OKLAHOMA STATE DEPARTMENT OF EDUCATION

Janet Barresi, State Superintendent of Public Instruction

Oklahoma
Modified Alternate Assessment Program

End-of-Instruction Assessments
2013 Technical Report

Submitted to
The Oklahoma State Department of Education
November 2013

Copyright

Introduction

The Oklahoma School Testing Program (OSTP) is a statewide assessment program that was established to improve academic achievement for all Oklahoma students. It also meets the requirements of the No Child Left Behind Act (NCLB) introduced by the federal government in 2001. The OSTP includes grades 3─8 and high school End-of-Instruction (EOI) assessments, for which students who complete an area of instruction must also take the corresponding statewide standardized assessment.

The OSTP developed three types of tests to assess the three groups of students defined by NCLB: the Oklahoma Core Curriculum Tests (OCCT) for the general student population, the Oklahoma Modified Alternate Assessment Program (OMAAP) for students who are instructed on grade level content and whose IEP indicates they need the modified assessment, and the Oklahoma Alternate Assessment Program (OAAP) for students with the significant cognitive disabilities. All three tests cover students in grades 3─8 and high school.

The *Oklahoma College, Career, and Citizen Ready Standards (OK C^3)* academic content standards are the foundation for all three tests. The *Curriculum Access Resource Guides* (*CARG*) describe access points to the *OK C*^{3} through scaffolding of skills. An alternate guide, the *CARG-A*, provides guidance for instruction and assessment of Oklahoma students with the most significant cognitive disabilities.

The Oklahoma Modified Alternate 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 Alternate Assessment Program (OAAP) portfolio assessments are inappropriate. The OMAAP EOI exams are based on modified blueprints and items from the corresponding OCCT EOI exams.

In 2013, CTB/McGraw-Hill was contracted by the Oklahoma State Department of Education (SDE) to develop, administer, and maintain the OMAAP EOI winter, trimester, spring, and summer administrations. Because the testing population is small for the winter, trimester, and summer tests, past spring tests are used for these administrations. Therefore, this technical report focuses on the details of work accomplished for the spring test only.

Purpose

The purpose of this technical report is to provide 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 test 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. Section 6 provides higher-level summaries of all the tests included in the OMAAP EOI testing program.

Table of Contents

Copyright i
Introduction ii
Purpose ii
Table of Contents iv
List of Tables and Figures vi
Acronyms and Abbreviations vii
Section 1 1
1.1 Overview of the OMAAP EOI Assessments 1
1.2 Content Standards 1
1.3 Blueprint 3
1.4 Universal Design and Modifications 3
1.5 Test Development and Content Validity 4
1.5.a. Item Development and Selection 4
1.5.b. Test Construction Guidelines 5
1.5.c. Configuration of the Tests 5
Section 2 9
2.1 Packaging and Shipping 9
2.2 Materials Return 9
2.3 Materials Discrepancies Process 10
Section 3 11
3.1 Data Quality Check and Clean-Up 11
3.2 Classical Item Analyses 13
3.2.a. Test-Level Summaries of Classical Item Analyses 13
3.3 Procedures for Detecting Item Bias 14
3.3.a. Differential Item Functioning Results 14
3.4 Test Reliability 15
3.4.a. Overall Test Reliability 15
3.4.b. Test Reliability by Subgroup 16
3.5 Inter-rater Reliability 17
Section 4 18
4.1 Item Response Theory (IRT) Models 18
4.1.a. Dichotomous Item Response Theory Model 18
4.1.b. Polytomous Item Response Theory Model 18
4.2 Assessment of Fit to the Model 18
4.3 Calibration and Equating 19
4.4 Anchor Item Stability Evaluation Methods 19
4.4.a Anchor Items for Spring 2013 21
4.5 Scaling and Scoring Results 22
Section 5 25
5.1 Classification Consistency and Accuracy 25
Section 6 28
6.1 Descriptive Statistics 28
6.2 Performance Level Distribution 29
6.3 Conditional Standard Error of Measurement 30
6.4 Standard Error of Measurement 31

References 32
Appendix A: Standards, Objectives/Skills, and Processes Assessed by Subject 33
Appendix B: Scale Score Distributions 39

List of Tables and Figures

Table 1.1. Oklahoma C^3 Standards by Subject 2
Table 1.2. OCCT and OMAAP EOI Item Count Comparison 3
Table 1.3. OMAAP Item Modification Rules 3
Table 1.4. Test Construction Guidelines 5
Table 1.5. Percentage of Items in Depth of Knowledge Levels 6
Table 1.6a. Number of Items by Content Standard for Algebra I 6
Table 1.6b. Number of Items by Content Standard for Biology I 7
Table 1.6c. Number of Items by Content Standard for English II 7
Table 1.6d. Number of Items by Content Standard for U.S. History 8
Table 3.1. Demographic Characteristics of Calibration and Equating Sample 11
Table 3.1.a. Demographic Characteristics of Calibration and Equating Sample (continued) 11
Table 3.2. Secondary Statistical Key Check Criteria 12
Table 3.3. Number of Item Removed and Maximum Score Point Possible 12
Table 3.4. Test-Level Summaries of Classical Item Analyses 13
Table 3.5. DIF Flag Incidence Across All OMAAP EOI Items 15
Table 3.6. Cronbach's Alpha by Subject 16
Table 3.7. Test Reliability by Subgroup 16
Table 3.7.a. Test Reliability by Subgroup (continued) 16
Table 3.8. Percentage of Students at Each Point Discrepancy Between the Two Raters 17
Table 3.8.a. Inter-rater Reliability for English II Operational Writing Prompts 17
Table 4.1. Number of Anchor Items per Subject 22
Table 4.2. OMAAP Scaling Constants, Scale Range, and Cut Scores by Subject 22
Table 4.3. Raw Score to Scale Score Conversion Table for Algebra I and Biology I 23
Table 4.4. Raw Score to Scale Score Conversion Table for English II and U.S. History 24
Table 5.1. Estimates of Accuracy and Consistency of Performance Classification 26
Table 5.2. Accuracy and Consistency Estimates by Cut Score: False Positive and False Negative Rates 27
Table 5.2.a. Accuracy and Consistency Estimates by Cut Score: (cont.) 27
Table 6.1. Scale Score Descriptive Statistics – Overall 28
Table 6.2. Scale Score Descriptive Statistics by Gender 28
Table 6.3. Scale Score Descriptive Statistics by Race/Ethnicity 28
Table 6.3.a. Scale Score Descriptive Statistics by Race/Ethnicity (cont.) 29
Table 6.3.b. Scale Score Descriptive Statistics by Race/Ethnicity (cont.) 29
Table 6.3.c. Scale Score Descriptive Statistics by Race/Ethnicity (cont.) 29
Table 6.4. Percentage of Students by Performance Level 29
Table 6.5. Overall Estimates of SEM by Subject 31

Acronyms and Abbreviations

2PPC Two Parameter Partial Credit model
3PL Three Parameter Logistic model
ACE Achieving Classroom Excellence
AERA American Educational Research Association
APA American Psychological Association
AYP Adequate Yearly Progress
BR Braille
BTC Building Test Coordinator
C^3 Oklahoma's Core curriculum, the College, Career and Citizen Ready
CCSSO Council of Chief State School Officers
CE Critical Element
CR Constructed-Response
CSEM Conditional Standard Error of Measurement
DIF Differential Item Functioning
DOK Depth of Knowledge
DTC District Test Coordinator
ELL English Language Learners
EOI End-of-Instruction
FP False Positive
EQ Equivalent
FN False Negative
GR Gridded-Response
GRT General Research Tape
HOSS Highest Obtainable Scale Score
ICC Item Characteristic Curve
IEP Individualized Education Program
IRT Item Response Theory
LIU Language in Use
LOSS Lowest Obtainable Scale Score
MC Multiple-Choice
MH Mantel-Haenszel
NCES National Center for Education Statistics
NCLB No Child Left Behind
NCME National Council on Measurement in Education
NGA National Governors Association Center
OAAP Oklahoma Alternate Assessment Program
OAC Oklahoma Administrative Code
OCCT Oklahoma Core Curriculum Test
OE Open-Ended
OK C^3 Oklahoma College, Career, and Citizen Ready Standards
OMAAP Oklahoma Modified Alternate Assessment Program
OP Operational
OSTP Oklahoma School Testing Program
PASS Priority Academic Student Skills
RT Retest
SAS Statistical Analysis System
SD Standard Deviation
SDE Oklahoma State department of Education
SEM Standard Error of Measurement
SS Scale Score
TA Test Administrator
TCC Test Characteristic Curve
TP Test Proctor
TPM Test Preparation Manual
US DOE United States Department of Education
WP Writing Prompt

Section 1
Overview of the Oklahoma Modified Alternate Assessment Program (OMAAP) End-of-Instruction (EOI) Assessments

1.1 Overview of the OMAAP EOI Assessments

The Oklahoma Modified Alternate Assessment Program (OMAAP) End-of-Instruction (EOI) assessments are a collection of tests that are mandated by the state that are secondary-level and criterion-referenced. The tests are given at the high school level to evaluate the proficiency level of the student. The OMAAP test is given to students for whom the Oklahoma Alternate Assessment Program (OAAP) portfolio assessment and the general Oklahoma Core Curriculum Tests (OCCT) are not appropriate.

