State of Delaware

End-of-Course (EOC)
2013–2014

Volume 4
Evidence of Reliability and Validity

American Institutes for Research

TABLE OF CONTENTS

APPENDICES

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION AND OVERVIEW OF RELIABILITY AND VALIDITY EVIDENCE

The State of Delaware implemented an online assessment for operational use during the 2010–2011 school year. This new test, referred to as the Delaware Comprehensive Assessment System (DCAS), replaced the paper-and-pencil test, referred to as the Delaware State Testing Program (DSTP). As in the previous school year, students who enrolled in various grades in public and charter schools were required to take the online assessment in school year 2011–2012, unless they had a special accommodation that called for a paper-and-pencil version of the assessment. The end-of-course (EOC) assessments were introduced for use as part of the DCAS beginning in the 2011–2012 school year. These assessments were offered at the end of the fall and spring semesters.

Both reliability and validity evidence are necessary to support appropriate inferences from the EOC test scores. This volume provides empirical evidence about the reliability and validity of the 2013–2014 EOC assessments, given its purported use.

Reliability refers to consistency in test scores. Reliability is directly tied to the standard errors of measurement—the smaller the standard error, the higher the precision of test scores and thus the greater the reliability. In item response theory (IRT), the standard errors of measurement differ across scores—that is, they are conditional on the observed test score. Because precision can be understood from various perspectives, this volume provides empirical data on precision in various ways, including marginal reliabilities, stratified alpha, mean standard errors by performance level, and mean standard errors by reporting subgroups.

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, American Psychological Association, & National Council on Measurement in Education, 1999). Whether sufficient evidence has been presented to support the validity of a test is subject to judgment. Thus, organizing such evidence requires, first, an explicit statement regarding the intended uses of the test scores and, subsequently, evidence that the scores can be used to support these inferences.

The purpose of this volume is to provide empirical evidence to support the following:

- Content validity: Evidence that test forms are comparable across students and that each form aligns with the prioritized state content standards
- Reliability: Evidence that examinees in end-of-course assessments are measured with a very high degree of precision and students are classified as either proficient or not proficient with a high degree of certainty

2. Purpose of Delaware's End-of-Course Assessment

Delaware has redesigned the student testing program to require specific end-of-course (EOC) tests that measure the Delaware content standards. The EOC assessments were introduced for use as part of the DCAS beginning in the 2011–2012 school year. These assessments were offered at the end of the fall and spring semesters. Most EOC tests are intended to measure Delaware content standards. However, Algebra II and Integrated Mathematics III are developed for and aligned to the common core state standards for mathematics (CCSSM).

The Biology assessment is a high school end-of-course exam for science. This is aligned with the expectations of the Delaware science content standards for 10th grade. Algebra II and Integrated Mathematics III are the first statewide assessments developed for and aligned to the CCSSM.

The EOC scores are indications of what students know and are able to do relative to the expectations by the end of the corresponding courses, rather than measurements of the breadth of the entire state content standards. However, no consequences of EOC test scores are linked to high-stakes uses at any level.

Post-equating was used in the first operational year. However, pre-equating has been used in subsequent years so that test scores can be provided to students immediately upon completion of a test, given the existing item parameters. Teachers can also have access to the scores for each student in an easy-to-access electronic data warehouse, including total scores and performance levels.

Unlike computer-adaptive tests (CAT), the end-of-course assessments are linear testing where test items are selected prior to the test administration so that matching to test blueprints is ensured. The multiple forms within a course were developed in such a way that they conform to the test blueprints. Matching test forms to the test blueprint is important to establish validity evidence that the measured construct is the same across the test forms for students.

3. RELIABILITY

Reliability refers to the consistency in test scores. To assess the level of precision of the scores in Delaware end-of-course assessments, a variety of test statistics were computed including reliability coefficient, standard error of measurement, and test characteristics curves. The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) defines reliability as “the consistency of [such] measurements when the testing procedure is repeated on a population of individuals or groups” (p. 25). Cronbach’s coefficient alpha (Cronbach, 1951) is widely used as an estimate of reliability that models internal consistency based on the average inter-item correlation.

The formula for Cronbach’s alpha is

$$\alpha = \frac{n}{n-1}\left[1 - \frac{\sum_{i=1}^{n} \sigma_i^2}{\sigma_x^2}\right]$$

where *n* is the number of items *i*, σ_i^2 is the variance of score on item *i*, and σ_x^2 is the variance of total score in the test.

