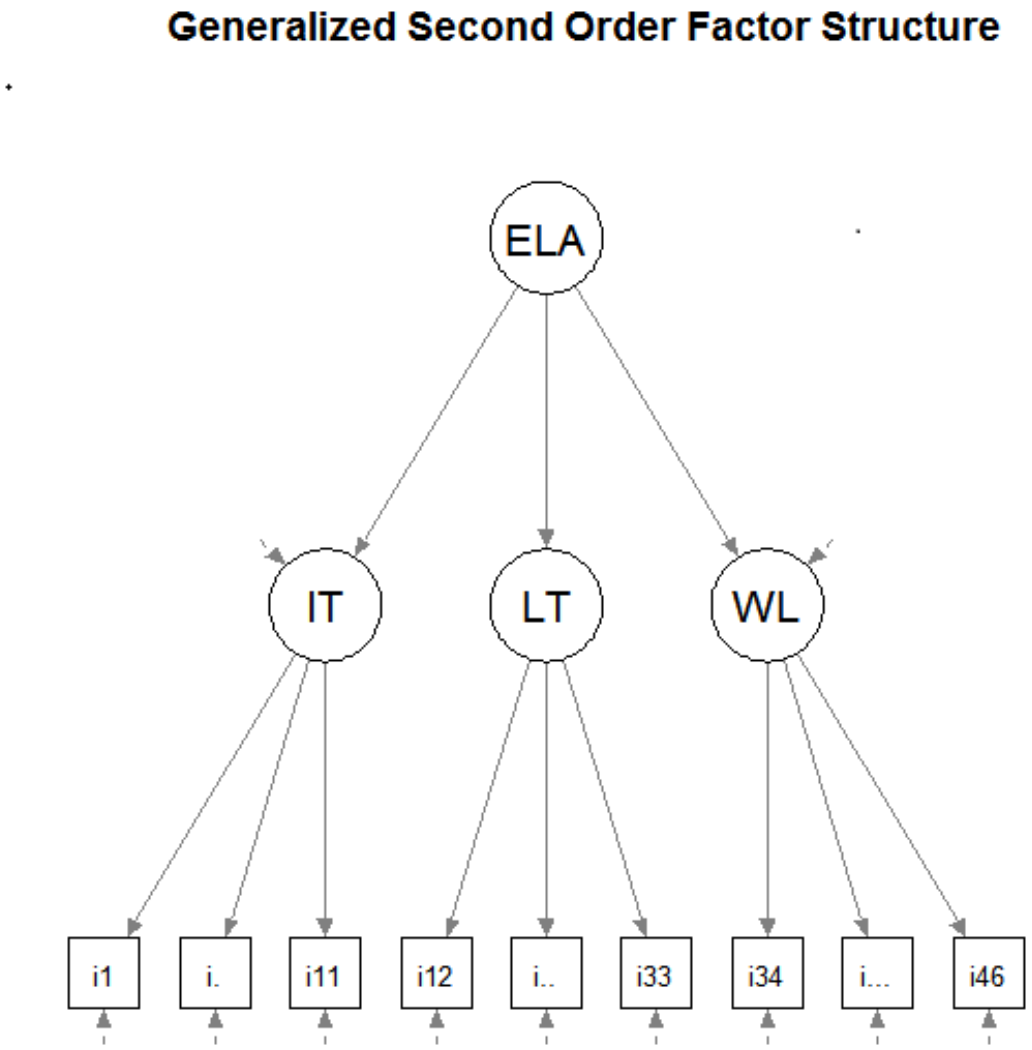
where $\hat{\boldsymbol{\Sigma}}$ is the implied correlation matrix among test items, $\boldsymbol{\Lambda}$ is the *p x k* matrix of the first-order factor loadings relating item scores to first-order factors, $\boldsymbol{\Gamma}$ is the *k x 1* matrix of the second-order factor loadings relating the first-order factors to the second-order factor with *k* denoting the number of factors, $\boldsymbol{\Phi}$ is the correlation matrix of the second-order factors, and $\boldsymbol{\Psi}$ is the matrix of the first-order factor residuals. All other notations are the same as in the first-order model. Note that the second-order model expands the first-order model such that $\boldsymbol{\Phi} \rightarrow \boldsymbol{\Gamma\Phi\Gamma}' + \boldsymbol{\Psi}$. Therefore, the first-order model is said to be nested within the second-order model.

There are three reporting categories for ELA and three to four categories for mathematics (refer to Table 26 and Table 27 for reporting category information). Therefore, the number of rows in $\boldsymbol{\Gamma}$ (*k)* differs between subjects, but the general structure of the factor analysis is consistent across ELA and mathematics.

The second-order factor model can also be represented graphically, and a sample of the generalized approaches is provided in Figure 5. This figure is representative of the factor analyses performed for all grades and subjects, with the understanding that the number of items within each reporting category could vary across the grades.

The purpose of conducting a CFA for the WVGSA was to provide evidence that each assessment in the WVGSA implied a second-order factor model: a single underlying second-order factor with the first-order factors defining each of the reporting categories.

Figure 5: Second-Order Factor Model, ELA



5.2.2. Results

Several goodness-of-fit statistics from each of the analyses are presented in Table 32 and Table 33. These tables present the summary results obtained from CFA. Three goodness-of-fit indices were used to evaluate the model fit of the item parameters to the way students responded to the items. The root mean square error of approximation (RMSEA) is referred to as a badness-of-fit index so that a value closer to 0 implies better fit and a value of 0 implies best fit. An RMSEA below 0.05 is generally considered good fit, and an RMSEA over 0.1 suggests poor fit (Browne & Cudeck, 1993). The Tucker–Lewis index (TLI) and the comparative fit index (CFI) are incremental goodness-of-fit indices. These indices compare the implied model to the baseline model in which no observed variables are correlated (i.e., there are no factors). Values greater than 0.9 are recognized as acceptable, and values over 0.95 are considered good fit (Hu & Bentler, 1999). As Hu and Bentler (1999) suggest, the selected cut-off values of the fit index should not be overgeneralized and should be interpreted with caution.

Based on the fit indices, the model showed good fit across content domains. For all tests, the RMSEA was equal to or less than 0.05, and the CFI and TLI were equal to or greater than 0.95 except for grade 4 mathematics Form B, which had a TLI of 0.94.

Table 32: Goodness-of-Fit Second-Order CFA, Spring 2018 ELA

Grade	Form	*df*	*RMSEA*	*CFI*	*TLI*	*Convergence*
3	A*	943	0.04	0.96	0.96	Yes
	B	986	0.04	0.95	0.95	Yes
4	A	986	0.03	0.97	0.97	Yes
	B	986	0.03	0.96	0.96	Yes
5	A*	943	0.03	0.97	0.97	Yes
	B*	943	0.03	0.96	0.96	Yes
6	A	986	0.03	0.97	0.97	Yes
	B*	987	0.03	0.96	0.96	Yes
7	A	942	0.03	0.97	0.97	Yes
	B	986	0.03	0.96	0.96	Yes
8	A*	987	0.03	0.98	0.98	Yes
	B	986	0.04	0.97	0.97	Yes

**For these tests, the second-order model was run by constraining the residual variance of a certain factor to zero due to non-significant negative residual variance.*

Table 33: Goodness-of-Fit Second-Order CFA, Spring 2018 Mathematics

Grade	Form	*df*	*RMSEA*	*CFI*	*TLI*	*Convergence*
3	A*	817	0.03	0.98	0.97	Yes
	B	816	0.03	0.98	0.98	Yes
4	A	816	0.03	0.97	0.97	Yes
	B	816	0.05	0.95	0.94	Yes
5	A	816	0.03	0.97	0.97	Yes
	B	816	0.03	0.97	0.97	Yes
6	A	816	0.03	0.97	0.97	Yes
	B	816	0.02	0.98	0.98	Yes
7	A	815	0.02	0.99	0.99	Yes
	B*	816	0.03	0.98	0.98	Yes
8	A	816	0.03	0.96	0.96	Yes
	B	816	0.03	0.96	0.96	Yes

**For these tests, the second-order model was run by constraining the residual variance of a certain factor to zero due to non-significant negative residual variance.*

The second-order factor model converged for all tests. However, the residual variance for one factor fell slightly below the boundary of 0 for ELA grade 3 Form A, grade 5 Forms A and B, grade 6 Form B, and grade 8 Form B; mathematics grade 3 Form A and grade 7 Form B when using the Mplus software package. This negative residual variance may be related to the computational implementation of the optimization approach in Mplus. It may also be a flag related to model misspecification, or it may be related to other causes (van Driel, 1978; Chen, Bollen, Paxton, Curran, & Kirby, 2001). The residual variance was constrained to 0 for these tests. This is equivalent to treating the parameter as fixed, which does not necessarily conform to our *a priori* hypothesis.

The estimated correlations between the reporting categories from the second-order factor model are presented in Table 34 and Table 35 for ELA and mathematics, respectively. In all cases, these correlations are very high. However, the results provide empirical evidence that there is some detectable dimensionality among reporting categories.

Table 34: Correlations among ELA Factors, Spring 2018

Grade	Reporting Category	Form A		Form B	
		IT	LT	IT	LT
3	LT	0.99	-	0.97	-
	WL	0.76	0.75	0.78	0.76
4	LT	0.96	-	0.98	-
	WL	0.73	0.76	0.80	0.80
5	LT	0.98	-	0.97	-
	WL	0.78	0.79	0.78	0.75
6	LT	0.96	-	0.98	-
	WL	0.78	0.80	0.84	0.82
7	LT	0.98	-	0.97	-
	WL	0.78	0.77	0.79	0.77
8	LT	0.96	-	0.95	-
	WL	0.77	0.80	0.82	0.80

IT: Reading Standards for Information/Nonfiction Text; LT: Reading Standards for Literature/Fiction Text; WL: Writing and Language Standards

Table 35: Correlations among Mathematics Factors, Spring 2018

Grade	Reporting Category	Form A		Form B	
		MDG	NBTF	MDG	NBTF
3	NBTF	1.00	-	0.98	-
	OAT	0.97	0.97	0.95	0.96
4	NBTF	0.96	-	0.89	-
	OAT	0.93	0.97	0.94	0.92
5	NBTF	0.94	-	0.97	-
	OAT	0.86	0.86	0.94	0.92

MDG: Measurement, Data, and Geometry; NBTF: Numbers and Operations in Base Ten & Fractions; OAT: Operations and Algebraic Thinking

Grade	Reporting Category	Form A		Form B	
		EENS	GSP	EENS	GSP
6	GSP	0.95	-	0.95	-
	RPNS	0.96	0.95	0.96	0.95

EENS: Expressions and Equations & Number System; GSP: Geometry & Statistics and Probability; RPNS: Ratios and Proportional Relationships & Number System

Grade	Reporting Category	Form A			Form B		
		EE	G	RPNS	EE	G	RPNS
7	G	0.93	-	-	0.94	-	-
	RPNS	0.98	0.93	-	0.97	0.97	-
	SP	0.98	0.93	0.98	0.92	0.92	0.95

EE: Expressions and Equations; G: Geometry; RPNS: Ratios and Proportional Relationships & Number System; SP: Statistics and Probability

Grade	Reporting Category	Form A		Form B	
		EENS	F	EENS	F
8	F	0.95	-	0.97	-
	GSP	0.96	0.93	0.94	0.94

EENS: Expressions and Equations & Number System; F: Functions; GSP: Geometry & Statistics and Probability

5.2.3. Discussion

In all scenarios, the empirical results suggest that the implied model fits the data well. These results indicate that reporting an overall score, in addition to separate scores for the individual reporting categories, is reasonable, as the inter-correlations among items suggest that there are detectable distinctions among reporting categories.

The correlations among the separate factors are high, which is reasonable. This finding supports the measurement model, given that the calibration of all items is performed concurrently. If the correlations among factors were very low, this could suggest that a different IRT model would be needed (e.g., multidimensional IRT) or that the IRT calibration should be performed separately for items measuring different factors. The high correlations among the factors suggest that these alternative methods are unnecessary and that our current approach is preferable.

Overall, these results provide empirical evidence and justify using our scoring and reporting methods. The results also justify the IRT model employed currently.

5.3. LOCAL INDEPENDENCE

The validity of the application of the IRT depends greatly on meeting the underlying assumptions of the models. One assumption is local independence, which means that for a given proficiency estimate, the (marginal) likelihood is maximized, assuming that the probability of correct responses is the product of independent probabilities over all items (Chen & Thissen, 1997):

$$\mathrm{L}(\theta) = \int \prod_{i=1}^{I} \Pr(z_i|\theta)\, f(\theta) \mathrm{d}\theta.$$

When local independence is not met, there are issues of multidimensionality that are unaccounted for in the modeling of the data (Bejar, 1980). In fact, Lord (1980) noted that "local independence follows automatically from unidimensionality" (Bejar, 1980, p.5). From a dimensionality perspective, there might be nuisance factors influencing the relationships among certain items after accounting for the intended construct of interest. These nuisance factors can be influenced by several testing features, such as speediness, fatigue, item chaining, and item or response formats (Yen, 1993).

Yen's Q_3 statistic (Yen, 1984) was used to measure local independence, which was derived from the correlation between the performances of two items. Simply, the Q_3 statistic is the correlation among IRT residuals and is computed using the following equations:

$$d_{ij} = u_{ij} - T_j(\hat{\theta}_i),$$

where u_{ij} is the item score of the *j*th test taker for item *i*; $T_i(\hat{\theta}_j)$ is the estimated true score for item *i* of test taker *j*, which is defined as

$$T_i(\hat{\theta}_j) = \sum_{l=1}^{m} y_{il} P_{il}(\hat{\theta}_j),$$

where y_{il} is the weight for response category *l*; *m* is the number of response categories; and $P_{il}(\hat{\theta}_j)$ is the probability of response category *l* to item *i* by test taker *j* with the ability estimate $\hat{\theta}_j$.

The pairwise index of local dependence Q_3 between item *i* and item *i'* is

$$Q_{3ii'} = r(d_i, d_{i'}),$$

where *r* refers to the Pearson product-moment correlation.

When there are *n* items, $n(n\text{-}1)/2$, Q_3 statistics will be produced. The Q_3 values are expected to be small. Table 36 and Table 37 present summaries of the distributions of Q_3 statistics: minimum, 5th

percentile; median, 95th percentile; and maximum values from each grade and subject. The results show that for all grades and subjects, about 90% of the items between the 5th and 95th percentiles were smaller than the critical value of 0.2 for $|Q_3|$ (Chen & Thissen, 1997).

Table 36: ELA Q_3 Statistic, Spring 2018

Grade	Form	Unconditional Observed Correlation	Q_3 Distribution					Within Passage Q_3**	
			Minimum	5th Percentile	Median	95th Percentile	Maximum*	Minimum	Maximum
3	A	0.26	-0.21	-0.12	-0.01	0.03	0.87	-0.15	0.17
	B	0.24	-0.22	-0.12	-0.01	0.03	0.88	-0.07	0.15
4	A	0.23	-0.20	-0.11	-0.02	0.02	0.60	-0.09	0.14
	B	0.26	-0.16	-0.10	-0.02	0.02	0.61	-0.07	0.10
5	A	0.25	-0.19	-0.11	-0.01	0.03	0.57	-0.04	0.12
	B	0.24	-0.18	-0.11	-0.01	0.03	0.58	-0.05	0.12
6	A	0.25	-0.21	-0.11	-0.02	0.02	0.53	-0.07	0.15
	B	0.29	-0.17	-0.11	-0.02	0.02	0.57	-0.12	0.18
7	A	0.23	-0.24	-0.11	-0.01	0.02	0.63	-0.07	0.11
	B	0.26	-0.22	-0.10	-0.02	0.02	0.66	-0.07	0.32
8	A	0.25	-0.23	-0.11	-0.02	0.02	0.79	-0.05	0.14
	B	0.29	-0.17	-0.11	-0.02	0.02	0.79	-0.07	0.13

**Maximum Q_3 values are from elaboration and organization dimensions of the writing prompt.*

***Within Passage Q_3, values are computed for each item pair within a passage.*

Table 37: Mathematics Q_3 Statistic, Spring 2018

Grade	Form	Unconditional Observed Correlation	Q_3 Distribution				
			Minimum	5th Percentile	Median	95th Percentile	Maximum
3	A	0.39	-0.11	-0.06	-0.02	0.02	0.33
	B	0.39	-0.13	-0.07	-0.02	0.02	0.21
4	A	0.40	-0.09	-0.06	-0.02	0.01	0.40
	B	0.42	-0.13	-0.07	-0.02	0.02	0.71
5	A	0.38	-0.09	-0.06	-0.02	0.01	0.40
	B	0.39	-0.13	-0.07	-0.02	0.02	0.71
6	A	0.38	-0.15	-0.06	-0.02	0.02	0.22
	B	0.37	-0.15	-0.06	-0.02	0.02	0.22
7	A	0.37	-0.11	-0.06	-0.02	0.02	0.21
	B	0.37	-0.12	-0.07	-0.02	0.02	0.31
8	A	0.35	-0.12	-0.07	-0.02	0.03	0.34
	B	0.35	-0.13	-0.07	-0.02	0.02	0.51

In the 2018–2019 school year, the WVGSA was administered as an adaptive test. When calculating the Q_3 statistics, pairwise deletion was used in the fixed-form tests. Therefore, Q_3 provides biased estimates under the computer-based test (CAT) administration. Due to this restriction, Q_3 statistics were not calculated for the spring 2019 and 2021 administrations.

5.4. CONVERGENT AND DISCRIMINANT VALIDITY

Collectively, Standard 1.16 through 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 constructs as ELA, mathematics, and science in West Virginia, which could easily permit a cross-test set of correlations, was not available. Therefore, the correlations between subscores within and across ELA, mathematics, and science were examined alternatively. The *a priori* expectation is that subscores within the same subject (e.g., mathematics) will correlate more positively than subscore correlations across subjects (e.g., mathematics and ELA). These correlations are based on a small number of items; consequently, the observed score correlations will be smaller in magnitude due to the very large measurement error at the subscore level. For this reason, the observed score and the disattenuated correlations are provided.

Observed and disattenuated subscore correlations were calculated both within and across subjects. The pattern is generally consistent with the *a priori* expectation that subscores within a test correlate higher than correlations between tests measuring a different construct with a few small notes on the writing dimensions.

Table 38 through Table 43 show the observed and disattenuated score correlations across ELA and mathematics subscores for grades 3–8, in which students took both subjects, including science for grades 5 and 8. Values in the lower triangle are observed correlations, and values in the upper triangle are disattenuated. Diagonals (highlighted in gray) are the reliability coefficient for ELA, mathematics, and science. The subscores are scale scores instead of raw scores, and the number of students used for the computation of the correlations are presented as well.