The OMAAP EOI tests are given to assess each student's proficiency in relation to the *Oklahoma C^3 Standards* that were instituted by educational committees in Oklahoma. SDE has officially required all students since 2009 to be tested through OCCT, OMAAP, or alternate EOI assessments in order to be eligible for graduation. It is required as part of this new policy that in order to graduate with a high school diploma from the State of Oklahoma, students have to be within the performance level of proficient or above in four subjects: Algebra I, English II, and two of the following: Algebra II, Biology I, English III, Geometry, or U.S. History. If a student fails to meet this standard of performance he or she is allowed to retake the test. Since OMAAP EOI only has Algebra I, Biology I, English II, and U.S. History subjects, students who take the test for OMAAP EOI must pass all four subjects to be eligible to receive a high school diploma.

Assessments for Spring 2013 OMAAP EOI were administered by CTB/McGraw-Hill with collaboration from the Oklahoma State Department of Education (SDE). Scoring, equating, and scaling for the OMAAP EOI assessments were also done by CTB/McGraw-Hill. Except for English II which has a writing prompt, all of the OMAAP EOI tests have multiple choice items only. One OMAAP EOI operation (OP) form was constructed in Spring 2013 for each subject, and CTB used the OP forms to generate the Braille forms. The equivalent form was constructed using operational forms from previous years. Students that were unable to take the operational form exam due to illness or a security breach took the equivalent form instead. The SDE Office of Accountability and Assessments determined which students would qualify for the need to take equivalent form on a case-by-case basis.

1.2 Content Standards

The OMAAP EOI assessments were developed to measure the *Oklahoma C^3 Standards* for high school. Table 1.1 outlines the *OK C^3* content standards by subject. Appendix A outlines the objectives of each content and/or process standard. The SDE provided *Curriculum Access Resource Guides* (*CARG*) which offers assistance to teachers by illustrating various methods of incorporating the *Oklahoma C^3 Standards* into classroom instruction through the appropriate development of skills.

Table 1.1. Oklahoma C^3 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.	Matter/Energy/Organization in Living Systems
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.	Post-Reconstruction to the Progressive Era, 1878-1900
Standard 2.	Expanding Role of the United States in International Affairs
Standard 3.	Cycles of Economic Boom and Bust in the 1920s and 1930s
Standard 4.	Role of the U.S. in International Affairs and World War II, 1933-1946
Standard 5.	U.S. Foreign and Domestic Policies during the Cold War, 1945-1975

1.3 Blueprint

OMAAP EOI assessments have fewer items than the OCCT EOI, but with the same or similar proportion of items across the various standards. To be a score reporting category a standard must have at least five items. The blueprint for OMAAP EOI was suggested by committees of teachers and administrators. These committees reviewed the *OK C*^{3} standards and the OCCT blueprint and came to the decision that the OMAAP EOI blueprints were appropriate for OMAAP student populations. In a final review, the SDE proposed the test blueprint and submitted it to the School Board of Education for its approval.

72-78% of the operational items on the OCCT exams were retained for the OMAAP forms. It should be noted that there were no field test items in the OMAAP EOI test forms. Table 1.2 shows a comparison of item counts of the OCCT and the OMAAP EOI tests.

Table 1.2. OCCT and OMAAP EOI Item Count Comparison

	OCCT		OMAAP	
Subject	**OP**	**FT**	**OP**	**FT**
Algebra I	55	20	45	0
Biology I	60	20	48	0
English II*	60/1	20/0	43/1	0
U.S. History	60	20	48	0

*OP=Operational, FT=Field test.
*The first number represents the count of multiple-choice items and the second number represents the count of constructed-response item.

1.4 Universal Design and Modifications

OMAAP EOI item and test formats follow the *Universal Design* guidelines, ensuring tests are appropriate for students with various needs. Subject specific modifications have also been applied to increase test suitability. Table 1.3 lists the *Universal Design* and subject specific modifications.

Table 1.3. OMAAP Item Modification Rules

Universal Design
Minimize the number of questions on the page (limit to 2)
Use a larger font size
Provide only three answer options instead of four
Highlight the main points in the question or passage by underlining and using boldface
Allow for the same accommodations as in the standard assessment
Avoid questions that require students to select the better/best answer
Eliminate answer choices that give students the option of making “no change” to the item

Continues on Next Page

Table 1.3. OMAAP Item Modification Rules (*Continued*)

Algebra I
Allow for read-aloud and calculators
Unless required by standard, avoid items with negative and positive answer choices that use the same number (e.g., −4 and +4)
Place any items with coordinate grids on one page
Be consistent with qualifiers in the stem and answer choices (e.g., use ml throughout or milliliters throughout)
Avoid questions that use "best" or "closest"
Avoid complicated art
Biology I
Reduce the amount of reading
Avoid complicated art
Simplify tables and charts by removing irrelevant rows or columns
Box formulas to make them stand out
English II – Reading Items
Display passages in a one column format
Break passages into smaller portions, and place the questions that pertain to the smaller portion
English II – Writing Prompt
Simplify the question
Update the checklist, simplifying it and describing the aspects that will be graded so it matches the new rubric
Use a three-point holistic writing rubric

1.5 Test Development and Content Validity

Development of a test relies upon test specifications to guide the construction process. Content validity is determined by specifications in content standards and test blueprints. Content standards address the knowledge and skills which are to be measured through the test, and test blueprints outline the number of items and item types to be included in each content area. The degree of content validity of a test is based on how closely it adheres to the specifications set forth. The closer the test is to meeting all specifications, the higher degree of content validity. This section describes the measures taken during the test construction process to ensure high content validity.

1.5.a. Item Development and Selection

The OMAAP EOI test design requires items to be pulled from two sources: anchor items to be selected from previously-used OMAAP items and non-anchor items to be selected from previously-used OCCT EOI items. Since anchor items are selected from previously-used OMAAP items, they can be used directly on the new test form. It should be noted that the previously-used OMAAP items were also modified from OCCT items when they were first used on the OMAAP test. The non-anchor items were modified following the *Universal Design* guidelines.

When items were developed for the OCCT test, the cognitive level each item measured was identified through Norman Webb's Depth of Knowledge (DOK) framework by teacher committees. Modification of OCCT items for the use of OMAAP tests only simplified the items. Their content standards or cognitive level were not changed. All modified items were reviewed by item review committees for content alignment to the *Oklahoma C^3 Standards* and item appropriateness. Only the items that passed the review could be added to the OMAAP item pool.

1.5.b. Test Construction Guidelines

Besides test blueprints, the OMAAP EOI has four additional test construction guidelines: categorical concurrence, range of knowledge, balance of representation, and source of challenge (see Table 1.4).

Table 1.4. Test Construction Guidelines

Type	Guidelines
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 using items from a variety of Depth of Knowledge levels that are consistent with the processes students need in order to demonstrate proficiency for each *OK C^3* objective.
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.

1.5.c. Configuration of the Tests

Table 1.5 through 1.6d show the test blueprint and DOK target for the OMAAP EOI operational (OP), Braille, and equivalent (EQ) tests. Content experts followed the test blueprint and DOK targets when selecting items during the assembly of the tests. Every effort was taken to meet the targets; however, some targets were not met exactly due to limitations within the item bank, especially on items associated with passages, as in the subject of English II.

Spring 2013 tests had been constructed to include the maximum number of items possible; there are 45 items for Algebra I, 48 items for Biology I, 44 items for English II, and 48 items for U.S.

History. Since non-anchor items had been modified from the OCCT items and used in the OMAAP operational form, it was not until after the test administration that the statistical properties of the OMAAP population became available. Because of this, these items were only screened after the test administration and prior to equating. CTB/McGraw-Hill staff and the SDE staff reviewed the items with problematic statistical findings. Some of the items were dropped/suppressed from score reporting before equating, after the review.

