OKLAHOMA STATE DEPARTMENT OF EDUCATION

Janet Barresi, State Superintendent of Public Instruction

# DRAFT

# Oklahoma Modified Alternative Assessment Program

# 2011 Technical Report

# End-of-Instruction Assessments

Submitted to
The Oklahoma State Department of Education
September 2011

PEARSON

# Executive Summary

## Introduction

The Oklahoma School Testing Program (OSTP) is a state-wide assessment program that includes End-of-Instruction (EOI) assessments, where students who complete an area of instruction must also take the corresponding state-wide, standardized assessment. The general Oklahoma Core Curriculum Tests (OCCT) include Algebra I, Algebra II, Geometry, Biology I, English II, English III, and U.S. History. Each test is a measure of a student's knowledge relative to the *Priority Academic Student Skills (PASS)*, Oklahoma's content standards. The Oklahoma Modified Alternative Assessment Program (OMAAP) EOI tests are used to assess student proficiency in Algebra I, Biology I, English II, and U.S. History. The OMAAP is intended for a population of students for whom the general OCCT exams and the Oklahoma Alternative Assessment Program (OAAP) portfolio assessments are inappropriate. The OCCT EOI tests are part of the Achieving Classroom Excellence (ACE) legislation passed in 2005 as amended in 2006, which outlines the curriculum, the competencies, and the testing requirements for students to receive a high school diploma from the state of Oklahoma. The OMAAP EOI exams are based on modified blueprints and items from the corresponding OCCT EOI exams. The OSTP was established to improve academic achievement for all Oklahoma students and it also meets the requirements of the No Child Left Behind Act (NCLB), which was introduced by the Federal Government in 2001. In 2010, Pearson was contracted by the Oklahoma State Department of Education (SDE) to develop, administer, and maintain the OMAAP EOI tests. This report provides technical details of work accomplished for the 2011 tests.

## Purpose

The purpose of this Technical Report is to provide objective information regarding technical aspects of the OMAAP EOI assessments. This volume is intended to be one source of information to Oklahoma K-12 educational stakeholders (including testing coordinators, educators, parents, and other interested citizens) about the development, implementation, scoring, and technical attributes of the OMAAP EOI assessments. Other sources of information regarding the OMAAP EOI tests include the administration manuals, interpretation manuals, student-, teacher-, and parent guides, implementation materials, and training materials.

The information provided here fulfills legal, professional, and scientific guidelines (AERA, APA, & NCME, 1999) for technical reports of large-scale educational assessments and is intended for use by qualified users within schools who use the OMAAP EOI assessments and interpret the results. Specifically, information was selected for inclusion in this report based on NCLB requirements and the following Standards for Educational and Psychological Testing:

- Standards 6.1 – 6.15 Supporting Documentation for Tests
- Standards 10.1—10.12 Testing Individuals with Disabilities
- Standards13.1—13.19 Educational Testing and Assessment

This technical report documented the OMAAP EOI development methods, data analysis, and results as is appropriate for use by qualified users and technical experts. Section 1 provides an overview of the test design, test content, and content standards. Section 2 provides summary information about the test administration. Section 3 details the classical item analyses and reliability results, and Section 4 details the calibration, equating, scaling analyses, and results. Section 5 provides the results of the classification accuracy and

classifications studies and Section 6 provides higher-level summaries of all the tests included in the OMAAP EOI testing program.

Information provided in this report included:

1. Content standards,
2. Content of the tests,
3. Test form design,
4. Administration of the tests,
5. Identification of ineffective items,
6. Detection of item bias,
7. Reliability of the tests,
8. Calibration of the tests,
9. Equating of tests,
10. Scaling and scoring of the tests, and
11. Decision accuracy and classification.

Each of these facets in the OMAAP EOI assessments development and administration cycle is critical to validity of test scores and interpretation of results. This technical report covers all of these topics for the 2011 testing year.

## Table of Contents

Section 1 ..... 1
1.1 Overview of the OMAAP EOI Assessments ..... 1
1.1.a Purpose ..... 1
1.1.b Modified *PASS* Content Standards ..... 2
1.2 Content Modifications ..... 2
1.2.a Modification of Blueprint and Item Specifications ..... 2
1.2.b Modification of Forms and Items ..... 3
1.3 Summary of Test Development and Content Validity ..... 4
1.3.a Aligning Test to Modified *PASS* Content Standards ..... 4
1.3.b Item Pool Development and Selection ..... 4
1.3.c Configuration of the Tests ..... 5
1.3.d Operational Items by Content Area ..... 5
Section 2 ..... 7
2.1 Packaging and Shipping ..... 7
2.2 Materials Return ..... 7
2.3 Materials Discrepancies Process ..... 8
2.4 Processing Assessment Materials Returned by Schools ..... 8
Section 3 ..... 9
3.1 Sampling Plan and Data Receipt ..... 9
3.1.a Sampling Plan ..... 9
3.1.b Data Receipt Activities ..... 9
3.2 Classical Item Analyses ..... 11
3.2.a Test-Level Summaries of Classical Item Analyses ..... 11
3.3 Procedures for Detecting Item Bias ..... 12
3.3.a Differential Item Functioning Results ..... 13
3.4 Test Reliability ..... 13
3.4.a Overall Test Reliability ..... 13
3.4.b Test Reliability by Subgroup ..... 14
3.5 Inter-rater Reliability ..... 15
Section 4 ..... 16
4.1 Item Response Theory (IRT) Models ..... 16
4.1.a Dichotomous Item Response Theory Model ..... 16
4.1.b Polytomous Item Response Theory Model ..... 16
4.2 Assessment of Fit to the Model ..... 16
4.3 Calibration and Equating ..... 17
4.4 Anchor Item Stability Evaluation Methods ..... 17
4.4.a Anchor Items for Spring 2011 ..... 18
4.4.b Results of the Anchor Item Stability Check ..... 18
4.5 Scaling and Scoring Results ..... 18

Section 5 ..... 24
5.1 Classification Consistency and Accuracy ..... 24
Section 6 ..... 27
6.1 Descriptive Statistics ..... 27
6.2 Performance Level Distribution ..... 28
6.3 Conditional Standard Error of Measurement ..... 28
6.4 Standard Error of Measurement ..... 29
References ..... 30
Appendix A ..... 31
Appendix B ..... 37

## List of Tables and Figures

Table 1.1. Oklahoma Content Standards by Subject ..... 2
Table 1.2. OMAAP Item Modification Rules ..... 3
Table 1.3. The OCCT and OMAAP EOI Item Count Comparison ..... 4
Table 1.4. Criteria for Aligning the Test with Modified *PASS* Standards and Objectives..... 4
Table 1.5. Percentage of Items in Depth of Knowledge Levels ..... 5
Table 1.6. Configuration of the OMAAP EOI Tests ..... 5
Table 1.7. Number of Items and Points by Content Standard for Algebra I..... 5
Table 1.8. Number of Items and Points by Content Standard for Biology I ..... 6
Table 1.9. Number of Items and Points by Content Standard for English II..... 6
Table 1.10. Number of Items and Points by Content Standard for U.S. History..... 6
Table 3.1. Demographic Characteristics of Calibration and Equating Sample..... 9
Table 3.2. Secondary Statistical Key Check Criteria ..... 10
Table 3.3. Number of Items Removed from Scoring by Subject ..... 11
Table 3.4. Total Possible Points after Item Review ..... 11
Table 3.5. Test-Level Summaries of Classical Item Analyses ..... 12
Table 3.6. DIF Flag Incidence Across All OMAAP EOI Items..... 13
Table 3.7. Cronbach's Alpha by Subject ..... 14
Table 3.8. Test Reliability by Subgroup ..... 14
Table 3.9. Inter-rater Reliability for English II Operational Writing Prompts..... 15
Table 4.1. Number of Anchor Items per Subject..... 18
Table 4.2: OMAAP Scaling Constants, Scale Range, and Cut Scores by Subject ..... 19
Table 4.3. Raw Score to Scale Score Conversion Table for Algebra I and Biology I..... 20
Table 4.4. Raw Score to Scale Score Conversion Table for English II and U.S. History..... 22
Table 5.1. Estimates of Accuracy and Consistency of Performance Classification..... 25
Table 5.2. Accuracy and Consistency Estimates by Cut Score: False Positive- and False Negative Rates ..... 25
Table 6.1. Descriptive Statistics of the Scale Scores - Overall ..... 27
Table 6.2. Descriptive Statistics of the Scale Scores by Gender ..... 27
Table 6.3. Descriptive Statistics of the Scale Scores by Race/Ethnicity ..... 27
Table 6.4. Percentage of Students by Performance Level ..... 28
Table 6.5. Overall Estimates of *SEM* by Subject ..... 29

# Section 1

## Overview of the Oklahoma Modified Alternative Assessment Program (OMAAP) End-of-Instruction (EOI) Assessments

### 1.1 Overview of the OMAAP EOI Assessments

The Oklahoma Modified Alternative Assessment Program (OMAAP) End-of-Instruction (EOI) assessments are a collection of state-mandated, secondary-level, criterion-referenced tests used to assess student proficiency at the End-of-Instruction in Algebra I, Biology I, English II, and U.S. History. The OMAAP is intended for a population of students for whom the Oklahoma Alternative Assessment Program (OAAP) portfolio assessment and the general Oklahoma Core Curriculum Tests (OCCT) are inappropriate.

The OMAAP EOI tests are used to assess student proficiency relative to the specific set of academic skills established by committees of Oklahoma educators. The OCCT Achieving Classroom Excellence (ACE) EOI exams are aligned with the *Priority Academic Student Skills* (*PASS*). The OMAAP EOI exams are aligned with a modified version of the *PASS* standards. Secondary-level students who have completed instruction in Algebra I, Algebra II, Geometry, Biology I, English II, English III, and U.S. History are mandated to take the corresponding Oklahoma ACE EOI tests; except those who are identified to be eligible for the modified or alternate assessments. Students who are eligible for taking the OMAAP assessments are tested on Algebra I, Biology I, English II, and U.S. History.

In 2009, SDE announced passing the OCCT, OMAAP, or alternate EOI assessments to be part of graduation requirement for the 2009 classes and forwards. The new policy dictated that to graduate with a high school diploma from the State of Oklahoma, students must score proficient or above in Algebra I and English II, and two of the following five: Algebra II, Biology I, English III, Geometry, or U.S. History. Students who fail to earn a proficient score are permitted to retake these tests. Students who take the OMAAP EOI tests must pass all the OMAAP EOI subjects in order to earn a high school diploma.

The OMAAP EOI tests consist exclusively multiple-choice items, except for English II, which includes one writing prompt. The Spring 2011 OMAAP EOI assessments were developed and administered by Pearson in collaboration with the Oklahoma State Department of Education (SDE). Pearson also scored, equated, and scaled the OMAAP EOI assessments. One OMAAP EOI form was administered in Spring 2011 for each subject. In addition, an equivalent form from one of the previous administrations was designated as a breach form and a Braille test was built for each subject. A student could receive an equivalent form for various reasons, including becoming ill during test administration or test security breach. The State Department of Education Office of Accountability and Assessments determines eligibility for an equivalent form on a case-by-case basis.

#### 1.1.a Purpose

The OMAAP EOI assessments were developed to measure the modified *PASS* content standards. The objectives associated with content and/or process standards tested are provided in Appendix A.

### 1.1.b Modified *PASS* Content Standards

The Oklahoma Modified content standards are shown in Table 1.1.

Table 1.1. Oklahoma Content Standards by Subject

| | |
|---|---|
| **Algebra I** | |
| Standard 1. | Number Sense and Algebraic Operations |
| Standard 2. | Relations and Functions |
| Standard 3. | Data Analysis, Probability & Statistics |
| **Biology I** | |
| Process Standards | |
| Process 1. | Observe and Measure |
| Process 2. | Classify |
| Process 3. | Experiment |
| Process 4. | Interpret and Communicate |
| Process 5. | Model |
| Content Standards | |
| Standard 1. | The Cell |
| Standard 2. | The Molecular Basis of Heredity |
| Standard 3. | Biological Diversity |
| Standard 4. | The Interdependence of Organisms |
| Standard 5. | Matter/Energy/Organization in Living Systems |
| Standard 6. | The Behavior of Organisms |
| **English II** | |
| Reading/Literature: | |
| Standard 1. | Vocabulary |
| Standard 2. | Comprehension |
| Standard 3. | Literature |
| Standard 4. | Research and Information |
| Writing/Grammar/Usage and Mechanics: | |
| Standard 1/2. | Writing |
| Standard 3. | Grammar/Usage and Mechanics |
| **U.S. History** | |
| Standard 1. | Civil War and Reconstruction Eras |
| Standard 2. | Impact of Immigration and Industrialization |
| Standard 3. | Imperialism, World War I, and Isolationism |
| Standard 4. | United States During the 1920s and 1930s |
| Standard 5. | World War II |
| Standard 6. | United States Since World War II |

## 1.2 Content Modifications

### 1.2.a Modification of Blueprint and Item Specifications

In comparison to the OCCT ACE EOI, the OMAAP EOI blueprints use similar proportions of items across standards. A standard must have at least six items to serve as a reporting category. Committees of teachers conducted an initial review the *PASS* item specifications and OMAAP EOI items and then the SDE conducted a final review.