Table 38: Correlations across Subjects, Grade 3

Subject	Number of Students	Reporting Category	ELA			Mathematics		
			IT	LT	WL	MDG	NBTF	OAT
ELA	17,515	Reading Informational Text (IT)	0.69	0.72	0.73	0.69	0.65	0.63
		Reading Literary Text (LT)	0.52	0.75	0.75	0.7	0.67	0.65
		Writing and Language (WL)	0.55	0.59	0.83	0.81	0.78	0.77
Mathematics		Measurement, Data, and Geometry (MDG)	0.49	0.52	0.63	0.73	0.98	0.92
		Numbers and Operations in Base Ten & Fractions (NBTF)	0.49	0.53	0.65	0.76	0.83	0.94
		Operations and Algebraic Thinking (OAT)	0.47	0.51	0.63	0.71	0.77	0.81

**Diagonal value (grey) represents the reliability coefficient of the reporting category and within-subject correlations are marked in blue. Observed correlations are below the diagonal, and disattenuated are above.*

Table 39: Correlations across Subjects, Grade 4

Subject	Number of Students	Reporting Category	ELA			Mathematics		
			IT	LT	WL	MDG	NBTF	OAT
ELA	17,303	Reading Informational Text (IT)	0.7	0.8	0.73	0.67	0.69	0.69
		Reading Literary Text (LT)	0.58	0.75	0.84	0.72	0.75	0.76
		Writing and Language (WL)	0.55	0.65	0.8	0.77	0.81	0.8
Mathematics		Measurement, Data, and Geometry (MDG)	0.48	0.54	0.59	0.74	0.95	0.87
		Numbers and Operations in Base Ten & Fractions (NBTF)	0.53	0.6	0.67	0.75	0.85	0.96
		Operations and Algebraic Thinking (OAT)	0.51	0.58	0.63	0.66	0.78	0.77

**Diagonal value (grey) represents the reliability coefficient of the reporting category and within-subject correlations are marked in blue. Observed correlations are below the diagonal, and disattenuated are above.*

Table 40: Correlations across Subjects, Grade 5

Subject	Number of Students	Reporting Category	ELA			Mathematics			Science		
			IT	LT	WL	MDG	NBTF	OAT	ESS	LS	PS
ELA	17,635	Reading Informational Text (IT)	0.71	0.80	0.76	0.72	0.70	0.74	0.81	0.82	0.79
		Reading Literary Text (LT)	0.58	0.75	0.75	0.70	0.68	0.72	0.82	0.83	0.79
		Writing and Language (WL)	0.58	0.59	0.82	0.79	0.78	0.82	0.81	0.81	0.80
Mathematics		Measurement, Data, and Geometry (MDG)	0.52	0.52	0.61	0.73	0.94	0.91	0.87	0.84	0.86
		Numbers and Operations in Base Ten & Fractions (NBTF)	0.53	0.53	0.64	0.73	0.82	0.92	0.83	0.79	0.81
		Operations and Algebraic Thinking (OAT)	0.54	0.54	0.64	0.68	0.72	0.75	0.85	0.82	0.83
Science		Earth and Space Science (ESS)	0.57	0.59	0.61	0.62	0.63	0.61	0.69	0.99	0.98
		Life Science (LS)	0.58	0.60	0.61	0.60	0.60	0.59	0.69	0.70	0.98
		Physical Science (PS)	0.55	0.57	0.60	0.61	0.61	0.60	0.68	0.68	0.69

**Diagonal value (grey) represents the reliability coefficient of the reporting category and within-subject correlations are marked in blue. Observed correlations are below the diagonal, and disattenuated are above.*

Table 41: Correlations across Subjects, Grade 6

Subject	Number of Students	Reporting Category	ELA			Mathematics		
			IT	LT	WL	EE	GSP	RPNS
ELA	17,658	Reading Informational Text (IT)	0.71	0.78	0.77	0.73	0.7	0.75
		Reading Literary Text (LT)	0.55	0.7	0.77	0.72	0.66	0.71
		Writing and Language (WL)	0.59	0.58	0.82	0.82	0.75	0.82
Mathematics		Expressions and Equations (EE)	0.53	0.52	0.64	0.75	0.87	0.99
		Geometry & Statistics and Probability (GSP)	0.46	0.43	0.53	0.59	0.61	0.88
		Ratios and Proportional Relationships & Number System (RPNS)	0.56	0.53	0.66	0.76	0.61	0.79

**Diagonal value (grey) represents the reliability coefficient of the reporting category and within-subject correlations are marked in blue. Observed correlations are below the diagonal, and disattenuated are above.*

Table 42: Correlations across Subjects, Grade 7

Subject	Number of Students	Reporting Category	ELA			Mathematics			
			IT	LT	WL	EE	G	RPNS	SP
ELA	18,193	Reading Informational Text (IT)	0.74	0.87	0.82	0.76	0.72	0.75	0.78
		Reading Literary Text (LT)	0.62	0.69	0.82	0.74	0.69	0.73	0.77
		Writing and Language (WL)	0.64	0.62	0.82	0.77	0.74	0.78	0.8
Mathematics		Expressions and Equations (EE)	0.54	0.51	0.58	0.69	0.88	0.92	0.9
		Geometry (G)	0.52	0.48	0.56	0.61	0.7	0.89	0.88
		Ratios and Proportional Relationships & Number System (RPNS)	0.57	0.54	0.63	0.68	0.66	0.79	0.93
		Statistics and Probability (SP)	0.56	0.53	0.6	0.62	0.61	0.69	0.69

**Diagonal value (grey) represents the reliability coefficient of the reporting category and within-subject correlations are marked in blue. Observed correlations are below the diagonal, and disattenuated are above.*

Table 43: Correlations across Subjects, Grade 8

Subject	Number of Students	Reporting Category	ELA			Mathematics			Science		
			IT	LT	WL	EENS	F	GSP	ESS	LS	PS
ELA	18,565	Reading Informational Text (IT)	0.73	0.77	0.80	0.73	0.72	0.75	0.80	0.82	0.82
		Reading Literary Text (LT)	0.57	0.73	0.75	0.66	0.66	0.67	0.74	0.76	0.75
		Writing and Language (WL)	0.63	0.58	0.84	0.79	0.77	0.80	0.80	0.80	0.80
Mathematics		Expressions and Equations & Number System (EENS)	0.55	0.50	0.64	0.78	0.93	0.98	0.85	0.84	0.85
		Functions (F)	0.50	0.45	0.57	0.66	0.65	0.94	0.86	0.84	0.84
		Geometry & Statistics and Probability (GSP)	0.57	0.51	0.64	0.76	0.66	0.77	0.87	0.86	0.85
Science		Earth and Space Science (ESS)	0.57	0.53	0.61	0.62	0.57	0.63	0.69	1.00	1.00
		Life Science (LS)	0.58	0.54	0.61	0.62	0.56	0.63	0.69	0.69	1.00
		Physical Science (PS)	0.58	0.53	0.61	0.62	0.56	0.62	0.69	0.69	0.68

**Diagonal value (grey) represents the reliability coefficient of the reporting category and within-subject correlations are marked in blue. Observed correlations are below the diagonal, and disattenuated are above.*

Additionally, the correlation was computed among the overall scores for the three tested subjects: ELA, mathematics, and science. The correlations are presented in Table 44 and are relatively high, with the observed correlation between 0.74–0.79 and disattenuated correlation between 0.83–0.89.

Table 44: Correlations across Spring 2022 ELA, Mathematics, and Science Scores

Grade	Subject	N	ELA	Mathematics	Science
3	ELA		0.89*	0.83	-
	Math		0.75	0.91*	-

Grade	Subject	N	ELA	Mathematics	Science
4	ELA		0.89*	0.83	-
	Math		0.75	0.92*	-
5	ELA		0.90*	0.84	0.89
	Math		0.76	0.90*	0.86
	Science		0.79	0.76	0.88*
6	ELA		0.89*	0.85	-
	Math		0.76	0.88*	-
7	ELA		0.90*	0.85	-
	Math		0.75	0.88*	-
8	ELA		0.91*	0.83	0.86
	Math		0.74	0.89*	0.86
	Science		0.77	0.76	0.87*

**Diagonal value represents the reliability coefficient of the subject. Observed correlations are below the diagonal, and disattenuated are above.*

In fall 2018, optional ELA and mathematics interim assessments were administered. These tests were also online and adaptive. Test takers who took both optional interim assessment and the summative assessment in spring 2022 were identified for conducting the cross-test set of correlations. Table 45 and Table 46 present the correlations between the summative spring 2022 and interim assessments from fall 2021. The means and standard deviations of scale scores reported in the tables are based on data that include test takers who took the summative assessment and the interim assessment. Across all tests, the observed correlations were medium to high, ranging from 0.76 to 0.82. Disattenuated correlations ranged from 0.89 to 0.97.

Table 45: Correlation between Summative and Interim Scores, ELA

Grade	Test	Scale Score Mean	Scale Score SD	Reliability Coefficient	Observed Correlation	Disattenuated Correlation	N
3	Summative	568.53	43.07	0.89	0.78	0.90	6,450
	Interim	555.45	50.36	0.84			
4	Summative	589.23	47.28	0.89	0.80	0.91	6,481
	Interim	575.27	48.91	0.86			
5	Summative	608.47	44.91	0.9	0.81	0.92	6,385
	Interim	599.66	49	0.86			
6	Summative	626.8	46.78	0.89	0.81	0.93	6,616
	Interim	611.2	48.33	0.86			
7	Summative	629.67	49.24	0.9	0.81	0.92	6,535
	Interim	623.84	50.53	0.87			
8	Summative	641.39	51.51	0.91	0.82	0.92	6,998
	Interim	630.28	51.67	0.87			

Table 46: Correlation between Summative and Interim Scores, Mathematics

Grade	Test	Scale Score Mean	Scale Score SD	Reliability Coefficient	Observed Correlation	Disattenuated Correlation	N
3	Summative	420.52	34.81	0.91	0.77	0.89	8,214
	Interim	400.77	36.59	0.83			
4	Summative	445.76	42.74	0.92	0.82	0.94	7,804
	Interim	427.95	42.64	0.82			
5	Summative	464.84	50.67	0.9	0.76	0.93	7,921
	Interim	443.76	50.73	0.75			
6	Summative	481.88	56.46	0.88	0.77	0.92	8,019

Grade	Test	Scale Score Mean	Scale Score SD	Reliability Coefficient	Observed Correlation	Disattenuated Correlation	N
	Interim	466.73	60.14	0.79			
7	Summative	513.78	62.74	0.88	0.78	0.95	8,132
	Interim	498.64	62.32	0.77			
8	Summative	541.56	77.73	0.89	0.80	0.97	8,542
	Interim	522.26	76.37	0.77			

5.5. Relationship of Test Scores to External Variables

The relationship of test scores to external variables, measuring the same or related constructs, is an important source of validity evidence. The WVGSA was first administered to students during spring 2018, replacing the Smarter Balanced Assessment Consortium (SBAC) assessments in ELA and mathematics. Ideally, we would correlate two different tests measuring a common construct administered within a similar time period. Here, we present correlations between two different tests measuring a common construct but measured one year apart. We expected the correlations to be high, suggesting that the WVGSA has a high relationship with an externally developed measure (the SBAC assessments), but the time gap between the two different assessments was also expected to cause correlations to be lower than if the two tests were measured within a similar testing window. Table 47 and Table 48 present correlations between WVGSA scores between spring 2021 and spring 2022. Correlations are between 0.78–0.83, which is relatively high compared to industry standards.

Table 47: Correlations between Spring 2021 and Spring 2022 Scores, ELA

Grade: Spring 2021 → Spring 2022	N	Spring 2021 Marginal Reliability	Spring 2022 Marginal Reliability	Correlations	Disattenuated Correlations
3 → 4	15,404	0.89	0.89	0.78	0.88
4 → 5	15,694	0.87	0.90	0.78	0.88
5 → 6	15,624	0.89	0.89	0.80	0.90
6 → 7	15,741	0.89	0.90	0.81	0.91
7 → 8	16,098	0.90	0.91	0.82	0.91

Table 48: Correlations between Spring 2021 and Spring 2022 Scores, Mathematics

Grade: Spring 2021 → Spring 2022	N	Spring 2021 Marginal Reliability	Spring 2022 Marginal Reliability	Correlations	Disattenuated Correlations
3 → 4	15,408	0.91	0.92	0.83	0.91

Grade: Spring 2021 → Spring 2022	N	Spring 2021 Marginal Reliability	Spring 2022 Marginal Reliability	Correlations	Disattenuated Correlations
4 → 5	15,732	0.91	0.90	0.83	0.92
5 → 6	15,691	0.89	0.88	0.80	0.90
6 → 7	15,855	0.88	0.88	0.81	0.92
7 → 8	16,206	0.88	0.89	0.80	0.90

The WVGSA for science was also first administered to grades 5 and 8 students during spring 2018, replacing the West Virginia Educational Standards Test (WESTEST). Students who took the test in the current school year were in grade 4 and grade 7 in the previous school year where no science test was administered. Therefore, correlations between the scores across two consecutive school years cannot be computed for science.

5.6. CLUSTER EFFECTS FOR SCIENCE

The science assessment is modeled with the Rasch testlet model (Wang & Wilson, 2005). Unlike the models for ELA and mathematics, the IRT model for science is a high-dimensional model, incorporating a nuisance dimension for each item cluster, in addition to an overall dimension representing the overall proficiency in science. Volume 1, Annual Technical Report, Section 5.2, Item Calibration and Linking for Science, presents a detailed description of the IRT model, which is illustrated using a directed graph in Figure 5. The psychometric approach for the science assessment is innovative and quite different from the traditional approach of ignoring local dependencies. The validity evidence on the internal structure presented in this section relates to the presence of cluster effects and how substantial they are.

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

We examined the distribution of cluster variances obtained from the 2018 IRT calibration. For elementary school, the estimated value of the cluster variances of all operational, scored items ranged from 0.11–4.46, with a median value of 0.53 and a mean value of 0.77. The median value was slightly smaller than the estimated variance parameter of the overall science dimension ($\hat{\sigma}_\theta^2 = 0.60$). For middle school, the estimated value of the cluster variances of all operational, scored items ranged from 0.06–5, with a median value of 0.56 and a mean value of 0.66. The median value was slightly larger than the estimated variance parameter of the overall science dimension ($\hat{\sigma}_\theta^2 = 0.53$). Figure 6 and Figure 7 present the histograms of the cluster variances expressed as the proportion of the total variance for all operational items for elementary and middle

school, respectively. For both grade bands, a wide range of cluster variances was observed. These results indicated that, for both 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 6: Cluster Variance Proportion for Science Operational Items in Elementary School

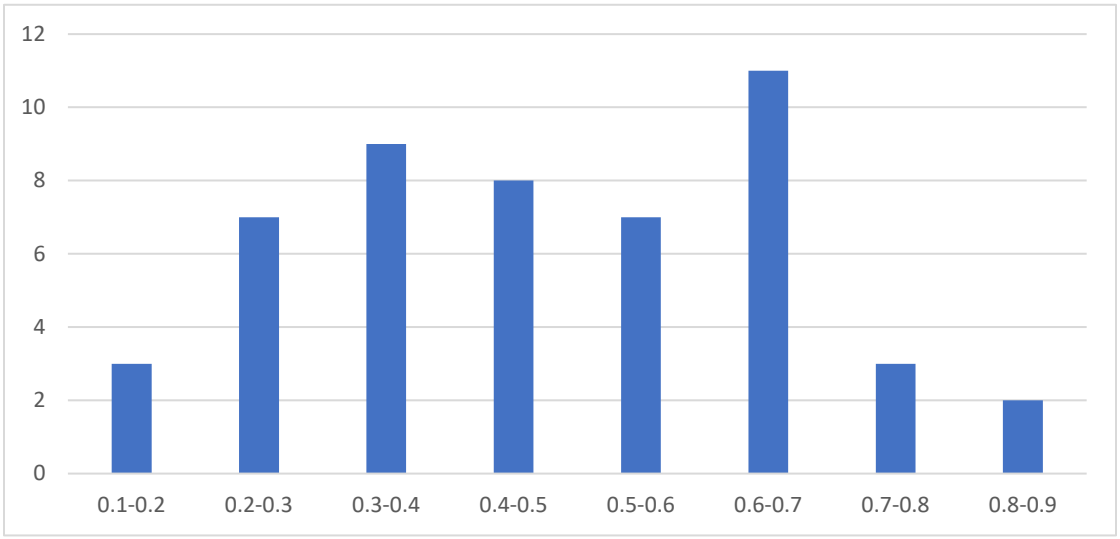
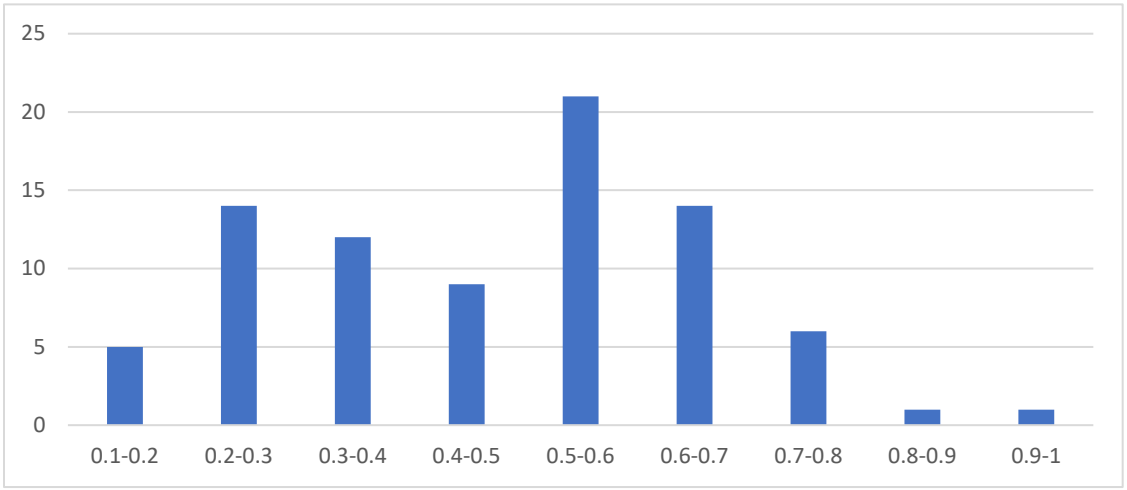


Figure 7: Cluster Variance Proportion for Science Operational Items in Middle School



5.7. CONFIRMATORY FACTOR ANALYSIS FOR SPRING 2018 UTAH SCIENCE

In Section 5.6, Cluster Effects for Science, 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 the CFA of discrete observed variables are not well suited for incomplete data collection designs where each case has data only on a subset of the set of observed variables. Both the linear-on-the-fly (LOFT) and the fully adaptive test design result in sparse data matrices. Because every student responded to only a small number of items relative to the size of the item pool, data are missing on most of the manifest variables for any given student. In year 2018 and beyond, a LOFT/adaptive test design was used for all operational science assessments inspired by the Next Generation Science Standards (NGSS) framework, except for Utah in 2018. As a result, the student responses from other states were not readily amenable to the application of CFA techniques.

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

In 2018, Utah's science assessments comprised a set of fixed-form tests per grade, and all items in those forms were clusters. The number of fixed-form tests varied by grade, but within each grade the total number of clusters was the same across forms. However, some items were rejected during rubric validation or data review and were removed from this analysis. All students with a "completed" status were included in the factor analysis. The percentage of students per grade that had a status other than "completed" was less than 0.85%. Table 49 summarizes the number of forms included in this analysis, the number of clusters per discipline (range across forms), the number of assertions (range across forms), and the number of students (range across forms) for each grade.

