DESE ACCESS Score Imputations

Sarah Jo Torgrimson, PhD

Table of Contents

[Overview 2](#_Toc171339040)

[Student Eligibility 2](#_Toc171339041)

[Imputing Methods 3](#_Toc171339042)

[Average Score Method 4](#_Toc171339043)

[Reweighting Method 5](#_Toc171339044)

[Method Comparison 5](#_Toc171339045)

[Bivariate correlations 7](#_Toc171339046)

[Validity statistics 7](#_Toc171339047)

[Results. 8](#_Toc171339048)

[Bivariate Correlations 8](#_Toc171339049)

[Predictive Validity Computations 9](#_Toc171339050)

[Conclusions 10](#_Toc171339051)

[Appendix A 11](#_Toc171339052)

[References 12](#_Toc171339055)

# Overview

In 2019, WIDA released a technical report detailing four methods states could use to impute scores for eligible students with IEPs or 504 plans (Porter, Cook, & Sahakyan, 2019). WIDA did not endorse any particular method at that time. DESE ran an internal investigation of the uses of two of the four methods, concluding that the average score method was easier to compute and similar in accuracy to the Z-score method (Flanagan & Wiener, 2019). Starting with the 2020 ACCESS administration, DESE adopted the average score method for imputing scores for eligible students. In 2020, WIDA released an additional report detailing a less computationally intensive method to impute scores, the reweighting method (Sahakyan, 2020). As of 2024, DESE still uses the average score method to impute scores, and WIDA offers a tool for states to impute scores based on the reweighting method.

The purpose of this paper is to revisit DESE’s imputation method by comparing our current average score method to the reweighting method provided by the WIDA consortium. Results indicate the reweighting method would have stronger validity and be computationally easier than the current average score method for imputing missing overall scores.

# Student Eligibility

DESE considers an eligible EL student with an assigned proficiency level to be proficient when their overall ACCESS proficiency level is 4.2 or higher. These eligible EL students do not need to meet the typical composite Literacy score of Level 3.9 in order to transition out of multilingual learner education (MLE), due to the methods used to assign missing scores.

EL students must meet **all four** eligibility criteria for DESE to impute a missing overall score:

* The EL student has an IEP or 504 plan that lists the accommodation(s) used by the student.
* The EL student’s disability status and nature of disability have been documented in the Student Information Management System (June SIMS).
* The untested domain test was designated with a “Do Not Score” code of “SPD” on the ACCESS test.
* The EL student did not participate in the domain test because they met one of the test exemption criteria shown in Table 1.

**Table 1. Test exemption criteria for ACCESS domain tests**

|  |  |
| --- | --- |
| **Test Exemption Criteria** | **Explanation** |
| The student’s primary disability is reported as “Sensory: Hard of Hearing or Deaf.” | The student is not required to take the Listening and Speaking tests because they are unable to listen to and/or respond verbally to test items. The student must take the Reading and Writing tests. |
| The student’s primary disability is reported as “Sensory: Vision Impairment or Blind.” | Students who read braille are not required to take the Speaking test, due to the use of graphics and other visual stimuli on the Speaking domain test. Students who do not read braille may take the ACCESS online or paper Writing test with in-person human reader and scribe accommodations, but they are not required to take the Speaking, Listening, or Reading domains. |
| The student meets the criteria to receive the ELA read-aloud “special access” accommodation on the MCAS ELA test, and this is already listed in the student’s IEP or 504 plan. | The student is not required to take the Reading test but must take the Listening, Speaking, and Writing tests. |
| The student is nonverbal (or selectively mute) and is therefore unable to participate in the Speaking test. | The student is not required to take the Speaking test if they do not use alternate and augmentative communication (AAC) but must take the Listening, Reading, and Writing tests. |

# Imputing Methods

In 2019, WIDA published a technical report detailing four methods to impute overall proficiency scores on ACCESS and Alternate ACCESS for students with IEPs and 504 plans missing one or two domains. A comparison of method assumptions, benefits, and drawbacks taken from their report is shown in Table 2.

**Table 2: Imputation Model Assumptions, Benefits, and Drawbacks (Porter, et al. 2019)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Assumptions** | **Benefits** | **Drawbacks** |
| **Designate exit score(s)** | Missing domain score should be assumed as met. | Easiest model to apply.  Easy concept to understand.  Provides assumed missing score benefit. | Likely to provide an inflated higher score. |
| **Apply average observed domain score(s)** | High correlation between domain scores. | Easy to apply.  Easy concept to understand.  State can easily calculate with its own data. | If correlations are not high, scores may be too high or low.  Requires lookup table. |
| **Assign average Z-score** | Normal observed domain score distribution and high correlation between domain scores | Frequently observed assumption about domain score distributions | Complex.  Requires multiple look-up tables. |
| **Conduct a standard setting** | State experts are better decision-makers than score distribution calculations. | Often accepted by peer reviewers.  Uses local, state-specific experts and criteria. | Requires standard setting expertise.  Labor intensive.  Requires substantial financial investment. |

In 2019, DESE conducted an internal report comparing the average score method and Z-score method of imputing scores and found that both methods had similar accuracies. The average score method had higher sensitivity and the Z-score method had higher correlations with observed (earned) scores. DESE concluded that the average score method provided similar predictive validities and decided to adopt it due to its relative computational ease (Flanagan & Wiener, 2019).