Table 1.5. Percentage of Items in Depth of Knowledge Levels

DOK Level	**1**	**2**	**¾**
Target DOK	**20-25%**	**60-65%**	**10-15%**
Subject	**Operational/Braille**		
Algebra I	27%	64%*	11%*
Biology I	23%	64%*	12%
English II	27%	61%	12%
U.S. History	15%	71%	14%
	Equivalent		
Algebra I	24%	67%	4%*
Biology I	24%	61%*	15%*
English II	20%*	60%	19%
U.S. History	19%*	67%	13%

*One DOK 2 item was suppressed in Spring 2013 Algebra I and Biology I OP form.
*One DOK 3 item was suppressed in Spring 2013 Algebra I OP and EQ form.
*Two DOK 2 items and one DOK 3 item were suppressed in the Biology I EQ form.
*One DOK 1 item was suppressed in Spring 2013 English II EQ form.
*One DOK 1 item was suppressed in Spring 2013 US History EQ form.

Table 1.6a. Number of Items by Content Standard for Algebra I

Content Standard	**Target**	**Operational**	**Equivalent**	**Braille**
Standard 1	10-12	14*	12	14*
Standard 2	21-23	23	22*	23
Standard 3	6-8	9	8	9
Total	***40-43***	**45**	**42**	**45**

*Two items were suppressed in Spring 2013 OP form.
*One item was suppressed in Spring 2013 EQ form.

Table 1.6b. Number of Items by Content Standard for Biology I

Content Standard	Target	Operational	Equivalent	Braille
Standard 1	9-12	9	10	9
Standard 2	9-12	9	6*	9
Standard 3	9-12	9	10	9
Standard 4	6-8	8	8	8
Standard 5	10	10*	8*	10*
Total	***43-46***	**45**	**42**	**45**
Process Standard	**Target**	**Operational**	**Equivalent**	**Braille**
Standard 1	6	5	6	5
Standard 2	6	7	7	7
Standard 3	13-16	14*	11*	14*
Standard 4	16-19	17	16*	17
Standard 5	6	5	6	5
Total	***46-49***	**48**	**46**	**48**

*Has item(s) suppressed in that standard category.

Table 1.6c. Number of Items by Content Standard for English II

Content Standard	Target	Operational	Equivalent	Braille
R1	6-7	7	6	1
R2	9-10	10	11	10
R3	12-13	14	12*	14
R4	6-7	5	6	5
W1/W2.	1(3pts)	1 (3 pts.)	1(3 pts.)	1 (3 pts.)
W3	7-8	7	7	7
Total	***41-44 (43-46pts)***	**44 (46 pts.)**	**43 (45 pts.)**	**44 (46 pts.)**

*Has suppressed item(s) in that standard category.

Table 1.6d. Number of Items by Content Standard for U.S. History

Content Standard	Target	Operational	Equivalent	Braille
Standard 1	8	9	11	9
Standard 2	6	5	6	5
Standard 3	8	8	7	8
Standard 4	8	8	9	8
Standard 5	18	18	12*	18
Total	***48***	**48**	**45**	**48**

*Has suppressed item(s) in that standard category.

Section 2

Administration of the OCCT EOI Assessments

To ensure a valid and reliable assessment, the OCCT EOI tests are first constructed in alignment with the *Oklahoma C^3 Standards* (now *Oklahoma Academic Standards*) by the Oklahoma State Department of Education in collaboration with CTB. The tests are then administered and scored according to sound measurement principles for the purpose of evaluating validity. Additionally, best practices require that the test administering and scoring entities perform their tasks in a consistent manner throughout the state so that all students have a fair and equitable opportunity for a score that reflects their achievement in each subject.

Schools play a key role in administering the OCCT 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 best practice procedures. School faculty members play a vital role in the success of OCCT EOI assessments by ensuring fairness in administration of the test.

2.1 Packaging and Shipping

In order to provide secure and dependable services for the shipping of the OCCT EOI assessment materials, CTB's Transportation Department maintains the quality and security of material distribution and return by hiring reputable carriers that possess the ability to trace shipments. CTB uses all available tracking capabilities to provide status information and early opportunities for corrective action.

Materials are packaged by the schools and delivered to the district test coordinators. Each shipment to a district contains a shipping document set that includes a packing list for each school's materials.

Materials are packaged using information provided by the test coordinators through the CTB Precode Utility (EOI) or the Oklahoma WAVE system (Grades 3-8). Oklahoma educators also use these systems to provide CTB with the precode information needed to print student barcode labels, which are affixed on answer documents or consumable test books. The bar-coding of all secure materials at the time of production allows for accurate tracking of these materials through the entire packing, delivery, and return process. This allows CTB to inventory all materials throughout the packaging and delivery process.

2.2 Materials Return

The Test Preparation Manual and Materials Return poster provide clear instructions on how to assemble, box, label, and return testing materials after test administration. CTB utilizes double-column boxes to distribute and collect test materials, and makes additional cartons available for order to meet the various return needs of the districts.

Stack cards and paper bands are provided to group and secure used student response booklets for scoring. Color-coded return labels with pre-printed return information are also provided. These labels facilitate the sorting of each carton and its contents upon receipt at CTB's Data Processing Facility.

2.3 Materials Discrepancies Process

The scanning process allows CTB to capture multiple-choice responses and student writing images. Test security form information is also captured electronically via a secure database. All scorable material discrepancies are captured, investigated by the CTB Oklahoma Help Desk, reported, and the results subsequently reported to the Oklahoma State Department of Education (SDE).

A pre-determined date is set by SDE and CTB to account for any materials that arrive after the scheduled deadline. Late arriving material is processed up to the agreed upon date, at which point the Oklahoma SDE must be notified of any late arriving documents and render a processing decision. Following an initial call campaign to all districts with outstanding secure material, the CTB Oklahoma Program Management team notifies the SDE regarding unresolved material discrepancies presented in a preliminary file. A subsequent call or email campaign may be conducted based on the results of the initial effort. Final missing inventory reports are then provided to the SDE. CTB takes test security seriously and makes every effort to recover missing material.

Section 3

Classical Item Analysis and Results

Students who complete a course with an EOI test associated with it must also take the test. Students who met the criteria were permitted to take the modified (OMAAP) exam for Algebra I, Biology I, English II, and U.S. History. The 2013 equivalent forms were reprints of previously-used operational forms and their characteristics were discussed in the past technical reports, they were not addressed in this technical report. Information presented in Sections 3, 4, and 5 focuses on the Spring 2013 operational forms.

3.1 Data Quality Check and Clean-Up

After all tests were scored, 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' data was used for classical item analyses, calibration, and equating.

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 is presented in Table 3.1 and 3.1.a.

Table 3.1. Demographic Characteristics of Calibration and Equating Sample

Subject	Total	Female	Male
Algebra I	2923	1107	1785
Biology I	2571	899	1656
English II	2404	817	1579
U.S. History	2894	1040	1829

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

Subject	African American	Native American	Hispanic	Asian	Pacific Islander	White	Other
Algebra I	330	624	269	8	11	1507	174
Biology I	258	573	235	7	5	1342	151
English II	264	532	229	8	7	1213	151
U.S. History	384	588	257	15	12	1476	162

Statistical Key Check. 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 for any of the followings:

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

Flagged operational items were submitted for key check and review by a CTB/McGraw-Hill content specialist. 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 2013 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. CTB/McGraw-Hill research scientists and content specialists reviewed the flagged items and proposed suggestions to the SDE. The SDE then evaluated and decided whether to drop any item from scoring.

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

Table 3.3 shows the number of items removed per SDE's decision and the number of items remained in each subject for final scoring. The items that were removed from operational scoring for the current operational and Braille forms were also excluded 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 also presented the final maximum possible raw score points of the 2013 OMAAP EOI exams, after the removal of items with poor statistics.

Table 3.3. Number of Item Removed and Maximum Score Point Possible

Subject	Number of Items Suppressed	Number of Items Remained	Number of Score Points
Algebra I	2	46	46
Biology I	1	48	48
English II	0	44	46
U.S. History	0	48	48

3.2 Classical Item Analyses

Following completion of the data cleaning activities and prior to calibration and equating, the following classical item analysis statistics were calculated for every item:

- Total case count
- Case count by student subgroup (e.g., males, females, African American, White, Hispanic, Asian, Pacific Islander, Native American, and Other)
- Response frequency distributions (overall and broken down by gender and ethnicity)
    - Distractor analysis for all multiple choice items
    - Score rating and condition code distribution for writing prompts
    -
- 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 by 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 data are shown in Table 3.4. 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 1.5% for all subjects.