### 1.2.b Modification of Forms and Items

Universal- and content-specific OMAAP EOI exam modifications are listed in Table 1.2.

Table 1.2. OMAAP Item Modification Rules

| **Universal** |
|---|
| • Minimize the number of questions on the page (limit to 2)<br>• Use a larger font size<br>• Provide only three answer options instead of four<br>• Highlight the main points in the question or passage by underlining and using boldface<br>• Allow for the same accommodations as in the standard assessment<br>• Avoid questions that require students to select the better/best answer<br>• Eliminate answer choices that give students the option of making "no change" to the item |
| **Algebra I** |
| • Allow for read-aloud and calculators<br>• Unless required by standard, avoid items with negative and positive answer choices that use the same number (e.g., −4 and +4)<br>• Place any items with coordinate grids on one page<br>• Be consistent with qualifiers in the stem and answer choices (e.g., use ml throughout or milliliters throughout)<br>• Avoid questions that use best or closest<br>• Avoid complicated art |
| **Biology I** |
| • Reduce the amount of reading<br>• Avoid complicated art<br>• Simplify tables and charts by removing irrelevant rows or columns<br>• Box formulas to make them stand out |
| **English II – Reading Items** |
| • Display passages in a one column format<br>• Break passages into smaller portions, and place the questions that pertain to the smaller portion |
| **English II – Writing Prompt** |
| • Simplify the question<br>• Update the checklist describing the aspects that will be graded so it matches the new rubric, and simplify it<br>• Use a three-point holistic writing rubric |

Furthermore, the number of items in the OMAAP forms was reduced to between 50-65% of the number of items on the OCCT exams. In addition, there are no field test items on the OMAAP EOI tests. Table 1.3 shows a comparison of item counts across the Spring 2011 OCCT and the OMAAP EOI forms.

Table 1.3. The OCCT and OMAAP EOI Item Count Comparison

| Subject | OCCT Items OP | OCCT Items FT | OMAAP Items (OP) |
|---|---|---|---|
| Algebra I | 55 | 20 | 43 |
| Biology I | 60 | 20 | 46 |
| English II* | 60/1 | 20/0 | 43/1 |
| U.S. History | 60 | 20 | 48 |

Note 1: OP = operational items, FT = field test items.
* The first number represents the count of multiple-choice items and the second number represents the count of constructed-response item.

## 1.3 Summary of Test Development and Content Validity

### 1.3.a Aligning Test to Modified *PASS* Content Standards

In addition to the test Blueprints provided by the SDE, Table 1.4 describes four criteria for test alignment with the modified *PASS* Standards and Objectives.

Table 1.4. Criteria for Aligning the Test with Modified *PASS* Standards and Objectives

| | |
|---|---|
| **1. Categorical Concurrence** | The test is constructed so that there are at least six items measuring each standard with the content category consistent with the related standard. The number of items, six, is based on estimating the number of items that could produce a reasonably reliable estimate of a student's mastery of the content measured. |
| **2. Range-of-Knowledge** | The test is constructed so that at least 50% of the objectives for a standard have at least one corresponding assessment item. |
| **3. Balance-of-Representation** | The test is constructed according to the alignment blueprint, which reflects the degree of representation given on the test to each standard and objective in terms of the percent of total test items measuring each standard and the number of test items measuring each objective. |
| **4. Source-of-Challenge** | Each test item is constructed in such a way that the major cognitive demand comes directly from the targeted skill or concept being assessed, not from specialized knowledge or cultural background that the test-taker may bring to the testing situation. |

### 1.3.b Item Pool Development and Selection

The OMAAP EOI tests were built using modified versions of previously-used OCCT ACE EOI items. To equate the forms across years, a set of items from the prior operational administrations were selected to serve as anchors for the Spring 2011 administration. During assembling of the tests, content experts followed two specifications; the test Blueprint and depth of knowledge (DOK) targets. Table 1.5 provides the DOK level percentages for the OMAAP EOI tests. Note that while all efforts were put forth to meet the specifications, some targets might not be met because of item bank limitation. The limitation becomes more noticeable when items are passage related, such as the English II.

Table 1.5. Percentage of Items in Depth of Knowledge Levels

| DOK Level | Target DOK | Actual | | | |
|---|---|---|---|---|---|
| | | Algebra I | Biology I | English II | U.S. History |
| 1 | 15-20% | 23.26% | 13.04% | 15.91% | 10.42% |
| 2 | 60-70% | 67.44% | 76.09% | 75.00% | 72.92% |
| 3/4 | 15-20% | 9.30% | 10.87% | 9.19% | 16.67% |

### 1.3.c Configuration of the Tests

Table 1.6 provides an overview of the number of test items and maximum possible points for the Spring 2011 OMAAP EOI assessments.

Table 1.6. Configuration of the OMAAP EOI Tests

| | Item Count | | Maximum Possible Points | |
|---|---|---|---|---|
| Content Area | MC | CR | MC | CR |
| Algebra I | 43 | 0 | 43 | 0 |
| Biology I | 46 | 0 | 46 | 0 |
| English II | 43 | 1 | 43 | 3 |
| U.S. History | 48 | 0 | 48 | 0 |

Note: MC = Multiple Choice; CR = Constructed Response.

### 1.3.d Operational Items by Content Area

*Algebra I*. The Algebra I form was comprised of 43 items and 14 of them were anchor items. The number of items and maximum points possible by content standard is shown in Table 1.7. Algebra I scores were reported by content standard and at the objective level. There were eight or more operational items in each reported category. Each item was mapped to one content standard and one objective.

Table 1.7. Number of Items and Points by Content Standard for Algebra I

| Standard | Items | Points |
|---|---|---|
| 1 | 12 | 12 |
| 2 | 23 | 23 |
| 3 | 8 | 8 |
| **Total** | **43** | **43** |

*Biology I*. The Biology I form was comprised of 46 items, of which 44 were mapped to a content standard. There were 16 anchor items. The number of items and the maximum number of points possible by content standard is shown in Table 1.8. Biology I scores were reported for content and process standards at the standard level. Each reported standard has six or more items. Unlike other subjects, all items in Biology I were primarily mapped to process standards. With the exception of safety items, all items were also mapped to content standards.

Table 1.8. Number of Items and Points by Content Standard for Biology I

| Standard | Items | Points |
|---|---|---|
| 1 | 7 | 7 |
| 2 | 7 | 7 |
| 3 | 7 | 7 |
| 4 | 11 | 11 |
| 5 | 6 | 6 |
| 6 | 6 | 6 |
| **Total** | **44** | **44** |

Note: Items classified by content standard do not sum to the total number of items on the form (46) because two items map to process standards, but not content standards.

*English II*. The English II form was comprised of 43 multiple-choice items and 1 open-ended writing prompt. There were 15 anchor items and all of them were multiple-choice items. The number of items and the maximum number of points possible by content standard is shown in Table 1.9. English II scores were reported at the content standard level. Each item was mapped to one content standard and one objective. The writing prompts in English II were scored analytically on five traits with a maximum of four score points per trait. The trait scores were weighted differentially to derive a composite score that ranged from 1 to 3. The composite scores contributed to the English II total score. The scores in the analytic traits were reported in the writing report.

Table 1.9. Number of Items and Points by Content Standard for English II

| Standard | Items | Points |
|---|---|---|
| R1 | 6 | 6 |
| R2 | 12 | 12 |
| R3 | 12 | 12 |
| R4 | 6 | 6 |
| W1/W2 | 1 | 3 |
| W3 | 7 | 7 |
| **Total** | **44** | **46** |

*U.S. History*. The U.S. History form was comprised of 48 items. There were 18 anchor items. The number of items and maximum points possible by content standard are shown in Table 1.10. U.S. History scores were reported at the content standard level and each reported standard had five or more items.

Table 1.10. Number of Items and Points by Content Standard for U.S. History

| Standard | Items | Points |
|---|---|---|
| 1 | 5 | 5 |
| 2 | 7 | 7 |
| 3 | 6 | 6 |
| 4 | 11 | 11 |
| 5 | 7 | 7 |
| 6 | 12 | 12 |
| **Total** | **48** | **48** |

## Section 2

### Administration of the OMAAP EOI Assessments

Valid and reliable assessment requires that assessments are first constructed in alignment with the Oklahoma content standards and then administered and scored according to sound measurement principles. Sound assessment practices require that schools administer all assessments in a consistent manner across the state so that all students have a fair and equitable opportunity for a score that reflects their achievement in each subject.

The schools play a key role in administering the OMAAP EOI assessments in a manner that is consistent with established procedures, monitoring the fair administration of the assessment, and working with the SDE office to address deviations from established assessment administration procedures. The role that district and school faculty members play is essential in the fair and equitable administration of successful OMAAP EOI assessments.

### 2.1 Packaging and Shipping

To provide secure and dependable services for the shipping of the OMAAP EOI assessment materials, Pearson's Warehousing and Transportation Department maintains the quality and security of material distribution and return by using such methods as sealed trailers and hiring reputable carriers with the ability to immediately trace shipments. Pearson uses all available tracking capabilities to provide status information and early opportunities for corrective action.

Materials are packaged by school and delivered to the district coordinators. Each shipment to a district contains a shipping document set that includes a packing list for each school's materials and a pallet map that shows the identity and pallet assignment of each carton.

Materials are packaged using information provided by the Assessment Coordinators through the Pearson Access website, and optionally with data received directly from Oklahoma. Oklahoma educators also use the Pearson Access site to provide Pearson with the Pre-Identification information needed to print the student identification section on answer documents. Bar-coding of all secure materials during the pre-packaging effort allows for accurate tracking of these materials through the entire packing, delivery, and return process. It also permits Pearson to inventory all materials throughout the packaging and delivery process along with the ability to provide the customer with status updates at any time. Use of handheld radio-frequency scanners in the packaging process help to eliminate the possibility of packing the wrong materials. The proprietary "pick-and-pack" process prompts packaging personnel as to what materials are to go in which shipping box. If the packer tries to pack the wrong item (or number of items into a shipping carton), the system signals an alert.

### 2.2 Materials Return

Test administration handbooks provide clear instructions on how to assemble, box, label, and return testing materials after test administration. Because of the criticality of used test materials and quantities often involved, safety is also a major concern, not only for the materials but for the people moving them. Only single-column boxes are used to distribute and collect test materials, so the weight of each carton is kept to a reasonable and manageable limit.

Paper bands are provided to group and secure used student response booklets for scoring. Color-coded return mailing labels with detailed return information (district address and code number, receipt address, box *x* of *y*, shipper's tracking number, etc.) are also provided. These labels facilitate accurate and efficient sorting of each carton and its contents upon receipt at Pearson.

### 2.3 Materials Discrepancies Process

The image scanning process enables Pearson to concurrently capture optical mark read (OMR) responses, images, and security information electronically. All scorable material discrepancies are captured, investigated by our Oklahoma Call Center team, reported, and resolved prior to a batch passing through a clean post edit and images being released for scoring.

As scanning of materials progresses, any discrepancies in materials received versus shipped are reported immediately to the SDE and scoring will begin. This system allows Pearson to proceed in scoring clean batches while any discrepant material issues are being resolved. As discrepant materials are received, they will be processed. Data from discrepant material receipts are captured in the same database as all other material receipts resulting in a complete record of materials for each school. As batches clear the clean post edit, clipped images are prepared and distributed for scoring. The Oklahoma Call Center Team notifies the SDE regarding unresolved material discrepancies within 24 hours after Pearson's initial attempt to contact the school principal. Within one week after materials are returned, Pearson's Service Center Team also notifies the SDE of any missing or incomplete shipments from schools that received testing materials.

*Resolution of missing secure test materials and used answer booklets.* Pearson provides updates on a daily basis to the initial discrepancy reports, in response to SDE specifications and requests. The Oklahoma Call Center team makes every attempt to resolve all discrepancies involving secure test books and used answer booklets in a timely manner. Using daily, updated discrepancy reports, Pearson is in constant contact with the respective districts/schools. Pearson and the SDE work out details on specific approaches to resolution of material return discrepancies, and what steps will be taken if "lost" secure test books and/or used answer documents are not found and remain unreturned to Pearson.

### 2.4 Processing Assessment Materials Returned by Schools

Pearson's receipt system provides for the logging of materials within 24 hours of receipt and the readiness of materials for scanning within 72 hours of receipt. District status is available from a web-based system accessible by the SDE. In addition, the Oklahoma Call Center is able to provide receipt status information if required. The receipt notification website's database is updated daily to allow for accurate information being presented to inquiring district/school personnel. As with initial shipping, the secure and accurate receipt of test materials is a priority with Pearson. Quality assurance procedures provide that all materials are checked in using pre-defined procedures. Materials are handled in a highly secure manner from the time of receipt until final storage and shredding. The receipt of all secure materials is verified through the scanning of barcodes and the comparison of this data to that in security files established during the initial shipment of Oklahoma test materials to the district assessment coordinators.

## Section 3

## Classical Item Analysis and Results

### 3.1 Sampling Plan and Data Receipt

#### 3.1.a Sampling Plan

Population data were used for classical and item response theory (IRT) analyses for all tests. All students who complete a course with an End-of-Instruction test associated with it must also take the test. Students who meet the criteria were permitted to take the modified (OMAAP) exam for Algebra I, Biology I, English II, and U.S. History for the Spring 2011 administration. The information presented below is for the Spring 2011 administration only.