## Average Score Method

The average score method is a three-step process:

1. An imputed score is estimated for up to two missing domains using an average of available scaled scores.
2. An overall score is imputed using WIDA’s standard weighting of domain scores (35% of Reading + 35% of Writing + 15% of Listening + 15% of Speaking).
3. An imputed overall proficiency level is assigned using a conversion table.

In 2020, WIDA released an additional report with a simple reweighting method to impute missing overall scores; this report was in response to state feedback detailing the computational complexity of imputation methods. The 2020 WIDA report showed that the reweighting method was highly correlated with observed scores and imputed scores from the Z-score method detailed in their first report (Sahakyan, 2020). WIDA now provides states with an Excel tool to calculate missing overall scores and levels simply by copying and pasting data.

## Reweighting Method

The reweighting method is a two-step process:

1. And overall score is imputed using a reweighting matrix (see Table 3 below; weights are presented as percents).
2. An imputed overall proficiency level is assigned using a conversion table.

**Table 3. Reweighting of domains for students with missing test scores in one or two domains**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Missing Domains** | | **Reading** | **Writing** | **Listening** | **Speaking** |
| **No Missing Domains (Standard Weighting)** | | **35** | **35** | **15** | **15** |
| **One Missing Domain** | **Reading** | N/A | 46 | 27 | 27 |
| **Writing** | 46 | N/A | 27 | 27 |
| **Listening** | 40 | 40 | N/A | 20 |
| **Speaking** | 40 | 40 | 20 | N/A |
| **Two Missing Domains** | **Reading and Writing** | N/A | N/A | 50 | 50 |
| **Reading and Listening** | N/A | 70 | N/A | 30 |
| **Reading and Speaking** | N/A | 70 | 30 | N/A |
| **Writing and Listening** | 70 | N/A | N/A | 30 |
| **Writing and Speaking** | 70 | N/A | 30 | N/A |
| **Listening and Speaking** | 50 | 50 | N/A | N/A |

# Method Comparison

In this paper, we compare DESE’s current average score method to the reweighting method in terms of validity and relation to actual observed scores; using imposed missing data. Missing domain(s) were imposed for students with complete data and imputed overall scores were compared to their actual earned overall English proficiency scaled score and level. A total of 115,765 students had scores in all four domains of the ACCESS test for the 2024 administration. Table 4 shows the count by grade.

**Table 4. Number of Students with Complete 2024 ACCESS Data by Grade**

|  |  |
| --- | --- |
| **Grade** | **Number of Students** |
| K | 12,739 |
| 1 | 13,022 |
| 2 | 13,270 |
| 3 | 11,906 |
| 4 | 10,486 |
| 5 | 7,877 |
| 6 | 6,533 |
| 7 | 6,914 |
| 8 | 6,857 |
| 9 | 8,208 |
| 10 | 7,166 |
| 11 | 6,365 |
| 12 | 4,422 |

We ran two data scenarios for imputation method comparison:

1. We imposed one missing domain (Reading).
2. We imposed two missing domains (Listening and Speaking).

Domains were chosen based on the highest incidence of missingness in 2024 (see Table 5).

**Table 5. Number of Students Missing One or Two Domains by Domain**

|  |  |  |
| --- | --- | --- |
| **Missing One Domain** | | |
| **Domain** | **Count** | **Percent** |
| Listening | 2 | 0.13% |
| Speaking | 28 | 1.86% |
| Reading | 1471 | 97.61% |
| Writing | 6 | 0.40% |
| **Total** | **1507** |  |
| **Missing Two Domains** | | |
| **Domains** | **Count** | **Percent** |
| Listening and Speaking | 50 | 65.79% |
| Listening and Reading | 5 | 6.58% |
| Speaking and Reading | 13 | 17.11% |
| Reading and Writing | 8 | 10.53% |
| **Total** | **76** |  |

For each scenario, both the average score method and reweighting method were applied to estimate an overall scaled score of English proficiency. To evaluate the validity of estimated scores, actual earned overall scaled scores were compared to those estimated by each imputation method. This was done using two different analytic methods, replicating those used by the 2019 internal DESE analysis (Flanagan & Wiener, 2019).