Table 3.4. Test-Level Summaries of Classical Item Analyses

Subject	Sample size	Mean Raw Score	Items/ Points	Mean *p*-value	Mean r_{pb}	Omit Min %	Omit Max %
Algebra I	2923	22.19	46	0.47	0.26	0.0	1.0
Biology I	2571	29.61	48	0.62	0.31	0.0	0.5
English II	2404	28.55	44/46	0.62	0.33	0.1	1.4
U.S. History	2894	24.48	48	0.51	0.31	0.0	0.3

*r_{pb} = point biserial correlation

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.

Because the OMAAP items were modified from the OCCT EOI test and used directly in the test, DIF analysis can only be conducted after the test is administered. In other words, this DIF analysis is conducted on operational data, not field test data. CTB/McGraw-Hill used the Mantel-Haenszel (MH) chi-square approach for detecting DIF in multiple choice and open-ended items. 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 for race were African American, Native American, and Hispanic students. The focal group for gender was female students.

Items were classified into three categories on the basis of the MH D-DIF chi-square statistics and the MH delta (Δ) value (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. 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):

- Classification C: $|\Delta| \geq 1.5$ and MH D-DIF chi-square < 0.05
- Classification B: $1 \leq |\Delta| \leq 1.5$ and MH D-DIF chi-square < 0.05
- Classification A: Otherwise

3.3.a. Differential Item Functioning Results

During field test stage, items flagged for DIF were reviewed by expert content specialists from CTB/McGraw-Hill 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.

As mentioned earlier, the OCCT items were modified and then used directly as OMAAP operational items. Item bias review was conducted when items were field tested under the OCCT item format. Only items that were free from bias could be used on the OMAAP tests. Because performances of these modified items on the OMAAP population were unknown, the DIF analysis, although typically conducted on field test items, was performed on the OMAAP operational items. This post-hoc analysis would help future test construction and item modification. Table 3.5 summarizes the number of items flagged for DIF in each subject.

Table 3.5. DIF Flag Incidence Across All OMAAP EOI Items

	Female/Male		**Race**	
Subject	**B**	**C**	**B**	**C**
Algebra I	1	0	3	0
Biology I	0	1	3	1
English II	2	0	6	1
U.S. History	1	1	4	0

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, assuming the repeated administrations are not affected by external factors. 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:

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

Table 3.6 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 produced relatively stable scores.

Table 3.6. Cronbach's Alpha by Subject

Subject	Alpha
Algebra I	0.69
Biology I	0.80
English II	0.81
U.S. History	0.79

3.4.b. Test Reliability by Subgroup

Table 3.7 shows the Cronbach's Alpha reliability coefficients for the various reporting subgroups on each subject of the OMAAP EOI assessments. The reliability coefficients range from 0.64 to 0.88.

Table 3.7. Test Reliability by Subgroup

Subject	Male	Female	African American	Native American	Hispanic	Asian	White
Algebra I	0.70	0.70	0.64	0.70	0.67	0.66	0.71
Biology I	0.82	0.78	0.77	0.79	0.81	0.91	0.81
English II	0.82	0.80	0.79	0.80	0.79	0.82	0.82
U.S. History	0.81	0.72	0.73	0.77	0.74	0.88	0.81

Table 3.7.a. Test Reliability by Subgroup (continued)

Subject	ELL	IEP	Free lunch
Algebra I	0.70	0.70	0.69
Biology I	0.81	0.81	0.80
English II	0.81	0.81	0.81
U.S. History	0.79	0.79	0.78

*ELL = English Language Learner; IEP = Individualized Education Program

3.5 Inter-rater Reliability

Inter-rater reliability refers to the degree of agreement among raters 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 differences between the two raters' assigned writing scores for all students ranged from 0 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 score zero.

The inter-rater reliability analysis results for the English II operational writing prompt are presented in Table 3.8 and 3.8.a. The results show that the two raters gave equal or adjacent (differed by 1 score point) scores on 99.82% of the students. The weighted Kappa statistic (Kraemer, 1982) is an indication of inter-rater reliability of ordinal data after correcting for chance. The Kappa values for the English II operational writing prompts are within the moderate range.

Table 3.8. Percentage of Students at Each Point Discrepancy Between the Two Raters

	Point Discrepancy						
Valid N	**-3**	**-2**	**-1**	**0**	**1**	**2**	**3**
3557	0.00	0.05	13.45	73.48	12.89	0.13	0.00

Table 3.8.a. Inter-rater Reliability for English II Operational Writing Prompts

Agreement Percentages			
Exact	**Adjacent**	**+/- 2 or more**	**Weighted Kappa**
73.48	26.34	0.18	0.65

Section 4

Calibration, Equating, and Scaling

As mentioned in Section 3, information presented in this section is based on operational forms only and the equating data is used for the analyses.

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 θ 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 writing prompt, the partial credit model (PCM; Masters, 1982) was used. In the partial credit model, the probability that a student with ability level θ will have a score in category h (indexed h = 1, 2, …, im) 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).

4.2 Assessment of Fit to the Model

Item fit was assessed using the WINSTEPS fit function. WINSTEPS provides two statistics on item model fit. Mean Square (MS) infit and MS outfit statistics, which indicate the degree of accuracy and predictability with which the data fit the IRT model (Linacre, 2002). Infit is sensitive to misfit on items targeted at the ability level of the person, whereas outfit is sensitive to misfit on items with difficulty far from the ability of the person (Linacre, 2002). Typically, values less than 0.7 and greater than 1.3 for MS infit indicate misfit, and values greater than 1.3

for MS outfit indicate misfit (Wright & Linacre, 1994). Given that the MS infit statistic best addresses item misfit for items that are well-targeted to the ability level of the student, CTB/McGraw-Hill traditionally prioritizes this statistic in evaluating the fit of the Rasch model for an assessment like the OMAAP. The expected value of infit 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 CTB/McGraw-Hill 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, the item fit plots were also generated in WINSTEPS output file. 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 design was used for all content areas to link the current test forms (i.e., Spring 2013) 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, CTB/McGraw-Hill evaluated changes in the item parameters from the previous operational administration to the Spring 2013 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. The 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.

Displacement Method. Each anchor item has two parameters, the parameter estimate from the baseline form ($b_{baseline}$) and the parameter estimate from current administration (b_{new}). Once an equating constant is computed and the b_{new} is brought to the base scale, b_{scaled}, the displacement value can be computed. The displacement value is the absolute value of the difference (D) between the $b_{baseline}$ and b_{scaled}. An anchor item is flagged when the D is larger than 0.30 logits. If, upon further review, it is decided that an item should be eliminated from the anchor item set then the equating constant is re-calculated, reapplied, and the stability check process is repeated until all D values are within desirable range or when 20% of anchor items are dropped.

$$Displacement = Abs(b_{baseline} - b_{scaled}). \quad (4)$$

Robust Z method. Robust *Z*-values, on the other hand, are computed by taking the difference between the pre-equated value and the post-equated value and subtracting the median value of all the differences and dividing this number by the interquartile range multiplied by 0.74 (SCDE, 2001):

$$\text{Robust } z_j = \frac{\left(bdif_i - MDdif_j\right)}{\left(IQ_j \times 0.74\right)}, \quad (5)$$

where $bdif_i$ = the difference between pre and post-equated parameters ($\hat{b}$) for item i, $_{MDdifj}$ = the median of the differences for anchor set j, and IQ_j = interquartile range for anchor set j. The steps for a given subject and grade is as follows:

1. Obtain the difference between new and old item parameters.
2. Calculate the mean, median, and interquartile range of the differences calculated in Step 1.
3. Calculate Robust *Z*.
4. Items with an absolute value of a Robust *Z* exceeding 1.645 for the Rasch item difficulty parameter are considered outliers.
5. Stopping rule: if the Robust *Z* for no items exceed the 1.645 criteria for either the Rasch item difficulty parameter, or fewer than 80% of the test remains in the anchor set. Since this is accomplished in a list-wise fashion it is possible that more items will be flagged than can be dropped. Items will be rank-ordered by magnitude of Robust *Z*, and those with the largest values were dropped.

The Displacement and Robust *Z* methods were used in conjunction to identify items with post-equated values large differences from the pre-equated values. The anchor item screening procedure was 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.

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. Do not 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.

The order for dropping items from the anchor set will occur based on collective rank ordering of the items change from the two approaches. Decisions about whether to keep or remove an item will be evaluated on a per item basis. If an item (note, only one item can be removed at a time) is removed from the set, then this process (beginning at the equating step) will be repeated until there are no further items to be removed. Even though an item may be removed for the purpose of equating, it will still contribute to student scores. Items flagged for removal during the anchor stability check, 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 too many items from a passage are eliminated, the entire passage and associated items may have to be removed)
- Request content review of an item for any modifications or edits since last operational use (should be none)

Once the equating item set is finalized, the equating constants obtained will be applied to all operational items for placing the items on the baseline scale for item banking and for computation of raw score to scale score tables.