Table 49. 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
		Physical Sciences	*Earth and Space Sciences*	*Life Sciences*		
6	3	2	2–3	2–3	74–83	6,804–6,881
7	6	2	2	5	83–89	3,822–3,890
8	3	6–7	2	2	93–100	5,061–5,104

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

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

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

Figure 8 through Figure 11 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 8. One-Factor Structural Model (Assertions-Overall): "Model 1"

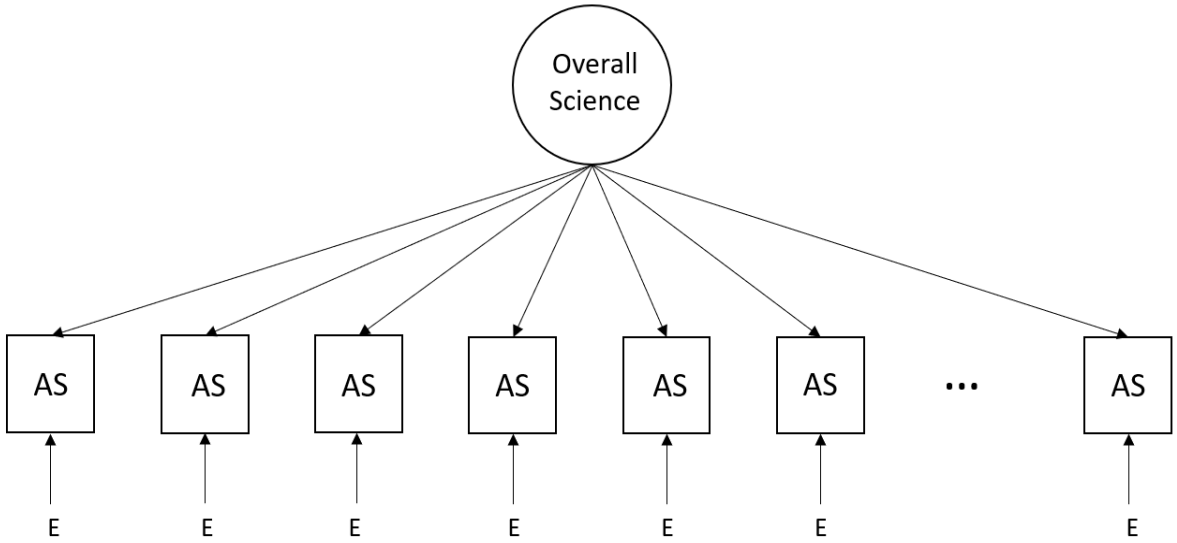


Figure 9. Second-Order Structural Model (Assertions-Disciplines-Overall): "Model 2"

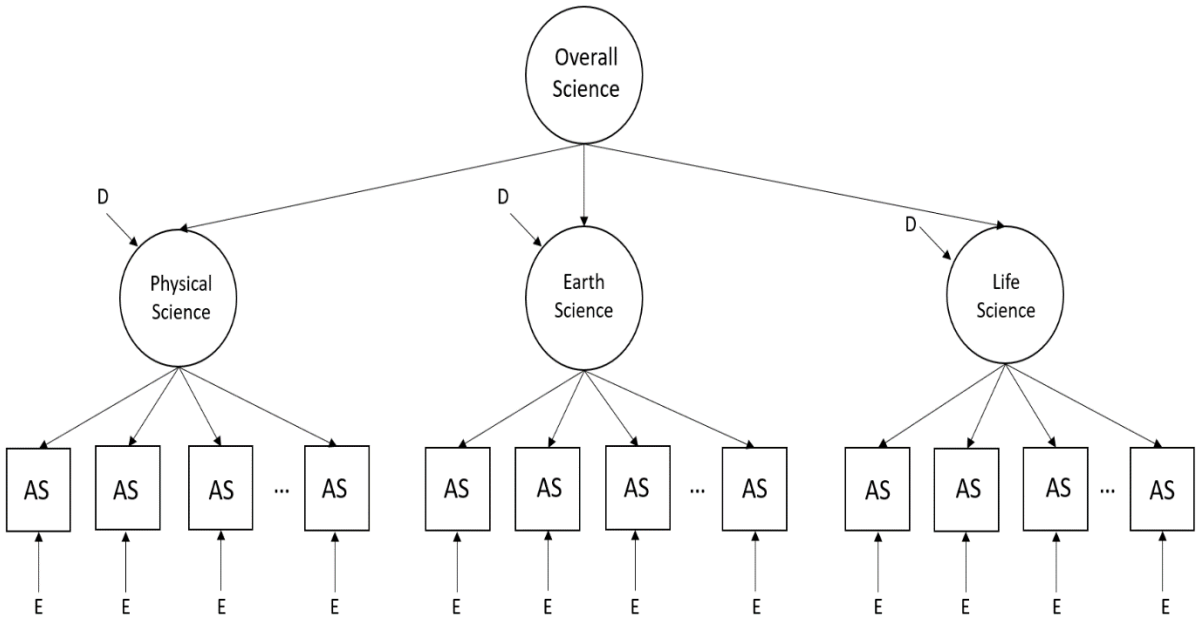


Figure 10. Second-Order Structural Model (Assertions-Clusters-Overall): "Model 3"

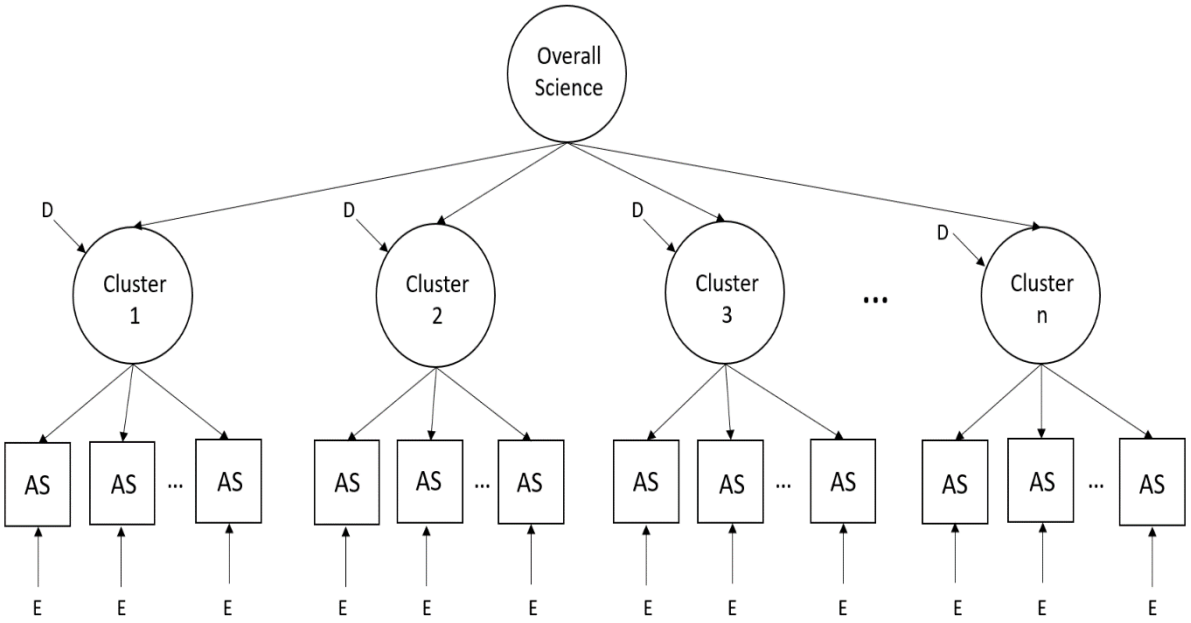
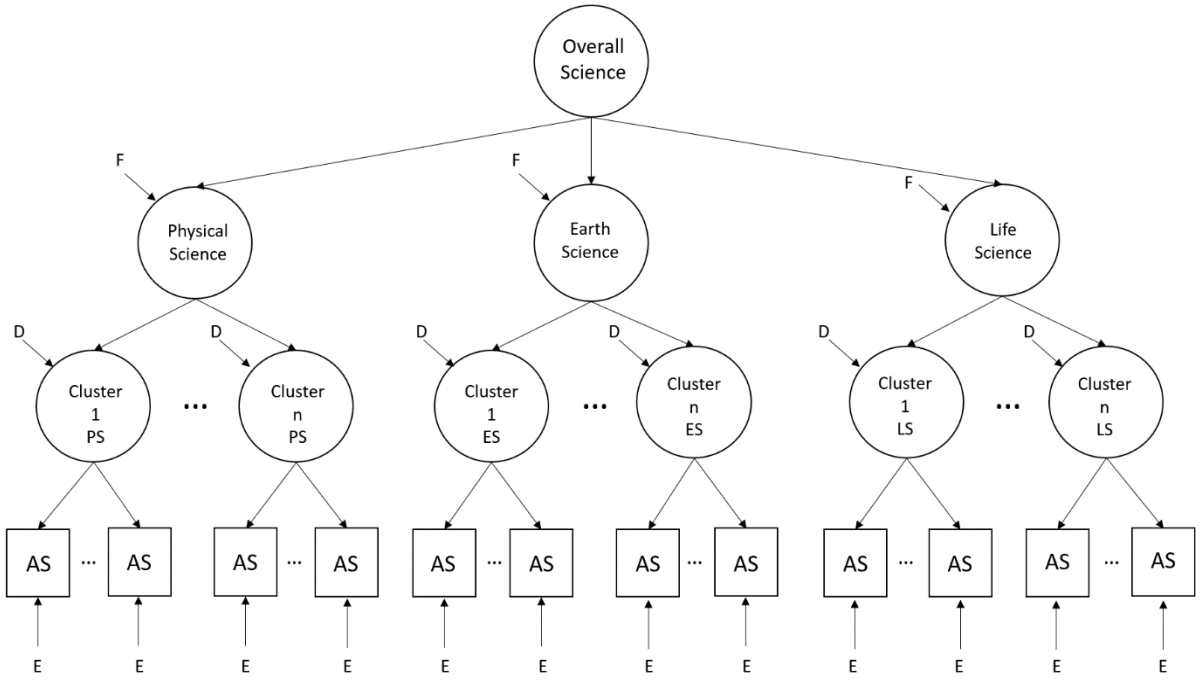


Figure 11. Third-Order Structural Model (Assertions-Clusters-Disciplines-Overall): "Model 4"



5.7.1. 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 CFI, TLI, the root mean square error of approximation (RMSEA), and the standardized root mean residual (SRMR). The CFI and TLI are relative fit indices, meaning they evaluate model fit by comparing the model of interest to a baseline model. The RMSEA and SRMR are indices of absolute fit. Table 50 provides a list of these measures along with the corresponding thresholds indicating a good fit.

*Table 50. 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, 1910; Hu & Bentler, 1999*

Table 51 through Table 53 show the goodness-of-fit statistics for grades 6 through 8, respectively.[1] Numbers in bold indicate those indices that did not meet the criteria established in Table 50. Across all grades and models, the following conclusions can be drawn:

- Model 1 shows the most misfit across grades and forms.
- Across forms, Model 3 generally shows more improvement in model fit relative to Model 1 than Model 2 does (i.e., higher values for the CFI and TLI and lower values for the RMSEA and SRMR). This means that accounting for the clusters resulted in a higher improvement in model fit over a single factor model than accounting for disciplines.
- Model 4 does not show improvement in model fit over Model 3. Fit measures remained the same (or had a difference of 0.001 or smaller in very few cases) across forms for Models 3 and 4. Hence, including the disciplines into the model (when clusters were considered) did not improve model fit.
- Overall model fit for Models 3 and 4 decreased with decreasing grades. For grade 8, all fit indices for Models 3 and 4 indicated good model fit for all three forms. For grade 7, all fit

[1] For very few assertions per form and models, some error variances for the assertions were slightly below 0. For grade 6, 1–2 assertions per form and model had error variance below 0, with the lowest error variance being -.027. For grade 7, Forms 1, 2, 5, and 6 had one negative error variance for one assertion in Models 3 and 4, with the lowest error variance being -0.099. Form 4 had 1–2 assertions with negative error variance in each model, and the lowest error variance was -0.102. For grade 8, there were no assertions with negative error variances for any of the forms and models.

indices for Models 3 and 4 indicated good fit for two out of the six forms, and the degree of misfit for the other four forms was small. For grade 6, all three forms had fit indices above the threshold values for at least one of the absolute fit indices for Models 3 and 4. The amount of misfit was small for the RMSEA but more substantial for the SRMR for two out of the three forms.

Table 51. Fit Measures per Model and Form, Grade 6

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

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

Table 52. Fit Measures per Model and Form, Grade 7

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

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

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

Table 53. Fit Measures per Model and Form, Grade 8

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

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

For Models 3 and 4, grade 6 showed some degree of misfit across all three forms according to the measures of absolute model fit, especially for the SRMR. Further examination indicated that the lack of fit could be attributed to a single item that was common to all three grade 6 forms that were part of this factor analysis study. After removing this item, there were only two forms that had two or more clusters per discipline. The fit for both forms improved drastically in Models 3 and 4, with all fit measures except the SRMR for one form meeting the criteria for model fit. The SRMR value that exceeded the threshold value did so barely, with a value of 0.083. Table 54 shows the fit measures for grade 6 after removing the item causing misfit. Note that, unlike Models 3 and 4, Models 1 and 2 still did not meet the criteria of model fit after removing the item.

Table 54. Fit Measures per Model and Form – Grade 6 – One Cluster Removed[2]

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

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

Table 55 shows the estimated correlations among disciplines for Model 4 (third-order model). The correlations were all very high, ranging between 0.913 and 1. The high correlations between the disciplines in Model 4 indicated that, after considering the cluster effects, the disciplines did not add much to the model. This may explain why Model 4 did not show an improvement in fit compared to Model 3. Overall, the findings supported the IRT model used for calibration.

Table 55. 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 variance 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

5.7.2. Conclusion

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

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

6. FAIRNESS IN CONTENT

The principles of universal design of assessments provide guidelines for test design to minimize the impact of construct-irrelevant factors in assessing student achievement. Universal design removes barriers to provide access for the widest range of students possible. The following seven principles of universal design are applied in the process of test development (Thompson, Johnstone, & Thurlow, 2002):

1. Inclusive assessment population
2. Precisely defined constructs
3. Accessible, non-biased items
4. Amenable to accommodations
5. Simple, clear, and intuitive instructions and procedures
6. Maximum readability and comprehensibility
7. Maximum legibility

Test development specialists have received extensive training on the principles of universal design and apply these principles in the development of all test materials. In the review process, West Virginia educators and stakeholders verified that the principles of universal design were adhered to.

6.1. STATISTICAL FAIRNESS IN ITEM STATISTICS

Due to the use of adaptive testing in spring 2022 English language arts (ELA) and mathematics, the number of West Virginia students who saw each item is relatively small. Differential item functioning (DIF) analysis for the WVGSA for ELA and mathematics was not available due to the small sample size for each demographic group. However, 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 4.5 of Volume 1, Annual Technical Report, along with the DIF analysis process for the WVGSA science.

6.2. COGNITIVE LABORATORY STUDIES FOR SCIENCE

In 2017, when the development of item clusters for the states that are part of the Memorandum of Understanding (MOU) started, cognitive lab studies were carried out to evaluate and refine the process of developing item clusters aligned to the Next Generation Science Standards (NGSS). The results of the cognitive lab studies confirmed the feasibility of the approach. Item clusters were completed within 12 minutes on average, and students reported being familiar with the format conventions and online tools used in the item clusters. They appeared to easily navigate the item clusters' interactive features and response formats. In general, students who received credit on a given item displayed a reasoning process that aligned with the skills that the item was intended to measure.

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

7. SUMMARY

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

- **Reliability.** Various measures of reliability are provided at the aggregate and subgroup levels, showing that the reliability of all tests is in line with acceptable industry standards.
- **Content Validity.** Evidence is provided to support the assertion that content coverage on each test was consistent with the test specifications of the blueprint across testing modes.
- **Internal Structural Validity.** Evidence is provided to support the selection of the measurement model, the tenability of model assumptions, and the reporting of an overall score and subscore at the reporting category levels.
- **Relationship of Test Scores to External Variables.** Evidence of convergent and discriminant validity is provided to support the relationship between the test and other measures intended to assess similar constructs, as well as between the test and other measures intended to assess different constructs.

8. REFERENCES

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing.*

Bejar, I. I. (1980). Biased assessment of program impact due to psychometric artifacts. *Psychological Bulletin*, *87*(3), 513–524. https://doi.org/10.1037/0033-2909.87.3.513

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*(2), 153–168. https://www.researchgate.net/publication/24063302_A_Bayesian_Random_Effects_Model_for_Testlets

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, *3*(3), 296–322. https://www.gwern.net/docs/statistics/1910-brown.pdf

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, *37*(1), 62–83. https://doi.org/10.1111/j.2044-8317.1984.tb00789.x

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen, & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Sage.

Chen, F., Bollen, K. A., Paxton, P., Curran, P. J., & Kirby, J. B. (2001). Improper solutions in structural equation models: Causes, consequences, and strategies. *Sociological Methods & Research*, *29*(4), 468–508. https://doi.org/10.1177/0049124101029004003

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265–289.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213–220. https://doi.org/10.1037/h0026256

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

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55. http://expsylab.psych.uoa.gr/fileadmin/expsylab.psych.uoa.gr/uploads/papers/Hu_Bentler_1999.pdf

Jöreskog, K. G. (1994). On the estimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika*, *59*(3), 381–389.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates.

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

Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*(1), 115–132. https://doi.org/10.1007/BF02294210

Muthén, B. O., du Toit, S. H. C., & Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. Conditionally accepted for publication in *Psychometrika*. https://www.statmodel.com/download/Article_075.pdf

Muthén, L. K., & Muthén, B. O. (2012). Mplus user's guide, 7th Edition.