#### 3.1.b Data Receipt Activities

After all tests were scored, a data file was provided for item analyses and calibration. A data clean-up process that removed invalid cases, ineligible responses, absent students, and repeat test-takers was completed. A statistical key check was also performed at this time. This 'cleaned' sample was used for classical item analyses, calibration, and equating. Upon receipt of data, a Pearson psychometric team inspected several data fields to determine if the data met expectations, including:

- Student ID
- Demographic fields
- Form identification fields
- Raw response fields
- Scored response fields
- Total score and subscore fields
- Fields used to implement exclusion from analysis rules

*Exclusion Rules.* Following data inspection and clean-up, exclusionary rules were applied to form the final sample that was used for classical item analyses, calibration, and equating. Any student who had attempted at least five responses was included in the data analyses; however, any student who took the Braille form, was a second time test-taker, had an invalidated code, or attended a private school was not included in the equating and scaling processes. The demographic breakdown of the students in the item analysis and calibration sample appears in Table 3.1.

Table 3.1. Demographic Characteristics of Calibration and Equating Sample

| Subject | Total | Male | Female |
|---|---|---|---|
| Algebra I | 4,410 | 2,838 | 1,555 |
| Biology I | 3,831 | 2,430 | 1,395 |
| English II | 3,778 | 2,415 | 1,353 |
| U.S. History | 3,168 | 2,033 | 1,130 |

Note: Gender and Ethnicity values may not add to the total due to missing responses.

Table 3.1. Demographic Characteristics of Calibration and Equating Sample (cont.)

| Subject | African American | Native American | Hispanic | Asian | Pacific Islander | White | Other |
|---|---|---|---|---|---|---|---|
| Algebra I | 613 | 899 | 372 | 26 | 12 | 2,499 | 114 |
| Biology I | 572 | 819 | 295 | 23 | 4 | 2,097 | 103 |
| English II | 549 | 804 | 267 | 18 | 5 | 2,098 | 91 |
| U.S. History | 422 | 736 | 212 | 18 | 4 | 1,747 | 81 |

Note: Gender and Ethnicity values may not add to the total due to missing responses.

*Statistical Key Check.* Administering items that have only one correct key and are correctly scored is critical for assessing student performance. To screen for potentially problematic items and to confirm multiple-choice items were accurately scored, a statistical key check was conducted and items were flagged that met any of the following criteria:

- Less than 200 students responded to the item
- Correct response *p*-value less than 0.20
- Correct response uncorrected point-biserial correlation less than 0.20
- Distractor *p*-value greater than or equal to 0.40
- Distractor point-biserial correlation greater than or equal to 0.05

Any flagged operational items were submitted for key review by an appropriate Pearson content specialist. Any flagged items that were identified by content experts as mis-keyed would be corrected prior to analysis. Once the keys were verified, a secondary statistical key check and evaluation of items was conducted for the potential of removing items from scoring. There were no items identified as having a key issue for the 2011 tests.

*Removal of Operational Items.* Once the statistical key check was complete, all items were screened using the statistical criteria defined in Table 3.2. This procedure identified items with poor statistics for potential removal from scoring. The Pearson research scientists and content specialists reviewed the flagged items and proposed suggestions to the SDE. SDE evaluated and decided the drop and keep status of each item.

Table 3.2. Secondary Statistical Key Check Criteria

| Key Validation Item-Flagging Criteria | Rationale |
|---|---|
| If *p*-value of keyed response < 0.35 | Difficult item |
| If *p*-value of keyed response < 0.05 or > 0.95 | Extreme item |
| If *p*-value of keyed response < p value of distractor | Possible mis-key |
| If *p*-value of distractor > 0.35 | Possible second correct option |
| If point-biserial correlation of keyed response < 0.20 | Poorly discriminating item |
| If point-biserial correlation of distractor > 0.05 | Possible second correct option |
| If point-biserial correlation of keyed response < point-biserial of distractor | Possible mis-key |

SDE's decision on the 2011 tests is listed by subject in Table 3.3. These items were removed from operational scoring for the current operational administration and also will be removed from the OMAAP EOI item bank. Once the final set of operationally-scored items were agreed upon, classical item analyses were conducted.

Table 3.3. Number of Items Removed from Scoring by Subject

| Subject | Item Number |
|---|---|
| Algebra I | N/A |
| Biology I | N/A |
| English II | N/A |
| U.S. History | 1 |

Table 3.4 provides the final number of possible points on the 2011 OMAAP EOI exams, following removal of items with poor statistics.

Table 3.4. Total Possible Points after Item Review

| Subject | Score Points |
|---|---|
| Algebra I | 43 |
| Biology I | 46 |
| English II | 46 |
| U.S. History | 47 |

### 3.2 Classical Item Analyses

Following completion of the data receipt activities, statistical key check and dropping operational items, the last analyses conducted prior to calibration and equating are the following classical item analyses:

- Total case count
- Summary demographic statistics (e.g., males, females, African American, White, Hispanic, Asian, Pacific Islander, Native American, and Other)
- Frequency distributions for all multiple choice items and frequency distributions of score ratings and condition codes for writing prompts
    - Percentage of students in different multiple choice categories and, for the writing prompt, in different score categories (overall and broken down by gender and ethnicity)
- Item *p*-value
    - Mean item *p*-value
- Item-test point-biserial correlation
    - Mean item-test point-biserial correlation
    - Point-biserial correlation by response option (overall and broken down by gender and ethnicity)
- Omit percentage per item
    - Not reached analysis results per item
- Mean score by response option (overall and broken down by gender and ethnicity)

#### 3.2.a Test-Level Summaries of Classical Item Analyses

The test-level raw score descriptive statistics for the calibration samples are shown in Table 3.5. Note that students whose tests were invalidated and those students taking the test for a second time were excluded. The test results indicate that the omit rates were smaller than 1% for all subjects, except the English II constructed-response item. The test p values (Mean *P*), indicate that the forms were relatively similar in difficulty for all content areas.

Table 3.5. Test-Level Summaries of Classical Item Analyses

| Subject and Administration | Sample Size | Mean | Items / Points | Mean *p* | Mean $r_{pb}$ | Omit Min | Omit Max |
|---|---|---|---|---|---|---|---|
| Algebra I | 4410 | 22.62 | 43 | 0.53 | 0.33 | 0.1 | 0.9 |
| Biology I | 3831 | 25.62 | 46 | 0.56 | 0.33 | 0.0 | 1.0 |
| English II | 3778 | 27.21 | 44/46 | 0.61 | 0.32 | 0.1 | 43.3* |
| U.S. History | 3168 | 26.18 | 47 | 0.55 | 0.34 | 0.1 | 0.8 |

Note 1: $r_{pb}$ = point biserial correlation.
Note 2: The constructed response item p value is computed using the average score of all students divided by the max score possible, 3.
Note 3: A high omit rate is found on the English II constructed-response item this year.

### 3.3 Procedures for Detecting Item Bias

One of the goals of the OMAAP EOI assessments is to assemble a set of items that provides a measure of a student's ability that is as fair and accurate as possible for all subgroups within the population. Differential item functioning (DIF) analysis refers to statistical procedures that assess whether items are differentially difficult for different groups of examinees when the subgroups are of the same ability. The ability is usually defined by the total test scores. If the item is differentially more difficult for an identifiable subgroup when conditioned on ability, the item may be measuring something different from the intended construct. However, it is important to recognize that DIF-flagged items might be related to actual differences in relevant knowledge or skills or statistical Type I error. As a result, DIF statistics are used only to identify potential sources of item bias. Subsequent review by content experts and bias committees are required to determine the source and meaning of performance differences. For the OMAAP EOI test DIF analyses, DIF statistics were estimated for race (African American and white) and gender (male and female). Items with statistically-significant differences in performance were flagged so that items could be carefully examined for possible biased or unfair content that was undetected in earlier fairness and bias content review meetings held prior to form construction.

Pearson used the Mantel-Haenszel (MH) chi-square approach for detecting DIF in multiple choice and open-ended items. Pearson calculated the Mantel-Haenszel statistic (MH D-DIF; Holland & Thayer 1988) to measure the degree and magnitude of DIF. The student group of interest is the *focal* group, and the group to which performance on the item is being compared is the *reference* group. The reference groups for these DIF analyses were white for race and male for gender. The focal groups were females and African Americans.

Items were separated into one of three categories on the basis of DIF statistics (Holland and Thayer 1988; Dorans and Holland 1993): negligible DIF (category A), intermediate DIF (category B), and large DIF (category C). The items in category C, which exhibit significant DIF, are of primary concern. The item classifications are based on the Mantel-Haenszel chi-square and the MH delta ($\Delta$) value. Positive values of delta indicate that the item is easier for the focal group, and a negative value of delta indicates that the item is more difficult for the focal group. The item classifications are made as follows (Michaelides, 2008):

- The item is classified as C category if the MH D-DIF is significantly greater than 1.0 in absolute value, and its absolute value is at least 1.5.
- The item is classified as B category if the MH D-DIF is significantly different from zero, its absolute value is at least 1.0, and its absolute value is either less than 1.5 or not significantly greater than 1.0.

- The item is classified as A category if the MH D-DIF is not significantly different from zero ($p \geq 0.05$), or if its absolute value is less than 1.0.

### 3.3.a Differential Item Functioning Results

The data in Table 3.6 summarize the number of items flagged for possible DIF in each subject. Items flagged for DIF were placed before expert content specialists from Pearson prior to inclusion as part of the final operational scored set. The content specialist reviewed the item content, the percentage of students selecting each response option and the point-biserial correlation for each response option by gender and race for all items flagged for DIF. The content specialist was then asked if there was context (for example, cultural barriers) or language in an item that might result in bias (i.e., an explanation for the existence of the statistical DIF flag). Items that were identified to be biased would be presented to SDE for confirmation and, if confirmed, could not be used on test or scoring. Items flagged by DIF, yet not biased, could be used for scoring for content coverage.

Table 3.6 presents the DIF analysis results using the data of 2011 administration.

Table 3.6. DIF Flag Incidence Across All OMAAP EOI Items

| Subject | Items | Race | | Gender | |
|---|---|---|---|---|---|
| | DIF Category | B | C | B | C |
| Algebra I | 43 | 0 | 0 | 8 | 0 |
| Biology I | 46 | 3 | 1 | 18 | 7 |
| English II | 44 | 2 | 1 | 7 | 0 |
| U.S. History | 48 | 9 | 0 | 12 | 3 |

## 3.4 Test Reliability

### 3.4.a Overall Test Reliability

The reliability of a test provides an estimate of the extent to which an assessment will yield the same results when administered in different times, locations, or samples, when the two administrations do not differ in relevant variables. The reliability coefficient is an index of consistency of test results. Reliability coefficients are usually forms of correlation coefficients and must be interpreted within the context and design of the assessment and of the reliability study. Cronbach's alpha is a commonly-used internal consistency measure, which is derived from analysis of the consistency of the performance of individuals on items in a test administration. Cronbach's alpha is calculated as shown in equation (1). In this formula, $s_i^2$ denotes the estimated variance for each item, with items indexed $i = 1, 2, \ldots, k$, and $s^2_{sum}$ denotes the variance for the sum of all $k$ items:

$$\alpha = \left(\frac{k}{k-1}\right)\left(1 - \frac{\sum_{i=1}^{k} s_i^2}{s_{sum}^2}\right). \quad (1)$$

Table 3.7 presents Cronbach's alpha for each of the OMAAP EOI tests. These reliability coefficients indicate that the OMAAP EOI assessments had adequate internal consistency and that the tests produce relatively stable scores.

Table 3.7. Cronbach's Alpha by Subject

| Subject | Alpha |
|---|---|
| Algebra I | 0.80 |
| Biology I | 0.82 |
| English II | 0.80 |
| U.S. History | 0.84 |

### 3.4.b Test Reliability by Subgroup

Table 3.8 addresses the reliability analysis results by the different reporting subgroups for the OMAAP EOI assessments. This table illustrates the subject, the subgroups, the number of students used in the analyses and the associated Cronbach's Alpha reliability coefficient for each subject and subgroup. In all instances, the reliability coefficients are well above the accepted lower limit of .70.

Table 3.8. Test Reliability by Subgroup

| Subject | Male | Female | African-American | Native American | Hispanic | Asian | White |
|---|---|---|---|---|---|---|---|
| Algebra I | 0.80 | 0.80 | 0.80 | 0.77 | 0.78 | 0.84 | 0.81 |
| Biology I | 0.83 | 0.77 | 0.76 | 0.82 | 0.80 | 0.76 | 0.82 |
| English II | 0.81 | 0.77 | 0.78 | 0.77 | 0.80 | 0.81 | 0.82 |
| U.S. History | 0.85 | 0.80 | 0.80 | 0.83 | 0.81 | 0.76 | 0.85 |

Note: N/A - insufficient sample size to compute statistic.

