

# Dimensionality Study Report for Expanded CSA

**Prepared for the California Department of Education by ETS**

**Presented January 15, 2025**



Table of Contents

[1. Executive Summary 1](#_Toc192839093)

[2. Introduction and Purposes 2](#_Toc192839094)

[3. Study Design 3](#_Toc192839095)

[4. Dimensionality Evaluation 3](#_Toc192839096)

[4.A. Overall Assessment 3](#_Toc192839097)

[4.B. Writing Domain 9](#_Toc192839098)

[5. Reliability of the Expanded CSA 10](#_Toc192839099)

[6. Evaluation of Subscore Reporting 11](#_Toc192839100)

[6.A. Augmented Subscore 12](#_Toc192839101)

[6.B. Reporting Achievement Levels at the Domain Level 15](#_Toc192839102)

[7. Calibration and Score Reporting Plan for Expanded CSA 17](#_Toc192839103)

[8. References 19](#_Toc192839104)

**List of Tables**

[Table 1. Blueprint of Original CSA 2](#_Toc192839105)

[Table 2. Blueprint of Expanded CSA 2](#_Toc192839106)

[Table 3. Sample Sizes for Analyses 3](#_Toc192839107)

[Table 4. CFA and MIRT Model Fit for Grades Three Through Eight 6](#_Toc192839108)

[Table 5. Factor Correlations from the CFA and MIRT for Grades Three Through Eight 7](#_Toc192839109)

[Table 6. CFA and MIRT Model Fit for High School 8](#_Toc192839110)

[Table 7. Factor Correlations from the CFA and MIRT for High School 9](#_Toc192839111)

[Table 8. Correlations Between Writing Mechanics Items and Full-Write CR Items 10](#_Toc192839112)

[Table 9. Cronbach’s Alpha of Expanded CSA 11](#_Toc192839113)

[Table 10. PRMSE of Augmented Subscores Based on Haberman’s Method 13](#_Toc192839114)

[Table 11. Reliability of Augmented Subscores Based on Wainer et al.’s Method 14](#_Toc192839115)

[Table 12. Predicted Decision Accuracy of Expanded CSA 15](#_Toc192839116)

[Table 13. Predicted Classification Consistency of Expanded CSA 16](#_Toc192839117)

[Table 14. Recommended Score Reporting Structure 18](#_Toc192839118)

**List of Figures**

[Figure 1. Models for dimensionality study for grades three through eight 4](#_Toc192839119)

[Figure 2. Models for dimensionality study for high school 5](#_Toc192839120)

[Figure 3. Model for dimensionality evaluation of writing domain 10](#_Toc192839121)

## Executive Summary

This report presents a summary of the dimensionality study conducted by ETS for the expanded California Spanish Assessment (CSA). The study involved several key analyses:

1. Evaluating competing models that represent the hypothesized factor structure of the expanded CSA
2. Assessing the content structure of the writing domain following the addition of a constructed-response (CR) writing item to the test blueprint
3. Summarizing the reliability estimates of the expanded CSA
4. Examining the validity and reliability of reporting subscores and exploring the feasibility of reporting augmented subscores

Based on the results from these analyses, ETS recommends using a unidimensional item response theory (IRT) model to calibrate items across all domains for grades three through eight. For high school (grades nine through twelve), dimensionality study results suggested that calibration should be conducted separately for the written literacy composite (encompassing items from the reading and writing domains) and the oral literacy composite (encompassing items from the listening and speaking domains). In addition to reporting the total test score and overall achievement levels, individual domain achievement levels should be reported for grade three through the high school grade band. For high school, in addition to domain achievement levels, two composite scores—written literacy and oral literacy—along with their associated achievement levels will be reported alongside the total test score and overall achievement levels.

## Introduction and Purposes

The CSA was initially developed in 2016 as a computer-based assessment for students in grades three through eight and high school to measure students’ literacy in Spanish in reading, writing mechanics, and listening. The CSA is now undergoing an expansion to include writing essay items and speaking items. The writing essay and speaking were field-tested for the first time in the 2023–24 test administration. The expanded CSA will be operationally administered starting with the 2024–25 test administration.

Originally, the CSA included only selected-response (SR) items. The expanded CSA has added CR items in the writing and speaking domains. Table 1 and table 2 provide the blueprints for both the original and expanded versions of the CSA.