4.4.a Anchor Items for Spring 2013

Table 4.1 presents the number and percentages of anchor items by subject for the Spring 2013 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.

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 2013, two anchor items from English II (see Table 4.1) were dropped.

Table 4.1. Number of Anchor Items per Subject

		Initial		Final	
Subject	**Total**	**Count**	**%**	**Count**	**%**
Algebra I	46	17	37	17	37
Biology I	48	15	31	15	31
English II	44	18	41	16	36
U.S. History	48	17	35	17	35

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:

$$\text{Scale Score} = \left(\hat{\tau} \times M1\right) + M2 \tag{6}$$

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

The raw score to scale score tables were generated using WINSTEPS. 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 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 (Perf. Level) 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 Constant		Scale Range		Cut Scores		
Subject	**M1**	**M2**	**LOSS**	**HOSS**	**Limited Knowledge**	**Satisfactory**	**Advanced**
Algebra I	19.07	254.60	100	350	237	250	269
Biology	22.34	237.94	100	350	237	250	273
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

	Algebra I			Biology I		
Raw Score	Scale Score	Perf. Level	SEM	Scale Score	Perf. Level	SEM
0	167	1	35	111	1	41
1	191	1	20	139	1	23
2	205	1	14	157	1	17
3	213	1	12	167	1	14
4	220	1	10	175	1	13
5	225	1	9	182	1	11
6	229	1	9	187	1	11
7	233	1	8	192	1	10
8	236	1	8	196	1	9
9	**239**	**2**	**7**	200	1	9
10	242	2	7	204	1	9
11	245	2	7	207	1	8
12	247	2	7	210	1	8
13	**250**	**3**	**7**	213	1	8
14	252	3	6	216	1	8
15	254	3	6	218	1	8
16	256	3	6	221	1	7
17	258	3	6	223	1	7
18	260	3	6	226	1	7
19	262	3	6	228	1	7
20	264	3	6	230	1	7
21	266	3	6	233	1	7
22	268	3	6	235	1	7
23	**269**	**4**	**6**	**237**	**2**	**7**
24	271	4	6	239	2	7
25	273	4	6	241	2	7
26	275	4	6	243	2	7
27	277	4	6	246	2	7
28	279	4	6	248	2	7
29	281	4	6	**250**	**3**	**7**
30	283	4	6	252	3	7
31	285	4	6	255	3	7
32	287	4	6	257	3	7
33	289	4	6	259	3	7
34	291	4	7	262	3	7
35	294	4	7	264	3	8
36	296	4	7	267	3	8
37	299	4	7	270	3	8
38	302	4	8	**273**	**4**	**8**
39	305	4	8	276	4	9
40	308	4	9	279	4	9
41	313	4	9	283	4	9
42	317	4	10	287	4	10
43	323	4	12	292	4	11
44	332	4	14	298	4	12
45	346	4	19	305	4	14
46	350	4	22	315	4	16
47				331	4	23
48				350	4	34

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	Perf. Level	SEM	Scale Score	Perf. Level	SEM
0	152	1	36	145	1	37
1	176	1	20	170	1	20
2	190	1	14	185	1	15
3	199	1	12	193	1	12
4	205	1	10	200	1	11
5	210	1	9	205	1	10
6	214	1	9	210	1	9
7	218	1	8	214	1	9
8	221	1	8	217	1	8
9	224	1	8	220	1	8
10	227	1	7	223	1	7
11	229	1	7	226	1	7
12	232	1	7	228	1	7
13	234	1	7	231	1	7
14	237	1	7	233	1	7
15	**239**	**2**	**7**	235	1	7
16	241	2	6	237	1	6
17	243	2	6	**239**	**2**	**6**
18	245	2	6	241	2	6
19	247	2	6	243	2	6
20	249	2	6	245	2	6
21	**251**	**3**	**6**	247	2	6
22	253	3	6	249	2	6
23	255	3	6	**251**	**3**	**6**
24	257	3	6	253	3	6
25	258	3	6	254	3	6
26	260	3	6	256	3	6
27	262	3	6	258	3	6
28	264	3	6	260	3	6
29	**266**	**4**	**6**	262	3	6
30	268	4	6	**264**	**4**	**6**
31	270	4	6	266	4	6
32	273	4	7	268	4	6
33	275	4	7	270	4	7
34	277	4	7	272	4	7
35	280	4	7	274	4	7
36	282	4	7	277	4	7
37	285	4	8	279	4	7
38	288	4	8	282	4	7
39	291	4	8	285	4	8
40	295	4	9	288	4	8
41	299	4	10	291	4	8
42	304	4	10	295	4	9
43	311	4	12	299	4	10
44	319	4	14	304	4	11
45	334	4	20	311	4	12
46	350	4	31	320	4	15
47				334	4	20
48				350	4	30

Section 5

Classification Consistency and Accuracy Studies

As mentioned in Section 3, information presented in this section is based on operational forms only and the equating data are used for the analyses.

5.1 Classification Consistency and Accuracy

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 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. The opposite situation could also happen, where a student's true ability is lower than the cut score but is classified as passing. As discussed in Section 6, 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 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. 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	False Positives	False Negatives	Kappa
Algebra I	0.76	0.66	0.11	0.13	0.40
Biology I	0.69	0.59	0.15	0.16	0.42
English II	0.76	0.67	0.10	0.14	0.45
U.S. History	0.66	0.55	0.16	0.18	0.39

*Source data is the final student data file, which has more cases than the equating file.

As shown in Table 5.1, the overall accuracy indices range between 66 and 76 percent and overall consistency ranges between 55 and 67 percent. Kappa coefficients range from 39 and 45 percent. The false positive rates range from 10 to 16 percent. The false negative rates range from 13 to 18 percent.

Tables 5.2 and 5.2.a provide 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 75 percent in all cases, with most above 80 percent. In general, the high rates of accuracy and consistency support the cut decisions made using these assessments. The false positive and false negative rates are lower in comparison to Table 5.1.

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 Tables 5.2 and 5.2.a Algebra I, for example, 94 percent of students are correctly classified as pass or fail based on their performance (scenario 1), 0 percent passed, but their true ability is below the standard (scenario 2), and 10 percent failed although their true ability is above the standard (scenario 3). Overall, the accuracy rate for accurate classification is 86% or above for all subjects—students

are appropriately (more than 86% of the time) categorized into pass/fail classifications based on their true ability using their observed score (raw score) as their classification score.

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	0.94	0.82	0.99	0.91	0.75
Biology I	0.90	0.86	0.93	0.86	0.80	0.90
English II	0.98	0.92	0.86	0.97	0.88	0.80
U.S. History	0.91	0.86	0.88	0.88	0.80	0.83

*U =Unsatisfactory; L = Limited Knowledge; S = Satisfactory; and A = Advanced.
*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.
*Source data is the final student data file, which has more cases than the equating file.

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	0.00	0.01	0.05	0.10	0.08
Biology I	0.04	0.06	0.07	0.07	0.05	0.02
English II	0.00	0.02	0.03	0.06	0.07	0.07
U.S. History	0.03	0.06	0.07	0.08	0.07	0.05

*U =Unsatisfactory; L = Limited Knowledge; S = Satisfactory; and A = Advanced.
*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.
*Source data is the final student data file, which has more cases than the equating file.

Section 6

Summary Statistics

Reports of this section are based on the student data file that is used for score reporting. This data file includes records received after equating from students who took Equivalent and Braille forms; therefore, *N* counts are greater than the *N* counts of previous sections. Invalid cases are excluded from the analyses.

6.1 Descriptive Statistics

Descriptive statistics of scale scores presented in the following tables are computed in various ways. First, the overall mean scale score of each subject is presented along with standard deviation and median. Then the same types of statistics are reported by gender and ethnicity subgroups. Results are suppressed if group *N* count is no more than ten.

Table 6.1. Scale Score Descriptive Statistics – Overall

Subject	N	Mean	SD	Med
Algebra I	4265	266.8	11.6	266
Biology	3644	252.2	17.1	250
English II	3686	266.3	15.8	266
U.S. History	3040	254.0	14.5	253

*N = Sample size; SD = Standard Deviation; Med. = Median.

Table 6.2. Scale Score Descriptive Statistics by Gender

	Female				Male			
Subject	**N**	**Mean**	**SD**	**Med**	**N**	**Mean**	**SD**	**Med**
Algebra I	1610	266.9	11.4	266	2634	266.8	11.7	266
Biology	1299	250.7	16.3	250	2333	253.0	17.6	252
English II	1265	268.2	15.4	268	2412	265.3	15.9	266
U.S. History	1098	250.1	12.2	249	1935	256.1	15.3	254

*N = Sample size; SD = Standard Deviation; Med. = Median.