Table 3.8. Test Reliability by Subgroup (cont.)

| Subject | English Language Learner | Individual Education Plan | Economically Disadvantaged |
|---|---|---|---|
| Algebra I | 0.85 | 0.80 | 0.79 |
| Biology I | 0.76 | 0.82 | 0.82 |
| English II | 0.78 | 0.80 | 0.80 |
| U.S. History | 0.80 | 0.84 | 0.84 |

### 3.5 Inter-rater Reliability

Inter-rater reliability is referred to as the degree of agreement among scorers that allows for the scores to be interpreted as reasonably intended by the test developer (AERA, APA and NCME, 1999). The English II test contained one writing prompt. Raters were trained to implement the scoring rubrics, anchor papers, check sets, and resolution reading. The items were holistically scored by two raters and the rounded average was the final score. The final writing score for a student ranged between 1 and 3 unless a non-score was specified, such as writing off-topic, illegible, written in another language, or blank, and these were assigned a zero.

The inter-rater reliability coefficients for the English II operational writing prompt are presented in Table 3.9. The results show that score differences between two raters were equal or less than one point. The weighted Kappa statistic (Kraemer, 1982) is an indication of inter-rater reliability after correcting for chance. The Kappa values for the English II operational writing prompts are within the moderate range.

Table 3.9. Inter-rater Reliability for English II Operational Writing Prompts

| | Max | Valid | Point Discrepancy Percentages | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Trait** | **Points** | **N** | **-3** | **-2** | **-1** | **0** | **1** | **2** | **3** |
| Rubric | 3 | 2129 | 0 | 0 | 12.73 | 75.81 | 11.46 | 0 | 0 |

Table 3.9. Inter-rater Reliability for English II Operational Writing Prompts (cont.)

| | Agreement Percentages | | | |
|---|---|---|---|---|
| **Trait** | **Exact** | **Adjacent** | **+/- 2 or more** | **Kappa** |
| Rubric | 75.81 | 24.19 | 0.00 | 0.56 |

## Section 4

## Calibration, Equating, and Scaling

### 4.1 Item Response Theory (IRT) Models

#### 4.1.a Dichotomous Item Response Theory Model

*Rasch Model*. The Rasch model (Rasch, 1960) was used for calibrating the dichotomously-scored multiple choice items. In the Rasch model, the probability that a student with an ability level of $\theta$ responds correctly to item $i$ is

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}}, \tag{2}$$

where $b_i$ is the item difficulty parameter.

#### 4.1.b Polytomous Item Response Theory Model

*Partial Credit Model*. For calibrating the polytomously-scored constructed response—or open-ended (OE)—writing prompt items, the partial credit model (PCM; Masters, 1982) was used. In the partial credit model, the probability that a student with ability level $\theta$ will have a score in category $h$ (indexed $h$ = 1, 2, …, $m_i$) on an item with difficulty parameter $b_i$ (indexed $b_{i1}$, …, $b_{ih}$, …, $b_{im_i}$ ) is given by

$$P_{ih}(\theta) = \frac{\exp\left(h\theta - \sum_{g=1}^{h} b_{ig}\right)}{1 + \sum_{k=1}^{m_i} \exp\left(k\theta - \sum_{g=1}^{k} b_{ig}\right)}. \tag{3}$$

The IRT models were implemented using WINSTEPS 3.61 (Linacre, 2006). All item and student ability calibrations were independently conducted and verified by at least two Pearson research scientists.

### 4.2 Assessment of Fit to the Model

Item fit was assessed using the WINSTEPS fit function. WINSTEPS provides two statistics for indicating how well the Rasch model fits test data. Infit is sensitive to aberrations in item response patterns and indicates overfit for “perfect” response patterns and underfit for items targeting different curricula. The infit statistic is expressed as a mean square (MS) statistic and was considered for assessing fit. The expected value is 1.0, with values greater than 1.0 indicating mis-fitting items. For the OMAAP EOI equating, infit values greater 1.3 were flagged and examined further.

Operational items flagged by the IRT fit that were not flagged by the classical item analyses and had reasonable estimated IRT parameters were not reviewed. Items that were also flagged by classical item analyses and/or had poor IRT parameter estimates (e.g., *b*

parameter estimate with absolute value greater than 3) were reviewed by Pearson content specialists. Any item that was potentially mis-keyed was presented to SDE to make a decision regarding whether to keep or remove the item.

In addition to the fit statistics provided by WINSTEPS, the item fit plot was also generated. The item fit plot presented the expected and observed item characteristics curves for each item. All item fit plots were examined by a research scientist during equating.

### 4.3 Calibration and Equating

The Rasch model was used for calibration of Algebra I, Biology I, and U.S. History because all of these tests consist of only multiple choice items. Because English II has multiple choice and constructed response items, a simultaneous calibration with the Rasch- and partial credit models was implemented.

A common item, non-equivalent groups (CINEG) design was used for all content areas to link the current test forms (i.e., Spring 2011) to the base scale. The horizontal linking items were selected to be representative of the test content in terms of difficulty and the test blueprint. With this equating design, the mean Rasch difficulty of the linking items on the current form is compared to their difficulty on the baseline form to derive an equating constant. The equating constant is then added to the *b*-parameter estimates for all items to put them on the same scale as the baseline form.

### 4.4 Anchor Item Stability Evaluation Methods

Despite the careful selection of anchor items, it is possible for the anchor items to perform differentially across administrations. Dramatic changes in anchor item parameter values can result in systematic errors in equating results (Kolen & Brennan, 2004). As a result, prior to finalizing the equating constant, Pearson evaluated changes in the item parameters from their prior operational administration to the Spring 2011 administration. The process used in this evaluation is referred to as an anchor item parameter stability check.

Two methods, displacement and robust Z, are used for the OMAAP EOI anchor stability check. Te procedure is iterative in that only one item is dropped at each screening step. Note that although anchor items may be dropped from anchoring function, they still contribute to student scores. Details of the anchor item screening procedure are as follows.

1. Compute equating constant of the anchor set (bank values vs. free-calibrated post-equating item parameters).
2. Compute displacement and robust z of each anchor items.
3. Flag items with displacement > 0.30 (flag = 1, else = blank).
4. Sort items in this order:
    - Displacement Flag (descending)
    - Absolute value of Robust Z (descending), and then
    - Item Sequence (ascending – for a tie-breaker)
5. Drop flagged item with the largest absolute value of Robust Z.
6. Recompute equating constant and displacement values based on the new anchor set. <u>Do not</u> recompute robust z.
7. Flag items with displacement > 0.30.
8. Sort items and drop one (if necessary) based on criteria outlined above.

9. Stop criteria – Stop if either of the following occur:
    - Anchor set is 20% of the total test, OR
    - No additional items are flagged based on displacement.

Besides anchor stability check methods, anchor items will also be evaluated using the following factors:

- Compare prior and current p-values and point biserials
- Compare prior and current IRT (Rasch item difficulty) values
- Compare prior and current item sequences
- Review Standard and objective/skill for item (make sure not eliminating too many items from one Standard)
- Review Passage ID/Title (if eliminating too many items from passage, may need to remove entire passage and associated items)
- Request content review of item for any modifications or edits since last operational use (should be none)

### 4.4.a Anchor Items for Spring 2011

Table 4.1 presents the number and percentages of anchor items by subject for the Spring 2011 administration. During test construction, the anchor set was determined to be 30% or more of the test. In addition, the anchor set was proportionally representative of the total test in terms of content assessed and mimicked the difficulty of the overall test as well.

### 4.4.b Results of the Anchor Item Stability Check

Once the anchor set was finalized, the equating constant was applied to the non-anchor items for computation of raw score to scale score tables. For Spring 2011, two anchor item was removed from Algebra I, one item from Biology I, and zero items from English II and U.S. History (see Table 4.1). Any item removed from the anchor set still contributed to student scores.

Table 4.1. Number of Anchor Items per Subject

| | | Initial | | Final | |
|---|---|---|---|---|---|
| Subject | Number of Items on Test | Number of Linking Items | Percent of Test | Number of Linking Items | Percent of Test |
| Algebra I | 43 | 14 | 33% | 12 | 28% |
| Biology I | 46 | 16 | 35% | 15 | 33% |
| English II | 44 | 15 | 34% | 15 | 34% |
| U.S. History | 48 | 15 | 31% | 15 | 31% |

Note: Prior to item dropping

### 4.5 Scaling and Scoring Results

The Lowest Obtainable Scale Score (LOSS), Highest Obtainable Scale Score (HOSS), and final scaling constants for each of the subjects are shown in Table 4.2. The scaling constants, *M*1 (multiplicative) and *M*2 (additive), place the true scores associated with each raw score point onto the reporting or operational scale using a linear transformation:

$$\text{Scale Score} = \left(\hat{\tau} \times M1\right) + M2 \qquad (4)$$

where, $\hat{\tau}$ = true score.

Each scale score on the assessment is associated with a performance level that describes the types of behavior, knowledge, and skill a student in this score level is likely to be able to do. For the OMAAP EOI assessments, there are three cut scores that divide scores into four performance levels: Unsatisfactory, Limited Knowledge, Satisfactory, and Advanced. The cut scores for each of the tests appear in Table 4.2. In addition, a conditional standard error of measurement (*CSEM*; please see Section 6.3 for computation of *CSEM*) was computed for each of the raw score points. The resulting raw score to scale score conversions, *CSEM*s, as well as the performance levels for Algebra I and Biology I are shown in Table 4.3 and English II and U.S. History are shown in Table 4.4.

Table 4.2: OMAAP Scaling Constants, Scale Range, and Cut Scores by Subject

| | **Scaling Constants** | | **Scale Range** | | **Cut Score** | | |
|---|---|---|---|---|---|---|---|
| **Subject** | **Multiplicative (Slope)** | **Additive (Mean)** | **LOSS** | **HOSS** | **Limited Knowledge** | **Satisfactory** | **Advanced** |
| Algebra I | 19.07 | 254.60 | 100 | 350 | 237 | 250 | 269 |
| Biology I | 19.68 | 247.55 | 100 | 350 | 233 | 250 | 265 |
| English II | 19.72 | 248.25 | 100 | 350 | 238 | 250 | 265 |
| U.S. History | 19.98 | 252.01 | 100 | 350 | 239 | 250 | 264 |

Table 4.3. Raw Score to Scale Score Conversion Table for Algebra I and Biology I

| Raw Score | Algebra I | | | Biology I | | |
|---|---|---|---|---|---|---|
| | Scale Score | *CSEM* | Perf. Level | Scale Score | *CSEM* | Perf. Level |
| 0 | 166 | 35 | 1 | 148 | 36 | 1 |
| 1 | 189 | 19 | 1 | 173 | 20 | 1 |
| 2 | 203 | 14 | 1 | 187 | 14 | 1 |
| 3 | 211 | 12 | 1 | 196 | 12 | 1 |
| 4 | 217 | 10 | 1 | 202 | 11 | 1 |
| 5 | 222 | 9 | 1 | 208 | 10 | 1 |
| 6 | 227 | 9 | 1 | 212 | 9 | 1 |
| 7 | 230 | 8 | 1 | 216 | 8 | 1 |
| 8 | 234 | 8 | 1 | 219 | 8 | 1 |
| 9 | 237 | 7 | 2 | 223 | 8 | 1 |
| 10 | 239 | 7 | 2 | 225 | 7 | 1 |
| 11 | 242 | 7 | 2 | 228 | 7 | 1 |
| 12 | 244 | 7 | 2 | 231 | 7 | 1 |
| 13 | 247 | 7 | 2 | 233 | 7 | 2 |
| 14 | 250 | 6 | 3 | 236 | 7 | 2 |
| 15 | 251 | 6 | 3 | 238 | 7 | 2 |
| 16 | 253 | 6 | 3 | 240 | 7 | 2 |
| 17 | 255 | 6 | 3 | 242 | 6 | 2 |
| 18 | 257 | 6 | 3 | 244 | 6 | 2 |
| 19 | 259 | 6 | 3 | 246 | 6 | 2 |
| 20 | 261 | 6 | 3 | 248 | 6 | 2 |
| 21 | 263 | 6 | 3 | 250 | 6 | 3 |
| 22 | 265 | 6 | 3 | 252 | 6 | 3 |
| 23 | 267 | 6 | 3 | 254 | 6 | 3 |
| 24 | 269 | 6 | 4 | 256 | 6 | 3 |
| 25 | 271 | 6 | 4 | 258 | 6 | 3 |
| 26 | 273 | 6 | 4 | 260 | 6 | 3 |
| 27 | 275 | 6 | 4 | 262 | 6 | 3 |
| 28 | 277 | 6 | 4 | 265 | 6 | 4 |
| 29 | 279 | 6 | 4 | 266 | 6 | 4 |
| 30 | 282 | 7 | 4 | 269 | 7 | 4 |
| 31 | 284 | 7 | 4 | 271 | 7 | 4 |
| 32 | 287 | 7 | 4 | 273 | 7 | 4 |
| 33 | 289 | 7 | 4 | 275 | 7 | 4 |
| 34 | 292 | 7 | 4 | 278 | 7 | 4 |
| 35 | 295 | 8 | 4 | 280 | 7 | 4 |
| 36 | 298 | 8 | 4 | 283 | 7 | 4 |
| 37 | 302 | 9 | 4 | 286 | 8 | 4 |
| 38 | 306 | 9 | 4 | 289 | 8 | 4 |
| 39 | 311 | 10 | 4 | 292 | 8 | 4 |
| 40 | 317 | 12 | 4 | 296 | 9 | 4 |
| 41 | 326 | 14 | 4 | 301 | 10 | 4 |
| 42 | 339 | 19 | 4 | 306 | 11 | 4 |
| 43 | 350 | 35 | 4 | 312 | 12 | 4 |