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, *44*(4), 443–460. https://doi.org/10.1007/BF02296207

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

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

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). University of Minnesota, National Center on Educational Outcomes. https://nceo.info/Resources/publications/onlinepubs/Synthesis44.html

van Driel, O. P. (1978). On various causes of improper solutions in maximum likelihood factor analysis. *Psychometrika*, *43*(2), 225–243. https://doi.org/10.1007/BF02293865

Wang, W. C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement*, *29*(2), 126–149. https://doi.org/10.1177/0146621604271053

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for the evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, *31*(1), 2–13. https://doi.org/10.1111/j.1745-3992.2011.00223.x

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*(2), 125–145. https://doi.org/10.1177/014662168400800201

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*(3), 187–213. https://doi.org/10.1111/j.1745-3984.1993.tb00423.x

Appendix A

Student Demographics and Reliability Coefficients

Table 1: Number of Students by Demographic Subgroup, ELA

Grade	Overall	Female	Male	African American	American Indian/ Native Alaskan*	Asian	Hispanic	Multi-Racial	Pacific Islander	White	Non-LEP	LEP
3	17,526	8,538	8,988	685	6	102	370	800	10	15,252	17,341	185
4	17,323	8,454	8,869	630	8	98	370	794	8	15,129	17,192	131
5	17,683	8,611	9,072	649	7	113	372	804	8	15,439	17,552	131
6	17,697	8,678	9,019	683	13	100	410	723	9	15,282	17,597	100
7	18,242	8,979	9,263	783	28	117	390	696	3	15,946	18,120	122
8	18,698	9,012	9,686	717	20	120	440	729	9	16,371	18,565	133

Table 2: Number of Students by Demographic Subgroup, Mathematics

Grade	Overall	Female	Male	African American	American Indian/ Native Alaskan*	Asian	Hispanic	Multi-Racial	Pacific Islander	White	Non-LEP	LEP
3	17,542	8,548	8,994	686	6	102	367	802	10	15,267	17,361	181
4	17,329	8,459	8,870	631	8	98	359	797	8	15,140	17,208	121
5	17,717	8,632	9,085	651	7	113	370	806	8	15,475	17,588	129
6	17,708	8,682	9,026	708	13	100	402	741	9	15,433	17,617	91
7	18,281	8,994	9,287	788	29	118	377	698	3	15,989	18,173	108
8	18,718	9,019	9,699	718	19	120	419	732	9	16,407	18,609	109

Table 3: Number of Students by Demographic Subgroup, Science

Grade	Overall	Female	Male	African American	American Indian/ Native Alaskan	Asian	Hispanic	Multi-Racial	Pacific Islander	White	LEP
5	17689	8621	9077	647	7	112	371	800	8	15305	131
8	18694	9013	9681	711	20	120	438	723	9	16265	134

Table 4: Reliability Coefficients by Demographic Subgroup, ELA

Grade	Overall	Female	Male	African American	American Indian/ Native Alaskan*	Asian	Hispanic	Multi-Racial	Pacific Islander*	White	Non-LEP	LEP
3	0.89	0.89	0.89	0.84	0.96	0.92	0.88	0.88	-	0.89	0.89	0.82
4	0.89	0.89	0.89	0.86	0.82	0.9	0.88	0.88	-	0.89	0.89	0.81
5	0.9	0.9	0.9	0.87	0.85	0.91	0.89	0.9	-	0.9	0.9	0.8
6	0.89	0.89	0.89	0.85	0.92	0.91	0.89	0.89	-	0.89	0.89	0.78
7	0.9	0.9	0.9	0.88	0.9	0.92	0.9	0.89	-	0.9	0.9	0.77
8	0.91	0.91	0.91	0.89	0.91	0.93	0.91	0.91	-	0.91	0.91	0.8

**Note: Subgroup reliability is not reported due to small sample size (sample size <30).*

Table 5: Reliability Coefficients by Demographic Subgroup, Mathematics

Grade	Overall	Female	Male	African American	American Indian/ Native Alaskan*	Asian	Hispanic	Multi-Racial	Pacific Islander*	White	Non-LEP	LEP
3	0.91	0.91	0.92	0.88	0.95	0.93	0.9	0.9	-	0.91	0.91	0.87
4	0.92	0.91	0.92	0.9	0.89	0.94	0.9	0.91	-	0.92	0.92	0.89
5	0.9	0.89	0.9	0.86	0.79	0.94	0.89	0.88	-	0.9	0.9	0.82
6	0.88	0.88	0.88	0.82	0.86	0.94	0.87	0.87	-	0.88	0.88	0.81
7	0.88	0.88	0.88	0.8	0.88	0.95	0.87	0.86	-	0.88	0.88	0.76
8	0.89	0.89	0.89	0.82	0.84	0.94	0.88	0.88	-	0.89	0.89	0.8

**Note: Subgroup reliability is not reported due to small sample size (sample size <30).*

Table 6: Reliability Coefficients by Demographic Subgroup, Science

Grade	Overall	Female	Male	African American	American Indian/ Native Alaskan	Asian	Hispanic	Multi-Racial	Pacific Islander	White	Non-LEP	LEP
5	0.83	0.82	0.85	0.86	0.79	0.85	0.84	0.82	0.83	0.83	0.83	0.76
8	0.84	0.83	0.85	0.92	0.80	0.89	0.84	0.83	0.87	0.84	0.84	0.69

**Note: Subgroup reliability is not reported due to small sample size (sample size <30).*

Table 7: Scale Score Summary by Reporting Category, ELA Grade 3

Reporting Category	Mean	SD	Min	Max	Reliability	SEM
Reading Informational Text	557.64	69.19	420	750	0.69	34.68

Reporting Category	Mean	SD	Min	Max	Reliability	SEM
Reading Literary Text	561.88	62.69	420	750	0.75	27.38
Writing and Language	565.03	49.48	420	750	0.83	19.23

Table 8: Scale Score Summary by Reporting Category, ELA Grade 4

Reporting Category	Mean	SD	Min	Max	Reliability	SEM
Reading Informational Text	586.39	68.28	431	787	0.7	34.14
Reading Literary Text	585.38	60.71	431	787	0.75	27.04
Writing and Language	584.49	57.72	431	787	0.8	24.8

Table 9: Scale Score Summary by Reporting Category, ELA Grade 5

Reporting Category	Mean	SD	Min	Max	Reliability	SEM
Reading Informational Text	607.99	66.55	450	806	0.71	33.13
Reading Literary Text	607.84	59.16	450	806	0.75	26.11
Writing and Language	603.68	54.86	450	806	0.82	22.36

Table 10: Scale Score Summary by Reporting Category, ELA Grade 6

Reporting Category	Mean	SD	Min	Max	Reliability	SEM
Reading Informational Text	618.31	64.33	464	824	0.71	30.88
Reading Literary Text	612.52	69.80	464	824	0.7	35.02
Writing and Language	629.11	57.91	464	824	0.82	23.7

Table 11: Scale Score Summary by Reporting Category, ELA Grade 7

Reporting Category	Mean	SD	Min	Max	Reliability	SEM
Reading Informational Text	624.43	63.23	470	839	0.74	29.77
Reading Literary Text	616.96	68.7	470	839	0.69	35.3
Writing and Language	631.45	59.71	470	839	0.82	24.69

Table 12: Scale Score Summary by Reporting Category, ELA Grade 8

Reporting Category	Mean	SD	Min	Max	Reliability	SEM
Reading Informational Text	632.60	69.94	480	862	0.73	33.84
Reading Literary Text	632.98	77.32	480	862	0.73	39.01
Writing and Language	638.67	58.96	480	862	0.84	22.72

Table 13: Scale Score Summary by Reporting Category, Mathematics Grade 3

Reporting Category	Mean	SD	Min	Max	Reliability	SEM
Measurement and Data and Geometry	415.55	43.01	308	548	0.73	20.77
Numbers and Operations—Base Ten and Fractions	421.3	37.9	308	548	0.83	14.31
Operations and Algebraic Thinking	413.74	49.45	308	548	0.81	19.17

Table 14: Scale Score Summary by Reporting Category, Mathematics Grade 4

Reporting Category	Mean	SD	Min	Max	Reliability	SEM
Measurement and Data and Geometry	438.27	55.4	312	608	0.74	26.39
Numbers and Operations—Base Ten and Fractions	443.7	46.82	312	608	0.85	15.87
Operations and Algebraic Thinking	442.46	54.91	312	608	0.77	23.94

Table 15: Scale Score Summary by Reporting Category, Mathematics Grade 5

Reporting Category	Mean	SD	Min	Max	Reliability	SEM
Measurement and Data and Geometry	460.34	63.51	320	659	0.73	30.2
Numbers and Operations—Base Ten and Fractions	462.76	56.22	320	659	0.82	21.26
Operations and Algebraic Thinking	463.5	66.47	320	659	0.75	31.07

Table 16: Scale Score Summary by Reporting Category, Mathematics Grade 6

Reporting Category	Mean	SD	Min	Max	Reliability	SEM
Expressions and Equations	475.95	70.15	330	720	0.75	31.18
Geometry and Statistics and Probability	475.05	74.13	330	720	0.61	44.39
Ratios and Proportional Relationships and Number System	478.77	62.18	330	720	0.79	25.24

Table 17: Scale Score Summary by Reporting Category, Mathematics Grade 7

Reporting Category	Mean	SD	Min	Max	Reliability	SEM
Expressions and Equations	498.81	83.67	348	750	0.69	42.05
Geometry	501.12	76.51	348	750	0.7	38.48
Ratios and Proportional Relationships and Number System	511.73	77.22	348	750	0.79	31.6
Statistics and Probability	504.96	82.52	348	750	0.69	42.58

Table 18: Scale Score Summary by Reporting Category, Mathematics Grade 8

Reporting Category	Mean	SD	Min	Max	Reliability	SEM
Expressions and Equations and Number System	535.98	88.67	350	830	0.78	37.82
Functions	532.33	92.22	350	830	0.65	52.77
Geometry and Statistics and Probability	530.27	86.67	350	830	0.77	37.16

Table 19: Scale Score Summary by Reporting Category, Science Grade 5

Reporting Category	Mean	SD	Min	Max	Reliability	SEM
Physical Science	545.75	18.27	500	600	0.69	10.03
Earth and Space Science	545.70	19.38	500	600	0.69	10.63
Life Science	545.34	19.02	500	600	0.70	10.32

Table 20: Scale Score Summary by Reporting Category, Science Grade 8

Reporting Category	Mean	SD	Min	Max	Reliability	SEM
Physical Science	843.94	17.27	800	900	0.68	9.67
Earth and Space Science	844.07	18.02	800	900	0.69	9.95
Life Science	843.42	18.82	800	900	0.69	10.29

Appendix B

Conditional Standard Error of Measurement

Table 1: CSEM at Each Scale Score, ELA Grade 3

ELA Grade 3		
Scale Score	**Achievement Level**	**CSEM**
420	1	62.11
424	1	59.95
425	1	53.49
427	1	55.77
429	1	57.63
430	1	51.17
432	1	50.50
434	1	49.58
435	1	54.84
436	1	46.83
437	1	52.88
438	1	50.68
439	1	48.32
440	1	49.72
442	1	47.04
443	1	49.82
444	1	44.34
445	1	47.95
446	1	48.99
447	1	45.42
448	1	46.42
449	1	42.85
450	1	43.45
451	1	43.09
452	1	45.55
453	1	43.43
454	1	44.71
455	1	43.26
456	1	41.49
457	1	41.26
458	1	41.47
459	1	37.30
460	1	38.71
461	1	38.69
462	1	38.62
463	1	36.93
464	1	36.84
465	1	37.86
466	1	36.30
467	1	35.29
468	1	35.01
469	1	34.39

ELA Grade 3

Scale Score	Achievement Level	CSEM
470	1	33.18
471	1	33.87
472	1	33.47
473	1	32.46
474	1	32.27
475	1	30.72
476	1	31.26
477	1	30.04
478	1	30.33
479	1	29.31
480	1	29.51
481	1	29.08
482	1	28.62
483	1	27.92
484	1	27.76
485	1	27.07
486	1	26.35
487	1	26.08
488	1	25.91
489	1	25.17
490	1	25.30
491	1	24.55
492	1	24.17
493	1	24.11
494	1	23.51
495	1	23.13
496	1	22.96
497	1	22.74
498	1	22.15
499	1	21.99
500	1	21.41
501	1	21.07
502	1	21.04
503	1	20.44
504	1	20.17
505	1	19.92
506	1	19.55
507	1	19.36
508	1	19.34
509	1	18.65
510	1	18.59
511	1	18.24
512	1	17.89
513	1	17.84

ELA Grade 3

Scale Score	Achievement Level	CSEM
514	1	17.42
515	1	17.19
516	1	17.06
517	1	16.77
518	1	16.45
519	1	16.28
520	1	15.99
521	1	15.94
522	1	15.64
523	1	15.40
524	1	15.37
525	1	14.96
526	1	14.89
527	1	14.73
528	1	14.50
529	1	14.44
530	1	14.37
531	1	14.03
532	1	13.84
533	1	13.62
534	1	13.58
535	1	13.56
536	1	13.38
537	1	13.31
538	1	13.13
539	1	12.81
540	1	12.88
541	1	12.82
542	1	12.64
543	1	12.58
544	1	12.53
545	1	12.45
546	1	12.35
547	1	12.14
548	1	12.08
549	1	12.09
550	2	11.97
551	2	11.84
552	2	11.79
553	2	11.80
554	2	11.59
555	2	11.56
556	2	11.53
557	2	11.52

ELA Grade 3

Scale Score	Achievement Level	CSEM
558	2	11.50
559	2	11.38
560	2	11.36
561	2	11.30
562	2	11.39
563	2	11.30
564	2	11.22
565	2	11.23
566	2	11.17
567	2	11.13
568	2	11.09
569	2	11.10
570	2	11.16
571	2	11.12
572	2	11.02
573	2	11.06
574	2	10.98
575	2	10.97
576	2	10.94
577	2	10.96
578	2	11.00
579	2	10.90
580	2	10.84
581	2	10.89
582	2	10.89
583	2	10.87
584	2	10.89
585	2	10.93
586	3	10.85
587	3	10.81
588	3	10.80
589	3	10.83
590	3	10.80
591	3	10.78
592	3	10.80
593	3	10.79
594	3	10.66
595	3	10.74
596	3	10.68
597	3	10.70
598	3	10.63
599	3	10.66
600	3	10.62
601	3	10.59

ELA Grade 3

Scale Score	Achievement Level	CSEM
602	3	10.57
603	3	10.61
604	3	10.51
605	3	10.55
606	3	10.53
607	3	10.46
608	3	10.51
609	3	10.45
610	3	10.47
611	3	10.50
612	3	10.49
613	3	10.55
614	3	10.51
615	3	10.52
616	4	10.50
617	4	10.51
618	4	10.51
619	4	10.53
620	4	10.51
621	4	10.53
622	4	10.55
623	4	10.55
624	4	10.59
625	4	10.57
626	4	10.60
627	4	10.78
628	4	10.66
629	4	10.71
630	4	10.73
631	4	10.81
632	4	10.77
633	4	10.88
634	4	10.81
635	4	10.91
636	4	10.93
637	4	10.96
638	4	11.08
639	4	11.07
640	4	11.11
641	4	11.11
642	4	11.14
643	4	11.37
644	4	11.07
645	4	11.33

ELA Grade 3

Scale Score	Achievement Level	CSEM
646	4	11.37
647	4	11.51
648	4	11.45
649	4	11.60
650	4	11.54
651	4	11.71
652	4	11.77
653	4	11.90
654	4	11.86
655	4	12.01
656	4	11.99
657	4	11.62
658	4	12.04
659	4	12.16
660	4	12.14
661	4	12.45
662	4	12.34
663	4	12.41
664	4	12.39
665	4	12.80
666	4	12.70
667	4	12.95
668	4	13.35
669	4	13.21
670	4	12.46
671	4	13.67
672	4	13.18
673	4	13.21
674	4	13.35
675	4	13.19
676	4	13.48
677	4	13.55
678	4	14.10
679	4	13.47
680	4	12.83
681	4	13.99
683	4	14.03
684	4	15.19
685	4	14.68
686	4	15.06
692	4	17.26
693	4	16.89
694	4	16.11
696	4	17.06

ELA Grade 3		
Scale Score	**Achievement Level**	**CSEM**
702	4	11.02
710	4	19.15

Table 2: CSEM at Each Scale Score, ELA Grade 4

ELA Grade 4		
Scale Score	**Achievement Level**	**CSEM**
430	1	60.76
434	1	60.48
435	1	56.48
436	1	58.28
437	1	54.67
438	1	63.97
439	1	58.53
440	1	55.69
442	1	52.16
443	1	50.69
445	1	50.78
446	1	51.76
447	1	52.30
448	1	49.23
449	1	49.43
450	1	52.28
451	1	47.26
452	1	48.60
453	1	51.19
454	1	45.92
455	1	45.38
456	1	45.26
457	1	46.24
458	1	46.2
459	1	40.78
460	1	41.53
461	1	41.74
462	1	44.00
463	1	43.73
464	1	40.84
465	1	41.37
466	1	37.91
467	1	40.21
468	1	37.12

ELA Grade 4

Scale Score	Achievement Level	CSEM
469	1	38.60
470	1	39.52
471	1	36.21
472	1	37.89
473	1	35.24
474	1	36.35
475	1	39.25
476	1	35.3
477	1	34.23
478	1	35.30
479	1	35.28
480	1	33.00
481	1	34.52
482	1	33.57
483	1	33.03
484	1	31.86
485	1	32.46
486	1	30.84
487	1	31.20
488	1	29.85
489	1	29.96
490	1	30.02
491	1	30.88
492	1	28.83
493	1	28.32
494	1	29.35
495	1	28.26
496	1	28.27
497	1	28.21
498	1	27.42
499	1	27.39
500	1	28.20
501	1	26.04
502	1	26.01
503	1	25.44
504	1	25.56
505	1	25.25
506	1	25.14
507	1	24.44
508	1	25.28
509	1	23.74
510	1	23.75
511	1	23.81
512	1	22.89

ELA Grade 4

Scale Score	Achievement Level	CSEM
513	1	23.2
514	1	22.61
515	1	22.40
516	1	22.11
517	1	21.90
518	1	21.53
519	1	21.00
520	1	20.77
521	1	21.14
522	1	20.52
523	1	19.99
524	1	19.84
525	1	19.46
526	1	19.44
527	1	19.10
528	1	18.65
529	1	18.43
530	1	18.13
531	1	18.09
532	1	18.01
533	1	17.82
534	1	17.48
535	1	17.15
536	1	17.64
537	1	17.03
538	1	16.89
539	1	16.65
540	1	16.24
541	1	16.35
542	1	16.25
543	1	16.05
544	1	15.88
545	1	16.11
546	1	15.53
547	1	15.67
548	1	15.45
549	1	15.16
550	1	15.13
551	1	15.44
552	1	15.30
553	1	15.03
554	1	14.90
555	1	14.69
556	1	14.78

ELA Grade 4

Scale Score	Achievement Level	CSEM
557	1	14.49
558	1	14.47
559	1	14.51
560	1	14.36
561	1	14.44
562	1	14.28
563	2	14.42
564	2	14.13
565	2	14.01
566	2	13.78
567	2	13.85
568	2	13.94
569	2	13.74
570	2	13.67
571	2	13.60
572	2	13.43
573	2	13.53
574	2	13.44
575	2	13.34
576	2	13.31
577	2	13.23
578	2	13.23
579	2	13.17
580	2	13.09
581	2	13.06
582	2	13.07
583	2	12.85
584	2	12.86
585	2	12.82
586	2	12.76
587	2	12.62
588	2	12.52
589	2	12.53
590	2	12.52
591	2	12.43
592	2	12.42
593	2	12.30
594	2	12.40
595	2	12.38
596	2	12.31
597	2	12.28
598	2	12.13
599	3	12.18
600	3	12.23

ELA Grade 4

Scale Score	Achievement Level	CSEM
601	3	12.15
602	3	12.14
603	3	12.02
604	3	12.14
605	3	12.00
606	3	12.11
607	3	12.04
608	3	12.00
609	3	12.06
610	3	11.86
611	3	11.92
612	3	12.05
613	3	12.04
614	3	12.05
615	3	11.92
616	3	11.94
617	3	11.85
618	3	11.96
619	3	12.03
620	3	12.06
621	3	12.14
622	3	12.15
623	3	12.04
624	3	12.08
625	3	12.12
626	3	12.15
627	3	12.13
628	3	12.22
629	4	12.32
630	4	12.26
631	4	12.24
632	4	12.35
633	4	12.32
634	4	12.47
635	4	12.36
636	4	12.46
637	4	12.52
638	4	12.52
639	4	12.65
640	4	12.70
641	4	12.80
642	4	12.78
643	4	12.82
644	4	12.95

ELA Grade 4

Scale Score	Achievement Level	CSEM
645	4	12.76
646	4	13.07
647	4	13.01
648	4	13.24
649	4	13.21
650	4	13.22
651	4	13.38
652	4	13.43
653	4	13.40
654	4	13.63
655	4	13.64
656	4	13.66
657	4	13.76
658	4	13.78
659	4	13.72
660	4	13.85
661	4	14.05
662	4	14.08
663	4	14.07
664	4	14.32
665	4	14.18
666	4	14.32
667	4	14.50
668	4	14.66
669	4	14.55
670	4	14.60
671	4	14.75
672	4	15.05
673	4	14.94
674	4	15.16
675	4	15.06
676	4	15.39
677	4	15.23
678	4	15.32
679	4	15.11
680	4	15.81
681	4	15.77
682	4	15.71
683	4	15.85
684	4	15.80
685	4	16.11
686	4	16.00
687	4	16.48
688	4	16.20