Table . Blueprint of Original CSA

|  |  |  |
| --- | --- | --- |
| **Domain** | **Number of Items** | **Number of Points** |
| Reading | 24 SRs | 27–35 |
| Writing | 16 SRs | 19–22 |
| Listening | 12 SRs | 15–17 |
| Speaking | N/A | N/A |
| **Total Form:** | **52 SRs** | **61–66** |

Table . Blueprint of Expanded CSA

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Domain** | **Grades 3–8 Number of Items** | **Grades 3–8 Number of Points** | **High School Number of Items** | **High School Number of Points** |
| Reading | 24 SRs | 27–35 | 24 SRs | 27–35 |
| Writing | 11 SRs + 1 CR | 19–22 | 11 SRs + 1 CR | 19–22 |
| Listening | 12 SRs | 15–17 | 12 SRs | 15–17 |
| Speaking | N/A | N/A | 8 CRs | 16 |
| **Total Form:** | **47 SRs + 1 CR** | **61–66** | **55 SRs + 9 CRs** | **76–82** |

The CSA consists of a fixed number of items: 48 for grades three through eight and 56 for high school. Items may vary in type, with each being worth either 1, 2, or 4 points. As a result, the total possible score of different CSA forms may not be the same but falls within a range.

The introduction of expanded writing prompts for all grade levels and a new speaking domain for the high school assessment can bring potential changes to various aspects of the CSA, such as the content standards coverage, the underlying construct structure, score interpretation, and achievement level descriptors. To address these changes, a series of psychometric activities, including a dimensionality study, equating and scaling, and standard setting, have been planned for the expanded CSA. The dimensionality study aims to assess the underlying construct measured by the expanded CSA and provide statistical evidence to inform the selection of IRT models and structure of score reporting. This report provides the results from the analyses of the dimensionality study for the expanded CSA.

## Study Design

Data from the 2023–24 test administration was used for this dimensionality study. For the purposes of this study, mock forms were created to align with the expanded blueprint. The listening and reading sections remained unchanged, using operational items from 2023–24. For writing, the mock forms incorporated a mix of 2023–24 operational items and a full-write item selected from the 2023–24 field test forms. Additionally, speaking field test items from one of the field test forms were included in the high school mock forms.

Since the mock forms consisted of both operational and field test items, only students who completed one of the field test forms were included in the analyses. The sample sizes used for the analyses in this report are presented in table 3. The sample data was compared to the population in terms of students’ demographic backgrounds and ability levels, and it was concluded that the samples are generally representative of the population.

Table . Sample Sizes for Analyses

|  |  |
| --- | --- |
| **Grade Level or Grade Band** | **Sample Size** |
| 3 | 798 |
| 4 | 801 |
| 5 | 800 |
| 6 | 1,933 |
| 7 | 1,400 |
| 8 | 1,145 |
| High school | 1,050 |

## Dimensionality Evaluation

### Overall Assessment

The dimensionality study aims to investigate the factor structure of the postexpansion CSA domains. The study results will be used in recommending the various scores for reporting. Confirmatory factor analysis (CFA) was conducted to examine the underlying structure of the expanded CSA. Results from the dimensionality study will inform the CSA scaling and reporting plan.

Four competing models were fitted to explore the construct structure of the expanded CSA.

The models for grades three through eight assessments included the following:

1. A single-factor model (labeled in figure 1 as Model 1: unidimensional model) in which three CSA domains (reading, writing, and listening) measure a single ability, a students’ literacy in Spanish
2. Model 2, a correlated three-factor model, in which reading, writing, and listening are unique but correlated Spanish language skills
3. Model 3, a correlated two-factor model, in which reading and writing measure students’ written skills and listening, is another unique language skill
4. Model 4, a correlated two-factor model, in which listening and reading measure students’ receptive language skills and writing is another unique language skill.

Figure 1 presents the models for grades three through eight.



Figure . Models for dimensionality study for grades three through eight

The models tested for high school included the following:

1. A single-factor model (labeled in figure 2 as Model 1: Unidimensional model) in which four CSA domains (reading, writing, listening, and speaking) measure a single ability, students’ literacy in Spanish
2. Model 2, a correlated four-factor model, in which reading, writing, listening, and speaking are unique Spanish language skills
3. Model 3, a correlated two-factor model, in which listening and speaking measure students’ oral skills and reading and writing measure students’ written skills
4. Model 4, a correlated two-factor model, in which listening and reading measure students’ receptive language skills and speaking and writing measure students’ expressive language skills

Figure 2 presents the models for high school.



Figure . Models for dimensionality study for high school

All models mentioned in the preceding section were fitted under two statistical frameworks: the CFA and multidimensional IRT (MIRT) model. The CFA was conducted using Mplus, and MIRT models were run by flexMIRT. Within each grade level or the high school grade band, the model-fit indices of the four models are compared to determine the best-fitting model.

In evaluating model fit for CFA on categorical data, Comparative Fit Index (CFI) (Bentler, 1990) and Tucker-Lewis Index (TLI) (Tucker & Lewis, 1973) are two commonly used indices. They are both relative model-fit indices that compare the specified model to a null or independent model with no correlations among variables. The CFI is a normed index that considers sample size, making it sensitive to smaller samples, while the TLI is nonnormed and includes a penalty for adding more parameters, helping to prevent overfitting. Additionally, the root mean square error of approximation (RMSEA) (Browne & Cudeck, 1992) provides an estimate of the average discrepancy per degree of freedom between observed and predicted covariances, adjusting for model complexity. Lower RMSEA values indicate better fit, as they suggest less difference between the observed and model-predicted data. Together, these indices provide a comprehensive evaluation of how well the model fits the data, with each index adding a different perspective on fit quality.