Table 4.3. Raw Score to Scale Score Conversion Table for Algebra I and Biology I (cont.)

| Raw Score | Algebra I | | | Biology I | | |
|---|---|---|---|---|---|---|
| | Scale Score | *CSEM* | Perf. Level | Scale Score | *CSEM* | Perf. Level |
| 44 | - | - | - | 321 | 14 | 4 |
| 45 | - | - | - | 335 | 20 | 4 |
| 46 | - | - | - | 350 | 36 | 4 |

Note: *CSEM* = Conditional Standard Error of Measure; Perf. Level = Performance Level; 1 = Unsatisfactory, 2 = Limited Knowledge, 3 = Satisfactory, 4 = Advanced

Table 4.4. Raw Score to Scale Score Conversion Table for English II and U.S. History

| Raw Score | English II | | | U.S. History | | |
|---|---|---|---|---|---|---|
| | Scale Score | *CSEM* | Perf. Level | Scale Score | *CSEM* | Perf. Level |
| 0 | 154 | 36 | 1 | 143 | 37 | 1 |
| 1 | 178 | 20 | 1 | 167 | 20 | 1 |
| 2 | 192 | 14 | 1 | 182 | 15 | 1 |
| 3 | 201 | 12 | 1 | 191 | 12 | 1 |
| 4 | 208 | 11 | 1 | 197 | 11 | 1 |
| 5 | 213 | 10 | 1 | 202 | 10 | 1 |
| 6 | 217 | 9 | 1 | 207 | 9 | 1 |
| 7 | 221 | 8 | 1 | 210 | 8 | 1 |
| 8 | 224 | 8 | 1 | 214 | 8 | 1 |
| 9 | 227 | 8 | 1 | 217 | 8 | 1 |
| 10 | 230 | 7 | 1 | 220 | 7 | 1 |
| 11 | 233 | 7 | 1 | 222 | 7 | 1 |
| 12 | 235 | 7 | 1 | 225 | 7 | 1 |
| 13 | 238 | 7 | 2 | 227 | 7 | 1 |
| 14 | 240 | 7 | 2 | 229 | 7 | 1 |
| 15 | 242 | 6 | 2 | 231 | 6 | 1 |
| 16 | 244 | 6 | 2 | 234 | 6 | 1 |
| 17 | 246 | 6 | 2 | 236 | 6 | 1 |
| 18 | 248 | 6 | 2 | 238 | 6 | 1 |
| 19 | 250 | 6 | 3 | 239 | 6 | 2 |
| 20 | 252 | 6 | 3 | 241 | 6 | 2 |
| 21 | 254 | 6 | 3 | 243 | 6 | 2 |
| 22 | 256 | 6 | 3 | 245 | 6 | 2 |
| 23 | 258 | 6 | 3 | 247 | 6 | 2 |
| 24 | 259 | 6 | 3 | 250 | 6 | 3 |
| 25 | 261 | 6 | 3 | 251 | 6 | 3 |
| 26 | 263 | 6 | 3 | 252 | 6 | 3 |
| 27 | 265 | 6 | 4 | 254 | 6 | 3 |
| 28 | 267 | 6 | 4 | 256 | 6 | 3 |
| 29 | 268 | 6 | 4 | 258 | 6 | 3 |
| 30 | 270 | 6 | 4 | 260 | 6 | 3 |
| 31 | 272 | 6 | 4 | 262 | 6 | 3 |
| 32 | 274 | 6 | 4 | 264 | 6 | 4 |
| 33 | 276 | 7 | 4 | 266 | 7 | 4 |
| 34 | 279 | 7 | 4 | 268 | 7 | 4 |
| 35 | 281 | 7 | 4 | 271 | 7 | 4 |
| 36 | 283 | 7 | 4 | 273 | 7 | 4 |
| 37 | 286 | 7 | 4 | 276 | 7 | 4 |
| 38 | 289 | 8 | 4 | 279 | 8 | 4 |
| 39 | 292 | 8 | 4 | 282 | 8 | 4 |
| 40 | 296 | 9 | 4 | 285 | 8 | 4 |
| 41 | 300 | 10 | 4 | 289 | 9 | 4 |
| 42 | 305 | 10 | 4 | 293 | 10 | 4 |
| 43 | 312 | 12 | 4 | 298 | 11 | 4 |

Table 4.4. Raw Score to Scale Score Conversion Table for English II and U.S. History (cont.)

| Raw Score | English II | | | U.S. History | | |
|---|---|---|---|---|---|---|
| | Scale Score | *CSEM* | Perf. Level | Scale Score | *CSEM* | Perf. Level |
| 44 | 320 | 14 | 4 | 304 | 12 | 4 |
| 45 | 335 | 20 | 4 | 313 | 15 | 4 |
| 46 | 350 | 36 | 4 | 327 | 20 | 4 |
| 47 | - | - | - | 350 | 37 | 4 |

Note: *CSEM* = Conditional Standard Error of Measure; Perf. Level = Performance Level; 1 = Unsatisfactory, 2 = Limited Knowledge, 3 = Satisfactory, 4 = Advanced

## Section 5

## Classification Consistency and Accuracy Studies

### 5.1 Classification Consistency and Accuracy

Every test administration will result in some error in classifying examinees. The concept of the standard error of measurement (SEM) has implications for the interpretation of cut scores used to classify students into different performance levels. For example, a given student may have a true performance level greater than a cut score; however, due to random variations (measurement error), the student's observed test score may be below the cut score. As a result, the student may be classified as having a lower performance level. As discussed in Section 6.4, a student's observed score is most likely to fall within a standard error band around his or her true score. Thus, the classification of students into different performance levels can be imperfect; especially for the borderline students whose true scores lie close to the performance level cut scores.

According to Livingston and Lewis (1995, p. 180), the accuracy of a classification is "the extent to which the actual classifications of the test takers… agree with those that would be made on the basis of their true score" and are calculated from cross-tabulations between "classifications based on an observable variable and classifications based on an unobservable variable." Since the unobservable variable—the true score—is not available, Livingston and Lewis provide a method to estimate the true score distribution of a test and create the cross-tabulation of the true score and observed variable (raw score) classifications. Consistency is "the agreement between classifications based on two non-overlapping, equally-difficult forms of the test" (p. 180). Consistency is estimated using actual response data from a test and the test's reliability to statistically model two parallel forms of the test and compare the classifications on those alternate forms. There are three types of accuracy and consistency indices that can be generated using Livingston and Lewis' approach: overall, conditional on level, and by cut score.

The overall accuracy of performance level classifications is computed as a sum of the proportions on the diagonal of the joint distribution of true score- and observed score levels. Essentially, overall accuracy is a proportion (or percentage) of correct classifications across all levels. The overall consistency index is computed as the sum of the diagonal cells in a consistency table. Another way to express overall consistency is to use the kappa coefficient, as used in the inter-rater reliability studies in Section 3.5. Like the inter-rater reliability studies, kappa provides an estimate of agreement or the proportion of consistent classifications between two different tests after taking into account chance.

Consistency conditional on performance level is computed as the ratio between the proportion of correct classifications at the selected performance level (for example, Satisfactory students who were classified as Satisfactory) and the proportion of all the students classified into that level (total proportion of students who were considered Satisfactory). Accuracy conditional on performance level is computed in a similar manner except that in the consistency table where both row and column marginal sums are the same, the accuracy table uses the sum based on estimated status as the total for computing accuracy conditional on performance level.

To evaluate decisions at specific cut scores, the joint distribution of all the performance levels are collapsed into dichotomized distributions around that specific cut score (for example collapsing Unsatisfactory and Limited Knowledge and then Satisfactory and Advanced to assess decisions at the Satisfactory cut score). The accuracy index at cut score is computed as the sum of the proportions of correct classifications around this selected cut score. The consistency at a specific cut score is obtained in a similar way, but by dichotomizing the distributions at the cut score performance level and between all other performance levels combined. Table 5.1 presents the overall accuracy and consistency indices for the OMAAP EOI tests.

Table 5.1. Estimates of Accuracy and Consistency of Performance Classification

| Subject | Accuracy | Consistency | Kappa | False Positives | False Negatives |
|---|---|---|---|---|---|
| Algebra I | 0.80 | 0.72 | 0.49 | 0.09 | 0.11 |
| Biology I | 0.74 | 0.64 | 0.47 | 0.13 | 0.13 |
| English II | 0.78 | 0.70 | 0.48 | 0.10 | 0.12 |
| U.S. History | 0.68 | 0.58 | 0.43 | 0.17 | 0.15 |

As shown in Table 5.1, the overall accuracy indices range between 68 and 80 percent and overall consistency ranges between 58 and 72 percent. Kappa coefficients range from 0.43 and 0.49. The false positive rates range from 9 to 17 percent. The false negative rates range from 11 to 15 percent.

Table 5.2 provides the accuracy-, consistency-, false positive-, and false negative rates by cut-score. The data in these tables reveal that the level of agreement for both accuracy and consistency is above 80 percent in all cases, with most above 85 percent. In general, the high rates of accuracy and consistency support the cut decisions made using these assessments. Similar to Table 5.1, the false positive- and false negative rates are quite low.

Table 5.2. Accuracy and Consistency Estimates by Cut Score: False Positive- and False Negative Rates

| | Accuracy | | | Consistency | | |
|---|---|---|---|---|---|---|
| Subject | U / L+S+A | U+L / S+A | U+L+S / A | U / L+S+A | U+L / S+A | U+L+S / A |
| Algebra I | 1.00 | 0.93 | 0.87 | 0.99 | 0.89 | 0.82 |
| Biology I | 0.98 | 0.87 | 0.88 | 0.97 | 0.82 | 0.84 |
| English II | 0.99 | 0.93 | 0.86 | 0.98 | 0.90 | 0.81 |
| U.S. History | 0.88 | 0.87 | 0.92 | 0.83 | 0.82 | 0.89 |

Note: U =Unsatisfactory; L = Limited Knowledge; S = Satisfactory; and A = Advanced.
Note: U / L+S+A = Unsatisfactory divided by Limited Knowledge plus Satisfactory plus Advanced; U+L / S+A = Unsatisfactory plus Limited Knowledge divided by Satisfactory plus Advanced; U+L+S / A = Unsatisfactory plus Limited Knowledge plus Satisfactory divided by Advanced.

Table 5.2. Accuracy and Consistency Estimates by Cut Score: False Positive- and False Negative Rates (cont.)

| | False Positives | | | False Negatives | | |
|---|---|---|---|---|---|---|
| Subject | U / L+S+A | U+L / S+A | U+L+S / A | U / L+S+A | U+L / S+A | U+L+S / A |
| Algebra I | 0.00 | 0.01 | 0.08 | 0.00 | 0.06 | 0.05 |
| Biology I | 0.00 | 0.06 | 0.07 | 0.02 | 0.06 | 0.05 |
| English II | 0.00 | 0.02 | 0.07 | 0.01 | 0.05 | 0.06 |
| U.S. History | 0.06 | 0.07 | 0.05 | 0.06 | 0.05 | 0.03 |

Note: U =Unsatisfactory; L = Limited Knowledge; S = Satisfactory; and A = Advanced.
Note: U / L+S+A = Unsatisfactory divided by Limited Knowledge plus Satisfactory plus Advanced; U+L / S+A = Unsatisfactory plus Limited Knowledge divided by Satisfactory plus Advanced; U+L+S / A = Unsatisfactory plus Limited Knowledge plus Satisfactory divided by Advanced.

The importance of the dichotomous categorization is particularly notable when they map onto pass/fail decisions for the assessments. For the OMAAP EOI tests, the U+L/S+A is the important dichotomization, because it directly translates to the pass/fail decision point. Similar to other dichotomization distinctions, there are three main scenarios at this cut point: 1) observed performance is accurately reflective of the true ability level (i.e., the examinee passed and should have passed); 2) the true ability level is below the standard, but the observed test score is above the standard (i.e., a false positive); and 3) the true ability level is above the standard, but the observed test score is below the standard (i.e., a false negative). In examining Table 5.2 Algebra I, for example, 93 percent of students are correctly classified as pass or fail based on their performance (scenario 1), 1 percent passed, but their true ability is below the standard (scenario 2), and 6 percent failed although their true ability is above the standard (scenario 3). Overall, the accuracy rate for accurate classification are 87% or above for all subjects – students are appropriately (more than 87% of the time) categorized into pass/fail classifications based on their true ability using their observed score (raw score) as their classification score.

## Section 6

## Summary Statistics

### 6.1 Descriptive Statistics

The summary descriptive statistics of the scale scores are presented in the following tables. The scales scores presented exclude invalid student cases and second time testers.