ELA Grade 4

Scale Score	Achievement Level	CSEM
689	4	16.74
690	4	15.97
691	4	16.15
692	4	16.20
693	4	16.51
694	4	16.49
695	4	15.90
696	4	17.06
697	4	16.22
698	4	16.73
699	4	16.98
700	4	16.18
701	4	16.36
702	4	16.55
703	4	17.18
704	4	16.96
705	4	17.05
706	4	17.59
707	4	17.04
708	4	16.59
709	4	16.37
710	4	17.47
711	4	17.21
712	4	17.32
713	4	17.21
714	4	17.32
715	4	17.98
716	4	17.99
717	4	17.24
718	4	17.48
719	4	17.81
720	4	17.95
722	4	18.78
724	4	18.30
725	4	16.98
726	4	17.43
730	4	18.86
734	4	19.11
736	4	18.58
746	4	19.95
747	4	19.38
762	4	22.59
786	4	34.30

Table 3: CSEM at Each Scale Score, ELA Grade 5

ELA Grade 5		
Scale Score	**Achievement Level**	**CSEM**
450	1	55.25
451	1	52.58
452	1	49.92
453	1	52.71
455	1	57.59
458	1	53.04
462	1	43.97
463	1	51.85
464	1	40.26
466	1	41.52
467	1	44.93
468	1	41.15
470	1	42.56
471	1	41.78
472	1	42.15
473	1	39.44
474	1	44.81
475	1	37.32
476	1	37.15
477	1	40.76
478	1	39.91
479	1	41.07
480	1	38.72
481	1	42.39
482	1	38.32
483	1	39.28
484	1	37.13
485	1	37.11
486	1	34.18
487	1	37.37
488	1	36.00
489	1	37.64
490	1	35.92
491	1	34.02
492	1	35.75
493	1	33.81
494	1	33.67
495	1	33.64
496	1	31.96
497	1	33.67
498	1	32.46

ELA Grade 5

Scale Score	Achievement Level	CSEM
499	1	30.67
500	1	31.66
501	1	31.65
502	1	31.14
503	1	30.15
504	1	30.26
505	1	29.43
506	1	29.25
507	1	28.60
508	1	28.60
509	1	27.94
510	1	28.21
511	1	27.89
512	1	26.70
513	1	27.03
514	1	25.61
515	1	26.01
516	1	25.87
517	1	25.42
518	1	24.27
519	1	24.90
520	1	24.35
521	1	24.34
522	1	24.52
523	1	23.68
524	1	23.61
525	1	23.12
526	1	22.71
527	1	22.49
528	1	22.31
529	1	22.33
530	1	21.86
531	1	21.47
532	1	21.18
533	1	21.18
534	1	21.37
535	1	20.64
536	1	20.43
537	1	20.23
538	1	20.13
539	1	19.90
540	1	19.57
541	1	19.70
542	1	19.14

ELA Grade 5

Scale Score	Achievement Level	CSEM
543	1	18.91
544	1	18.60
545	1	18.61
546	1	18.24
547	1	18.09
548	1	18.12
549	1	17.96
550	1	17.75
551	1	17.54
552	1	17.35
553	1	17.16
554	1	16.83
555	1	16.76
556	1	16.69
557	1	16.30
558	1	16.36
559	1	16.02
560	1	15.83
561	1	15.91
562	1	15.68
563	1	15.57
564	1	15.57
565	1	15.21
566	1	15.30
567	1	14.92
568	1	14.94
569	1	14.74
570	1	14.69
571	1	14.44
572	1	14.46
573	1	14.16
574	1	14.23
575	1	14.22
576	1	14.22
577	1	13.88
578	1	13.75
579	1	13.80
580	1	13.55
581	1	13.52
582	1	13.54
583	1	13.32
584	1	13.40
585	1	13.15
586	1	13.01

ELA Grade 5

Scale Score	Achievement Level	CSEM
587	1	13.10
588	2	12.94
589	2	13.11
590	2	12.84
591	2	12.83
592	2	12.72
593	2	12.67
594	2	12.54
595	2	12.51
596	2	12.60
597	2	12.50
598	2	12.33
599	2	12.31
600	2	12.27
601	2	12.15
602	2	12.16
603	2	12.22
604	2	12.14
605	2	12.02
606	2	11.98
607	2	11.92
608	2	11.86
609	2	11.84
610	2	11.78
611	2	11.73
612	2	11.71
613	2	11.65
614	2	11.59
615	2	11.52
616	2	11.57
617	2	11.47
618	2	11.43
619	2	11.37
620	2	11.30
621	2	11.29
622	3	11.25
623	3	11.12
624	3	11.13
625	3	11.12
626	3	11.15
627	3	11.10
628	3	11.10
629	3	10.99
630	3	11.12

ELA Grade 5

Scale Score	Achievement Level	CSEM
631	3	11.05
632	3	10.95
633	3	11.08
634	3	10.90
635	3	11.15
636	3	10.99
637	3	11.07
638	3	11.10
639	3	11.03
640	3	11.03
641	3	11.12
642	3	11.06
643	3	11.01
644	3	10.89
645	3	10.99
646	3	11.21
647	3	11.25
648	3	11.30
649	3	11.25
650	3	11.14
651	3	11.24
652	3	11.37
653	3	11.45
654	3	11.38
655	4	11.37
656	4	11.37
657	4	11.35
658	4	11.65
659	4	11.52
660	4	11.62
661	4	11.56
662	4	11.66
663	4	11.81
664	4	11.70
665	4	11.90
666	4	11.72
667	4	11.72
668	4	12.00
669	4	11.92
670	4	12.05
671	4	12.21
672	4	12.07
673	4	12.34
674	4	12.36

ELA Grade 5

Scale Score	Achievement Level	CSEM
675	4	12.21
676	4	12.29
677	4	12.51
678	4	12.55
679	4	12.52
680	4	12.73
681	4	12.74
682	4	12.57
683	4	12.74
684	4	13.16
685	4	12.96
686	4	13.16
687	4	13.30
688	4	13.15
689	4	13.13
690	4	13.54
691	4	13.28
692	4	13.31
693	4	13.43
694	4	13.55
695	4	13.67
696	4	13.67
697	4	13.76
698	4	13.93
699	4	13.95
700	4	14.10
701	4	13.96
702	4	14.32
703	4	14.14
704	4	14.21
705	4	14.33
706	4	14.69
707	4	14.72
708	4	14.83
709	4	14.87
710	4	14.95
711	4	15.05
712	4	15.16
713	4	14.96
714	4	15.00
715	4	15.40
716	4	15.32
717	4	15.23
718	4	15.72

ELA Grade 5		
Scale Score	**Achievement Level**	**CSEM**
719	4	15.42
720	4	15.96
721	4	16.12
722	4	15.32
723	4	15.34
724	4	16.21
725	4	15.37
726	4	16.28
727	4	15.82
728	4	16.01
729	4	16.44
730	4	16.58
731	4	16.95
732	4	16.76
733	4	16.09
735	4	16.69
736	4	16.73
737	4	17.19
739	4	17.98
742	4	17.58
746	4	18.07
750	4	18.20

Table 4: CSEM at Each Scale Score, ELA Grade 6

ELA Grade 6		
Scale Score	**Achievement Level**	**CSEM**
464	1	53.52
465	1	50.74
468	1	51.67
469	1	50.35
470	1	53.69
471	1	51.21
473	1	47.33
474	1	47.12
475	1	47.01
476	1	50.00
477	1	51.68
478	1	48.97
479	1	46.64
480	1	45.51

ELA Grade 6

Scale Score	Achievement Level	CSEM
481	1	46.41
482	1	48.72
483	1	49.37
484	1	43.38
485	1	45.56
486	1	49.79
487	1	43.35
488	1	41.75
489	1	41.97
490	1	42.94
491	1	42.66
492	1	40.49
493	1	41.48
494	1	40.86
495	1	40.66
496	1	40.68
497	1	39.95
498	1	39.70
499	1	37.68
500	1	37.68
501	1	37.60
502	1	37.15
503	1	38.61
504	1	36.23
505	1	35.98
506	1	37.43
507	1	35.87
508	1	35.17
509	1	35.13
510	1	35.22
511	1	33.51
512	1	34.59
513	1	32.85
514	1	32.17
515	1	32.73
516	1	32.27
517	1	32.21
518	1	31.80
519	1	31.93
520	1	31.49
521	1	30.86
522	1	30.87
523	1	29.37
524	1	29.99

ELA Grade 6

Scale Score	Achievement Level	CSEM
525	1	29.27
526	1	29.20
527	1	29.13
528	1	28.63
529	1	27.80
530	1	28.18
531	1	28.02
532	1	27.06
533	1	26.79
534	1	27.00
535	1	26.52
536	1	25.90
537	1	25.61
538	1	25.97
539	1	25.40
540	1	24.97
541	1	25.41
542	1	24.99
543	1	24.66
544	1	24.42
545	1	23.82
546	1	23.81
547	1	23.40
548	1	23.35
549	1	22.64
550	1	23.35
551	1	22.59
552	1	22.51
553	1	22.30
554	1	22.24
555	1	22.00
556	1	21.56
557	1	21.35
558	1	20.96
559	1	20.66
560	1	20.69
561	1	20.49
562	1	20.33
563	1	20.17
564	1	20.23
565	1	19.91
566	1	20.07
567	1	19.63
568	1	19.40

ELA Grade 6

Scale Score	Achievement Level	CSEM
569	1	19.50
570	1	18.95
571	1	18.93
572	1	18.54
573	1	18.67
574	1	18.51
575	1	18.14
576	1	18.18
577	1	18.06
578	1	17.82
579	1	17.69
580	1	17.54
581	1	17.41
582	1	17.50
583	1	17.10
584	1	16.87
585	1	16.72
586	1	16.63
587	1	16.35
588	1	16.46
589	1	16.08
590	1	16.21
591	1	16.05
592	1	15.78
593	1	15.47
594	1	15.67
595	1	15.63
596	1	15.26
597	2	15.15
598	2	14.80
599	2	14.88
600	2	14.92
601	2	14.65
602	2	14.46
603	2	14.35
604	2	14.13
605	2	14.15
606	2	14.11
607	2	13.92
608	2	14.03
609	2	13.80
610	2	13.70
611	2	13.68
612	2	13.61

ELA Grade 6

Scale Score	Achievement Level	CSEM
613	2	13.29
614	2	13.28
615	2	13.08
616	2	13.14
617	2	12.93
618	2	12.88
619	2	12.80
620	2	12.88
621	2	12.86
622	2	12.77
623	2	12.51
624	2	12.54
625	2	12.52
626	2	12.57
627	2	12.47
628	2	12.44
629	2	12.39
630	2	12.29
631	2	12.48
632	2	12.26
633	2	12.34
634	2	12.28
635	2	12.22
636	2	12.16
637	2	12.27
638	2	12.26
639	3	12.20
640	3	12.19
641	3	12.29
642	3	12.19
643	3	12.21
644	3	12.19
645	3	12.22
646	3	12.24
647	3	12.21
648	3	12.25
649	3	12.23
650	3	12.29
651	3	12.33
652	3	12.25
653	3	12.24
654	3	12.29
655	3	12.33
656	3	12.29

ELA Grade 6

Scale Score	Achievement Level	CSEM
657	3	12.29
658	3	12.37
659	3	12.40
660	3	12.31
661	3	12.53
662	3	12.42
663	3	12.41
664	3	12.51
665	3	12.47
666	3	12.55
667	3	12.47
668	3	12.52
669	3	12.59
670	3	12.52
671	3	12.52
672	3	12.53
673	3	12.52
674	3	12.59
675	3	12.51
676	3	12.59
677	3	12.58
678	3	12.52
679	3	12.54
680	4	12.72
681	4	12.50
682	4	12.69
683	4	12.67
684	4	12.57
685	4	12.81
686	4	12.61
687	4	12.66
688	4	12.93
689	4	12.84
690	4	12.96
691	4	12.90
692	4	12.95
693	4	12.85
694	4	13.13
695	4	13.26
696	4	13.25
697	4	13.30
698	4	13.22
699	4	13.07
700	4	13.11

ELA Grade 6

Scale Score	Achievement Level	CSEM
701	4	13.50
702	4	13.45
703	4	13.56
704	4	13.58
705	4	13.80
706	4	13.80
707	4	14.06
708	4	13.82
709	4	13.80
710	4	13.86
711	4	13.91
712	4	14.25
713	4	14.24
714	4	14.64
715	4	14.33
716	4	14.45
717	4	14.38
718	4	14.46
719	4	14.59
720	4	14.76
721	4	14.72
722	4	14.89
723	4	14.84
724	4	14.84
725	4	14.91
726	4	15.07
727	4	15.28
728	4	15.25
729	4	15.53
730	4	15.12
731	4	15.87
732	4	15.56
733	4	15.85
734	4	15.64
735	4	15.20
736	4	15.53
737	4	15.84
738	4	16.05
739	4	16.33
740	4	15.88
741	4	15.71
742	4	15.97
743	4	16.30
744	4	16.22

ELA Grade 6		
Scale Score	**Achievement Level**	**CSEM**
745	4	15.97
746	4	16.20
747	4	16.21
748	4	17.36
749	4	17.32
750	4	16.54
751	4	16.92
752	4	17.15
754	4	16.02
756	4	17.19
757	4	16.37
758	4	17.47
760	4	17.49
761	4	17.10
762	4	17.67
764	4	17.62
767	4	18.49
768	4	19.64
770	4	18.65
773	4	18.87
776	4	17.87
789	4	21.54
806	4	25.67

Table 5: CSEM at Each Scale Score, ELA Grade 7

ELA Grade 7		
Scale Score	**Achievement Level**	**CSEM**
470	1	53.71
472	1	49.65
473	1	45.83
474	1	46.36
476	1	46.25
477	1	42.31
478	1	47.25
479	1	45.48
480	1	43.56
481	1	44.80
482	1	43.19
483	1	41.22
484	1	41.95

ELA Grade 7

Scale Score	Achievement Level	CSEM
485	1	44.44
487	1	42.33
488	1	39.57
489	1	39.84
490	1	39.48
491	1	39.90
492	1	36.24
493	1	37.16
494	1	39.00
495	1	36.70
496	1	36.68
497	1	37.02
498	1	34.69
499	1	35.98
500	1	34.24
501	1	35.00
502	1	33.83
503	1	35.04
504	1	34.06
505	1	32.30
506	1	32.55
507	1	32.44
508	1	32.49
509	1	32.34
510	1	31.70
511	1	32.43
512	1	30.13
513	1	31.03
514	1	30.75
515	1	30.54
516	1	33.06
517	1	29.45
518	1	28.79
519	1	28.85
520	1	28.19
521	1	28.32
522	1	27.70
523	1	27.74
524	1	27.13
525	1	26.59
526	1	27.05
527	1	26.22
528	1	26.26
529	1	25.28

ELA Grade 7

Scale Score	Achievement Level	CSEM
530	1	25.94
531	1	25.26
532	1	24.91
533	1	24.59
534	1	24.76
535	1	24.17
536	1	23.82
537	1	23.54
538	1	23.45
539	1	23.53
540	1	23.15
541	1	23.59
542	1	22.87
543	1	22.69
544	1	22.14
545	1	22.12
546	1	22.12
547	1	22.00
548	1	21.72
549	1	21.44
550	1	21.77
551	1	21.25
552	1	20.97
553	1	21.03
554	1	20.79
555	1	20.60
556	1	20.57
557	1	20.38
558	1	20.35
559	1	20.15
560	1	19.94
561	1	19.69
562	1	19.76
563	1	19.51
564	1	19.44
565	1	19.18
566	1	19.11
567	1	18.94
568	1	18.67
569	1	18.88
570	1	18.41
571	1	18.19
572	1	18.24
573	1	17.96

ELA Grade 7

Scale Score	Achievement Level	CSEM
574	1	17.84
575	1	17.66
576	1	17.51
577	1	17.50
578	1	17.24
579	1	17.11
580	1	17.05
581	1	16.80
582	1	16.69
583	1	16.68
584	1	16.47
585	1	16.29
586	1	16.35
587	1	16.00
588	1	15.95
589	1	16.02
590	1	15.64
591	1	15.71
592	1	15.52
593	1	15.35
594	1	15.19
595	1	15.02
596	1	15.02
597	1	15.06
598	1	14.90
599	1	14.63
600	1	14.58
601	1	14.31
602	2	14.46
603	2	14.27
604	2	14.20
605	2	14.15
606	2	14.16
607	2	14.00
608	2	13.94
609	2	13.81
610	2	13.80
611	2	13.71
612	2	13.70
613	2	13.52
614	2	13.48
615	2	13.42
616	2	13.44
617	2	13.20

ELA Grade 7

Scale Score	Achievement Level	CSEM
618	2	13.29
619	2	13.18
620	2	13.24
621	2	13.09
622	2	13.03
623	2	13.05
624	2	13.00
625	2	12.97
626	2	12.98
627	2	12.83
628	2	12.83
629	2	12.88
630	2	12.77
631	2	12.79
632	2	12.73
633	2	12.67
634	2	12.73
635	2	12.73
636	2	12.75
637	2	12.65
638	2	12.62
639	2	12.68
640	2	12.61
641	2	12.57
642	2	12.54
643	2	12.60
644	3	12.63
645	3	12.64
646	3	12.56
647	3	12.60
648	3	12.65
649	3	12.60
650	3	12.65
651	3	12.64
652	3	12.68
653	3	12.66
654	3	12.65
655	3	12.63
656	3	12.64
657	3	12.68
658	3	12.66
659	3	12.76
660	3	12.80
661	3	12.81

ELA Grade 7

Scale Score	Achievement Level	CSEM
662	3	12.70
663	3	12.87
664	3	12.83
665	3	12.91
666	3	12.92
667	3	12.92
668	3	12.97
669	3	12.95
670	3	12.97
671	3	12.96
672	3	13.00
673	3	13.04
674	3	13.31
675	3	13.06
676	3	13.20
677	3	13.11
678	3	13.30
679	3	13.34
680	3	13.36
681	3	13.31
682	3	13.35
683	3	13.45
684	3	13.39
685	4	13.52
686	4	13.38
687	4	13.43
688	4	13.47
689	4	13.42
690	4	13.44
691	4	13.40
692	4	13.63
693	4	13.67
694	4	13.65
695	4	13.73
696	4	13.75
697	4	13.51
698	4	13.74
699	4	13.56
700	4	13.76
701	4	13.61
702	4	13.68
703	4	13.98
704	4	14.01
705	4	13.83

ELA Grade 7

Scale Score	Achievement Level	CSEM
706	4	14.10
707	4	14.01
708	4	13.86
709	4	13.88
710	4	14.06
711	4	14.18
712	4	14.22
713	4	14.15
714	4	14.22
715	4	14.09
716	4	14.30
717	4	14.06
718	4	14.27
719	4	14.41
720	4	14.45
721	4	14.52
722	4	14.39
723	4	14.64
724	4	14.84
725	4	14.85
726	4	14.75
727	4	14.61
728	4	15.21
729	4	15.27
730	4	14.95
731	4	14.78
732	4	14.96
733	4	14.68
734	4	15.29
735	4	15.37
736	4	15.47
737	4	15.64
738	4	15.37
739	4	15.92
740	4	15.32
741	4	15.82
742	4	16.42
743	4	15.51
744	4	16.24
745	4	15.59
746	4	16.01
747	4	15.76
748	4	15.92
749	4	15.81