In evaluating the model fit of MIRT models, the Akaike information criterion (AIC) and Bayesian information criterion (BIC) are widely used metrics. Both AIC and BIC test the relative quality of a model by estimating the amount of information loss, striking a balance between model fit and complexity. A lower AIC or BIC value by 10 points provides strong evidence of a preferred model, considering how well the model fits the data and how many parameters are used in the model. However, the BIC imposes a stronger penalty for adding parameters compared to the AIC, making it more conservative when selecting a model. This makes BIC especially useful in avoiding overfitting in complex models with multiple dimensions, while AIC can sometimes favor models with a slightly better fit at the cost of additional parameters. Both indices provide valuable insight into the trade-off between model accuracy and parsimony in multidimensional IRT analyses. If the AIC or BIC of a more complex model is not at least 10 points lower than that of a simpler model, the simpler model is generally preferred, as the added complexity does not significantly improve the model fit.

Both the CFA and MIRT analyses indicate that a unidimensional model is preferred over more complex models for grades three through eight. In the CFA results, the CFI, TLI, and RMSEA values are similar across all proposed models, suggesting that the more complex models (i.e., Model 2, Model 3, and Model 4) do not provide a better fit than the unidimensional model (i.e., Model 1). For MIRT, the AIC and BIC values are the smallest for the unidimensional model or Model 1 (that is, the single-factor model), which further supports the unidimensional IRT model or Model 1 as the preferred model. For more detailed information, please refer to table 4.

Table . CFA and MIRT Model Fit for Grades Three Through Eight

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Grade Level** | **Model** | **CFI** | **TLI** | **RMSEA** | **AIC** | **BIC** |
| 3 | 1-Factor | 0.954 | 0.952 | 0.021 | 52435 | 52735 |
| 3 | 3-Factor | 0.959 | 0.957 | 0.020 | 52511 | 52820 |
| 3 | Receptive & Writing | 0.954 | 0.952 | 0.021 | 52551 | 52850 |
| 3 | Written & Listening | 0.959 | 0.957 | 0.020 | 52520 | 52820 |
| 4 | 1-Factor | 0.975 | 0.973 | 0.014 | 54236 | 54536 |
| 4 | 3-Factor | 0.976 | 0.975 | 0.013 | 54437 | 54747 |
| 4 | Receptive & Writing | 0.975 | 0.973 | 0.014 | 54453 | 54753 |
| 4 | Written & Listening | 0.976 | 0.975 | 0.013 | 54440 | 54740 |
| 5 | 1-Factor | 0.968 | 0.967 | 0.015 | 55223 | 55532 |
| 5 | 3-Factor | 0.969 | 0.968 | 0.015 | 55395 | 55714 |
| 5 | Receptive & Writing | 0.968 | 0.967 | 0.015 | 55400 | 55709 |
| 5 | Written & Listening | 0.969 | 0.968 | 0.015 | 55409 | 55718 |
| 6 | 1-Factor | 0.968 | 0.967 | 0.017 | 136118 | 136486 |
| 6 | 3-Factor | 0.969 | 0.968 | 0.016 | 136463 | 136842 |
| 6 | Receptive & Writing | 0.969 | 0.967 | 0.017 | 136505 | 136872 |
| 6 | Written & Listening | 0.969 | 0.968 | 0.017 | 136489 | 136857 |
| 7 | 1-Factor | 0.973 | 0.972 | 0.013 | 98279 | 98620 |
| 7 | 3-Factor | 0.975 | 0.974 | 0.013 | 98691 | 99043 |
| 7 | Receptive & Writing | 0.973 | 0.972 | 0.013 | 98743 | 99084 |
| 7 | Written & Listening | 0.974 | 0.973 | 0.013 | 98732 | 99073 |
| 8 | 1-Factor | 0.977 | 0.976 | 0.015 | 79909 | 80237 |
| 8 | 3-Factor | 0.978 | 0.977 | 0.014 | 80044 | 80381 |
| 8 | Receptive & Writing | 0.977 | 0.976 | 0.015 | 80088 | 80416 |
| 8 | Written & Listening | 0.978 | 0.977 | 0.015 | 80058 | 80385 |

Additionally, high correlations among domain factors in both CFA and MIRT results confirmed the unidimensionality of the construct measured by the CSA, or Model 1, for grades three through eight. Typically, a correlation above 0.8 indicates a very strong relationship between variables. As shown in table 5, all factor correlations for grades three through eight are close to or exceed 0.9, indicating a very strong or close-to-perfect relationship between the factors. Note that in table 5, factor correlations in the CFA—presented in the *Correlation from the CFA* column—can occasionally exceed 1 because of improper solutions, often resulting from issues such as overextraction of factors, multicollinearity, or model misspecification.