Table 6.3. Scale Score Descriptive Statistics by Race/Ethnicity

	African American				Native American			
Subject	**N**	**Mean**	**SD**	**Med**	**N**	**Mean**	**SD**	**Med**
Algebra I	582	263.5	10.6	262	811	266.5	11.6	266
Biology	461	244.9	15.2	243	741	252.4	16.3	252
English II	477	261.1	14.3	260	760	265.8	15.4	266
U.S. History	431	248.4	12.9	247	606	254.0	13.7	253

*N = Sample size; SD = Standard Deviation; Med. = Median.

Table 6.3.a. Scale Score Descriptive Statistics by Race/Ethnicity (cont.)

	Hispanic				Asian			
Subject	**N**	**Mean**	**SD**	**Med**	**N**	**Mean**	**SD**	**Med**
Algebra I	443	267.1	11.2	266	10	266.1	8.9	268
Biology	368	248.6	18	246	9	-	-	-
English II	378	264.3	15	264	13	260.4	16.3	264
U.S. History	273	250.6	12.7	249	13	258.2	21.4	254

*N = Sample size; SD = Standard Deviation; Med. = Median

Table 6.3.b. Scale Score Descriptive Statistics by Race/Ethnicity (cont.)

	White				Pacific Islander			
Subject	**N**	**Mean**	**SD**	**Med**	**N**	**Mean**	**SD**	**Med**
Algebra I	2158	267.7	11.7	268	14	273.2	10.2	274
Biology	1860	254.5	17.1	255	9	-	-	-
English II	1852	268.1	15.9	268	9	-	-	-
U.S. History	1548	255.8	14.9	254	12	254.8	14.9	254

*N = Sample size; SD = Standard Deviation; Med. = Median.

Table 6.3.c. Scale Score Descriptive Statistics by Race/Ethnicity (cont.)

	Other			
Subject	**N**	**Mean**	**SD**	**Med**
Algebra I	247	266.9	12.3	266
Biology	196	253.1	16.3	252
English II	197	267.7	17.3	268
U.S. History	157	256.6	15.5	254

*N = Sample size; SD = Standard Deviation; Med. = Median.

6.2 Performance Level Distribution

Distributions of performance level are presented in Table 6.4 (See Appendix B for scale score distributions). The values are reported in percentages. Table 6.4 shows that 96.3% of students passed (in the Satisfactory and Above performance levels) Algebra I, 54.8% of students passed Biology I, 84.1% of students passed English II, and 56.8% of students passed U.S. History.

Table 6.4. Percentage of Students by Performance Level

Subject	N	Unsatisfactory	Limited Knowledge	Satisfactory	Advanced	Satisfactory and Above
Algebra I	4265	0.3	3.4	55.9	40.4	96.3
Biology	3644	17.9	27.3	40.9	13.9	54.8
English II	3686	2.4	13.6	31.7	52.4	84.1
U.S. History	3040	11.8	31.3	32.1	24.7	56.8

6.3 Conditional Standard Error of Measurement

The Rasch model standard error (SE) for ability estimate ($\hat{\beta}$) is as follows (Andrich & Luo, 2004):

$$\sigma_{\hat{\beta}=} \frac{1}{\sqrt{\sum_{i=1}^{L} p_{vi}(1-p_{vi})}} \quad , \tag{7}$$

where

v = subscript for a person,
i = subscript for an item,
L = length of the test,
$\hat{\beta}$ = ability estimate, and
p_{vi} = the probability that a person answers an item correctly and defined as follows:

$$p_{vi} = \frac{e^{\beta_v - \delta_i}}{1+e^{\beta_v - \delta_i}} \quad , \tag{8}$$

where β_v is person's ability and δ_i is item's difficulty.

A confidence band can be found for use in interpreting the ability estimate. For example, an approximate 68% confidence interval for $\hat{\beta}$ is given by $\hat{\beta}$ ± SE. Because different ability estimates ($\hat{\beta}$s) has different SE, Rasch SE it is generally referred to as the conditional standard error of measurement (CSEM) to separate from the standard error of measurement of the classical measurement model. The *CSEM*s by subject are reported 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 using test reliability:

$$SEM = SD\sqrt{(1-r)} \qquad (9)$$

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

Content	SEM
Algebra I	3.21
Biology	3.06
English II	3.03
U.S. History	3.19

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.

Andrich, A., & Luo, G. (2004). Modern measurement and analysis in social science. Murdoch University, Perth, Western Australia.

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.

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.

Appendix A

Standards, Objectives/Skills, and Processes Assessed by Subject

OMAAP Test Blueprint and Actual Item Counts: Algebra I

***OKC*^{3} Standard and Objective**	**Ideal Number of Items for Alignment to *OKC*^{3}**	**Actual Number of Items on 2013 Test**
Number Sense and Algebraic Operations	**10-12**	**14**
Equations and Formulas (1.1)	4-6	7*
Expression (1.2)	5-7	7*
Relations and Functions	**21-23**	**23**
Relations and Functions (2.1)	2-3	3
Linear Equations and Graphs (2.2)	12-14	11
Linear Inequalities and Graphs (2.3)	3-5	5
Systems of Equations (2.4)	2-3	4
Data Analysis, Probability, and Statistics	**6–8**	**9**
Data Analysis (3.1)	4-6	6
Line of Best Fit (3.2)	11-3	3
Total Test	**40–43**	**46**

*Suppressed item in this reporting category.

OMAAP Test Blueprint and Actual Item Counts: English II

OKC[3] Standard and Objective	Ideal Number of Items for Alignment to *OKC*[3]	Actual Number of Items on 2013 Test
Reading/Literature		
Vocabulary (1.0)	**6-7**	**7**
Comprehension	**9-10**	**10**
Literal Understanding (2.1)	1-3	1
Inferences and Interpretation (2.2)	2-4	4
Summary and Generalization (2.3)	2-4	3
Analysis and Examination (2.4)	1-3	2
Literature	**12-13**	**14**
Literary Genres (3.1)	2-3	4
Literary Elements (3.2)	3-5	5
Figurative Language and Sound Devices (3.3)	3-5	5
Literary Works (3.4)	2-3	0
Research and Information	6-7	5
Accessing Information (4.1)	2-4	2
Interpreting Information (4.2)	2-4	3
Writing/Grammar/Usage and Mechanics		
Writing (1.0, 2.0)	**1 (3 points)**	**1**
Writing Prompt	1	
Grammar/Usage and Mechanics	**7-8**	**7**
Standard English Usage (3.1)	2-3	2
Mechanics and Spelling (3.2)	2-3	2
Sentence Structure (3.3)	2-3	3
Total Test	**41-44 (43-46 points)**	**44 (46 pts.)**

OMAAP Test Blueprint and Actual Item Counts: Biology I

OKC[3] Standard and Objective	Ideal Number of Items for Alignment to *OKC*[3]	Actual Number of Items on 2013 Test
Process Standards		
Observe and Measure	**6**	**5**
Qualitative/quantitative observations and changes (P1.1)	4	3
Use appropriate tools & (P1.2)	2	2
Use appropriate SI units (P1.3)		
Classify	**6**	**7**
Use observable properties to classify (P2.1)	2-4	4
Identify properties of a classification system (P2.2)	2-4	3
Experimental Design	**13-16**	**14**
Evaluate the design of investigations (P3.1)	3-4	3*
Hazards/practice safety (P3.2) & Identify a testable hypothesis in a biology investigation (P3.4)	3-4	4
Use mathematics to show relationships (P3.3)	3-4	4
Identify potential hazards and practice safety procedures in all science activities (P3.5)	3-4	3
Interpret and Communicate	**16-19**	**17**
Select predictions based on observed patterns of evidence (P4.1)	3-4	6
Interpret line, bar, trend, and circle graphs (P4.3)	3-4	3
Accept or reject a hypothesis (P4.4)	3	3
Make logical conclusions based on experimental data	3-4	3
Identify an appropriate graph or chart (4.8)	3-4	2
Translate quantitative information expressed in words into visual form (4.8a)		
Translate information expressed visually or mathematically (4.8b)		
Model	**66616-19**	
Interpret a model which explains a given set of observations (P5.1)	3-4	3
Select predictions based on models using mathematics when appropriate (P5.2)	2	2
Total Test	**46-49**	**48**
Content Standards		
The Cell	**9-12**	**9**
Cells structures and functions (C1.1)	3-5	5
Differentiation of cells (C1.2)	2-4	1