ELA Grade 7		
Scale Score	**Achievement Level**	**CSEM**
750	4	15.87
751	4	16.13
752	4	16.76
753	4	16.14
754	4	16.97
755	4	16.46
756	4	16.85
758	4	16.83
759	4	17.52
761	4	16.71
762	4	17.45
763	4	16.63
764	4	17.38
766	4	17.65
767	4	17.95
769	4	17.83
770	4	19.06
771	4	18.32
773	4	17.44
774	4	17.83
775	4	18.28
777	4	18.23
778	4	19.03
779	4	20.15
780	4	18.43
783	4	19.08
784	4	19.42
788	4	19.05
789	4	18.90
797	4	18.33
798	4	20.17
811	4	20.47
824	4	28.31

Table 6: CSEM at Each Scale Score, ELA Grade 8

ELA Grade 8		
Scale Score	**Achievement Level**	**CSEM**
480	1	52.70
482	1	48.79
483	1	42.65

ELA Grade 8

Scale Score	Achievement Level	CSEM
484	1	44.78
485	1	46.01
486	1	42.48
487	1	44.87
488	1	45.33
489	1	43.81
490	1	37.41
491	1	43.58
492	1	38.23
493	1	41.45
494	1	39.50
495	1	39.00
496	1	36.62
497	1	37.69
498	1	38.28
499	1	38.83
500	1	38.85
501	1	38.86
502	1	36.68
503	1	37.82
504	1	34.94
505	1	34.58
506	1	36.73
507	1	34.29
508	1	35.60
509	1	34.93
510	1	34.22
511	1	34.29
512	1	32.29
513	1	33.19
514	1	32.52
515	1	32.05
516	1	31.49
517	1	31.18
518	1	30.86
519	1	30.42
520	1	29.88
521	1	30.11
522	1	29.77
523	1	29.61
524	1	29.56
525	1	28.68
526	1	28.31
527	1	28.41

ELA Grade 8

Scale Score	Achievement Level	CSEM
528	1	28.42
529	1	27.55
530	1	27.68
531	1	26.87
532	1	26.55
533	1	26.44
534	1	26.40
535	1	26.11
536	1	25.61
537	1	25.23
538	1	25.10
539	1	25.27
540	1	24.71
541	1	24.34
542	1	23.79
543	1	23.92
544	1	23.45
545	1	23.41
546	1	23.23
547	1	22.84
548	1	22.70
549	1	22.54
550	1	22.41
551	1	22.12
552	1	21.85
553	1	21.66
554	1	21.33
555	1	20.94
556	1	20.98
557	1	20.66
558	1	20.51
559	1	20.47
560	1	20.15
561	1	19.99
562	1	19.69
563	1	19.71
564	1	19.62
565	1	19.37
566	1	19.22
567	1	18.92
568	1	18.77
569	1	18.57
570	1	18.51
571	1	18.43

ELA Grade 8

Scale Score	Achievement Level	CSEM
572	1	18.15
573	1	18.14
574	1	18.14
575	1	17.92
576	1	17.77
577	1	17.73
578	1	17.51
579	1	17.39
580	1	17.31
581	1	17.16
582	1	16.90
583	1	16.94
584	1	16.74
585	1	16.75
586	1	16.50
587	1	16.37
588	1	16.43
589	1	16.39
590	1	16.20
591	1	16.05
592	1	15.92
593	1	15.85
594	1	15.84
595	1	15.72
596	1	15.70
597	1	15.43
598	1	15.43
599	1	15.29
600	1	15.30
601	1	15.24
602	1	15.04
603	1	15.08
604	1	15.00
605	1	14.84
606	1	14.91
607	1	14.80
608	1	14.65
609	1	14.66
610	1	14.57
611	1	14.52
612	1	14.60
613	2	14.42
614	2	14.34
615	2	14.22

ELA Grade 8

Scale Score	Achievement Level	CSEM
616	2	14.30
617	2	14.15
618	2	14.14
619	2	14.07
620	2	14.04
621	2	13.97
622	2	13.81
623	2	13.89
624	2	13.71
625	2	13.71
626	2	13.61
627	2	13.79
628	2	13.68
629	2	13.65
630	2	13.62
631	2	13.52
632	2	13.45
633	2	13.56
634	2	13.38
635	2	13.41
636	2	13.23
637	2	13.37
638	2	13.20
639	2	13.16
640	2	13.07
641	2	13.22
642	2	13.14
643	2	13.03
644	2	13.19
645	2	12.94
646	2	13.12
647	2	13.01
648	2	13.11
649	2	13.08
650	2	12.96
651	2	12.87
652	2	13.01
653	2	12.88
654	2	12.83
655	2	12.87
656	3	12.77
657	3	12.99
658	3	12.99
659	3	12.85

ELA Grade 8

Scale Score	Achievement Level	CSEM
660	3	12.93
661	3	13.02
662	3	13.06
663	3	12.96
664	3	12.86
665	3	12.96
666	3	12.98
667	3	13.06
668	3	12.92
669	3	12.95
670	3	13.02
671	3	13.00
672	3	13.04
673	3	13.05
674	3	13.07
675	3	13.12
676	3	13.19
677	3	13.13
678	3	13.26
679	3	13.15
680	3	13.24
681	3	13.13
682	3	13.11
683	3	13.22
684	3	13.27
685	3	13.32
686	3	13.28
687	3	13.45
688	3	13.39
689	3	13.50
690	3	13.36
691	3	13.56
692	3	13.64
693	3	13.66
694	3	13.67
695	3	13.60
696	3	13.62
697	3	13.71
698	4	13.76
699	4	13.80
700	4	13.76
701	4	13.80
702	4	13.82
703	4	13.99

ELA Grade 8

Scale Score	Achievement Level	CSEM
704	4	13.90
705	4	14.16
706	4	14.15
707	4	14.17
708	4	14.36
709	4	14.23
710	4	14.32
711	4	14.40
712	4	14.36
713	4	14.37
714	4	14.58
715	4	14.72
716	4	14.70
717	4	14.62
718	4	14.83
719	4	14.70
720	4	14.90
721	4	15.07
722	4	15.14
723	4	14.99
724	4	15.37
725	4	15.02
726	4	15.10
727	4	15.52
728	4	15.64
729	4	15.57
730	4	15.49
731	4	15.66
732	4	16.05
733	4	15.94
734	4	15.39
735	4	16.02
736	4	15.63
737	4	15.70
738	4	15.76
739	4	16.05
740	4	16.20
741	4	15.94
742	4	16.56
743	4	15.65
744	4	16.62
745	4	16.47
746	4	16.47
747	4	16.83

ELA Grade 8

Scale Score	Achievement Level	CSEM
748	4	17.01
749	4	17.02
750	4	16.70
751	4	17.16
752	4	17.38
753	4	17.79
754	4	17.14
755	4	17.97
756	4	17.80
757	4	17.81
758	4	18.31
759	4	18.21
760	4	17.65
761	4	18.51
762	4	19.03
763	4	18.88
764	4	18.36
765	4	19.23
766	4	18.72
767	4	18.76
768	4	17.18
769	4	18.98
770	4	18.77
771	4	19.96
772	4	20.89
773	4	19.21
774	4	19.69
775	4	21.12
776	4	18.88
777	4	20.25
778	4	20.43
779	4	19.87
780	4	21.30
785	4	22.63
786	4	20.29
787	4	22.25
788	4	23.12
789	4	22.78
790	4	23.29
791	4	26.05
794	4	23.92
795	4	24.01
798	4	24.56
803	4	25.35

ELA Grade 8		
Scale Score	**Achievement Level**	**CSEM**
804	4	26.81
814	4	30.57

Table 7: CSEM at Each Scale Score, Mathematics Grade 3

Mathematics Grade 3		
Scale Score	**Achievement Level**	**CSEM**
308	1	41.76
309	1	37.36
310	1	37.22
311	1	39.32
312	1	38.46
313	1	36.23
314	1	37.94
315	1	36.13
316	1	35.36
317	1	34.47
318	1	34.97
319	1	32.36
320	1	31.30
321	1	31.29
322	1	31.15
323	1	29.87
324	1	29.42
325	1	29.00
326	1	28.82
327	1	28.02
328	1	28.97
329	1	26.40
330	1	26.45
331	1	25.54
332	1	25.62
333	1	25.14
334	1	24.62
335	1	24.02
336	1	23.31
337	1	23.50
338	1	23.09
339	1	22.60
340	1	22.17
341	1	21.51

Mathematics Grade 3

Scale Score	Achievement Level	CSEM
342	1	21.01
343	1	20.51
344	1	20.02
345	1	20.04
346	1	18.85
347	1	18.79
348	1	18.60
349	1	18.20
350	1	17.78
351	1	17.36
352	1	17.29
353	1	16.81
354	1	16.68
355	1	16.15
356	1	16.00
357	1	15.62
358	1	15.33
359	1	14.97
360	1	14.58
361	1	14.40
362	1	14.23
363	1	13.74
364	1	14.01
365	1	13.38
366	1	13.16
367	1	12.89
368	1	12.71
369	1	12.45
370	1	12.49
371	1	11.89
372	1	11.68
373	1	11.66
374	1	11.19
375	1	11.25
376	1	11.00
377	1	10.85
378	1	10.69
379	1	10.51
380	1	10.27
381	1	10.20
382	1	10.15
383	1	10.00
384	1	9.82
385	1	9.75

Mathematics Grade 3

Scale Score	Achievement Level	CSEM
386	1	9.58
387	1	9.41
388	1	9.35
389	1	9.25
390	1	9.18
391	1	9.04
392	1	9.05
393	1	8.88
394	1	8.82
395	1	8.66
396	1	8.65
397	1	8.59
398	1	8.46
399	1	8.42
400	1	8.42
401	2	8.30
402	2	8.24
403	2	8.19
404	2	8.18
405	2	8.16
406	2	8.11
407	2	8.05
408	2	8.01
409	2	7.95
410	2	7.90
411	2	7.93
412	2	7.90
413	2	7.86
414	2	7.82
415	2	7.79
416	2	7.72
417	2	7.70
418	2	7.72
419	2	7.71
420	2	7.73
421	2	7.65
422	2	7.66
423	2	7.68
424	2	7.65
425	2	7.64
426	3	7.61
427	3	7.62
428	3	7.65
429	3	7.60

Mathematics Grade 3

Scale Score	Achievement Level	CSEM
430	3	7.62
431	3	7.56
432	3	7.64
433	3	7.64
434	3	7.60
435	3	7.63
436	3	7.67
437	3	7.65
438	3	7.67
439	3	7.74
440	3	7.70
441	3	7.74
442	3	7.71
443	3	7.72
444	3	7.74
445	3	7.73
446	3	7.77
447	3	7.73
448	4	7.81
449	4	7.84
450	4	7.84
451	4	7.86
452	4	7.87
453	4	7.94
454	4	7.96
455	4	7.96
456	4	8.02
457	4	8.04
458	4	8.03
459	4	8.02
460	4	8.17
461	4	8.23
462	4	8.15
463	4	8.26
464	4	8.26
465	4	8.26
466	4	8.42
467	4	8.41
468	4	8.38
469	4	8.46
470	4	8.42
471	4	8.59
472	4	8.57
473	4	8.65

Mathematics Grade 3

Scale Score	Achievement Level	CSEM
474	4	8.65
475	4	8.85
476	4	8.91
477	4	8.93
478	4	8.89
479	4	9.17
480	4	8.95
481	4	9.13
482	4	9.25
483	4	9.25
484	4	9.38
485	4	9.45
486	4	9.35
487	4	9.67
488	4	9.56
489	4	9.73
490	4	9.88
491	4	10.06
492	4	9.97
493	4	10.36
494	4	10.38
495	4	10.49
496	4	10.49
497	4	11.00
498	4	11.19
499	4	10.73
500	4	11.14
501	4	11.71
502	4	12.09
503	4	11.70
504	4	11.83
505	4	11.87
506	4	12.13
507	4	13.23
508	4	12.51
509	4	12.30
510	4	13.36
511	4	13.25
513	4	13.88
514	4	14.50
515	4	14.30
516	4	15.01
517	4	14.59
519	4	15.63

Mathematics Grade 3		
Scale Score	**Achievement Level**	**CSEM**
520	4	16.14
522	4	16.16
523	4	16.4
527	4	17.90
529	4	17.17
530	4	18.05
532	4	18.93
535	4	20.58
540	4	21.01
541	4	25.18
543	4	25.10
548	4	27.00

Table 8: CSEM at Each Scale Score, Mathematics Grade 4

Mathematics Grade 4		
Scale Score	**Achievement Level**	**CSEM**
312	1	46.05
313	1	44.65
314	1	41.44
315	1	36.34
316	1	39.70
317	1	41.90
318	1	39.49
319	1	36.91
320	1	38.42
321	1	42.78
322	1	36.01
323	1	35.29
324	1	35.84
325	1	35.78
326	1	37.04
327	1	34.64
328	1	33.66
329	1	33.65
330	1	35.43
331	1	31.75
332	1	32.34
333	1	31.04
334	1	30.56
335	1	30.07

Mathematics Grade 4

Scale Score	Achievement Level	CSEM
336	1	31.40
337	1	31.51
338	1	29.83
339	1	28.25
340	1	28.88
341	1	28.17
342	1	27.20
343	1	27.04
344	1	25.40
345	1	27.29
346	1	25.90
347	1	25.02
348	1	24.20
349	1	24.21
350	1	23.03
351	1	22.52
352	1	22.48
353	1	23.41
354	1	22.07
355	1	21.29
356	1	21.57
357	1	21.87
358	1	21.68
359	1	20.93
360	1	19.87
361	1	20.39
362	1	19.80
363	1	19.19
364	1	19.32
365	1	19.08
366	1	18.32
367	1	18.08
368	1	17.69
369	1	17.66
370	1	17.15
371	1	17.21
372	1	17.02
373	1	16.50
374	1	16.51
375	1	16.04
376	1	16.02
377	1	15.79
378	1	15.06
379	1	15.25

Mathematics Grade 4

Scale Score	Achievement Level	CSEM
380	1	14.91
381	1	14.77
382	1	14.64
383	1	14.29
384	1	14.19
385	1	14.07
386	1	13.94
387	1	13.62
388	1	13.58
389	1	13.57
390	1	13.41
391	1	13.26
392	1	12.83
393	1	12.98
394	1	12.80
395	1	12.70
396	1	12.70
397	1	12.66
398	1	12.33
399	1	12.19
400	1	12.09
401	1	12.14
402	1	12.13
403	1	11.83
404	1	11.81
405	1	11.64
406	1	11.65
407	1	11.48
408	1	11.39
409	1	11.40
410	1	11.40
411	1	11.29
412	1	11.22
413	1	11.08
414	1	10.87
415	1	10.97
416	1	10.85
417	1	10.76
418	1	10.72
419	1	10.79
420	1	10.77
421	1	10.59
422	2	10.41
423	2	10.49

Mathematics Grade 4

Scale Score	Achievement Level	CSEM
424	2	10.39
425	2	10.42
426	2	10.51
427	2	10.23
428	2	10.19
429	2	10.28
430	2	10.13
431	2	10.09
432	2	10.04
433	2	9.94
434	2	9.94
435	2	9.90
436	2	9.83
437	2	9.75
438	2	9.82
439	2	9.71
440	2	9.70
441	2	9.56
442	2	9.57
443	2	9.50
444	2	9.49
445	2	9.47
446	2	9.48
447	2	9.38
448	2	9.40
449	2	9.33
450	2	9.30
451	2	9.34
452	2	9.23
453	2	9.17
454	2	9.19
455	2	9.15
456	3	9.10
457	3	9.16
458	3	9.07
459	3	9.09
460	3	9.12
461	3	8.98
462	3	8.96
463	3	8.95
464	3	8.90
465	3	8.99
466	3	8.93
467	3	8.89

Mathematics Grade 4

Scale Score	Achievement Level	CSEM
468	3	8.88
469	3	8.85
470	3	8.94
471	3	8.85
472	3	8.91
473	3	8.82
474	3	8.86
475	3	8.89
476	3	8.92
477	3	8.85
478	4	8.89
479	4	8.75
480	4	8.83
481	4	8.84
482	4	8.89
483	4	8.86
484	4	8.92
485	4	8.83
486	4	8.94
487	4	8.97
488	4	8.98
489	4	8.88
490	4	8.95
491	4	8.98
492	4	9.03
493	4	9.02
494	4	9.01
495	4	9.15
496	4	9.08
497	4	9.19
498	4	9.13
499	4	9.13
500	4	9.10
501	4	9.25
502	4	9.20
503	4	9.19
504	4	9.27
505	4	9.25
506	4	9.31
507	4	9.41
508	4	9.33
509	4	9.38
510	4	9.29
511	4	9.50

Mathematics Grade 4

Scale Score	Achievement Level	CSEM
512	4	9.51
513	4	9.47
514	4	9.55
515	4	9.56
516	4	9.70
517	4	9.67
518	4	9.80
519	4	9.79
520	4	9.89
521	4	10.03
522	4	9.82
523	4	10.11
524	4	10.07
525	4	10.09
526	4	10.30
527	4	10.28
528	4	10.45
529	4	10.30
530	4	10.64
531	4	10.74
532	4	10.68
533	4	10.71
534	4	10.68
535	4	10.75
536	4	10.80
537	4	11.01
538	4	11.09
539	4	11.26
540	4	11.47
541	4	11.77
542	4	11.67
543	4	11.48
544	4	12.06
545	4	12.35
546	4	12.52
547	4	12.23
548	4	12.56
549	4	12.73
550	4	13.24
551	4	12.70
552	4	13.28
553	4	13.97
554	4	13.32
555	4	13.76

Mathematics Grade 4		
Scale Score	**Achievement Level**	**CSEM**
556	4	13.52
557	4	12.78
558	4	14.13
559	4	14.34
560	4	14.35
561	4	14.14
562	4	15.18
563	4	16.20
564	4	14.34
565	4	15.49
566	4	16.40
571	4	16.23
573	4	15.91
575	4	16.19
576	4	17.08
577	4	18.36
580	4	18.09
583	4	21.27
584	4	19.21
586	4	20.29
587	4	18.48
599	4	27.12
608	4	28.67

Table 9: CSEM at Each Scale Score, Mathematics Grade 5

Mathematics Grade 5		
Scale Score	**Achievement Level**	**CSEM**
320	1	55.08
321	1	57.30
322	1	54.52
323	1	51.23
324	1	48.21
325	1	50.95
326	1	46.72
327	1	47.30
328	1	49.86
329	1	50.18
330	1	44.64
331	1	46.00
332	1	46.13

Mathematics Grade 5

Scale Score	Achievement Level	CSEM
333	1	40.66
334	1	42.95
335	1	40.98
336	1	41.15
337	1	43.55
338	1	42.97
339	1	39.38
340	1	42.80
341	1	40.44
342	1	38.32
343	1	42.08
344	1	39.65
345	1	40.22
346	1	38.48
347	1	38.40
348	1	37.24
349	1	37.12
350	1	35.74
351	1	36.50
352	1	36.51
353	1	33.79
354	1	33.69
355	1	33.83
356	1	34.18
357	1	33.14
358	1	32.31
359	1	32.52
360	1	30.59
361	1	32.18
362	1	31.31
363	1	30.91
364	1	29.83
365	1	28.50
366	1	29.51
367	1	29.14
368	1	29.36
369	1	28.15
370	1	27.79
371	1	27.63
372	1	27.48
373	1	25.72
374	1	26.04
375	1	26.36
376	1	25.40