The fixed forms in all EOC tests contain mixed item types—machine-scored constructed-response and multiple-choice. In these cases, it is appropriate to report the stratified Cronbach alpha coefficient as an estimate of reliability. Each item type component (constructed-response, multiple-choice) is treated as a subtest when the stratified coefficient alpha is computed as an estimate of reliability. The reliability is computed separately for each component and combined as follows:

$$\text{stratified } \alpha = 1 - \frac{\sum_{i=1}^{k} \sigma_i^2 (1 - \alpha_i)}{\sigma_x^2}$$

where α_i is the reliability of the *i*th strata, σ_i^2 is the variance between items in the *i*th strata, and σ_x^2 is the variance between all items on the test.

Table 1 presents the stratified alpha coefficients for each EOC fixed-form test in year 2013–2014. Due to very small sample size taking each form in Biology, reliability coefficients are not reported here. The reliability coefficients across forms in each course are similar, ranging from 0.87 to 0.90. These reliabilities suggest that the error variance accounts for about 10% to 13% of the total variance in test scores.

Table 1: Stratified Alpha Coefficients in EOC Assessments

Course	Year 2013–2014		
	Form A	**Form B**	**Form C**
U.S. History	0.90	0.89	0.90
Algebra II	0.87	0.89	0.89
Integrated Mathematics III	0.87	0.87	0.88

3.1 STANDARD ERROR OF MEASUREMENT

The standard error of measurement (SEM) provides the information regarding the accuracy of an examinee's obtained score. Since no test provides a perfect measure of an examinee's ability, the SEM represents the amount of expected variability in an examinee's test score due to imprecision of the test. For example, if a scale score of 400 has an SEM of 10, it can be interpreted as follows: If the examinee were tested again, he or she would be likely to obtain the scale score in the range of 390–410 about 68% of the time.

Figures 1 through 4 are plots of the conditional standard error of measurement for EOC scale scores obtained based on 2013–2014 operational data. The vertical lines denote the cut scores for the *Below Standard*, *Meets Standard*, and *Advanced* performance categories on each test.

Figure 1:
Conditional Standard Errors of Measurement in EOC Biology (2013–2014)