Table . Factor Correlations from the CFA and MIRT for Grades Three Through Eight

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Grade Level** | **Model** | **Factor 1** | **Factor 2** | **Correlation from the CFA** | **Correlation from MIRT** |
| 3 | 3-Factor | Reading | Listening | 0.85 | 0.89 |
| 3 | 3-Factor | Reading | Writing | 1.00 | 0.98 |
| 3 | 3-Factor | Listening | Writing | 0.92 | 0.95 |
| 3 | Receptive & Writing | Receptive | Writing | 1.01 | 1.00 |
| 3 | Written & Listening | Written | Listening | 0.88 | 0.93 |
| 4 | 3-Factor | Reading | Listening | 0.91 | 0.93 |
| 4 | 3-Factor | Reading | Writing | 0.98 | 0.97 |
| 4 | 3-Factor | Listening | Writing | 0.92 | 0.94 |
| 4 | Receptive & Writing | Receptive | Writing | 0.98 | 0.97 |
| 4 | Written & Listening | Written | Listening | 0.91 | 0.94 |
| 5 | 3-Factor | Reading | Listening | 0.92 | 0.96 |
| 5 | 3-Factor | Reading | Writing | 0.97 | 0.95 |
| 5 | 3-Factor | Listening | Writing | 0.93 | 0.94 |
| 5 | Receptive & Writing | Receptive | Writing | 0.97 | 0.96 |
| 5 | Written & Listening | Written | Listening | 0.94 | 0.97 |
| 6 | 3-Factor | Reading | Listening | 0.92 | 0.94 |
| 6 | 3-Factor | Reading | Writing | 0.96 | 0.96 |
| 6 | 3-Factor | Listening | Writing | 0.90 | 0.92 |
| 6 | Receptive & Writing | Receptive | Writing | 0.96 | 0.96 |
| 6 | Written & Listening | Written | Listening | 0.92 | 0.95 |
| 7 | 3-Factor | Reading | Listening | 0.89 | 0.92 |
| 7 | 3-Factor | Reading | Writing | 0.96 | 0.95 |
| 7 | 3-Factor | Listening | Writing | 0.94 | 0.96 |
| 7 | Receptive & Writing | Receptive | Writing | 0.98 | 0.97 |
| 7 | Written & Listening | Written | Listening | 0.92 | 0.96 |
| 8 | 3-Factor | Reading | Listening | 0.90 | 0.91 |
| 8 | 3-Factor | Reading | Writing | 0.96 | 0.96 |
| 8 | 3-Factor | Listening | Writing | 0.95 | 0.93 |
| 8 | Receptive & Writing | Receptive | Writing | 0.98 | 0.97 |
| 8 | Written & Listening | Written | Listening | 0.93 | 0.93 |

For the high school grade band, the CFA indicated that Model 2, a four-factor model, provided a better fit compared to other models. However, Model 2, the four-factor model, calibration under MIRT framework failed to converge using the flexMIRT and R “mirt” packages. Nonconvergence frequently occurs in IRT calibration when extracting more factors than the data can support. For example, if the software attempts to calibrate a four-factor IRT model when the true model is a two-factor model, the estimation process may fail to reach a stable solution because of insufficient data to reliably estimate four distinct factors, resulting in nonconvergence.

A two-factor model, Model 3, which combined written components and oral components, also showed a better fit than Model 1, a single-factor model, as indicated by higher CFI and TLI values, suggesting distinct factors for written skills and oral skills. Notably, the CFA and MIRT analyses yielded different factor correlations for the two-factor model (refer to table 7), highlighting potential differences in how these methods capture the relationships between the components. In general, MIRT produced much higher factor correlations for the two-factor models [i.e., Model 3: Two-Factor Model A (Written & Oral) and Model 4: Two-Factor Model B (Receptive & Expressive)] than the CFA.

As presented in table 6 and table 7, AIC and BIC from MIRT are similar after rounding, suggesting similar model fit in the MIRT framework. However, the written literacy and oral literacy composite model has slightly better fit than the receptive and expressive model in CFA, as indicated by higher CFI and TLI values. Lower factor correlations between the written literacy composite and oral literacy composite indicate a greater distinction between the two factors.

