

Oklahoma School Testing Program

2009 Technical Report

Achieving Classroom Excellence

End-of-Instruction

Assessments

Submitted to
The Oklahoma State Department of Education
October, 2009



Executive Summary

Introduction

The Oklahoma School Testing Program (OSTP) is a state-wide assessment program which includes the End-of-Instruction (EOI) assessments where students who complete an area of instruction must also take the corresponding state-wide, standardized assessment. The subjects included within this testing program are 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. These 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. Algebra I, English II, Biology I, and U.S. History were existing tests in the program with Algebra II, Geometry, and English III added as operational tests for the 2007-2008 testing cycle. These End-of-Instruction tests are administered in Winter, Trimester, Spring, and Summer. 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) since its introduction by the Federal Government in 2001. In 2006, Pearson was contracted by the Oklahoma State Department of Education (SDE) to develop, administer, and maintain the OSTP-ACE EOI tests. This report provides technical details of work accomplished through the end of 2009 on these tests.

Purpose

The purpose of this 2009 OSTP Technical Report is to provide objective information regarding technical aspects of the OSTP-ACE 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 OSTP-ACE EOI assessments. Other sources of information regarding the OSTP-ACE EOI tests, administered mostly online, with some paper formatted tests available, include the administration manuals, interpretation manuals, student, teacher, and parent guides, implementation material, 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 OSTP-ACE 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
- Standards 13.1—13.19 Educational Testing and Assessment

This technical report provides accurate, complete, current, and clear documentation of the OSTP-ACE 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 overviews the procedures and results of the standard setting completed for Biology I, English II, and U.S. History. Finally, Section 7 provides higher-level summaries of all the tests included in the OSTP-ACE EOI testing program.

Information provided in this report presents valuable information about the OSTP-ACE EOI assessments regarding:

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;
11. Decision accuracy and classification; and
12. Setting performance standard cut scores.

Each of these facets in the OSTP-ACE EOI assessments development and use cycle is critical to validity of test scores and interpretation of results. This technical report covers all of these topics for the 2008-2009 testing year.

Table of Contents

Section 1	1
1.1 Overview of the OSTP ACE/EOI Assessments	1
1.1.a Purpose	2
1.1.b PASS Content Standards	2
1.2 Summary of Test Development and Content Validity	3
1.2.a Aligning Test to PASS Content Standards	4
1.2.b Item Pool Development and Selection	4
1.2.c Configuration of the Seven Tests	5
1.2.d Operational and Field-Test Items by Content Area	9
Section 2	17
2.1 Packaging and Shipping.....	17
2.2 Materials Return	18
2.3. Materials Discrepancies Process	18
2.4 Processing Assessment Materials Returned by Schools	19
Section 3	20
3.1 Sampling Plan and Field Test Design	20
3.1.a Sampling Plan.....	20
3.1.b Field-Test Design	20
3.1.c Data Receipt Activities	20
<i>Statistical Key Check</i>	22
3.2 Classical Item Analyses	22
3.2.a Test-levels summaries of classical item analyses	23
3.3 Procedures for Detecting Item Bias	23
3.3.a Different Item Functioning Results	24
3.4 Data Review	25
3.4.a Results of Data Review	26
3.5 Test Reliability	27
3.6 Test Reliability by Subgroup	28
3.5 Inter-rater Reliability	29
Section 4	31
4.1 Item Response Theory (IRT) models	31
<i>Dichotomous Item Response Theory Model</i>	31
<i>Polytomous Item Response Theory Model</i>	31
4.2 Assessment of IRT Fit to the model	32
4.2.a Calibration and IRT Fit Results	33
4.2.a.i Winter/Trimester 2008-2009	33
4.2.a.ii Spring 2009	34
4.3 Calibration and Equating	35
4.4 Anchor Item Stability Evaluation Methods	36
4.4.a Anchor Items for Winter/Trimester 2008-2009 and Spring 2009	37