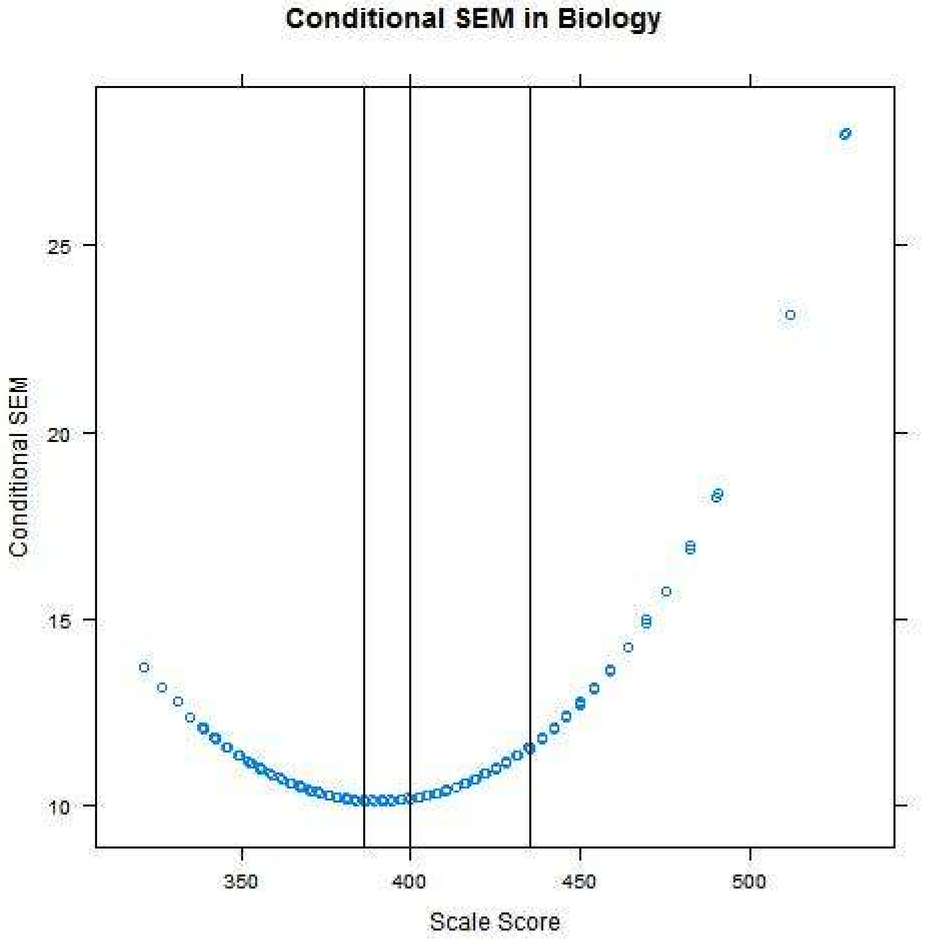


Figure 2:
Conditional Standard Errors of Measurement in EOC U.S. History (2013–2014)

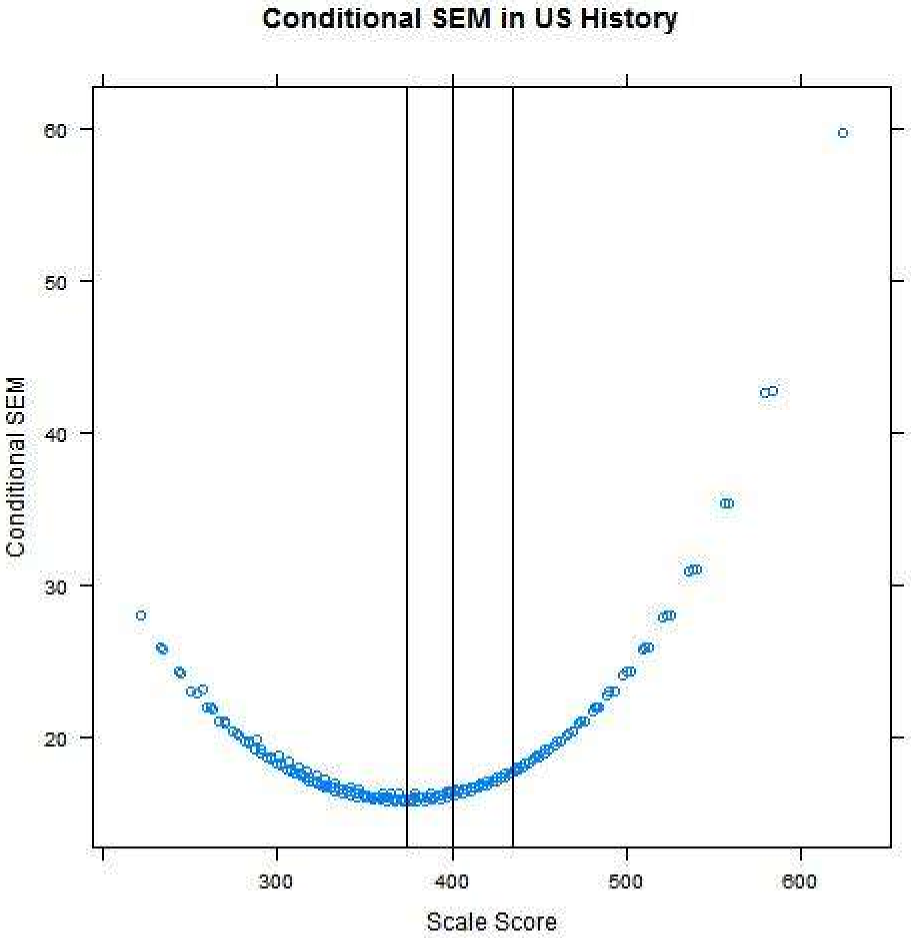


Figure 3:
Conditional Standard Errors of Measurement in EOC Algebra II (2013–2014)