Table 6.1. Descriptive Statistics of the Scale Scores - Overall

| | Total | | | |
|---|---|---|---|---|
| **Subject** | ***N*** | **Mean** | ***SD*** | **Med.** |
| Algebra I | 4,410 | 266.9 | 15.1 | 265 |
| Biology I | 3,831 | 260.2 | 16.1 | 258 |
| English II | 3,778 | 265.7 | 13.7 | 267 |
| U.S. History | 3,168 | 253.9 | 17.2 | 251 |

Note: N = Sample size; SD = Standard Deviation; Med. = Median.

Table 6.2. Descriptive Statistics of the Scale Scores by Gender

| | Female | | | | Male | | | |
|---|---|---|---|---|---|---|---|---|
| **Subject** | ***N*** | **Mean** | ***SD*** | **Med.** | ***N*** | **Mean** | ***SD*** | **Med.** |
| Algebra I | 1,555 | 266.6 | 15.0 | 265 | 2,838 | 267.0 | 15.2 | 265 |
| Biology I | 1,395 | 257.3 | 14.0 | 256 | 2,430 | 261.9 | 17.0 | 260 |
| English II | 1,353 | 266.6 | 12.8 | 267 | 2,415 | 265.1 | 14.1 | 265 |
| U.S. History | 1,130 | 250.1 | 15.0 | 247 | 2,033 | 256.1 | 18.0 | 254 |

Note: N = Sample size; SD = Standard Deviation; Med. = Median.

Table 6.3. Descriptive Statistics of the Scale Scores by Race/Ethnicity

| | African-American | | | | Native American | | | |
|---|---|---|---|---|---|---|---|---|
| **Subject** | ***N*** | **Mean** | ***SD*** | **Med.** | ***N*** | **Mean** | ***SD*** | **Med.** |
| Algebra I | 613 | 262.8 | 14.8 | 259 | 899 | 266.8 | 14.1 | 265 |
| Biology I | 572 | 253.1 | 14.0 | 252 | 819 | 262.2 | 16.8 | 260 |
| English II | 549 | 262.5 | 13.1 | 263 | 804 | 266.9 | 13.2 | 267 |
| U.S. History | 422 | 248.6 | 14.6 | 247 | 736 | 254.6 | 16.8 | 252 |

Note: N = Sample size; SD = Standard Deviation; Med. = Median.

Table 6.3. Descriptive Statistics of the Scale Scores by Race/Ethnicity (cont.)

| | Hispanic | | | | Asian | | | |
|---|---|---|---|---|---|---|---|---|
| **Subject** | ***N*** | **Mean** | ***SD*** | **Med.** | ***N*** | **Mean** | ***SD*** | **Med.** |
| Algebra I | 372 | 264.4 | 14.5 | 263 | 26 | 271.5 | 16.8 | 270 |
| Biology I | 295 | 257.3 | 14.8 | 256 | 23 | 262.6 | 12.9 | 265 |
| English II | 267 | 262.9 | 13.6 | 263 | 18 | 265.5 | 14.6 | 267 |
| U.S. History | 212 | 249.6 | 15.5 | 247 | 18 | 246.6 | 15.4 | 245 |

Note: N = Sample size; SD = Standard Deviation; Med. = Median.

Table 6.3. Descriptive Statistics of the Scale Scores (cont.)

| | White | | | | Pacific Islander | | | |
|---|---|---|---|---|---|---|---|---|
| **Subject** | ***N*** | **Mean** | ***SD*** | **Med.** | ***N*** | **Mean** | ***SD*** | **Med.** |
| Algebra I | 2,499 | 267.4 | 15.3 | 265 | 12 | 271.2 | 14.3 | 270 |
| Biology I | 2,097 | 261.5 | 16.3 | 260 | - | - | - | - |
| English II | 2,098 | 266.5 | 14.2 | 267 | - | - | - | - |
| U.S. History | 1,747 | 255.3 | 18.0 | 252 | - | - | - | - |

Note: Pacific Islander *N* counts on Biology I, English II, and U.S. History are less then 10. Their information is suppressed for confidentiality reason.

Table 6.3. Descriptive Statistics of the Scale Scores by Race/Ethnicity (cont.)

| | Other | | | |
|---|---|---|---|---|
| **Subject** | ***N*** | **Mean** | ***SD*** | **Med.** |
| Algebra I | 114 | 267.5 | 15.0 | 263 |
| Biology I | 103 | 262.4 | 17.2 | 262 |
| English II | 91 | 268.5 | 14.2 | 268 |
| U.S. History | 81 | 253.7 | 17.9 | 250 |

Note: N = Sample size; SD = Standard Deviation; Med. = Median.

### 6.2 Performance Level Distribution

The distributions of students in the four performance levels are presented in Table 6.4 (please see Appendix B for distributions by scale score). As above, these percentages exclude invalid student data and second-time test-takers. The passing rate (Satisfactory and Above) of all subjects increased compared to previous administration.

Table 6.4. Percentage of Students by Performance Level

| **Subject** | ***N*** | **Unsatisfactory** | **Limited Knowledge** | **Satisfactory** | **Advanced** | **Satisfactory and Above** |
|---|---|---|---|---|---|---|
| Algebra I | 4,410 | 0.3% | 6.4% | 51.5% | 41.8% | 92.3% |
| Biology I | 3,831 | 1.5% | 24.6% | 35.7% | 38.2% | 73.9% |
| English II | 3,778 | 1.5% | 10.6% | 30.8% | 57.0% | 87.9% |
| U.S. History | 3,168 | 17.8% | 24.1% | 32.9% | 25.3% | 58.2% |

### 6.3 Conditional Standard Error of Measurement

Measurement error is associated with every test score. The conditional standard error of measurement (*CSEM*) was computed for each reported scale score. *CSEM* was computed using an IRT-based approach using on the following formula:

$$CSEM\,(O_X \mid \theta) = \sqrt{\left[\sum_{X=0}^{MaxX} O_X^2\, p(X \mid \theta)\right] - \left[\sum_{X=0}^{MaxX} O_X \cdot p(X \mid \theta)\right]^2} \qquad (5)$$

where $O_X$ is the observed scaled score for a particular number-correct score $X$, $\theta$ is the IRT ability scale value conditioned on, and $p(\bullet)$ is the probability function. Pearson has implemented a computational approach for estimating $CSEM(O_x \mid \theta)$ in which $p(X \mid \theta)$ is computed using a recursive algorithm given by Thissen, Pommerich, Billeaud, and Williams (1995). This algorithm is a polytomous generalization of the algorithm for dichotomous items

given by Lord and Wingersky (1984). The values of $\theta$ used with the algorithm are obtained through the true score equating process (i.e., by solving for $\theta$ through the test characteristic curve for each number-correct score, $X$). There is one *CSEM* per number-correct score. The *CSEM*s by subject appear in Tables 4.3 and Table 4.4.

### 6.4 Standard Error of Measurement

From the classical measurement theory aspect, the observed score (raw score) has two components; true score and error. A student's true score is the hypothetical average score that would result if the student took the test repeatedly under similar conditions. The error is the difference between true score and observed score. Among the three scores, only the observed score is known; the true score and error are derived from theory.

The standard error of measurement (*SEM*), as an overall test-level measure of error, is the average of all errors associated with student scores. Instead of using errors of student scores, the classical SEM is derived uses test reliability:

$$SEM = SD\sqrt{(1-r)} \tag{6}$$

where,

*SEM* = test Standard Error of Measurement of classical theory
*SD* = standard deviation of raw score
*r* = test reliability, Cronbach's Alpha in this case

The equation indicates that test reliability and *SEM* are in reverse relation; while test reliability increases, the SEM decreases. Table 6.5 presents the overall estimates of *SEM* for each of the content areas.

Table 6.5. Overall Estimates of *SEM* by Subject

| Subject | SEM |
|---|---|
| Algebra I | 3.01 |
| Biology I | 3.06 |
| English II | 3.05 |
| U.S. History | 3.17 |

Note: *SEM* = Standard Error of Measurement.

## References

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

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum.

Holland, P. W., & Thayer, D.T. (1988). Differential Item Performance and the Mantel-Haenszel Procedure. (ETS RR-86-31). Princeton, NJ: Educational Testing Service.

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

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings." *Applied Psychological Measurement, 8,* 453-461.

Kim, S. & Kolen, M. J. (2004). STUIRT: A computer program. Iowa City, IA: The University of Iowa. (Available from the web address: http://www.uiowa.edu/~casma).

Kolen, M.J. (2004). POLYEQUATE: A computer program. Iowa City, IA: The University of Iowa. (Available from the web address: http://www.uiowa.edu/~casma).

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: methods and practices (2nd ed.)*. New York: Springer.

Kraemer, H. C. (1982). Kappa coefficient. Encyclopedia of Statistical Sciences. Wiley.

Linacre, J.M. (2009). Winsteps® (Version 3.61) [Computer Software]. Beaverton, Oregon: Winsteps.com.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32,* 179–197.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149-174.

Michaelides, M. P. (2008). An Illustration of a Mantel-Haenszel Procedure to Flag Misbehaving Common Items in Test Equating. *Practical Assessment Research & Evaluation, 13(7).* Available online: http://pareonline.net/pdf/v13n7.pdf

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

Stocking, M.L., & Lord, F.M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201-210.

Thissen, D., Pommerich, M., Billeaud, K., & Williams, V.S.L. (1995). Item response theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement, 19,* 39-49.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5,* 245–262.

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,* 125-145.

# Appendix A

## Standards, Objectives/Skills, and Processes Assessed by Subject

| Algebra I | |
|---|---|
| **Standard 1: Number Sense and Algebraic Operations** | |
| Standard 1.1 | Equations and Formulas |
| | 1.1a Translate |
| | 1.1b Literal Equations |
| | 1.1c Problem Solving with Formulas |
| | 1.1d Problem Solving |
| Standard 1.2 | Expressions |
| | 1.2a Simplify expressions… |
| | 1.2b Compute with polynomials… |
| | 1.2c Factor polynomials |
| | |
| **Standard 2: Relations and Functions** | |
| Standard 2.1 | Relations/Functions |
| | 2.1a Distinguish linear and nonlinear |
| | 2.1b Distinguish between relations… |
| | 2.1c Dependent, Independent, Domain, Range |
| | 2.1d Evaluate a function… |
| Standard 2.2 | Linear Equations and Graphs |
| | 2.2a Solve linear equations |
| | 2.2b Graph Transformations |
| | 2.2c Slope |
| | 2.2d Equation of a Line |
| | 2.2e Match to a graph, table, etc. |
| Standard 2.3 | Linear Inequalities and Graphs |
| | 2.3a Solve linear inequalities |
| | 2.3b Match to a table, graph, etc. |
| Standard 2.4 | Systems of Equations |
| | |
| **Standard 3: Data Analysis, Probability & Statistics** | |
| Standard 3.1 | Data Analysis |
| | 3.1a Data Representations |
| | 3.1b Data Predictions |
| | 3.1c Problem Solving |
| Standard 3.2 | Line of Best Fit |

| Biology I | |
|---|---|
| ***PASS* Process/Inquiry Standards and Objectives** | |
| **Process 1 Observe and Measure** | |
| P1.1 | Qualitative/quantitative observations and changes |
| P1.2<br>P1.3 | Use appropriate System International (SI) units and tools |
| **Process 2 Classify** | |
| P2.1 | Use observable properties to classify |
| P2.2 | Identify properties of a classification system |
| **Process 3 Experiment** | |
| P3.1 | Evaluate the design of investigations |
| P3.2<br>P3.4 | Identify a testable hypothesis, variables, and control in an experiment |
| P3.3 | Use mathematics to show relationships |
| P3.5 | Identify potential hazards and practice safety procedures in all science activities |
| **Process 4 Interpret and Communicate** | |
| P4.1 | Select predictions based on observed patterns of evidence |
| P4.3 | Interpret line, bar, trend, and circle graphs |
| P4.4 | Accept or reject a hypothesis |
| P4.5 | Make logical conclusions based on experimental data |
| P4.8 | Identify an appropriate graph or chart |
| **Process 5 Model** | |
| P5.1 | Interpret a model which explains a given set of observations |
| P5.2 | Select predictions based on models |
| | |
| ***PASS* Content Standards** | |
| **Standard 1 The Cell** | |
| 1.1 | Cell structures and functions |
| 1.2 | Differentiation of cells |
| **Standard 2 The Molecular Basis of Heredity** | |
| 2.1 | DNA structure and function in heredity |
| 2.2 | Sorting and recombination of genes |
| **Standard 3 Biological Diversity** | |
| 3.1 | Variation among organisms |
| 3.2 | Natural selection and biological adaptations |
| **Standard 4 The Interdependence of Organisms** | |
| 4.1 | Earth cycles including abiotic and biotic factors |
| 4.2 | Organisms both cooperate and compete |
| 4.3 | Population dynamics |
| **Standard 5 Matter/Energy/Organization in Living Systems** | |
| 5.1 | Complexity and organization used for survival |
| 5.2 | Matter and energy flow in living and nonliving systems |

| Biology I continued | |
|---|---|
| **Standard 6 The Behavior of Organisms** | |
| 6.1 | Specialized cells |
| 6.2 | Behavior patterns can be used to ensure reproductive success |

| English II | |
|---|---|
| **Reading/Literature** | |
| **Standard 1 Vocabulary** | |
| **Standard 2 Comprehension** | |
| 2.1 | Literal Understanding |
| 2.2 | Inferences and Interpretation |
| 2.3 | Summary and Generalization |
| 2.4 | Analysis and Evaluation |
| **Standard 3 Literature** | |
| 3.1 | Literary Genres |
| 3.2 | Literary Elements |
| 3.3 | Figurative Language |
| 3.4 | Literary Works |
| **Standard 4 Research and Information** | |
| | |
| **Writing/Grammar/Usage and Mechanics** | |
| **Standard 1/2 Writing** | |
| | Writing Prompt |
| **Standard 3 Grammar/Usage and Mechanics** | |
| 3.1 | Standard Usage |
| 3.2 | Mechanics and Spelling |
| 3.3 | Sentence Structure |