Mathematics Grade 5

Scale Score	Achievement Level	CSEM
377	1	25.30
378	1	25.01
379	1	24.71
380	1	25.09
381	1	23.83
382	1	24.54
383	1	22.77
384	1	23.16
385	1	22.24
386	1	22.76
387	1	22.44
388	1	21.29
389	1	21.40
390	1	21.11
391	1	21.06
392	1	21.04
393	1	20.27
394	1	20.12
395	1	20.31
396	1	20.39
397	1	19.62
398	1	18.90
399	1	19.15
400	1	18.53
401	1	18.72
402	1	18.54
403	1	18.33
404	1	18.20
405	1	17.85
406	1	17.56
407	1	17.02
408	1	17.02
409	1	17.22
410	1	16.76
411	1	16.50
412	1	16.44
413	1	15.91
414	1	16.05
415	1	15.79
416	1	15.70
417	1	15.72
418	1	15.34
419	1	15.22
420	1	15.17

Mathematics Grade 5

Scale Score	Achievement Level	CSEM
421	1	14.97
422	1	14.95
423	1	14.96
424	1	14.52
425	1	14.60
426	1	14.48
427	1	14.59
428	1	14.05
429	1	14.27
430	1	13.96
431	1	14.24
432	1	13.78
433	1	13.90
434	1	13.58
435	1	13.80
436	1	13.44
437	1	13.58
438	1	13.33
439	1	13.32
440	1	13.17
441	1	13.06
442	1	13.08
443	1	12.98
444	1	13.00
445	1	12.79
446	1	12.65
447	1	12.91
448	1	12.72
449	2	12.72
450	2	12.57
451	2	12.56
452	2	12.44
453	2	12.46
454	2	12.31
455	2	12.25
456	2	12.16
457	2	12.17
458	2	12.15
459	2	12.06
460	2	12.10
461	2	11.93
462	2	11.91
463	2	11.81
464	2	11.80

Mathematics Grade 5

Scale Score	Achievement Level	CSEM
465	2	11.71
466	2	11.66
467	2	11.72
468	2	11.61
469	2	11.53
470	2	11.55
471	2	11.49
472	2	11.49
473	2	11.53
474	2	11.36
475	2	11.30
476	2	11.38
477	2	11.28
478	2	11.32
479	2	11.25
480	2	11.20
481	2	11.27
482	2	11.25
483	2	11.13
484	2	11.17
485	2	11.15
486	2	11.08
487	3	11.11
488	3	11.07
489	3	11.01
490	3	11.06
491	3	11.04
492	3	10.94
493	3	10.94
494	3	11.04
495	3	10.94
496	3	10.92
497	3	11.01
498	3	10.94
499	3	10.99
500	3	10.93
501	3	10.94
502	3	10.94
503	3	11.00
504	3	10.95
505	3	10.95
506	3	10.91
507	3	10.94
508	3	11.00

Mathematics Grade 5

Scale Score	Achievement Level	CSEM
509	3	10.96
510	3	10.88
511	3	10.96
512	3	10.92
513	4	10.94
514	4	11.03
515	4	10.99
516	4	10.96
517	4	11.01
518	4	11.05
519	4	10.95
520	4	10.96
521	4	11.03
522	4	10.96
523	4	11.09
524	4	11.00
525	4	10.95
526	4	11.08
527	4	11.23
528	4	11.15
529	4	11.13
530	4	11.16
531	4	11.30
532	4	11.17
533	4	11.20
534	4	11.24
535	4	11.29
536	4	11.39
537	4	11.37
538	4	11.42
539	4	11.40
540	4	11.40
541	4	11.28
542	4	11.43
543	4	11.55
544	4	11.69
545	4	11.45
546	4	11.58
547	4	11.63
548	4	11.66
549	4	11.58
550	4	11.78
551	4	11.77
552	4	11.75

Mathematics Grade 5

Scale Score	Achievement Level	CSEM
553	4	11.98
554	4	11.94
555	4	12.24
556	4	11.71
557	4	12.06
558	4	12.19
559	4	12.33
560	4	12.16
561	4	12.29
562	4	12.67
563	4	12.51
564	4	12.43
565	4	12.43
566	4	12.55
567	4	12.99
568	4	12.86
569	4	12.95
570	4	12.85
571	4	13.09
572	4	13.44
573	4	13.51
574	4	13.77
575	4	12.96
576	4	13.29
577	4	13.65
578	4	13.87
579	4	14.41
580	4	14.22
581	4	13.98
582	4	14.41
583	4	15.42
584	4	14.85
585	4	14.68
586	4	14.69
587	4	14.75
588	4	15.82
589	4	15.39
590	4	15.16
591	4	16.69
592	4	15.38
593	4	16.05
594	4	16.14
595	4	17.09
596	4	17.54

Mathematics Grade 5		
Scale Score	**Achievement Level**	**CSEM**
597	4	16.16
598	4	16.75
599	4	17.02
600	4	18.31
601	4	16.33
602	4	17.39
604	4	18.23
605	4	20.14
606	4	18.22
608	4	18.38
609	4	19.30
610	4	20.07
611	4	19.05
613	4	20.31
614	4	21.79
615	4	21.84
616	4	20.41
617	4	19.41
618	4	22.14
621	4	21.56
623	4	23.19
624	4	23.08
625	4	24.53
627	4	24.38
630	4	22.20
632	4	24.62
639	4	30.12
640	4	26.09
647	4	30.53
648	4	32.37
658	4	35.29

Table 10: CSEM at Each Scale Score, Mathematics Grade 6

Mathematics Grade 6		
Scale Score	**Achievement Level**	**CSEM**
330	1	67.26
331	1	50.90
332	1	56.38
333	1	56.18
334	1	59.01

Mathematics Grade 6

Scale Score	Achievement Level	CSEM
335	1	61.85
336	1	55.87
337	1	51.55
338	1	51.38
339	1	57.65
340	1	58.34
341	1	51.32
342	1	49.29
343	1	64.11
344	1	51.28
345	1	49.72
346	1	47.50
347	1	46.94
348	1	44.52
349	1	47.39
350	1	48.66
351	1	47.35
352	1	44.99
353	1	47.57
354	1	45.05
355	1	43.97
356	1	44.24
357	1	43.71
358	1	43.84
359	1	44.11
360	1	40.66
361	1	42.15
362	1	40.37
363	1	43.23
364	1	38.25
365	1	38.65
366	1	39.00
367	1	37.58
368	1	38.45
369	1	38.91
370	1	37.58
371	1	36.92
372	1	37.32
373	1	33.57
374	1	34.56
375	1	34.11
376	1	34.19
377	1	34.19
378	1	33.41

Mathematics Grade 6

Scale Score	Achievement Level	CSEM
379	1	32.76
380	1	32.43
381	1	32.26
382	1	31.18
383	1	31.63
384	1	31.99
385	1	28.98
386	1	29.86
387	1	29.71
388	1	29.65
389	1	28.42
390	1	32.38
391	1	28.42
392	1	29.47
393	1	28.25
394	1	27.02
395	1	27.53
396	1	26.82
397	1	28.21
398	1	26.72
399	1	25.56
400	1	26.43
401	1	26.11
402	1	26.07
403	1	25.40
404	1	24.52
405	1	23.70
406	1	24.27
407	1	24.65
408	1	23.70
409	1	23.24
410	1	23.39
411	1	23.75
412	1	22.54
413	1	22.49
414	1	22.63
415	1	22.66
416	1	21.54
417	1	21.80
418	1	21.61
419	1	21.07
420	1	21.43
421	1	21.23
422	1	20.96

Mathematics Grade 6

Scale Score	Achievement Level	CSEM
423	1	20.48
424	1	20.38
425	1	20.34
426	1	20.81
427	1	19.91
428	1	20.12
429	1	19.72
430	1	19.91
431	1	19.16
432	1	19.11
433	1	18.95
434	1	18.85
435	1	18.71
436	1	18.52
437	1	18.34
438	1	18.28
439	1	17.87
440	1	18.15
441	1	17.83
442	1	17.96
443	1	17.84
444	1	17.57
445	1	17.29
446	1	17.44
447	1	17.40
448	1	17.38
449	1	16.95
450	1	17.05
451	1	17.02
452	1	16.62
453	1	16.84
454	1	16.76
455	1	16.57
456	1	16.65
457	1	16.35
458	1	16.20
459	1	16.02
460	1	16.17
461	1	15.98
462	1	15.72
463	1	15.67
464	1	15.65
465	1	15.66
466	1	15.53

Mathematics Grade 6

Scale Score	Achievement Level	CSEM
467	1	15.41
468	1	15.41
469	1	15.26
470	1	15.22
471	1	15.17
472	1	15.15
473	1	15.18
474	2	15.03
475	2	15.01
476	2	14.75
477	2	14.74
478	2	14.82
479	2	14.73
480	2	14.67
481	2	14.51
482	2	14.48
483	2	14.43
484	2	14.38
485	2	14.28
486	2	14.25
487	2	14.27
488	2	14.16
489	2	14.12
490	2	14.07
491	2	13.90
492	2	13.83
493	2	13.77
494	2	13.81
495	2	13.67
496	2	13.66
497	2	13.65
498	2	13.59
499	2	13.45
500	2	13.34
501	2	13.29
502	2	13.31
503	2	13.28
504	2	13.20
505	2	13.18
506	2	13.00
507	2	13.00
508	2	12.88
509	2	12.86
510	2	12.86

Mathematics Grade 6

Scale Score	Achievement Level	CSEM
511	2	12.75
512	2	12.78
513	2	12.67
514	2	12.61
515	2	12.57
516	2	12.49
517	2	12.42
518	3	12.40
519	3	12.45
520	3	12.25
521	3	12.26
522	3	12.20
523	3	12.15
524	3	12.11
525	3	12.11
526	3	12.06
527	3	11.98
528	3	11.99
529	3	11.87
530	3	11.85
531	3	11.76
532	3	11.75
533	3	11.78
534	3	11.72
535	3	11.72
536	3	11.64
537	3	11.57
538	3	11.55
539	3	11.68
540	3	11.55
541	3	11.54
542	3	11.38
543	3	11.32
544	3	11.33
545	3	11.43
546	3	11.25
547	3	11.25
548	3	11.14
549	3	11.33
550	4	11.38
551	4	11.16
552	4	11.17
553	4	11.33
554	4	11.22

Mathematics Grade 6

Scale Score	Achievement Level	CSEM
555	4	11.15
556	4	11.18
557	4	11.09
558	4	11.18
559	4	11.23
560	4	11.23
561	4	11.37
562	4	11.26
563	4	11.08
564	4	11.12
565	4	11.14
566	4	11.28
567	4	11.14
568	4	11.19
569	4	11.10
570	4	11.12
571	4	11.21
572	4	11.20
573	4	11.24
574	4	11.38
575	4	11.10
576	4	11.40
577	4	11.43
578	4	11.35
579	4	11.33
580	4	11.31
581	4	11.32
582	4	11.40
583	4	11.42
584	4	11.47
585	4	11.23
586	4	11.49
587	4	11.52
588	4	11.50
589	4	11.50
590	4	11.50
591	4	11.63
592	4	11.74
593	4	11.89
594	4	11.52
595	4	11.42
596	4	11.68
597	4	11.69
598	4	11.81

Mathematics Grade 6

Scale Score	Achievement Level	CSEM
599	4	11.53
600	4	12.00
601	4	11.90
602	4	11.89
603	4	11.81
604	4	12.27
605	4	12.46
606	4	12.35
607	4	12.24
608	4	12.18
609	4	12.75
610	4	13.36
611	4	12.62
612	4	12.58
613	4	12.86
614	4	12.97
615	4	12.70
616	4	13.26
617	4	13.02
618	4	13.46
619	4	13.00
620	4	13.94
621	4	13.64
622	4	12.30
623	4	13.46
624	4	13.42
625	4	14.70
626	4	13.73
628	4	14.39
629	4	15.22
631	4	15.18
632	4	15.11
633	4	14.54
634	4	15.82
635	4	14.81
636	4	14.97
637	4	16.00
638	4	16.98
640	4	16.88
643	4	17.46
647	4	18.13
651	4	18.24
652	4	19.06
653	4	18.49

Mathematics Grade 6		
Scale Score	**Achievement Level**	**CSEM**
663	4	20.86
666	4	21.00

Table 11: CSEM at Each Scale Score, Mathematics Grade 7

Mathematics Grade 7		
Scale Score	**Achievement Level**	**CSEM**
348	1	74.58
349	1	76.00
350	1	70.12
351	1	66.19
352	1	72.28
353	1	67.84
354	1	67.22
355	1	68.68
356	1	70.22
357	1	63.84
358	1	65.96
359	1	65.93
360	1	64.77
361	1	62.62
362	1	66.33
363	1	63.36
364	1	65.47
365	1	62.27
366	1	60.65
367	1	58.44
368	1	59.50
369	1	58.70
370	1	55.45
371	1	57.50
372	1	57.07
373	1	56.21
374	1	55.06
375	1	55.16
376	1	52.65
377	1	54.72
378	1	52.16
379	1	52.60
380	1	50.68
381	1	50.18

Mathematics Grade 7

Scale Score	Achievement Level	CSEM
382	1	48.46
383	1	49.57
384	1	49.72
385	1	48.82
386	1	48.20
387	1	47.18
388	1	45.19
389	1	45.67
390	1	44.47
391	1	43.51
392	1	44.04
393	1	44.33
394	1	42.68
395	1	41.52
396	1	40.50
397	1	40.56
398	1	40.72
399	1	40.14
400	1	39.99
401	1	38.98
402	1	39.05
403	1	38.62
404	1	37.41
405	1	37.67
406	1	37.12
407	1	36.75
408	1	35.77
409	1	36.16
410	1	35.59
411	1	34.51
412	1	33.87
413	1	34.20
414	1	33.25
415	1	33.32
416	1	32.74
417	1	32.53
418	1	31.91
419	1	31.51
420	1	31.22
421	1	30.14
422	1	31.11
423	1	29.91
424	1	30.02
425	1	29.51

Mathematics Grade 7

Scale Score	Achievement Level	CSEM
426	1	29.07
427	1	28.70
428	1	28.21
429	1	28.28
430	1	27.56
431	1	27.36
432	1	26.82
433	1	26.45
434	1	26.37
435	1	26.00
436	1	25.62
437	1	25.85
438	1	25.18
439	1	24.91
440	1	24.06
441	1	23.96
442	1	23.89
443	1	23.69
444	1	23.77
445	1	23.02
446	1	22.76
447	1	22.56
448	1	22.19
449	1	22.14
450	1	21.78
451	1	21.18
452	1	21.42
453	1	20.92
454	1	21.16
455	1	20.59
456	1	20.26
457	1	20.11
458	1	19.88
459	1	19.77
460	1	19.66
461	1	19.42
462	1	19.13
463	1	19.16
464	1	18.78
465	1	18.62
466	1	18.42
467	1	18.49
468	1	18.23
469	1	17.89

Mathematics Grade 7

Scale Score	Achievement Level	CSEM
470	1	17.67
471	1	17.64
472	1	17.58
473	1	17.35
474	1	17.30
475	1	17.15
476	1	16.95
477	1	16.80
478	1	16.98
479	1	16.44
480	1	16.57
481	1	16.42
482	1	16.35
483	1	16.23
484	1	16.20
485	1	15.86
486	1	15.77
487	1	15.77
488	1	15.61
489	1	15.49
490	1	15.48
491	1	15.31
492	1	15.30
493	1	15.20
494	1	15.24
495	1	14.96
496	1	14.90
497	1	15.03
498	1	14.78
499	1	14.66
500	1	14.64
501	1	14.77
502	1	14.49
503	2	14.59
504	2	14.61
505	2	14.45
506	2	14.40
507	2	14.31
508	2	14.34
509	2	14.36
510	2	14.05
511	2	14.02
512	2	14.06
513	2	13.94

Mathematics Grade 7

Scale Score	Achievement Level	CSEM
514	2	13.89
515	2	13.94
516	2	13.82
517	2	13.71
518	2	13.69
519	2	13.85
520	2	13.65
521	2	13.68
522	2	13.63
523	2	13.60
524	2	13.61
525	2	13.54
526	2	13.62
527	2	13.52
528	2	13.45
529	2	13.39
530	2	13.46
531	2	13.45
532	2	13.32
533	2	13.32
534	2	13.48
535	2	13.31
536	2	13.19
537	2	13.25
538	2	13.20
539	2	13.24
540	2	13.14
541	2	13.13
542	2	13.05
543	2	13.07
544	2	12.99
545	2	13.04
546	2	13.00
547	2	13.02
548	3	12.96
549	3	12.90
550	3	13.09
551	3	12.94
552	3	12.96
553	3	12.83
554	3	12.92
555	3	13.06
556	3	12.97
557	3	13.00

Mathematics Grade 7

Scale Score	Achievement Level	CSEM
558	3	12.90
559	3	12.87
560	3	13.00
561	3	12.87
562	3	12.91
563	3	13.02
564	3	12.79
565	3	12.99
566	3	12.95
567	3	13.00
568	3	12.99
569	3	12.93
570	3	12.83
571	3	13.10
572	3	12.92
573	3	13.09
574	3	12.96
575	3	13.03
576	3	12.97
577	3	12.91
578	3	13.03
579	3	12.92
580	3	12.71
581	3	12.95
582	3	12.79
583	4	12.89
584	4	12.92
585	4	12.91
586	4	12.92
587	4	12.90
588	4	12.84
589	4	12.77
590	4	12.83
591	4	12.61
592	4	12.66
593	4	12.78
594	4	12.81
595	4	12.80
596	4	12.81
597	4	12.74
598	4	12.62
599	4	12.82
600	4	12.72
601	4	12.56

Mathematics Grade 7

Scale Score	Achievement Level	CSEM
602	4	12.57
603	4	12.90
604	4	12.73
605	4	12.41
606	4	12.54
607	4	12.58
608	4	12.70
609	4	12.46
610	4	12.62
611	4	12.78
612	4	12.61
613	4	12.36
614	4	12.61
615	4	12.55
616	4	12.52
617	4	12.37
618	4	12.50
619	4	12.74
620	4	12.45
621	4	12.66
622	4	12.34
623	4	12.61
624	4	12.46
625	4	12.24
626	4	12.38
627	4	12.54
628	4	12.54
629	4	12.66
630	4	12.34
631	4	12.60
632	4	12.78
633	4	12.70
634	4	12.34
635	4	12.30
636	4	12.56
637	4	12.50
638	4	13.16
639	4	12.86
640	4	12.99
641	4	12.64
642	4	12.76
643	4	12.73
644	4	13.09
645	4	12.97

Mathematics Grade 7

Scale Score	Achievement Level	CSEM
646	4	13.15
647	4	13.02
648	4	12.96
649	4	13.70
650	4	13.18
651	4	13.05
652	4	13.64
653	4	13.40
654	4	13.96
655	4	13.24
656	4	13.42
657	4	13.80
658	4	12.23
659	4	12.57
660	4	14.47
661	4	13.98
662	4	13.53
663	4	13.94
664	4	14.71
665	4	13.58
666	4	13.98
667	4	15.24
668	4	13.88
669	4	13.72
670	4	15.12
671	4	14.15
672	4	14.22
674	4	14.13
675	4	15.16
676	4	14.34
677	4	15.07
678	4	15.33
679	4	15.90
680	4	14.63
682	4	16.65
686	4	14.89
688	4	12.28
690	4	15.34
692	4	15.81
694	4	15.72
696	4	16.98
699	4	21.33
703	4	15.86
705	4	19.44