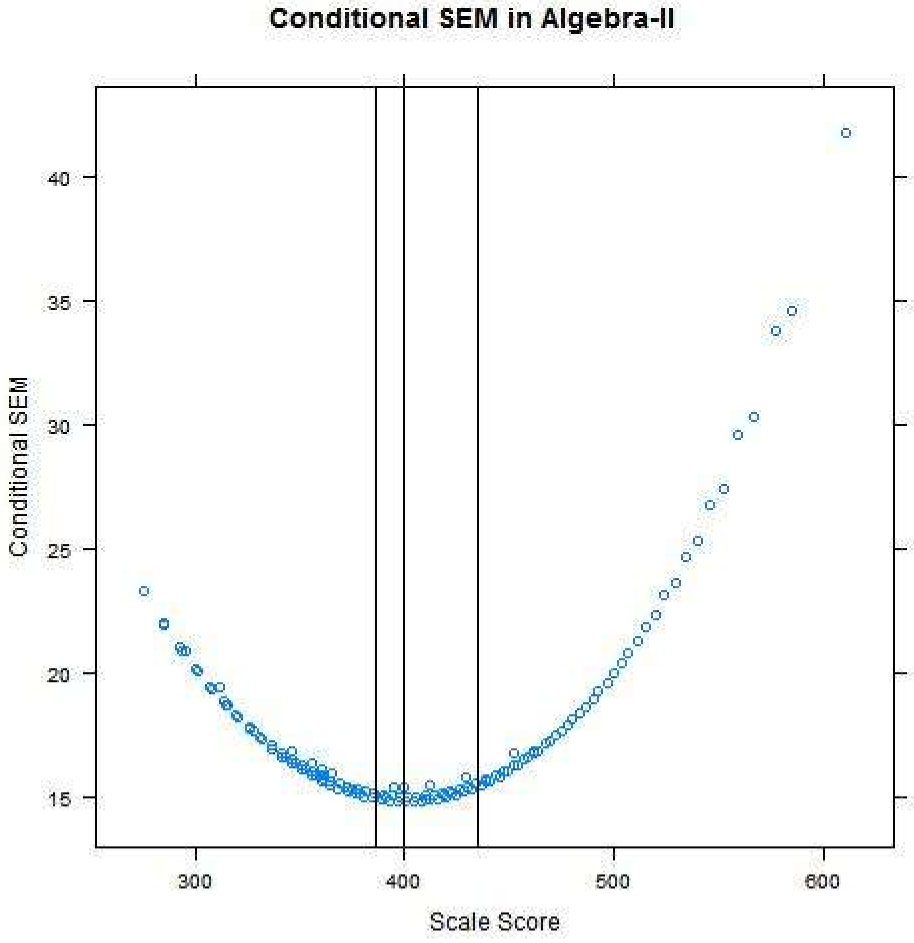
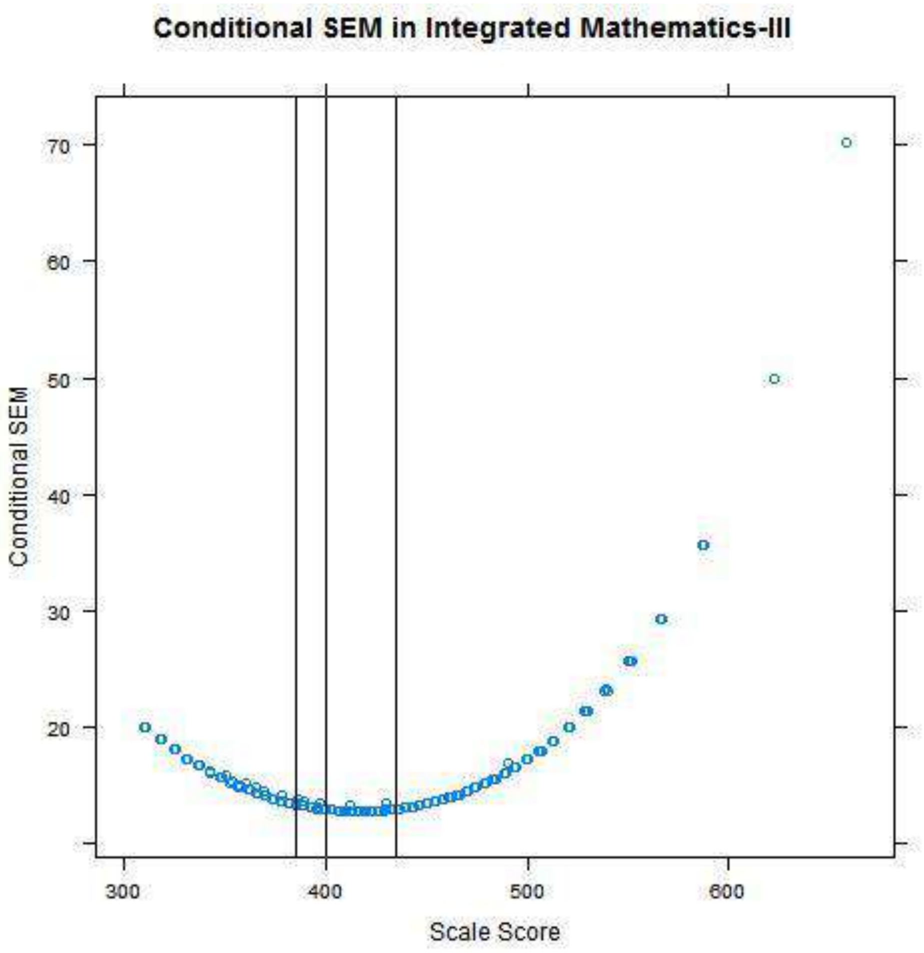