4.4.b Results of the Anchor Item Stability Check	38
4.5 Scaling and Scoring Results	38
<i>Section 5</i>	52
5.1 Classification Consistency and Accuracy	52
<i>Section 6</i>	56
6.1 Overview and Standard Setting Process.....	56
6.2 Results – Biology I, English II, and U.S. History Cut Scores	57
<i>Section 7</i>	60
7.1 Means and Standard Deviations	60
7.2 Performance Level Distribution	60
7.3 Conditional Standard Error of Measurement.....	61
7.4 Standard Error of Measurement	61
<i>References</i>	63
<i>Appendix A</i>	65
<i>Appendix B</i>	76
<i>Appendix C</i>	91

Table of Tables and Figures

Table 1.1. Oklahoma Content Standards by Subject	2
Table 1.2. Criteria for Aligning the Test with PASS Standards and Objectives.	Error!
Bookmark not defined.	
Table 1.3. Percentage of Items in Depth of Knowledge Levels	5
Table 1.4. Configuration of the OSTP-ACE/EOI tests for Winter/Trimester 2008-2009..	6
Table 1.5. Configuration of the OSTP-ACE/EOI tests for Spring 2009	7
Table 1.6a. Number of item and points by Content Standard for Algebra I.....	10
Table 1.6b. Number of item and points by Content Standard for Algebra II	11
Table 1.6c. Number of item and points by Content Standard for Geometry	12
Table 1.6d. Number of item and points by Content Standard for Biology I.....	13
Table 1.6e. Number of item and points by Content Standard for English II.....	14
Table 1.6f. Number of item and points by Content Standard for English III	15
Table 1.6g. Number of item and points by Content Standard for U.S. History.....	16
Table 3.1. Demographic characteristics of calibration and equating sample for Winter/Trimester 2008-2009	21
Table 3.2. Demographic characteristics of calibration and equating sample for Spring 2009.....	21
Table 3.3. Test level summaries of classical item analyses for Winter/Trimester 2008- 2009 and Spring 2009	23
Table 3.4. DIF flag incidence across all OSTP-ACE/EOI field test items for Winter/Trimester 2008-2009 and Spring 2009	25
Table 3.5. Number of items per subject flagged and rejected during Winter/Trimester 2008-2009 and Spring 2009 field test data review	27
Table 3.6. Cronbach's alpha for Winter/Trimester 2008-2009 and Spring 2009 Administration by Subject	28
Table 3.7. Test Reliability by Subgroup for Spring 2009.....	28
Table 3.8. Inter-rater reliability for English II operational writing prompts for Winter/Trimester 2008-2009 and Spring 2009.....	30
Table 3.9. Inter-rater reliability for English III operational writing prompts for Winter/Trimester 2008-2009 and Spring 2009.....	30
Table 4.1. Number of anchor items per subject	38
Table 4.2. LOSS, HOSS, and Scaling Constants by Subject.....	39
Table 4.3. Performance Level Cut Scores by Content A.....	39
Table 4.4. Raw Score to Scale Score Conversion Tables for Winter/Trimester 2008-2009	40
Table 4.5. Raw Score to Scale Score Conversion Tables for Spring 2009.....	46
Table 5.1. Estimates of Accuracy and Consistency of Performance Classification for Winter/Trimester 2008-2009	53
Table 5.2. Estimates of Accuracy and Consistency of Performance Classification for Spring 2009.....	53
Table 5.3. Accuracy and Consistency estimates by cut-score: False positives and false negatives rates for Winter/Trimester 2008-2009	55
Table 5.4. Accuracy and Consistency estimates by cut-score: False positives and false negatives rates for Spring 2009.....	55

Table 6.1. OIB Cut Scores After the Final Round of Rating by Subject.....	58
Table 6.2. Raw Score and Scale Score Cut Scores After the Round 3 Final Rating	58
Figure 6.1. The percentage of students in each performance level using the cut scores after the Round 3 final rating for Biology I, English II, and U.S. History.	59
Table 7.1. Descriptive Statistics of the Scale Scores for Winter/Trimester 2008-2009 ...	60
Table 7.2. Descriptive Statistics of the Scale Scores for Spring 2009.....	60
Table 7.3. Percentage of Students by Performance Level for Winter/Trimester 2008-2009 and Spring 2009	61
Table 7.4. Overall Estimates of SEM by Subject	62