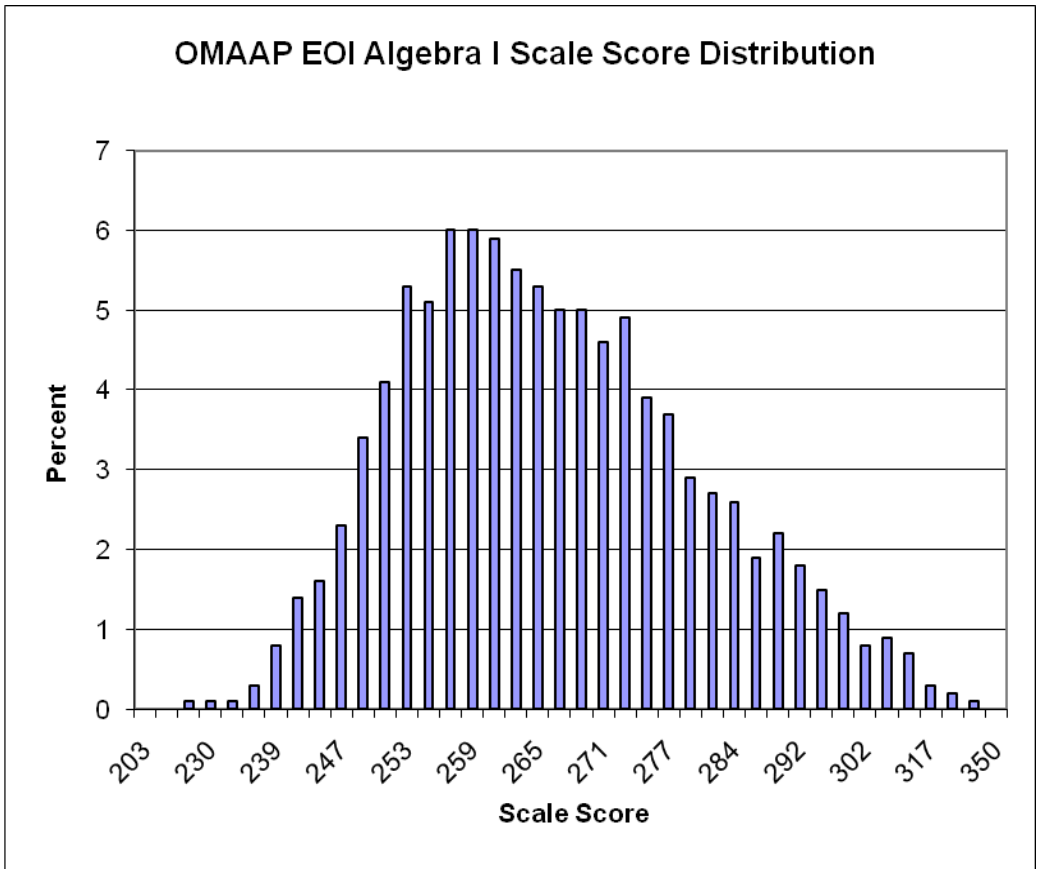
| U.S. History | |
|---|---|
| **Standard 1 Civil War/Reconstruction Era** | |
| | |
| **Standard 2 Impact of Immigration and Industrialization** | |
| 2.1 | Immigration and Impact on Native Americans |
| 2.2 | Industrialization |
| **Standard 3 Imperialism, World War I, and Isolationism** | |
| 3.1 | American Imperialism |
| 3.2 | World War I and Isolationism |
| **Standard 4 United States During the 1920s and 1930s** | |
| 4.1 | Cultural Life Between the Wars |
| 4.2 | Economic Destabilization |
| 4.3 | The Great Depression, the Dust Bowl, and the New Deal |
| **Standard 5 World War II** | |
| 5.1 | Preparing for War |
| 5.2 | World War II |
| **Standard 6 United States Since World War II** | |
| 6.1 | Post War Foreign Policies and Events |
| 6.2 | Events Changing Domestic and Foreign Policies and Events |
| 6.3 | Post War Domestic Policies and Events |

# Appendix B

## Scale Score Distributions

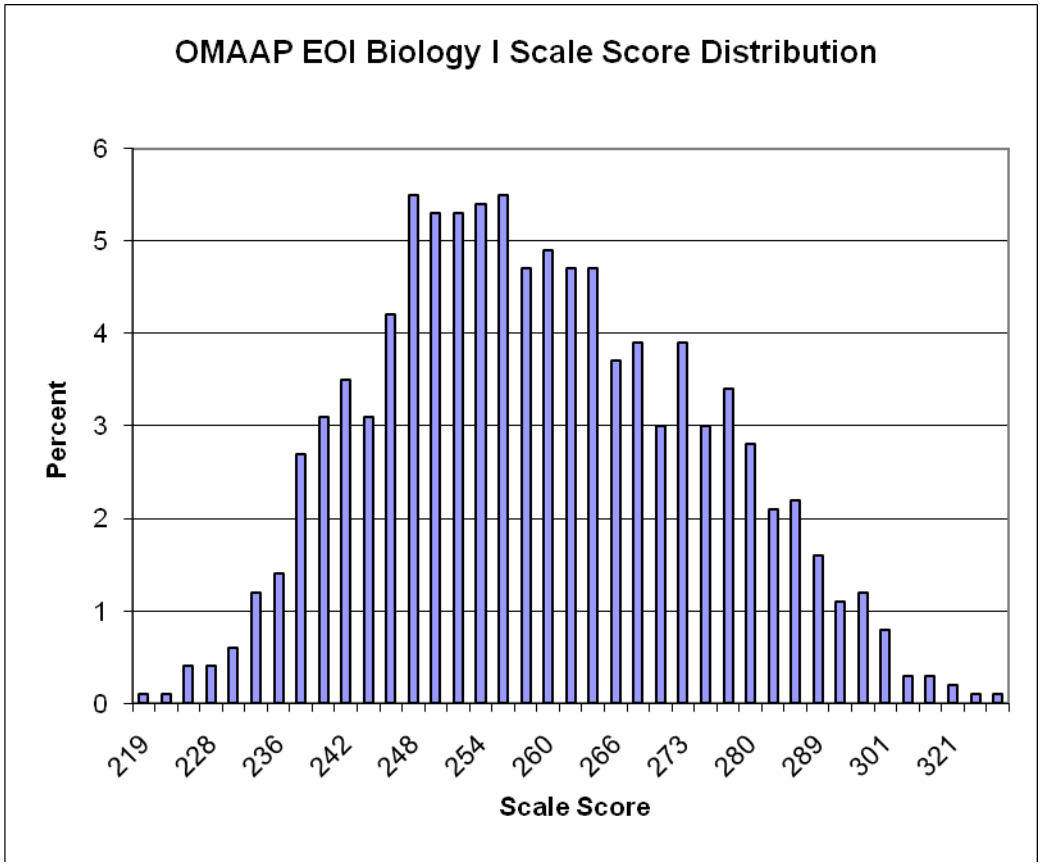
Algebra I Scale Score Distribution

| Raw Score | Scale Score | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|---|
| 2 | 203 | 1 | 0.0 | 1 | 0.0 |
| 4 | 217 | 1 | 0.0 | 2 | 0.1 |
| 6 | 227 | 3 | 0.1 | 5 | 0.1 |
| 7 | 230 | 4 | 0.1 | 9 | 0.2 |
| 8 | 234 | 4 | 0.1 | 13 | 0.3 |
| 9 | 237 | 13 | 0.3 | 26 | 0.6 |
| 10 | 239 | 37 | 0.8 | 63 | 1.4 |
| 11 | 242 | 63 | 1.4 | 126 | 2.9 |
| 12 | 244 | 70 | 1.6 | 196 | 4.4 |
| 13 | 247 | 101 | 2.3 | 297 | 6.7 |
| 14 | 250 | 151 | 3.4 | 448 | 10.2 |
| 15 | 251 | 182 | 4.1 | 630 | 14.3 |
| 16 | 253 | 232 | 5.3 | 862 | 19.6 |
| 17 | 255 | 224 | 5.1 | 1,086 | 24.6 |
| 18 | 257 | 263 | 6.0 | 1,349 | 30.6 |
| 19 | 259 | 263 | 6.0 | 1,612 | 36.6 |
| 20 | 261 | 258 | 5.9 | 1,870 | 42.4 |
| 21 | 263 | 244 | 5.5 | 2,114 | 47.9 |
| 22 | 265 | 235 | 5.3 | 2,349 | 53.3 |
| 23 | 267 | 219 | 5.0 | 2,568 | 58.2 |
| 24 | 269 | 222 | 5.0 | 2,790 | 63.3 |
| 25 | 271 | 203 | 4.6 | 2,993 | 67.9 |
| 26 | 273 | 214 | 4.9 | 3,207 | 72.7 |
| 27 | 275 | 171 | 3.9 | 3,378 | 76.6 |
| 28 | 277 | 162 | 3.7 | 3,540 | 80.3 |
| 29 | 279 | 127 | 2.9 | 3,667 | 83.2 |
| 30 | 282 | 121 | 2.7 | 3,788 | 85.9 |
| 31 | 284 | 115 | 2.6 | 3,903 | 88.5 |
| 32 | 287 | 82 | 1.9 | 3,985 | 90.4 |
| 33 | 289 | 99 | 2.2 | 4,084 | 92.6 |
| 34 | 292 | 77 | 1.8 | 4,161 | 94.4 |
| 35 | 295 | 64 | 1.5 | 4,225 | 95.8 |
| 36 | 298 | 54 | 1.2 | 4,279 | 97.0 |
| 37 | 302 | 33 | 0.8 | 4,312 | 97.8 |
| 38 | 306 | 40 | 0.9 | 4,352 | 98.7 |
| 39 | 311 | 29 | 0.7 | 4,381 | 99.3 |
| 40 | 317 | 14 | 0.3 | 4,395 | 99.7 |
| 41 | 326 | 9 | 0.2 | 4,404 | 99.9 |
| 42 | 339 | 5 | 0.1 | 4,409 | 100.0 |
| 43 | 350 | 1 | 0.0 | 4,410 | 100.0 |


OMAAP EOI Algebra I Scale Score Distribution
Percent
0
1
2
3
4
5
6
7
203
230
239
247
253
259
265
271
277
284
292
302
317
350
Scale Score