Figure 4:
Conditional Standard Errors of Measurement in EOC Integrated Mathematics III (2013–2014)



The standard error plots suggest that students are measured with a high degree of precision. In fact, the typical standard error for all tests approaches the theoretical lower limit for each test. However, we do observe larger standard errors at the higher ends of the score distribution relative to the lower ends. This occurs because, due to a limited item bank, the items selected for the end-of-course forms currently lack items that are better targeted toward these higher-achieving individuals. Content experts use this information to consider how to further target and populate item pools for any new form development.

Table 2 provides the standard deviation of scale score and mean standard error of measurement for all EOC tests in 2013–2014.

Table 2: Standard Deviation and Mean SEM of Scale Scores in EOC

Course	**Year 2013–2014**	
	Standard Deviation	**SEM**
Biology	34.1	11.0
U.S. History	52.0	17.4
Algebra II	48.7	16.2
Integrated Mathematics III	39.0	14.1

Table 3 shows the mean standard error of measurement of scale scores for students scoring within each of the EOC performance levels by course. The results are presented for all operational EOC tests in 2013–2014.

Table 3: Mean Standard Errors of Scale Scores in EOC

Course	**Well Below Standard**	**Below Standard**	**Meets Standard**	**Advanced**
Year 2013–2014				
Biology	10.8	10.1	10.6	14.3
U.S. History	16.8	16.1	16.9	20.3
Algebra II	16.4	15.0	15.1	17.5
Integrated Mathematics III	14.5	13.2	12.9	15.8

The test reliability information is also provided at the reporting subgroup levels. Appendix A of this volume provides the stratified reliability disaggregated by the subgroups for the courses with adequate sample size.

3.2 Reliability of Achievement Classification

Student performance on EOC tests is also reported in four performance categories: Well Below Standard, Below Standard, Meets Standard, and Advanced. Volume 3 of the standard-setting technical report from years 2011–2012 and 2012–2013 provide detailed information about the standard-setting process, methodology, and results. These cut scores are used to classify student scores into different performance levels.

The precision of classifying students as either proficient or not proficient (i.e., misclassification probabilities) is reported in Volume 1, Section 7.1.

4. Evidence of Content Validity

This section demonstrates that the knowledge and skills covered by Delaware EOC assessments are representative of the prioritized content standards of the larger knowledge domain. We describe the process of setting the content standards for EOC and the test development process that maps the EOC tests to these prioritized standards.

4.1 Content Standards

The EOC assessments were introduced for use as part of the DCAS beginning in the 2011–2012 school year. The terminology used in referring to the content standard varies by subjects. For example, the highest level in the content hierarchy is referred to as "Standard" in Algebra II, Integrated Mathematics III, and U.S. History; "Reporting Category" is used in Biology. For DCAS, these levels are also used for reporting purposes. Since the EOC tests focus on certain content standards, no subscores are reported. The content categories for each subject for the 2013–2014 EOC are presented in Table 4.