Bivariate correlations between scores estimated by methods (average score, reweighting) and earned scores were conducted using Pearson’s r.

Validity statistics for each method were calculated using a model developed for evaluating diagnostic screeners (Akobeng, 2006). Validity statistics include sensitivity, specificity, positive predictive validity (PPV), and negative predictive validity (NPV). These statistics generally represent accuracy and error rates associated with using a screener to diagnose a condition. In this paper, we use validity statistics to represent accuracy and error rates associated with using an imputation method to predict earned scores. Table 6 shows the relation between imputed proficiency levels and earned proficiency levels (a dichotomy of proficient or not proficient as established by DESE’s criterion of 4.2 overall level). Within the matrix, students are represented by the alignment between earned and imputed proficiency levels. For example, box A would be the number of students that had an imputed proficiency level at or above 4.2 and an earned proficiency level at or above 4.2, indicating a “true” positive.

**Table 6. Model of Imputation Accuracy adapted from Trevethan, 2017**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | **Earned Overall Level** | |  |
| **Proficient** | **Not Proficient** | ***Validity Statistic*** |
| **Imputation Method** | **Proficient** | True Positive (A) | False Positive (B) | ***PPV*** |
| **Not Proficient** | False Negative (C) | True Negative (D) | ***NPV*** |
|  | ***Validity Statistic*** | ***Sensitivity*** | ***Specificity*** |  |

The validity statistics were then calculated based on values in boxes A, B, C, and D using the following formulas:

* PPV=a/(a+b)\*100
* NPV=d/(d+c)\*100
* Sensitivity=a/(a+c)\*100
* Specificity=d/(d+b)\*100
* Accuracy=(a+d)/(a+b+c+d)\*100

Higher values indicate greater validity. PPV and NPV indicate the likelihood that the imputed method accurately identifies students that are or are not proficient in English, whereas sensitivity and specificity indicate the extent to which an imputation method is aligned to a reference criterion (in this case overall score of 4.2). Accuracy provides a holistic metric for understanding test validity. These metrics should be considered together when evaluating the validity of an imputation method. However, PPV and NPV are more appropriate metrics to consider in terms of making decisions about individual student’s English proficiency; compared to sensitivity and specificity (Trevethan, 2017).

# Results

## Bivariate Correlations

Both the average score method and the reweighting method produced estimated overall scaled scores that were highly correlated with each other and with the actual earned overall scores (see Table 7). The reweighting method had a slightly higher correlation with earned scores than the average score method did. Similarly, when correlating estimated overall proficiency levels with the earned proficiency level, the reweighting method had a slightly stronger correlation with earned overall proficiency level (see Table 8). These trends were similar for both estimating overall scaled scores with one or two missing domains.

**Table 7. Correlation Matrix between Earned and Imputed Overall Scaled Scores**

|  |  |  |  |
| --- | --- | --- | --- |
| **One Missing Domain** | | | |
|  | **Average Score** | **Reweighting** | **Earned** |
| **Average Score** | 1.000 | 0.988 | 0.957 |
| **Reweighting** | 0.988 | 1.000 | 0.964 |
| **Earned** | 0.957 | 0.964 | 1.000 |
| **Two Missing Domains** | | | |
|  | **Average Score** | **Reweighting** | **Earned** |
| **Average Score** | 1.000 | 0.995 | 0.966 |
| **Reweighting** | 0.995 | 1.000 | 0.976 |
| **Earned** | 0.966 | 0.976 | 1.000 |

**Table 8. Correlation Matrix between Earned and Imputed Overall Proficiency Levels**

|  |  |  |  |
| --- | --- | --- | --- |
| **One Missing Domain** | | | |
|  | **Average Score** | **Reweighting** | **Earned** |
| **Average Score** | 1.000 | 0.974 | 0.929 |
| **Reweighting** | 0.974 | 1.000 | 0.944 |
| **Earned** | 0.929 | 0.944 | 1.000 |
| **Two Missing Domains** | | | |
|  | **Average Score** | **Reweighting** | **Earned** |
| **Average Score** | 1.000 | 0.990 | 0.941 |
| **Reweighting** | 0.990 | 1.000 | 0.955 |
| **Earned** | 0.941 | 0.955 | 1.000 |

## Predictive Validity Computations

Counts of students in each imputation method alignment category are outlined in Table 9. Rows indicate imputation methods. Columns indicate number of missing domains. Four values are in each outlined box, corresponding to the imputation method and the number of missing domains (e.g., average score, one missing domain).