Table . CFA and MIRT Model Fit for High School

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **CFI** | **TLI** | **RMSEA** | **AIC** | **BIC** |
| 1-Factor | 0.905 | 0.901 | 0.032 | 84104 | 84511 |
| 4-Factor | 0.959 | 0.957 | 0.021 | N/A | N/A |
| Receptive & Expressive | 0.920 | 0.917 | 0.029 | 83466 | 83942 |
| Written & Oral | 0.928 | 0.926 | 0.028 | 83466 | 83942 |

Table . Factor Correlations from the CFA and MIRT for High School

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Factor 1** | **Factor 2** | **Correlation from the CFA** | **Correlation from MIRT** |
| 4-Factor | Listening | Reading | 0.877 | N/A |
| 4-Factor | Writing | Reading | 0.932 | N/A |
| 4-Factor | Writing | Listening | 0.894 | N/A |
| 4-Factor | Speaking | Reading | 0.618 | N/A |
| 4-Factor | Speaking | Listening | 0.652 | N/A |
| 4-Factor | Speaking | Writing | 0.676 | N/A |
| Receptive & Expressive | Receptive | Expressive | 0.820 | 0.980 |
| Written & Oral | Written | Oral | 0.779 | 0.981 |

### Writing Domain

Additional dimensionality analyses were conducted to explore whether writing mechanics items (selected-response items) and full-write CR items measure the same construct within the writing domain. Using a structural equation model, as depicted in figure 3, the relationship between students’ writing mechanics ability and essay writing ability was examined. The results, summarized in table 8, indicate moderate correlations across various assessments, with values ranging from 0.47 (grade four) to 0.70 (grade three). These findings suggest that while writing mechanics and full-write abilities are related, they also represent somewhat distinct skills. This supports the decision to create a new scale and conduct standard setting for the expanded CSA, highlighting the importance of integrating both skills for a comprehensive evaluation of students’ writing abilities.

Figure 3 presents the model for dimensionality evaluation of the writing domain.



Figure . Model for dimensionality evaluation of writing domain

Table . Correlations Between Writing Mechanics Items and Full-Write CR Items

|  |  |
| --- | --- |
| **Grade Level or Grade Band** | **Correlation** |
| 3 | 0.70 |
| 4 | 0.47 |
| 5 | 0.59 |
| 6 | 0.59 |
| 7 | 0.54 |
| 8 | 0.56 |
| High school | 0.59 |

## Reliability of the Expanded CSA

Checking reliability before reporting test scores is essential to ensure that the scores, whether overall, composite, or domain level, are consistent, accurate, and dependable. By evaluating reliability metrics, such as Cronbach’s alpha (Cronbach, 1951), assessment developers and interest holders can determine whether it is appropriate to report subscores and ensure that the assessment provides meaningful, actionable information for educational decisions.

Table 9 presents Cronbach’s alpha of total assessments, composites, and domains of the expanded CSA. At the overall test level, reliability values are acceptable, ranging from 0.82 (grade seven) to 0.89 (high school). However, due to fewer items within each domain, reliability at the domain level has lower values observed in listening (0.56–0.69) and writing (0.61–0.73), suggesting variations in reliabilities across grade levels and the high school grade band in these domains. Receptive and expressive composite scores show moderate reliability (0.78–0.86), while written scores exhibit slightly higher values (0.78–0.82). The speaking domain for high school shows the highest domain reliability at 0.85. Overall, while the CSA demonstrates strong reliability at the test level and acceptable reliability at the composite level, variability at the domain levels highlights the need for further evaluation before reporting subscores, particularly in the listening and writing domains.

Table . Cronbach’s Alpha of Expanded CSA

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Grade Level or Grade Band** | **Overall** | **Receptive (Listening and Reading)** | **Expressive (Speaking and Writing)** | **Written (Reading and Writing)** | **Oral (Listening and Speaking)** | **Reading** | **Listening** | **Writing** | **Speaking** |
| 3 | 0.87 | 0.82 | N/A | 0.82 | N/A | 0.73 | 0.69 | 0.67 | N/A |
| 4 | 0.85 | 0.81 | N/A | 0.81 | N/A | 0.73 | 0.64 | 0.61 | N/A |
| 5 | 0.83 | 0.78 | N/A | 0.79 | N/A | 0.69 | 0.61 | 0.61 | N/A |
| 6 | 0.86 | 0.79 | N/A | 0.84 | N/A | 0.73 | 0.56 | 0.73 | N/A |
| 7 | 0.82 | 0.74 | N/A | 0.78 | N/A | 0.64 | 0.58 | 0.66 | N/A |
| 8 | 0.87 | 0.82 | N/A | 0.86 | N/A | 0.79 | 0.56 | 0.72 | N/A |
| High school | 0.89 | 0.81 | 0.82 | 0.83 | 0.81 | 0.75 | 0.60 | 0.65 | 0.85 |

## Evaluation of Subscore Reporting

Haberman et al. (2024) recommend evaluating subscore reporting on the basis of three key factors. The first factor, subtest length, requires more than 24 items to ensure sufficient measurement precision. The second factor, reliability, necessitates a reliability coefficient greater than 0.7 to ensure consistent and dependable scores. The third factor, disattenuated correlations among subscores, should be less than 0.9 to confirm that subscores provide distinct information from the total test score and are not overly redundant. Together, these criteria determine whether subscores are meaningful, reliable, and add value beyond the overall test score.