Table 4: Content Standards for Delaware EOC Assessments

Subject	Content Standards
Biology	Life Science Earth Science Physical Science
U.S. History	Civics Economics Geography History
Algebra II	Algebra Functions Geometry Number and Quantity Statistics and Probability
Integrated Mathematics III	Algebra Functions Geometry Statistics and Probability

4.2 Test Specifications

Test specifications are blueprints developed to ensure that the test and the items are aligned to the prioritized standards they are intended to measure. The EOC tests measure the corresponding course specifications. Multiple-choice (MC) and machine-scored constructed-response (MSCR) items are used. For both the linear tests and the fixed forms, the blueprint specifies the percentage of operational items that must be administered. The blueprints also include the minimum and maximum number of items for each content standard. The test specifications for each EOC test are presented in Volume 2, Section 1.1 of school year 2012-2013

4.3 TEST DEVELOPMENT

4.3.1 Item Sources

The operational items used in end-of-course tests came from the field-test items from previous years. Items for the 2011 independent field tests and spring 2012 embedded field tests came from three different sources: items that had been used in the previous assessment program (DSTP), newly developed items, and some items that were shared with other states based on the contract. The embedded field-test items of spring 2013 were newly developed items. The embedded field-test items of spring 2014 came from two sources. In EOC mathematics, they came either from the previous year (due to no field-test slots available during the 2013 field test window) or from items that did not survive spring 2013 data review and were suggested to re-field test with or without edits. In U.S. History, the field-test items were either newly developed items or were re-field tested for a variety of reasons.

To develop new items, content experts from the contractor, American Institutes for Research (AIR), Delaware Department of Education (DDOE), and Delaware educators worked together to review and select items for the field test and then add new items to the item pool. This section provides an overview of the test development process. See Section 2 in Volume 2 of 2012–2013 technical reports for a more detailed description of the new item development procedure, including the criteria used in passage selection and item writing and the quality control and review process.

4.3.2 Development of New Items

New items are originally developed by content specialists at AIR. Those items are reviewed and revised as needed and pass through an extensive review process, including various AIR content and editorial reviews, a DDOE content review, a bias review by the Fairness and Sensitivity Committee, and a content review by the Content Advisory Committee (CAC) before they can be included in the field-test pools.

Items that survive AIR internal reviews are sent to DDOE for review. DDOE content and assessment experts review each item and render a decision: items are accepted or sent back for revisions and then reviewed again, or rejected. AIR content experts and DDOE staff discuss suggested revisions and come to an agreement for revision.

Following the completion of the AIR and DDOE internal reviews, the items are reviewed by two Delaware committees: the CAC and the Fairness and Sensitivity Committee.

The CAC consists of Delaware classroom teachers for each subject area. The primary responsibility of the committee members is to ensure that the items are based on defensible content and are free from such flaws as inappropriate readability level, ambiguity, multiple answer keys, and unclear instructions. These items are accepted in content, approved with modifications, or rejected.

Items that have passed through CAC review are then reviewed by the Fairness and Sensitivity Committee. This committee specifically reviews items for potential bias and controversial content and attempts to identify any items that are likely to present problems for specific groups of Delaware students. The Fairness and Sensitivity Committees consist of Delaware educators who are selected to ensure geographic and ethnic diversity. The committee ensures that items

- present racial, ethnic, and cultural groups in a positive light;
- do not contain controversial, offensive, or potentially upsetting content;
- avoid content familiar only to specific groups of students because of gender, race or ethnicity, class, and/or geographic location;
- aid in the elimination of stereotypes; and
- avoid using words or phrases that have multiple meanings.

DDOE and AIR reject or edit items based on the committee feedback. Items that are approved by both of these committees will advance to be field-tested.

After the field test is completed, members of the rubric validation committee review the responses provided to every machine-scored constructed-response (MSCR) item and either approve the scoring rubric or suggest a revision based on their professional judgments. More details on the review process of these various committees are provided in Volume 2, Section 2.

4.4 Summary of Alignment Studies

Item alignment is an integral component of test development. An alignment study reviews and determines the degree to which the test and the standards set are in agreement and support student learning of intended expectations. The section summarizes the same information presented in last year’s technical reports regarding the alignment studies. To maintain objectivity, the alignment study that evaluates the alignment of EOC (Biology and U.S. History) item banks to the content standards was completed by an independent contractor, Dr. John L. Smithson from the University of Wisconsin-Madison.

The EOC assessments were compared to the Delaware Assessment Framework, which was drawn from relevant state content standards and grade-level expectations. In order to determine form-to-form alignment within the same course, two forms of each EOC assessment were analyzed for each course. Additionally, an alignment study was conducted for Algebra II and Integrated Mathematics III in summer 2013.

In summary, all EOC assessments appear to be well aligned to the EOC framework. Similarly, the English II EOC assessments also appear to be well aligned to the CCSS. Both Algebra and Integrated Mathematics EOC assessments are more narrowly focused than the CCSS, though these assessments are well aligned to Algebra content described in the CCSS.