Specialized cells (C1.3)	2-4	3
The Molecular Basis of Heredity	**9–12**	**9**
DNA structure and function in heredity (C2.1)	3–6	3
Sorting and recombination of genes (C2.2)	4–7	6
Biological Diversity	**9–12**	**9**
Variation among organisms (C3.1)	2-4	3
Natural selection and biological adaptations (C3.2)	3-5	4
Behavior patterns can be used to ensure reproductive success (C3.3)	2-4	2
The Interdependence of Organisms	**6-8**	**8**
Organisms both cooperate and compete (C4.1)	3-5	5
Population dynamics (C4.2)	3-5	3
Matter/Energy/Organization in Living Systems	**10**	**10**
Complexity and organization used for survival (C5.1)	3-4	4
Matter and energy flow in living and nonliving systems (C5.2)	3-4	5
Earth cycles including abiotic and biotic factors (C5.3)	3-4	1*
Total Test	**43-46**	**45**

** Items from the Safety Objective (P3.5) are not dual aligned to a content standard
*Suppressed item

OMAAP Test Blueprint and Actual Item Counts: U.S. History

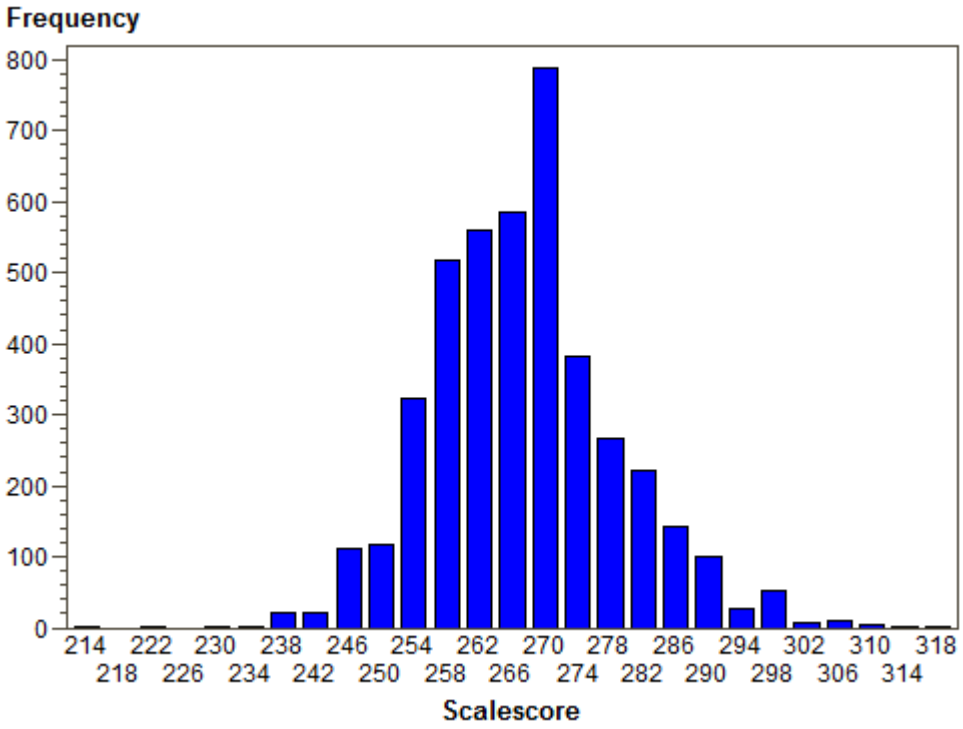
***OKC*^{3} Standard and Objective**	**Ideal Number of Items for Alignment to *OKC*^{3}**	**Actual Number of Items on 2013 Test**
Post-Reconstruction to the Progressive Era, 1878-1900	**8**	**9**
Post Reconstruction Amendments (1.1)	2-4	2
Immigration, Westward Movement, and Native American Experiences (1.2)	2-4	4
Impact of Industrialization on Society, Economics, and Politics (1.3)	2-4	3
Expanding Role of the United States in International Affairs	**6**	**5**
Cycles of Economic Boom and Bust in the 1920s and 1930s	**8**	**8**
Economic, Political, & Social Transformation Between the World Wars (3.1)	3-5	3
Economic Destabilization and the Great Depression/New Deal (3.2, 3.3)	3-5	5
Role of the U.S. in International Affairs and World War II, 1933-1946	**8**	**8**
Mobilization of World War II (4.1)	3-5	3
World War II and U.S. Reaction to the Holocaust (4.2, 4.3)	3-5	5
Foreign and Domestic Policies during the Cold War, 1945-1975	**18**	**18**
The Cold War – Foreign and Domestic (5.1, 5.2)	4-6	5
The Vietnam War Era (5.3)	4-6	4
The African American Civil Rights Movement (5.4)	4-6	5
Social Political Transformation (5.5)	4-6	4
Total Test	**48**	**48**

Appendix B: Scale Score Distributions

Algebra I Scale Score Distribution

Raw Score	Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
3	**213**	1	0.02	1	0.02
4	**220**	1	0.02	2	0.05
6	**229**	1	0.02	3	0.07
7	**233**	2	0.05	5	0.12
8	**236**	9	0.21	14	0.33
9	**239**	11	0.26	25	0.59
10	**242**	20	0.47	45	1.06
11	**245**	39	0.91	84	1.97
12	**247**	73	1.71	157	3.68
13	**250**	118	2.77	275	6.45
14	**252**	140	3.28	415	9.73
15	**254**	183	4.29	598	14.02
16	**256**	233	5.46	831	19.48
17	**258**	284	6.66	1115	26.14
18	**260**	271	6.35	1386	32.50
19	**262**	290	6.80	1676	39.30
20	**264**	307	7.20	1983	46.49
21	**266**	279	6.54	2262	53.04
22	**268**	280	6.57	2542	59.60
23	**269**	266	6.24	2808	65.84
24	**271**	241	5.65	3049	71.49
25	**273**	203	4.76	3252	76.25
26	**275**	179	4.20	3431	80.45
27	**277**	152	3.56	3583	84.01
28	**279**	116	2.72	3699	86.73
29	**281**	129	3.02	3828	89.75
30	**283**	92	2.16	3920	91.91
31	**285**	79	1.85	3999	93.76
32	**287**	63	1.48	4062	95.24
33	**289**	58	1.36	4120	96.60
34	**291**	42	0.98	4162	97.58
35	**294**	27	0.63	4189	98.22
36	**296**	24	0.56	4213	98.78
37	**299**	28	0.66	4241	99.44
38	**302**	7	0.16	4248	99.60
39	**305**	10	0.23	4258	99.84
40	**308**	4	0.09	4262	99.93
41	**313**	1	0.02	4263	99.95
42	**317**	2	0.05	4265	100.00