Section 1

Overview of the Oklahoma School Testing Program (OSTP) Achieving Classroom Excellence (ACE) End-of-Instruction (EOI) Assessments

1.1 Overview of the OSTP ACE EOI Assessments

The Achieving Classroom Excellence End-of-Instruction (hereafter, ACE and EOI, respectively) is a state-mandated, secondary level, criterion-referenced testing program used to assess student proficiency at the End-of-Instruction in Algebra I, Algebra II, Geometry, Biology I, English II, English III, and U.S. History. The Oklahoma ACE EOI tests are used to assess student proficiency relative to a specific set of academic skills established by committees of Oklahoma educators. This special set of skills is referred to as the *Priority Academic Student Skills*, or *PASS*, which represents skills that students are expected to master by the End-of-Instruction for each subject. All secondary level students, who have completed instruction in Algebra I, Algebra II, Geometry, Biology I, English II, English III, and U.S. History, must take the corresponding Oklahoma ACE EOI tests in order to graduate from high school. The Spring 2009 administration was the first administration with graduation requirements attached to them for the incoming freshman students. For these students, and future students, in order 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 are permitted to retake these tests. All *PASS* standards and objectives are measured by multiple-choice items except for English II and English III, which include one writing prompt. The Winter/Trimester 2008-2009 and Spring 2009 OSTP-ACE EOI Algebra I, Algebra II, Geometry, Biology I, English II, English III, and U.S. History assessments were developed by Pearson in collaboration with the Oklahoma State Department of Education (SDE) and administered by SDE.

Pearson scored, equated, and scaled the assessments. There was one form administered in Winter/Trimester 2008-2009 for each subject. In the Spring 2009 administration, there were eleven forms in Algebra I, Algebra II, Biology I, and Geometry, thirteen forms in English II, sixteen forms in English III, and fourteen forms in U.S. History. Each test form was embedded with field test items to enhance the item pool. 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 using the Winter/Trimester 2008-2009 test forms and then used again in the Spring 2009 administration. A student could receive an Equivalent test for various reasons, including becoming ill during test administration or experiencing some kind of security breach. The State Department of Education Office of Accountability and Assessments determines eligibility for an Equivalent test on a case-by-case basis. These students' responses were scored and reported using the scoring tables from the previous administration.

1.1.a Purpose

Pearson developed the 2008-2009 OSTP-ACE EOI assessments to measure the Oklahoma content standards listed in the following pages below. The objectives associated with content and/or process standards tested are provided in Appendix A.

1.1.b PASS Content Standards

The Oklahoma Content Standards by subject appears in Table 1.1.

Table 1.1. Oklahoma Content Standards by Subject

Algebra I	
Standard 1.	Number Systems and Algebraic Operations
Standard 2.	Relations and Functions
Standard 3.	Data Analysis, Probability & Statistics
Algebra II	
Standard 1.	Number Sense and Algebraic Operations
Standard 2.	Relations and Functions
Standard 3.	Data Analysis, Probability, & Statistics
Geometry	
Standard 1.	Logical Reasoning
Standard 2.	Properties of 2-Dimensional Figures
Standard 3.	Triangles and Trigonometric Ratios
Standard 4.	Properties of 3-Dimensional Figures
Standard 5.	Coordinate Geometry
Biology I	
PASS Process/Inquiry Standards and Objectives	
Process 1.	Observe and Measure
Process 2.	Classify
Process 3.	Experiment
Process 4.	Interpret and Communicate
Process 5.	Model
PASS 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

Table 1.1 cont. Oklahoma Content Standards by Subject

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
English III	
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.	Social Studies Process Skills
Standard 2.	Civil War/Reconstruction Era
Standard 3.	Immigration/Westward Movement
Standard 4.	Industrial Revolution
Standard 5.	Imperialism/Isolationism
Standard 6.	Twenties Culture/Change
Standard 7.	Great Depression
Standard 8.	World War II
Standard 9.	Post-War Foreign Policy
Standard 10.	Post-War Domestic Policy