Mathematics Grade 7		
Scale Score	**Achievement Level**	**CSEM**
706	4	17.23
708	4	15.96
709	4	18.98
713	4	17.70
717	4	19.84
718	4	17.99
719	4	17.99
721	4	20.97
724	4	21.35
728	4	23.22
736	4	21.42
747	4	28.39
750	4	30.51

Table 12: CSEM at Each Scale Score, Mathematics Grade 8

Mathematics Grade 8		
Scale Score	**Achievement Level**	**CSEM**
350	1	83.02
351	1	69.74
352	1	75.04
353	1	67.95
354	1	73.25
355	1	71.07
356	1	66.92
357	1	66.93
358	1	67.18
359	1	67.08
360	1	64.88
361	1	66.64
362	1	65.42
363	1	64.68
364	1	62.84
365	1	68.13
366	1	67.23
367	1	60.51
368	1	62.64
369	1	63.62
370	1	61.56
371	1	57.53
372	1	61.35

Mathematics Grade 8

Scale Score	Achievement Level	CSEM
373	1	57.00
374	1	59.36
375	1	56.77
376	1	57.54
377	1	55.67
378	1	57.40
379	1	53.16
380	1	54.65
381	1	54.18
382	1	57.73
383	1	53.33
384	1	54.51
385	1	52.77
386	1	52.41
387	1	50.74
388	1	51.93
389	1	51.81
390	1	50.56
391	1	49.76
392	1	50.76
393	1	50.22
394	1	47.68
395	1	48.94
396	1	47.95
397	1	46.11
398	1	46.06
399	1	47.09
400	1	43.90
401	1	44.30
402	1	44.14
403	1	44.08
404	1	42.80
405	1	44.19
406	1	43.84
407	1	42.44
408	1	43.18
409	1	41.39
410	1	43.06
411	1	41.94
412	1	40.42
413	1	41.12
414	1	40.94
415	1	40.65
416	1	41.39

Mathematics Grade 8

Scale Score	Achievement Level	CSEM
417	1	39.09
418	1	38.05
419	1	39.10
420	1	37.38
421	1	37.86
422	1	38.09
423	1	37.36
424	1	38.05
425	1	36.84
426	1	36.12
427	1	35.92
428	1	35.64
429	1	36.01
430	1	34.54
431	1	35.07
432	1	34.78
433	1	34.67
434	1	33.83
435	1	33.90
436	1	33.97
437	1	33.18
438	1	33.90
439	1	32.62
440	1	33.54
441	1	32.54
442	1	32.09
443	1	32.28
444	1	31.80
445	1	31.73
446	1	31.58
447	1	30.76
448	1	31.09
449	1	30.29
450	1	30.28
451	1	30.24
452	1	29.11
453	1	29.39
454	1	28.98
455	1	28.34
456	1	28.24
457	1	28.60
458	1	28.61
459	1	27.95
460	1	28.26

Mathematics Grade 8

Scale Score	Achievement Level	CSEM
461	1	27.56
462	1	27.56
463	1	27.14
464	1	26.67
465	1	27.30
466	1	26.67
467	1	26.85
468	1	26.39
469	1	25.81
470	1	25.67
471	1	25.82
472	1	25.21
473	1	25.41
474	1	25.63
475	1	24.99
476	1	25.12
477	1	24.77
478	1	24.52
479	1	24.57
480	1	24.38
481	1	24.38
482	1	23.99
483	1	23.71
484	1	23.81
485	1	23.37
486	1	23.52
487	1	23.28
488	1	23.05
489	1	23.12
490	1	23.05
491	1	22.81
492	1	22.63
493	1	22.48
494	1	22.46
495	1	22.34
496	1	22.47
497	1	21.95
498	1	21.79
499	1	21.60
500	1	21.64
501	1	21.43
502	1	21.42
503	1	21.23
504	1	21.33

Mathematics Grade 8

Scale Score	Achievement Level	CSEM
505	1	21.03
506	1	20.94
507	1	20.78
508	1	20.96
509	1	20.69
510	1	20.30
511	1	20.42
512	1	20.23
513	1	20.33
514	1	19.97
515	1	20.15
516	1	20.02
517	1	19.92
518	1	19.89
519	1	19.68
520	1	19.91
521	1	19.65
522	1	19.55
523	1	19.45
524	1	19.25
525	1	19.47
526	1	19.25
527	1	19.05
528	1	19.04
529	2	18.97
530	2	18.70
531	2	18.83
532	2	18.91
533	2	18.74
534	2	18.71
535	2	18.77
536	2	18.62
537	2	18.54
538	2	18.31
539	2	18.36
540	2	18.26
541	2	18.35
542	2	18.19
543	2	18.28
544	2	17.96
545	2	18.10
546	2	17.91
547	2	17.97
548	2	17.97

Mathematics Grade 8

Scale Score	Achievement Level	CSEM
549	2	17.79
550	2	17.88
551	2	17.75
552	2	17.60
553	2	17.66
554	2	17.57
555	2	17.45
556	2	17.48
557	2	17.50
558	2	17.36
559	2	17.41
560	2	17.23
561	2	17.27
562	2	17.26
563	2	17.08
564	2	17.06
565	2	17.07
566	2	17.09
567	2	16.92
568	2	16.83
569	2	16.90
570	2	16.74
571	2	16.68
572	2	16.66
573	2	16.53
574	2	16.47
575	2	16.61
576	2	16.67
577	2	16.44
578	2	16.51
579	2	16.33
580	2	16.37
581	2	16.26
582	2	16.03
583	2	16.08
584	2	16.04
585	2	16.08
586	2	16.08
587	3	16.07
588	3	15.96
589	3	15.99
590	3	15.91
591	3	15.99
592	3	15.74

Mathematics Grade 8

Scale Score	Achievement Level	CSEM
593	3	15.87
594	3	15.68
595	3	15.64
596	3	15.65
597	3	15.59
598	3	15.60
599	3	15.55
600	3	15.49
601	3	15.47
602	3	15.38
603	3	15.46
604	3	15.56
605	3	15.27
606	3	15.41
607	3	15.14
608	3	15.30
609	3	15.31
610	3	15.17
611	3	15.19
612	3	15.32
613	3	15.31
614	3	15.13
615	3	15.20
616	3	15.27
617	4	15.19
618	4	15.02
619	4	15.04
620	4	15.15
621	4	15.13
622	4	15.06
623	4	14.86
624	4	15.05
625	4	14.97
626	4	15.07
627	4	15.11
628	4	15.06
629	4	15.05
630	4	14.93
631	4	15.12
632	4	14.93
633	4	14.95
634	4	15.15
635	4	15.00
636	4	14.74

Mathematics Grade 8

Scale Score	Achievement Level	CSEM
637	4	14.50
638	4	14.87
639	4	14.92
640	4	14.73
641	4	14.91
642	4	14.90
643	4	14.77
644	4	14.72
645	4	14.74
646	4	14.77
647	4	14.81
648	4	14.83
649	4	14.94
650	4	14.71
651	4	14.72
652	4	14.79
653	4	14.60
654	4	15.07
655	4	14.72
656	4	14.61
657	4	14.76
658	4	15.00
659	4	14.53
660	4	14.84
661	4	14.90
662	4	14.84
663	4	14.90
664	4	15.13
665	4	14.69
666	4	15.07
667	4	15.21
668	4	14.88
669	4	15.00
670	4	15.32
671	4	15.26
672	4	14.99
673	4	15.23
674	4	15.21
675	4	15.10
676	4	15.24
677	4	15.31
678	4	15.44
679	4	15.23
680	4	15.34

Mathematics Grade 8

Scale Score	Achievement Level	CSEM
681	4	15.27
682	4	15.60
683	4	16.01
684	4	15.86
685	4	15.82
686	4	15.82
687	4	15.33
688	4	15.94
689	4	16.00
690	4	16.00
691	4	16.54
692	4	15.62
693	4	16.08
694	4	15.74
695	4	16.70
696	4	16.51
697	4	16.96
698	4	16.03
699	4	16.31
700	4	17.04
701	4	17.10
702	4	16.81
703	4	16.74
704	4	17.49
705	4	16.43
706	4	16.50
707	4	18.06
708	4	17.05
709	4	17.40
710	4	17.93
711	4	16.65
712	4	15.66
713	4	16.72
714	4	16.79
715	4	17.26
716	4	17.03
717	4	17.74
718	4	16.87
719	4	17.70
720	4	18.01
721	4	18.01
722	4	17.50
723	4	16.88
724	4	17.18

Mathematics Grade 8

Scale Score	Achievement Level	CSEM
725	4	19.24
726	4	19.34
727	4	17.96
728	4	17.20
729	4	18.35
730	4	18.45
731	4	17.56
732	4	19.36
733	4	19.47
734	4	19.66
735	4	19.16
736	4	19.85
737	4	18.39
738	4	17.54
739	4	18.51
740	4	24.16
741	4	19.49
743	4	20.81
744	4	19.63
746	4	19.92
747	4	19.02
748	4	21.53
751	4	21.77
752	4	18.66
753	4	18.00
754	4	22.24
755	4	22.81
756	4	24.88
757	4	21.11
758	4	20.01
759	4	24.26
760	4	20.65
762	4	23.50
763	4	21.82
764	4	20.78
766	4	24.97
768	4	21.61
769	4	21.12
773	4	22.35
776	4	23.52
777	4	23.67
780	4	29.80
785	4	26.23
791	4	29.62

Mathematics Grade 8		
Scale Score	**Achievement Level**	**CSEM**
801	4	32.90
806	4	38.94
820	4	37.62
822	4	41.17
828	4	39.68
830	4	42.09

Table 13: CSEM at Each Scale Score, Science Grade 5

Science Grade 5		
Scale Score	**Performance Level**	**CSEM**
500	1	8.59
501	1	7.76
502	1	7.89
503	1	7.73
504	1	7.56
505	1	7.71
506	1	7.45
507	1	7.31
508	1	7.01
509	1	6.93
510	1	6.94
511	1	6.83
512	1	6.71
513	1	6.55
514	1	6.59
515	1	6.57
516	1	6.45
517	1	6.37
518	1	6.33
519	1	6.29
520	1	6.15
521	1	6.10
522	1	5.99
523	1	5.96
524	1	5.93

Science Grade 5

Scale Score	Performance Level	CSEM
525	1	5.82
526	1	5.81
527	1	5.77
528	1	5.69
529	1	5.66
530	1	5.64
531	1	5.64
532	1	5.61
533	1	5.58
534	1	5.59
535	1	5.57
536	1	5.57
537	2	5.59
538	2	5.55
539	2	5.56
540	2	5.53
541	2	5.55
542	2	5.55
543	2	5.55
544	2	5.53
545	2	5.52
546	2	5.54
547	2	5.56
548	2	5.55
549	2	5.55
550	2	5.58
551	2	5.61
552	2	5.62
553	2	5.61
554	2	5.63
555	3	5.66
556	3	5.66
557	3	5.69
558	3	5.70
559	3	5.73
560	3	5.74
561	3	5.73

Science Grade 5

Scale Score	Performance Level	CSEM
562	3	5.77
563	3	5.74
564	3	5.79
565	3	5.78
566	3	5.81
567	3	5.83
568	4	5.86
569	4	5.88
570	4	5.90
571	4	5.93
572	4	5.97
573	4	5.97
574	4	6.06
575	4	6.06
576	4	6.13
577	4	6.22
578	4	6.24
579	4	6.27
580	4	6.24
581	4	6.24
582	4	6.45
583	4	6.43
584	4	6.50
585	4	6.59
586	4	6.59
587	4	6.74
588	4	6.83
589	4	6.79
590	4	6.81
591	4	6.84
592	4	6.95
593	4	7.02
594	4	7.14
595	4	6.85
596	4	7.75
597	4	7.29
598	4	7.50

Science Grade 5

Scale Score	Performance Level	CSEM
599	4	7.94
600	4	9.62

Table 14: CSEM at Each Scale Score, Science Grade 8

Science Grade 8

Scale Score	Performance Level	CSEM
800	1	7.64
801	1	7.75
802	1	7.99
803	1	7.72
804	1	7.71
805	1	7.30
806	1	7.50
807	1	7.08
808	1	7.19
809	1	6.82
810	1	7.01
811	1	6.92
812	1	6.77
813	1	6.60
814	1	6.47
815	1	6.52
816	1	6.34
817	1	6.35
818	1	6.22
819	1	6.23
820	1	6.16
821	1	6.10
822	1	6.03
823	1	6.02
824	1	5.93
825	1	5.93
826	1	5.91
827	1	5.81
828	1	5.79
829	1	5.77
830	1	5.71
831	1	5.74
832	1	5.66

Science Grade 8

Scale Score	Performance Level	CSEM
833	1	5.65
834	1	5.63
835	1	5.62
836	1	5.59
837	2	5.53
838	2	5.53
839	2	5.52
840	2	5.47
841	2	5.47
842	2	5.45
843	2	5.44
844	2	5.40
845	2	5.37
846	2	5.36
847	2	5.35
848	2	5.34
849	2	5.34
850	2	5.32
851	2	5.28
852	2	5.28
853	2	5.29
854	2	5.31
855	3	5.25
856	3	5.28
857	3	5.28
858	3	5.30
859	3	5.33
860	3	5.37
861	3	5.39
862	3	5.39
863	3	5.42
864	3	5.47
865	3	5.48
866	3	5.51
867	4	5.56
868	4	5.57
869	4	5.59
870	4	5.62
871	4	5.64
872	4	5.68
873	4	5.67
874	4	5.71
875	4	5.75
876	4	5.78

Science Grade 8

Scale Score	Performance Level	CSEM
877	4	5.75
878	4	5.97
879	4	5.95
880	4	5.91
881	4	5.77
882	4	6.04
883	4	6.00
884	4	6.29
885	4	6.10
886	4	6.07
887	4	6.08
888	4	5.97
889	4	6.03
890	4	6.19
891	4	6.17
892	4	6.35
893	4	6.38
894	4	6.47
895	4	6.08
896	4	6.99
897	4	6.63
898	4	6.46
899	4	6.47
900	4	7.34

Appendix F

Conditional Standard Error of Measurement by Subgroups

Table 1: Mean and Standard Deviation of Conditional Standard Error of Measurement by Subgroup, ELA

Grade	Group	All Students	Female	Male	African American	American Indian/ Native Alaskan	Asian	Hispanic	Multi-Racial	Pacific Islander	White	LEP
3	*N*	17,526	8,538	8,988	685	6	102	370	800	10	15,252	185
	Mean	13.27	12.91	13.6	14.89	-	12.66	13.15	13.47	-	13.2	14.73
	SD	5.38	4.96	5.73	7.09	-	4.27	4.21	5.26	-	5.35	6.16
4	*N*	17,323	8,454	8,869	630	8	98	370	794	8	15,129	131
	Mean	14.89	14.55	15.22	15.85	-	13.73	15.46	15.13	-	14.84	16.85
	SD	5.28	4.88	5.62	6.29	-	2.24	6.48	5.21	-	5.22	5.72
5	*N*	17,683	8,611	9,072	649	7	113	372	804	8	15,439	131
	Mean	13.67	13.32	14	14.8	-	12.68	14.1	13.89	-	13.62	15.8
	SD	4.3	3.89	4.64	5.54	-	2.42	4.57	4.09	-	4.27	5.17
6	*N*	17,697	8,678	9,019	683	13	100	410	723	9	15,282	100
	Mean	15.04	14.5	15.57	16.59	15.69	13.91	15.62	15.27	-	14.97	19.59
	SD	5.02	4.31	5.57	6.14	4.82	3.95	5.61	5.17	-	4.92	7.65
7	*N*	18,242	8,979	9,263	783	28	117	390	696	3	15,946	122
	Mean	15.08	14.53	15.61	15.84	17.05	14.56	15.43	15.41	-	15.03	18.69
	SD	4.45	3.9	4.86	4.6	4.65	3.8	5.19	5.2	-	4.4	7.95
8	*N*	18,698	9,012	9,686	717	20	120	440	729	9	16,371	133
	Mean	15.22	14.78	15.64	15.97	15.66	15.13	15.51	15.05	-	15.19	18.08
	SD	4.01	3.5	4.4	4.36	3.27	3.09	4.58	3.14	-	4.03	6.38

* The descriptive statistics are not provided when the number of students for given group is 10 or less than 10.

Table 2: : Mean and Standard Deviation of Conditional Standard Error of Measurement by Subgroup, Mathematics

Grade	Group	All Students	Female	Male	African American	American Indian/ Native Alaskan	Asian	Hispanic	Multi-Racial	Pacific Islander	White	LEP
3	*N*	17,542	8548	8,994	686	6	102	367	802	10	15,267	181
	Mean	9.41	9.33	9.49	10.42	-	9.14	9.72	9.66	-	9.36	10.19
	SD	4.77	4.5	5.01	6.09	-	4.13	5.39	5.21	-	4.69	6.25
4	*N*	17,329	8459	8,870	631	8	98	359	797	8	15,140	121
	Mean	11.15	10.98	11.3	12.37	-	10.07	11.38	11.31	-	11.1	12.45
	SD	5.46	5.02	5.85	6.89	-	1.86	5.65	5.08	-	5.45	6.94
5	*N*	17,717	8632	9,085	651	7	113	370	806	8	15,475	129
	Mean	14.32	14.13	14.5	15.99	-	13.03	14.92	15.09	-	14.23	16.36
	SD	7.82	7.46	8.15	8.8	-	4.22	8.08	8.74	-	7.77	9.42
6	*N*	17,708	8682	9,026	708	13	100	402	741	9	15,433	91
	Mean	17.07	16.62	17.5	20.06	22.24	13.49	17.56	17.95	-	16.89	22.25
	SD	9.9	8.85	10.79	12.47	17.44	4.31	9.47	10.6	-	9.74	12.81
7	*N*	18,281	8994	9,287	788	29	118	377	698	3	15,989	108
	Mean	18.12	17.61	18.62	22.36	24.82	14.78	18.36	19.81	-	17.86	23.62
	SD	11.72	10.73	12.58	15.45	17.48	5.62	11.1	13.31	-	11.43	15.81
8	*N*	18,718	9019	9,699	718	19	120	419	732	9	16,407	109
	Mean	22.59	21.81	23.32	27.01	26.08	19.42	23.07	23.97	-	22.32	26.57
	SD	12.68	11.38	13.74	16.07	16.37	10.17	12.2	13.61	-	12.43	15.07

* The descriptive statistics are not provided when the number of students for given group is 10 or less than 10.

Table 3: Mean and Standard Deviation of Conditional Standard Error of Measurement by Subgroup, Science

Grade	Group	All Students	Female	Male	African American	American Indian/ Native Alaskan	Asian	Hispanic	Multiple race	Pacific Islander	White	Declined to Report	LEP
5	*N*	17698	8621	9077	647	7	112	371	800	8	15305	448	131
	Mean	5.75	5.73	5.77	5.80	-	5.86	5.77	5.73	-	5.75	5.74	5.77
	SD	0.48	0.46	0.51	0.53	-	0.82	0.47	0.45	-	0.48	0.48	0.52
8	*N*	18694	9013	9681	711	20	120	438	723	9	16265	408	134
	Mean	5.58	5.55	5.61	5.65	5.61	5.64	5.62	5.59	-	5.58	5.57	5.79
	SD	0.44	0.43	0.45	0.46	0.54	0.60	0.48	0.46	-	0.44	0.45	0.50

*The descriptive statistics are not provided when the number of students for given group is 10 or less than 10.