The results from the alignment study on Algebra II and Integrated Mathematics III suggest that the tests forms were found to possess an adequate representation to the CCSSM and the College Board High School Mathematics Standards (CBHSM). In addition, the study also found potential areas for improvement. Researchers recommended reconsidering current test specifications and implementing a stronger emphasis on statistics and less emphasis on functions.

Section 2.4 in Volume 2 of 2012-2013 EOC Technical Report provides details on two alignment studies. For detailed information about these studies, refer to the corresponding alignment study report included in Appendix A of Volume 2.

5. Evidence of Comparability

When multiple test forms are constructed, it is important to provide evidence of comparability across test forms. If the content between forms varies, then it will be difficult to justify score comparability. The parallelism across multiple forms depends on content, statistical, and psychometric indicators.

Student scores should not depend on the mode of administration or the type of test form. EOC assessments are primarily online assessment. To improve the accessibility of the statewide assessment, alternate assessments are provided for students with special needs. Thus, the comparability of scores obtained via alternate means of administration must be established and evaluated.

5.1 Comparability of EOC Test Scores Over Time

At the beginning of EOC implementation, all test items were initially calibrated and scaled using the Rasch and generalized partial credit models through an independent field test in spring 2011. The post-equating was conducted for the first operational administration in winter/spring 2012 to update the item pool. In subsequent years, pre-equating has been used so that test scores can be provided immediately to students upon conclusion of a test, and those scores will be derived based on the existing item parameters.

Since all selected item parameters used to construct scores are drawn from a single item bank and are on the same scale of measurement, this ensures that scores between students are also on the same scale and directly comparable. The score comparability exists because each form is equated to every other form through the use of pre-equated items, commonly scaled, from a common item bank.

5.2 Match with Test Blueprints for Both Paper-and-Pencil and Online Tests

For the 2013–2014 EOC tests, the same paper-and-pencil version of the test from the previous year was used. The paper-and-pencil version with a fixed form was developed according to the same test specifications used for the online fixed-form versions. The procedures used to establish comparable fixed forms are provided in Volume 2, Test Development, in the 2012–2013 Technical Reports for End of Course Tests.

One of the important analyses is to determine whether all test forms conform to the same test blueprint, thus providing evidence of content comparability. This is an important validity and fairness issue when test scores are used for high-stake accountability decisions.

For 2013–2014, the paper-and-pencil version of the EOC is a fixed-form test developed according to the test specifications. Since the paper-and-pencil version of the test was constructed by removal of machine-scored constructed-response items from Form A of the online version, some strands do not meet the blueprint. Thus, match-to-blueprint is not 100%. However, the blueprint originally constructed for the EOC assessments did not make such a distinction. Both AIR and DDOE acknowledge the need to provide some details on blueprint specification.

5.3 Comparability of Online and Paper-and-Pencil Test Scores

The paper form is offered as a special accommodation and not as a typical mode of delivery. Both the paper form and online versions of the tests are fixed with respect to the items, and both align to the same test specifications and the same item parameters. Machine-scored constructed-response items do not exist in the paper forms, which is the only difference between these two versions. Consequently, the scores from both forms are on the same scale of measurement and are directly comparable. Standard error plots overlaying the online tests with the paper-and-pencil test for EOC assessments are presented in Appendix B. Note that each EOC test comprises three online fixed forms (Forms A, B, and C) and one paper-and-pencil form. These plots show that the SEM curves are virtually superimposed with each other, showing the comparability of the forms. Also as expected, each test has larger standard errors on the upper and lower end of the scale scores.

5.4 Translation Accuracy from English to Spanish

The test items used in end-of-course assessments were originally developed in English. To meet the needs of some Hispanic students, test items are translated into Spanish by a professional language translation firm for the alternate version in Algebra I, Integrated Mathematics I, Biology, and U.S. History. Three steps were stipulated in the process of translating the items:

1. Translation: All content, including text-related graphics are translated into Spanish.
2. Vendor quality control (QC): The translated item is reviewed and edited for content and/or style, if necessary, by a second translator who did not do the original translation. The English text in graphics is translated at this phase.
3. Vendor sign-off: The translated item, including the translated text in graphics, the format, and the display, goes through one more review with final sign-off by the vendor.