In evaluating these criteria, subtest length is satisfied for the reading domain and all composite scores across all grade levels but not for other domains. Reliability (>0.7) is achieved for the reading domain at all grade levels except grades five and seven and is also met for all composite scores. Reliability (>0.7) is not achieved for the other domains, specifically, the listening and writing domains. Finally, disattenuated correlations among subscores (<0.9), as indicated in table 5 and table 7, are not satisfied for grades three through eight. At the high school level, most domain and composite correlations are lower than 0.9. Correlations between speaking and other domains are below 0.7, while correlations among listening, reading, and writing are higher but remain under 0.9. The correlation between receptive and expressive composites was below 0.9, whereas the correlation between oral and written composites was below 0.8.

These results highlight significant limitations in subscore reporting, especially for lower grade levels and the domains of listening and writing, while providing stronger support for composite scores of the high school assessment. As a result, ETS conducted further research to assess the feasibility of reporting augmented subscores, aiming to enhance the value of subscore reporting for the expanded CSA.

### Augmented Subscore

An augmented subscore is a refined measure that enhances the reliability and interpretability of a specific domain score by incorporating information from other related scores and the overall test score. Unlike raw subscores, which may suffer from low reliability due to limited data or variability within a single domain, augmented subscores use statistical techniques such as regression to “borrow strength” from correlations across domains. This approach leverages the relationships between a target domain and other domains, as well as the total test score, to produce a more stable and accurate estimate of achievement. Various methods have been proposed to calculate augmented subscores (Haberman, 2008; Haberman & Sinharay, 2010; Wainer et al., 2001; Yen, 1987; Yen et al., 1997). Given the nature of the expanded CSA and the practicality of reporting subscores, three methods were evaluated: Haberman’s observed score method (Haberman, 2008) and the observed score and IRT score methods proposed by Wainer et al. (2001).

Despite the similarities between Haberman’s method and Wainer et al.’s method, they proposed different criteria for evaluating augmented scores. Haberman used the proportional reduction of mean squared error (PRMSE), while Wainer et al. relied on subscore reliability for this purpose. Augmented subscores are considered valuable for score reporting if their PRMSE or reliability exceeds that of the total test score by at least 0.01. Conversely, if this threshold is not met, reporting augmented subscores would introduce more errors, reducing their usefulness in score reporting.

Table 10 and table 11 provide the PRMSE and the reliability of the augmented subscores, respectively. The analysis of augmented subscores was performed at the domain level for grades three through eight, while for high school, it was conducted at the composite level, which aligns with the focus on composite-level score reporting.

Table . PRMSE of Augmented Subscores Based on Haberman’s Method

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Grade Level or Grade Band** | **Domain or Composite** | **PRMSE of Total Score** | **PRMSE of Augmented Score** | **Difference >0.01?** |
| 3 | Reading | 0.837 | 0.839 | No |
| 3 | Writing | 0.887 | 0.890 | No |
| 3 | Listening | 0.773 | 0.803 | Yes |
| 4 | Reading | 0.815 | 0.816 | No |
| 4 | Writing | 0.828 | 0.828 | No |
| 4 | Listening | 0.753 | 0.776 | Yes |
| 5 | Reading | 0.823 | 0.824 | No |
| 5 | Writing | 0.810 | 0.813 | No |
| 5 | Listening | 0.803 | 0.809 | No |
| 6 | Reading | 0.838 | 0.838 | No |
| 6 | Writing | 0.830 | 0.836 | No |
| 6 | Listening | 0.793 | 0.797 | No |
| 7 | Reading | 0.780 | 0.781 | No |
| 7 | Writing | 0.810 | 0.812 | No |
| 7 | Listening | 0.750 | 0.765 | Yes |
| 8 | Reading | 0.837 | 0.843 | No |
| 8 | Writing | 0.851 | 0.851 | No |
| 8 | Listening | 0.783 | 0.792 | No |
| High school | Written | 0.804 | 0.843 | Yes |
| High school | Oral | 0.792 | 0.825 | Yes |