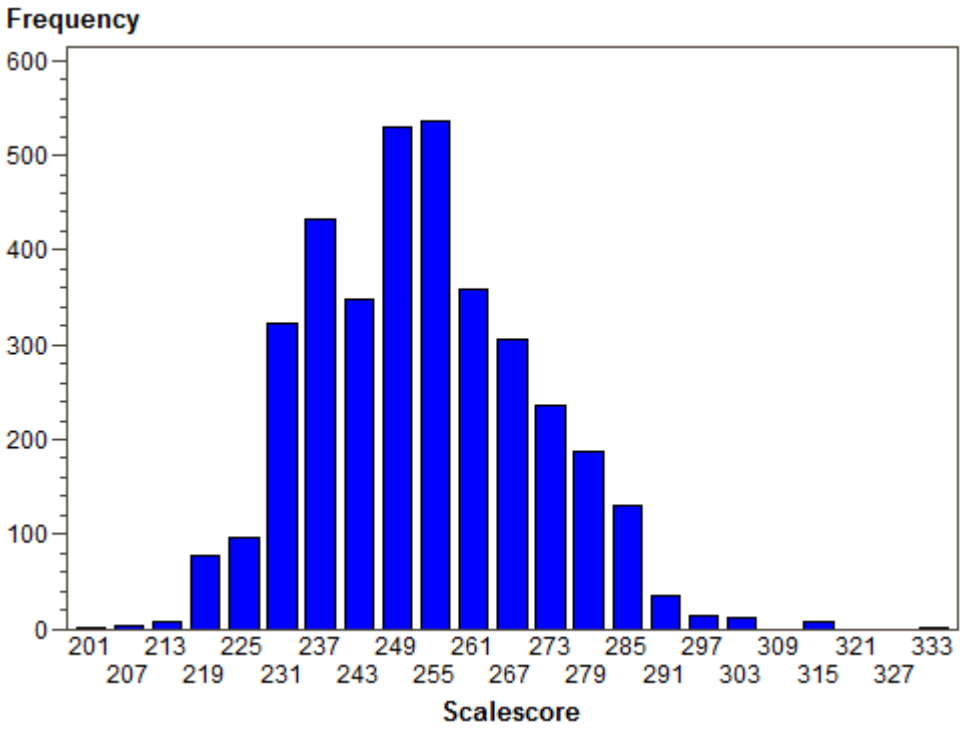
Algebra I Scale Score Distribution



Biology I Scale Score Distribution

Raw Score	Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
9	**200**	2	0.05	2	0.05
10	**204**	1	0.03	3	0.08
11	**207**	3	0.08	6	0.16
12	**210**	3	0.08	9	0.25
13	**213**	5	0.14	14	0.38
14	**216**	16	0.44	30	0.82
15	**218**	26	0.71	56	1.54
16	**221**	36	0.99	92	2.52
17	**223**	34	0.93	126	3.46
18	**226**	63	1.73	189	5.19
19	**228**	92	2.52	281	7.71
20	**230**	98	2.69	379	10.40
21	**233**	133	3.65	512	14.05
22	**235**	141	3.87	653	17.92
23	**237**	147	4.03	800	21.95
24	**239**	145	3.98	945	25.93
25	**241**	164	4.50	1109	30.43
26	**243**	185	5.08	1294	35.51
27	**246**	197	5.41	1491	40.92
28	**248**	155	4.25	1646	45.17
29	**250**	177	4.86	1823	50.03
30	**252**	166	4.56	1989	54.58
31	**255**	186	5.10	2175	59.69
32	**257**	184	5.05	2359	64.74
33	**259**	163	4.47	2522	69.21
34	**262**	196	5.38	2718	74.59
35	**264**	154	4.23	2872	78.81
36	**267**	151	4.14	3023	82.96
37	**270**	114	3.13	3137	86.09
38	**273**	121	3.32	3258	89.41
39	**276**	91	2.50	3349	91.90
40	**279**	96	2.63	3445	94.54
41	**283**	75	2.06	3520	96.60
42	**287**	55	1.51	3575	98.11
43	**292**	36	0.99	3611	99.09
44	**298**	13	0.36	3624	99.45
45	**305**	12	0.33	3636	99.78
46	**315**	7	0.19	3643	99.97
47	**331**	1	0.03	3644	100.00

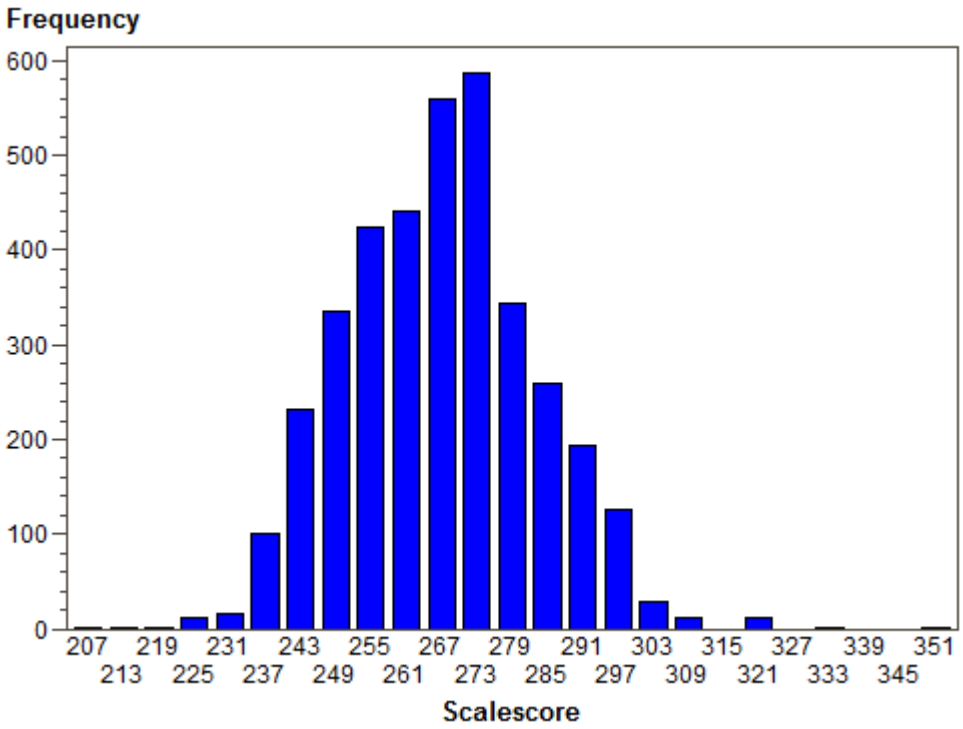
Biology I Scale Score Distribution



English II Scale Score Distribution

Raw Score	Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
4	**205**	1	0.03	1	0.03
6	**214**	1	0.03	2	0.05
7	**218**	1	0.03	3	0.08
9	**224**	3	0.08	6	0.16
10	**227**	9	0.24	15	0.41
11	**229**	4	0.11	19	0.52
12	**232**	13	0.35	32	0.87
13	**234**	22	0.60	54	1.47
14	**237**	34	0.92	88	2.39
15	**239**	45	1.22	133	3.61
16	**241**	71	1.93	204	5.53
17	**243**	75	2.03	279	7.57
18	**245**	85	2.31	364	9.88
19	**247**	108	2.93	472	12.81
20	**249**	116	3.15	588	15.95
21	**251**	111	3.01	699	18.96
22	**253**	121	3.28	820	22.25
23	**255**	142	3.85	962	26.10
24	**257**	162	4.40	1124	30.49
25	**258**	142	3.85	1266	34.35
26	**260**	147	3.99	1413	38.33
27	**262**	151	4.10	1564	42.43
28	**264**	192	5.21	1756	47.64
29	**266**	178	4.83	1934	52.47
30	**268**	189	5.13	2123	57.60
31	**270**	190	5.15	2313	62.75
32	**273**	208	5.64	2521	68.39
33	**275**	189	5.13	2710	73.52
34	**277**	178	4.83	2888	78.35
35	**280**	165	4.48	3053	82.83
36	**282**	141	3.83	3194	86.65
37	**285**	118	3.20	3312	89.85
38	**288**	127	3.45	3439	93.30
39	**291**	67	1.82	3506	95.12
40	**295**	74	2.01	3580	97.12
41	**299**	51	1.38	3631	98.51
42	**304**	29	0.79	3660	99.29
43	**311**	12	0.33	3672	99.62
44	**319**	11	0.30	3683	99.92
45	**334**	2	0.05	3685	99.97
46	**350**	1	0.03	3686	100.00

English II Scale Score Distribution



U.S. History Scale Score Distribution

Raw Score	Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
2	**185**	1	0.03	1	0.03
9	**220**	2	0.07	3	0.10
10	**223**	2	0.07	5	0.16
11	**226**	20	0.66	25	0.82
12	**228**	36	1.18	61	2.01
13	**231**	42	1.38	103	3.39
14	**233**	60	1.97	163	5.36
15	**235**	82	2.70	245	8.06
16	**237**	114	3.75	359	11.81
17	**239**	133	4.38	492	16.18
18	**241**	161	5.30	653	21.48
19	**243**	168	5.53	821	27.01
20	**245**	166	5.46	987	32.47
21	**247**	174	5.72	1161	38.19
22	**249**	151	4.97	1312	43.16
23	**251**	168	5.53	1480	48.68
24	**253**	152	5.00	1632	53.68
25	**254**	134	4.41	1766	58.09
26	**256**	151	4.97	1917	63.06
27	**258**	145	4.77	2062	67.83
28	**260**	123	4.05	2185	71.88
29	**262**	104	3.42	2289	75.30
30	**264**	100	3.29	2389	78.59
31	**266**	114	3.75	2503	82.34
32	**268**	83	2.73	2586	85.07
33	**270**	76	2.50	2662	87.57
34	**272**	66	2.17	2728	89.74
35	**274**	61	2.01	2789	91.74
36	**277**	49	1.61	2838	93.36
37	**279**	45	1.48	2883	94.84
38	**282**	47	1.55	2930	96.38
39	**285**	29	0.95	2959	97.34
40	**288**	25	0.82	2984	98.16
41	**291**	22	0.72	3006	98.88
42	**295**	11	0.36	3017	99.24
43	**299**	10	0.33	3027	99.57
44	**304**	6	0.20	3033	99.77
45	**311**	7	0.23	3040	100.00

U.S. History Scale Score Distribution