**Table 9. Imputation Accuracy by Missing Domain**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | | | **Earned Overall Level** | | | |
| **Proficient** | | **Not Proficient** | |
|  |  |  | One Domain | Two Domains | One Domain | Two Domains |
| **Imputation Method** | **Proficient** | **Average Score** | 15,528 | 12,886 | 12,386 | 3,416 |
| **Reweighting** | 13,720 | 12,229 | 3,785 | 2,327 |
| **Not Proficient** | **Average Score** | 785 | 3,427 | 87,851 | 99,463 |
| **Reweighting** | 2,593 | 4,084 | 98, 260 | 101,209 |

Table 10 shows validity statistics for each imputation method by the number of missing domains . In terms of overall accuracy, the average score and reweighting methods were similar in terms of predicting earned scores for two missing domains. However, the reweighting method was more accurate for predicting overall scores for one missing domain. For both one and two missing domains, reweighting showed a higher PPV value. This difference was larger for the one missing domain analysis than the two missing domains analysis. NPV values were slighting higher for average score method in the one missing domain analysis but similar for the two missing domains analysis. Sensitivity was higher in the average score method, but specificity was higher in the reweighting method.

**Table 10. Predictive Validity Statistics by Imputation Method and Missing Domain**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **One Missing Domain** | | **Two Missing Domains** | |
| **Statistic** | **Average Score** | **Reweighting** | **Average Score** | **Reweighting** |
| PPV | 0.56 | 0.78 | 0.79 | 0.84 |
| NPV | 0.99 | 0.97 | 0.97 | 0.96 |
| Sensitivity | 0.95 | 0.84 | 0.79 | 0.75 |
| Specificity | 0.88 | 0.96 | 0.97 | 0.98 |
| Accuracy | 0.89 | 0.95 | 0.94 | 0.95 |

# Conclusions

The average score and reweighting methods were highly correlated and relatively similar in predictive validity for the two missing domain analysis. However, the reweighting method had slightly stronger correlations with earned overall scores and showed higher overall accuracy in the one missing domain analysis (representative of 93% of data missingness in 2023). Results suggest an adoption of the reweighting method would be more valid for imputing missing overall scores than the current average score method. In addition, the reweighting method is computationally easier than the average score method and is aligned with the practices of the WIDA consortium.

# Appendix A

## Correlations between Earned and Imputed Overall Score by Method and Grade: One Missing Domain

|  |  |  |
| --- | --- | --- |
| **Grade** | **Average Score** | **Reweighting** |
| K | 0.939 | 0.958 |
| 1 | 0.935 | 0.965 |
| 2 | 0.950 | 0.971 |
| 3 | 0.958 | 0.975 |
| 4 | 0.965 | 0.979 |
| 5 | 0.968 | 0.980 |
| 6 | 0.958 | 0.973 |
| 7 | 0.961 | 0.977 |
| 8 | 0.968 | 0.980 |
| 9 | 0.959 | 0.974 |
| 10 | 0.958 | 0.974 |
| 11 | 0.959 | 0.974 |
| 12 | 0.951 | 0.969 |

## Correlations between Earned and Imputed Overall Score by Method and Grade: Two Missing Domains

|  |  |  |
| --- | --- | --- |
| **Grade** | **Average Score** | **Reweighting** |
| K | 0.930 | 0.943 |
| 1 | 0.916 | 0.937 |
| 2 | 0.934 | 0.952 |
| 3 | 0.952 | 0.964 |
| 4 | 0.958 | 0.967 |
| 5 | 0.959 | 0.967 |
| 6 | 0.949 | 0.963 |
| 7 | 0.956 | 0.966 |
| 8 | 0.962 | 0.970 |
| 9 | 0.957 | 0.970 |
| 10 | 0.958 | 0.966 |
| 11 | 0.961 | 0.967 |
| 12 | 0.955 | 0.960 |

# References

Akobeng, A.K (2006). Understanding diagnostic tests 1: sensitivity, specificity and predictive values. *Acta Paediatrica, 96*, 338-341. doi:10.1111/j.1651-2227.2006.00180

Flanagan, K. & Wiener, D. (2019). *Estimating scores for missing subdomains on ACCESS for ELLs Tests (Two Methods)*. DESE.

Porter, T., Cook, H.G., & Sahakyan, N. (2019). Less than four domains: Creating an overall composite score for English learners with individualized education plans. *WIDA Technical Report.*

Sahakyan, N. (2020). Generating imputed\* overall composite scale scores for English learners with disabilities who are missing domain scores in the ACCESS for ELLs Assessment. *WIDA Technical Report.*

Trevethan, R. (2017). Sensitivity, specificity, and predictive values: Foundations, pliabilities, and pitfalls in research and practice. *Frontiers in Public Health, 5(307).* doi:10.3389/fpubh.2017.00307