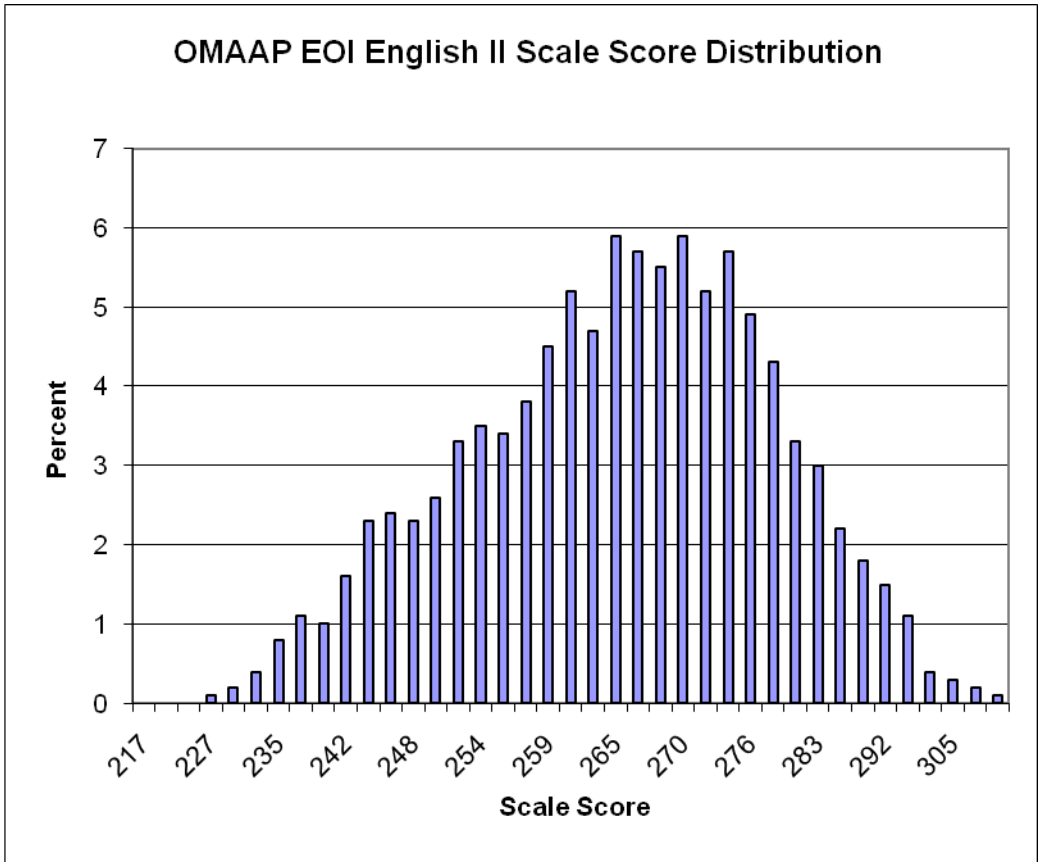
Biology I Scale Score Distribution

| Raw Score | Scale Score | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|---|
| 8 | 219 | 3 | 0.1 | 3 | 0.1 |
| 9 | 223 | 3 | 0.1 | 6 | 0.2 |
| 10 | 225 | 15 | 0.4 | 21 | 0.6 |
| 11 | 228 | 16 | 0.4 | 37 | 1.0 |
| 12 | 231 | 22 | 0.6 | 59 | 1.5 |
| 13 | 233 | 45 | 1.2 | 104 | 2.7 |
| 14 | 236 | 55 | 1.4 | 159 | 4.2 |
| 15 | 238 | 103 | 2.7 | 262 | 6.8 |
| 16 | 240 | 117 | 3.1 | 379 | 9.9 |
| 17 | 242 | 132 | 3.5 | 511 | 13.3 |
| 18 | 244 | 117 | 3.1 | 628 | 16.4 |
| 19 | 246 | 161 | 4.2 | 789 | 20.6 |
| 20 | 248 | 212 | 5.5 | 1,001 | 26.1 |
| 21 | 250 | 201 | 5.3 | 1,202 | 31.4 |
| 22 | 252 | 203 | 5.3 | 1,405 | 36.7 |
| 23 | 254 | 205 | 5.4 | 1,610 | 42.0 |
| 24 | 256 | 211 | 5.5 | 1,821 | 47.5 |
| 25 | 258 | 180 | 4.7 | 2,001 | 52.2 |
| 26 | 260 | 189 | 4.9 | 2,190 | 57.2 |
| 27 | 262 | 179 | 4.7 | 2,369 | 61.8 |
| 28 | 265 | 178 | 4.7 | 2,547 | 66.5 |
| 29 | 266 | 143 | 3.7 | 2,690 | 70.2 |
| 30 | 269 | 148 | 3.9 | 2,838 | 74.1 |
| 31 | 271 | 116 | 3.0 | 2,954 | 77.1 |
| 32 | 273 | 150 | 3.9 | 3,104 | 81.0 |
| 33 | 275 | 113 | 3.0 | 3,217 | 84.0 |
| 34 | 278 | 129 | 3.4 | 3,346 | 87.3 |
| 35 | 280 | 107 | 2.8 | 3,453 | 90.1 |
| 36 | 283 | 81 | 2.1 | 3,534 | 92.3 |
| 37 | 286 | 83 | 2.2 | 3,617 | 94.4 |
| 38 | 289 | 61 | 1.6 | 3,678 | 96.0 |
| 39 | 292 | 42 | 1.1 | 3,720 | 97.1 |
| 40 | 296 | 46 | 1.2 | 3,766 | 98.3 |
| 41 | 301 | 29 | 0.8 | 3,795 | 99.1 |
| 42 | 306 | 10 | 0.3 | 3,805 | 99.3 |
| 43 | 312 | 12 | 0.3 | 3,817 | 99.6 |
| 44 | 321 | 9 | 0.2 | 3,826 | 99.9 |
| 45 | 335 | 3 | 0.1 | 3,829 | 100.0 |
| 46 | 350 | 2 | 0.1 | 3,831 | 100.0 |


OMAAP EOI Biology I Scale Score Distribution
Percent
0
1
2
3
4
5
6
219
228
236
242
248
254
260
266
273
280
289
301
321
Scale Score

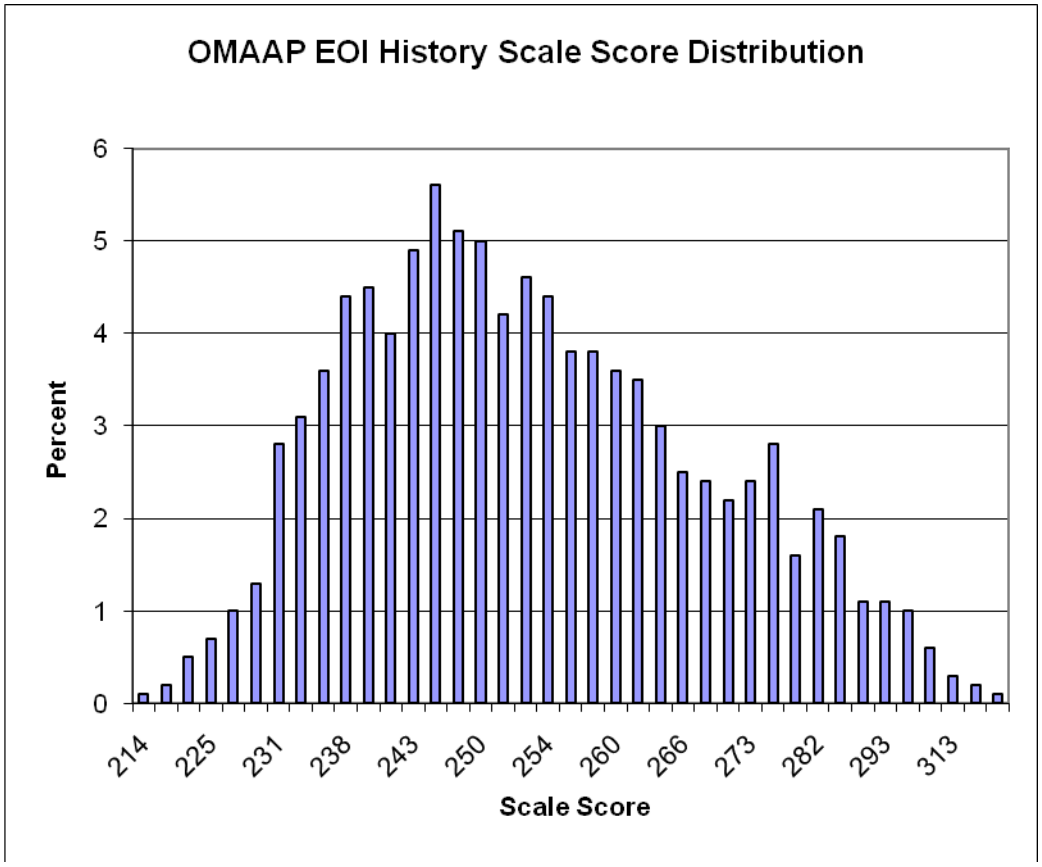
English II Scale Score Distribution

| Raw Score | Scale Score | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|---|
| 6 | 217 | 1 | 0.0 | 1 | 0.0 |
| 7 | 221 | 1 | 0.0 | 2 | 0.1 |
| 8 | 224 | 1 | 0.0 | 3 | 0.1 |
| 9 | 227 | 2 | 0.1 | 5 | 0.1 |
| 10 | 230 | 6 | 0.2 | 11 | 0.3 |
| 11 | 233 | 16 | 0.4 | 27 | 0.7 |
| 12 | 235 | 29 | 0.8 | 56 | 1.5 |
| 13 | 238 | 40 | 1.1 | 96 | 2.5 |
| 14 | 240 | 39 | 1.0 | 135 | 3.6 |
| 15 | 242 | 61 | 1.6 | 196 | 5.2 |
| 16 | 244 | 86 | 2.3 | 282 | 7.5 |
| 17 | 246 | 90 | 2.4 | 372 | 9.9 |
| 18 | 248 | 86 | 2.3 | 458 | 12.1 |
| 19 | 250 | 97 | 2.6 | 555 | 14.7 |
| 20 | 252 | 126 | 3.3 | 681 | 18.0 |
| 21 | 254 | 131 | 3.5 | 812 | 21.5 |
| 22 | 256 | 128 | 3.4 | 940 | 24.9 |
| 23 | 258 | 142 | 3.8 | 1,082 | 28.6 |
| 24 | 259 | 169 | 4.5 | 1,251 | 33.1 |
| 25 | 261 | 196 | 5.2 | 1,447 | 38.3 |
| 26 | 263 | 176 | 4.7 | 1,623 | 43.0 |
| 27 | 265 | 224 | 5.9 | 1,847 | 48.9 |
| 28 | 267 | 215 | 5.7 | 2,062 | 54.6 |
| 29 | 268 | 206 | 5.5 | 2,268 | 60.0 |
| 30 | 270 | 224 | 5.9 | 2,492 | 66.0 |
| 31 | 272 | 196 | 5.2 | 2,688 | 71.2 |
| 32 | 274 | 217 | 5.7 | 2,905 | 76.9 |
| 33 | 276 | 184 | 4.9 | 3,089 | 81.8 |
| 34 | 279 | 161 | 4.3 | 3,250 | 86.0 |
| 35 | 281 | 126 | 3.3 | 3,376 | 89.4 |
| 36 | 283 | 114 | 3.0 | 3,490 | 92.4 |
| 37 | 286 | 83 | 2.2 | 3,573 | 94.6 |
| 38 | 289 | 69 | 1.8 | 3,642 | 96.4 |
| 39 | 292 | 55 | 1.5 | 3,697 | 97.9 |
| 40 | 296 | 40 | 1.1 | 3,737 | 98.9 |
| 41 | 300 | 16 | 0.4 | 3,753 | 99.3 |
| 42 | 305 | 13 | 0.3 | 3,766 | 99.7 |
| 43 | 312 | 8 | 0.2 | 3,774 | 99.9 |
| 44 | 320 | 4 | 0.1 | 3,778 | 100.0 |


OMAAP EOI English II Scale Score Distribution
Percent
0
1
2
3
4
5
6
7
217
227
235
242
248
254
259
265
270
276
283
292
305
Scale Score

U.S. History Scale Score Distribution

| Raw Score | Scale Score | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|---|
| 8 | 214 | 3 | 0.1 | 3 | 0.1 |
| 10 | 220 | 6 | 0.2 | 9 | 0.3 |
| 11 | 222 | 17 | 0.5 | 26 | 0.8 |
| 12 | 225 | 23 | 0.7 | 49 | 1.6 |
| 13 | 227 | 31 | 1.0 | 80 | 2.5 |
| 14 | 229 | 40 | 1.3 | 120 | 3.8 |
| 15 | 231 | 90 | 2.8 | 210 | 6.6 |
| 16 | 234 | 99 | 3.1 | 309 | 9.8 |
| 17 | 236 | 115 | 3.6 | 424 | 13.4 |
| 18 | 238 | 139 | 4.4 | 563 | 17.8 |
| 19 | 239 | 141 | 4.5 | 704 | 22.2 |
| 20 | 241 | 128 | 4.0 | 832 | 26.3 |
| 21 | 243 | 156 | 4.9 | 988 | 31.2 |
| 22 | 245 | 178 | 5.6 | 1,166 | 36.8 |
| 23 | 247 | 160 | 5.1 | 1,326 | 41.9 |
| 24 | 250 | 159 | 5.0 | 1,485 | 46.9 |
| 25 | 251 | 133 | 4.2 | 1,618 | 51.1 |
| 26 | 252 | 145 | 4.6 | 1,763 | 55.7 |
| 27 | 254 | 138 | 4.4 | 1,901 | 60.0 |
| 28 | 256 | 121 | 3.8 | 2,022 | 63.8 |
| 29 | 258 | 119 | 3.8 | 2,141 | 67.6 |
| 30 | 260 | 115 | 3.6 | 2,256 | 71.2 |
| 31 | 262 | 112 | 3.5 | 2,368 | 74.8 |
| 32 | 264 | 95 | 3.0 | 2,463 | 77.8 |
| 33 | 266 | 79 | 2.5 | 2,542 | 80.2 |
| 34 | 268 | 76 | 2.4 | 2,618 | 82.6 |
| 35 | 271 | 71 | 2.2 | 2,689 | 84.9 |
| 36 | 273 | 75 | 2.4 | 2,764 | 87.3 |
| 37 | 276 | 89 | 2.8 | 2,853 | 90.1 |
| 38 | 279 | 52 | 1.6 | 2,905 | 91.7 |
| 39 | 282 | 65 | 2.1 | 2,970 | 93.8 |
| 40 | 285 | 57 | 1.8 | 3,027 | 95.6 |
| 41 | 289 | 36 | 1.1 | 3,063 | 96.7 |
| 42 | 293 | 34 | 1.1 | 3,097 | 97.8 |
| 43 | 298 | 32 | 1.0 | 3,129 | 98.8 |
| 44 | 304 | 20 | 0.6 | 3,149 | 99.4 |
| 45 | 313 | 10 | 0.3 | 3,159 | 99.7 |
| 46 | 327 | 5 | 0.2 | 3,164 | 99.9 |
| 47 | 350 | 4 | 0.1 | 3,168 | 100.0 |


OMAAP EOI History Scale Score Distribution
Percent
0
1
2
3
4
5
6
214
225
231
238
243
250
254
260
266
273
282
293
313
Scale Score