Table . Reliability of Augmented Subscores Based on Wainer et al.’s Method

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Grade Level or Grade Band** | **Domain or Composite** | **Raw Score–Based Augmented Score Reliability** | **Theta-Based Augmented Score Reliability** | **Total Test Score Reliability** | **Did Raw Score–Based Augmented Score Add Value?** | **Did Theta-Based Augmented Score Add Value?** |
| 3 | Reading | 0.855 | 0.829 | 0.865 | No | No |
| 3 | Writing | 0.861 | 0.869 | 0.865 | No | No |
| 3 | Listening | 0.846 | 0.805 | 0.865 | No | No |
| 4 | Reading | 0.827 | 0.832 | 0.852 | No | No |
| 4 | Writing | 0.829 | 0.931 | 0.852 | No | Yes |
| 4 | Listening | 0.819 | 0.824 | 0.852 | No | No |
| 5 | Reading | 0.839 | 0.832 | 0.833 | No | No |
| 5 | Writing | 0.838 | 0.931 | 0.833 | No | Yes |
| 5 | Listening | 0.837 | 0.824 | 0.833 | No | No |
| 6 | Reading | 0.851 | 0.829 | 0.862 | No | No |
| 6 | Writing | 0.850 | 0.853 | 0.862 | No | No |
| 6 | Listening | 0.845 | 0.849 | 0.862 | No | No |
| 7 | Reading | 0.811 | 0.853 | 0.822 | No | Yes |
| 7 | Writing | 0.818 | 0.880 | 0.822 | No | Yes |
| 7 | Listening | 0.809 | 0.845 | 0.822 | No | Yes |
| 8 | Reading | 0.859 | 0.833 | 0.873 | No | No |
| 8 | Writing | 0.861 | 0.871 | 0.873 | No | No |
| 8 | Listening | 0.849 | 0.850 | 0.873 | No | No |
| High school | Written | 0.860 | 0.838 | 0.888 | No | No |
| High school | Oral | 0.851 | 0.860 | 0.888 | No | No |

Overall, augmented subscores derived from raw scores did not provide additional value for reporting student achievement in subdomains compared to the total test score. However, augmented subscores based on IRT theta estimates demonstrated a significant improvement in reliability for reporting achievement in the writing domain for grades four and five. A moderate improvement in subscore reliability was also observed across all three domains for grade seven, but no such improvements were evident for other grade levels or subscores. Considering the added complexity to the system and reporting processes, implementing augmented subscores does not appear to be a practical or beneficial approach.

### Reporting Achievement Levels at the Domain Level

Although the lower reliability of domain scores does not support reporting scale scores at the domain level, and augmented scores did not offer better alternatives, the conditional standard errors of measurement from the conversion tables of mock forms indicate that reporting achievement levels at the domain level remains feasible. To predict decision accuracy and classification consistency, equated IRT parameters from the 2023–24 test administration were used to develop conversion tables for each domain and for each grade level and the high school grade band. The original CSA achievement range cuts were then applied to assign students to appropriate achievement levels. Furthermore, student distribution data from the dimensionality study and mock form reliability were analyzed to estimate decision accuracy and classification consistency (refer to table 12 and table 13).

Note that because the cut scores for the expanded CSA have yet to be established, these results are estimated on the basis of the original cut scores. For the two achievement levels, ETS evaluated either using the original Level 2 cut score or the original Level 3 cut score (as shown in the last two columns of table 12 and table 13). The values reported are approximations and may change once the final cut scores are established. Assuming some slight changes of the cut scores, these values, even though only approximations, are sufficient to inform whether two or three levels can be supported. The results show that for grades three through eight, predicted classification accuracy and consistency are moderate when reporting three achievement levels and high when reporting two levels. For the high school composite scores, classification accuracy and consistency are high for both two and three achievement levels.

The decision accuracy values in table 12 exceed 0.7 for both three-level and two-level classifications across all grade levels and the high school grade band, indicating moderate to high classification accuracy. Similarly, the classification consistency values in table 13 are all above 0.6, aligning with the current CSA classification consistency. On the basis of these results, while reporting a scale score at the domain level may not be reliable, reporting students’ performance levels for each domain should be both feasible and informative.

Table . Predicted Decision Accuracy of Expanded CSA

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Grade Level or Grade Band** | **Domain or Composite** | **Decision Accuracy (3 Levels)** | **Decision Accuracy on Second Cut (2 Levels)** | **Decision Accuracy on First Cut (2 Levels)** |
| 3 | Reading | 0.86 | 0.96 | 0.90 |
| 3 | Listening | 0.79 | 0.91 | 0.86 |
| 3 | Writing | 0.77 | 0.93 | 0.83 |
| 4 | Reading | 0.88 | 0.96 | 0.92 |
| 4 | Listening | 0.80 | 0.92 | 0.86 |
| 4 | Writing | 0.80 | 0.94 | 0.85 |
| 5 | Reading | 0.78 | 0.95 | 0.83 |
| 5 | Listening | 0.73 | 0.91 | 0.80 |
| 5 | Writing | 0.71 | 0.91 | 0.79 |
| 6 | Reading | 0.78 | 0.94 | 0.84 |
| 6 | Listening | 0.69 | 0.90 | 0.79 |
| 6 | Writing | 0.76 | 0.90 | 0.85 |
| 7 | Reading | 0.83 | 0.96 | 0.87 |
| 7 | Listening | 0.70 | 0.91 | 0.78 |
| 7 | Writing | 0.78 | 0.93 | 0.85 |
| 8 | Reading | 0.81 | 0.92 | 0.88 |
| 8 | Listening | 0.73 | 0.90 | 0.81 |
| 8 | Writing | 0.77 | 0.91 | 0.85 |
| High school | Written literacy | 0.86 | 0.95 | 0.90 |
| High school | Oral literacy | 0.83 | 0.94 | 0.89 |