At the QC and sign-off phases, the reviewers have a checklist to follow, including these items:

- All text that needs to be translated has been translated.
- The graphics display properly and are well adapted to the translation.
- The fonts and uppercase/lowercase are consistent.
- The tables/charts are still aligned properly.
- The spacing and disposition of the text are the same as in the English version.

After vendor sign-off, AIR implements an internal review process for the translated items with various review levels. For example, in the Spanish Web Preview/Approval stage, the final review and approval level for the Spanish translation, if a discrepancy is identified, the item is moved back to Spanish Translation step. In other stages, the initial Spanish descriptions (tags) are entered for Spanish text-to-speech (TTS) by AIR bilingual professionals. Then, all Spanish content is reviewed for accuracy to ensure that each item accurately conveys the intent of the English text. If a discrepancy is identified, the item is held at this level for consultation with the Spanish vendor. The Spanish tags are reviewed and modified accordingly before translated items appear in the test.

6. Fairness and Accessibility

6.1 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. Applying the principles of universal design to the development of computer-based testing systems, such as end-of-course assessments, has the potential to increase test validity by removing the barriers to the accurate measure of achievement for students with disabilities and other special needs.

Universal design removes barriers to access for the widest range of students possible. Seven principles of universal design are applied in the process of test development (Thompson, Johnstone, & Thurlow, 2002):

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

Test development specialists receive extensive training on the principles of universal design and apply these principles in the development of all test materials, including tasks, items, and manipulatives. In the review process, adherence to the principles of universal design is verified.

6.2 Statistical Fairness in Item Statistics (DIF)

Differential item functioning (DIF) analysis is an important element in the evaluation of the fairness and validity of educational tests. The operational items used in end-of-course assessments were originally field-tested as an independent field test in spring 2011. Similarly, additional items were field-tested by embedding the items in the operational tests in spring 2012. Following the field test in each of those administrations, DIF analysis was conducted for all field-tested items to detect potential item bias from a statistical perspective across major ethnic and gender groups. DIF analyses were performed for the following groups:

- Male/female
- White/African-American
- Special education/not special education

Items are classified into three categories (A, B, or C) for DIF, ranging from no evidence of DIF to severe DIF according to the DIF classification convention illustrated in Table 5. Items are also categorized as positive DIF (i.e., +A, +B, or +C), signifying that the item favors the focal group (e.g., African-American/black, Hispanic, or female), or negative DIF (i.e.,–A,–B, or–C), signifying that the item favors the reference group (e.g., white or male). Items are flagged if their DIF statistics fall into the "C" category for any group. A DIF classification of "C" indicates that the item shows significant differential item functioning and should be reviewed for potential content bias, differential validity, or other issues that may reduce item fairness, whether the DIF statistic favors the focal or the reference group.

Table 5: DIF Classification Rules

Dichotomous Items	
Category	**Rule**
C	$MH\chi^2$ is significant and $\lvert\hat{\Delta}_{MH}\rvert \geq 1$.5
B	$MH\chi^2$ is significant and $\lvert\hat{\Delta}_{MH}\rvert < 1$.5
A	$MH\chi^2$ is not significant.
Polytomous Items	
Category	**Rule**
C	$MH\chi^2$ is significant and $\lvert SMD\rvert / \lvert SD\rvert \geq .25$.
B	$MH\chi^2$ is significant and $\lvert SMD\rvert / \lvert SD\rvert < .25$.
A	$MH\chi^2$ is not significant.

A detailed description of the DIF analysis performed is presented in Volume 1, Section 5.2. Flagged items went through a two-stage data review process in content—AIR internal review and DDOE review—before they were included in the final item pool for operation. The results from the field-test analysis, including the number of items flagged, and the results from the item data review meeting are also presented in Volume 1, Section 5.

7. SUMMARY

Both reliability evidence and validity evidence are necessary to support appropriate inferences from the test scores. This volume focused on presenting the empirical evidence of the reliability and validity of the 2013–2014 EOC assessments, given its purported use.

The data presented in this report provide empirical evidence on content validity and reliability of test scores.

- Content validity: All the EOC forms were developed with the goal of matching-to-blueprint rates ensuring that content coverage on each form is consistent with test specifications.
- Reliability: The standard error curves show that students are measured with a very high degree of precision, although larger standard errors are observed at the higher ends of the score than at the lower end. Classification accuracy analysis also shows that students are classified as either proficient or not proficient with a high degree of certainty.

8. References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: Author.

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

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments.* (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.