Table . Predicted Classification Consistency of Expanded CSA

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Grade Level or Grade Band** | **Domain or Composite** | **Classification Consistency (3 Levels)** | **Classification Consistency on Second Cut (2 Levels)** | **Classification Consistency on First Cut (2 Levels)** |
| 3 | Reading | 0.80 | 0.94 | 0.85 |
| 3 | Listening | 0.72 | 0.87 | 0.81 |
| 3 | Writing | 0.68 | 0.90 | 0.76 |
| 4 | Reading | 0.83 | 0.95 | 0.88 |
| 4 | Listening | 0.73 | 0.88 | 0.80 |
| 4 | Writing | 0.72 | 0.91 | 0.79 |
| 5 | Reading | 0.69 | 0.93 | 0.75 |
| 5 | Listening | 0.63 | 0.88 | 0.73 |
| 5 | Writing | 0.62 | 0.87 | 0.72 |
| 6 | Reading | 0.70 | 0.91 | 0.77 |
| 6 | Listening | 0.59 | 0.85 | 0.70 |
| 6 | Writing | 0.68 | 0.86 | 0.79 |
| 7 | Reading | 0.75 | 0.95 | 0.80 |
| 7 | Listening | 0.60 | 0.87 | 0.71 |
| 7 | Writing | 0.70 | 0.91 | 0.78 |
| 8 | Reading | 0.74 | 0.88 | 0.83 |
| 8 | Listening | 0.65 | 0.85 | 0.73 |
| 8 | Writing | 0.69 | 0.87 | 0.80 |
| High school | Written literacy | 0.80 | 0.93 | 0.86 |
| High school | Oral literacy | 0.77 | 0.92 | 0.84 |

## Calibration and Score Reporting Plan for Expanded CSA

ETS recommends using a unidimensional model for grades three through eight IRT calibration, as results from both the CFA and MIRT support the unidimensionality of the assessments. For high school, two unidimensional IRT calibrations are proposed for written components (including items from the writing and reading domains) and oral components (including items from the speaking and listening domains). While CFA model-fit indices support a four-factor structure, it is deemed impractical to calibrate individual domains because of low reliabilities. Instead, a two-factor model, distinguishing the written literacy composite and oral literacy composite, provides a better fit compared to a unidimensional model. Additionally, the total test score for high school is calculated as a weighted sum of written scores and oral scores. Weights of 0.6 for written points and 0.4 for oral points reflect their relative contributions of score points to the total assessment.

Given the low reliability of domain scale scores for grades three through eight and the high classification accuracy and consistency for two achievement levels, ETS recommends reporting two achievement levels but not reporting scale scores at the domain level. For high school, ETS recommends reporting both scale scores and achievement levels at the composite level. The proposed composites—written literacy and oral literacy—are supported by model-fit indices, which show a better fit for a two-factor model with these composites treated as distinct factors. The high correlation between the reading and writing domains (0.932) further supports this grouping. The correlation between speaking and listening is moderate (0.652); however, calibrating or reporting these two domains separately is not feasible because of low reliability. Combining the two domains is the best option. This approach simplifies the logic of domain grouping, making it more accessible and understandable to the public. Additionally, if preferred, it is feasible to report domain-specific achievement levels for high school as well.

The score reporting structure of the expanded CSA is listed in table 14.

Table . Recommended Score Reporting Structure

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Score** | **Grades 3–8 Achievement Levels** | **High School Achievement Levels** | **Grades 3–8 Scale Scores** | **High School Scale Scores** |
| Domain | 2 levels | 2 levels | No | No |
| Composite | N/A | 3 levels | N/A | Yes |
| Overall | 3 levels | 3 levels | Yes | Yes |

## References

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238–46.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, *21*(2), 230–58.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.

Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics*, *33*, 204–29.

Haberman, S. J., & Sinharay, S. (2010). Reporting of subscores using multidimensional item response theory. *Psychometrika*, *75*, 209–27.

Haberman, S. J., Sinharay, S., Feinberg, R. A., & Wainer, H. (2024). *Subscores: A practical guide to their production and consumption*. Cambridge University Press.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*(1), 1–10.

Wainer, H., Vevea, J. L., Camacho, F., Reeve, B. B., Rosa, K., Nelson, L., Thissen, D. (2001). Augmented scores: “Borrowing strength” to compute scores based on small numbers of items. In D. Thissen & H. Wainer (Eds.), *Test Scoring* (pp. 343–87). Mahwah, NJ: Lawrence Erlbaum.

Yen, W. M. (1987, June). *A Bayesian/IRT index of objective performance*. Paper presented at the annual meeting of the Psychometric Society, Montreal, QC.

Yen, W. M., Sykes, R. C., Ito, K., & Julian, M. (1997, April). *A Bayesian/IRT index of objective performance for a test with mixed-item types*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.