1.2 Summary of Test Development and Content Validity

In order to obtain adequate content validity of the Oklahoma ACE/EOI tests, Pearson content experts closely study the *Oklahoma Priority Academic Student Skills (PASS)* and work with Oklahoma content area specialists, teachers, and assessment experts, to develop a pool of items that measured Oklahoma's Assessment Frameworks (*PASS*) for each subject. Once the need for field test items was determined, based on the availability of items for future test construction, a pool of items that measured Oklahoma's *PASS* in each subject was developed. These items were developed under universal design guidelines set by the SDE and carefully reviewed and discussed by Content and Bias/Sensitivity Review Committees to evaluate not only content validity, but also plain language, and the quality and appropriateness of the items. These committees were comprised of Oklahoma teachers and SDE staff. The committees' recommendations were

used to select and/or revise items from the item pool used to construct the field test portions of the Winter/Trimester 2008-2009 and the Spring 2009 assessments.

1.2.a Aligning Test to *PASS* Content Standards

In addition to the test Blueprints provided by SDE, Table 1.2 describes four criteria for test alignment with the *PASS* Standards and Objectives.

Table 1.2. Criteria for Aligning the Test with *PASS* Standards and Objectives.

1. Categorical Concurrence	The test is constructed so that there are at least six items measuring each <i>PASS</i> 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 <i>PASS</i> standard have at least one corresponding assessment items.
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 <i>PASS</i> 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 <i>PASS</i> skill or concept being assessed, not from specialized knowledge or cultural background that the test-taker may bring to the testing situation.

1.2.b Item Pool Development and Selection

The source of the operational items included a pool of previously field-tested or operationally administered items ranging from the Spring 2005 to the Spring 2008 administration for Algebra I, Biology I, English II, and U.S. History and from the census Spring 2007 field test to the Spring 2008 embedded field test for Algebra II, Geometry, and English III. Note that the items were calibrated live using data from the operational administration in order to estimate parameters for these items.

The ACE EOI tests for the Winter/Trimester 2008-2009 and Spring 2009 cycle were built by including previously field tested and operational items around the anchor sets. In order to equate the forms across years, a set of field test and operational items from the Spring 2008 administration served as anchors for Winter/Trimester 2008-2009 and Spring 2009 administrations. Equating is necessary to account for slight year-to-year differences in test difficulty and to maintain comparability across years. Details of the equating procedures applied are provided in a later section in this document. Content experts also targeted the percentage of items measuring various Depth of Knowledge (DOK) levels for assembling the tests. Table 1.3 provides the DOK level percentages for the Winter/Trimester 2008-2009 and Spring 2009 operational assessments. Notice that the actual percentage is close but not exactly within the target percentages in the operational test for some content areas. These targets are expected to be met in future tests.

Table 1.3. Percentage of Items in Depth of Knowledge Levels

Test Session	DOK	Target	Actual						
	Level	DOK	Alg. I	Alg. II	Geo.	Bio. I	Eng. II	Eng. III	U.S. His.
Winter/ Trimester 2008-2009	1	15%-20%	20.00%	16.36%	21.82%	20.00%	4.92%	12.70%	21.67%
	2	60%-70%	61.82%	67.27%	60.00%	65.00%	75.41%	66.67%	58.33%
	3/4	15%-20%	18.18%	16.36%	18.18%	15.00%	19.67%	20.64%	20.00%
Spring 2009	1	15%-20%	20.00%	18.18%	20.00%	20.00%	9.84%	4.76%	18.33%
	2	60%-70%	61.82%	65.45%	61.82%	65.00%	68.85%	80.95%	63.33%
	3/4	15%-20%	18.18%	16.36%	18.18%	15.00%	21.31%	14.29%	18.33%

Note: Alg. I = Algebra I, Alg. II = Algebra II, Geo. = Geometry, Bio. I = Biology I, Eng. II = English II, Eng. III = English III, and U.S. His. = U.S. History.

1.2.c Configuration of the Seven Tests

Tables 1.4 and 1.5 provide overviews of the number of operational and field test items for the Winter/Trimester 2008-2009 and Spring 2009 OSTP-ACE EOI assessments. Field test items were embedded in the operational test forms for all content areas in order to build the item bank for future use. The forms in the Spring 2009 assessments were randomly assigned within classrooms in order to obtain equivalent samples of examinees for the field test items. Table 1.4 provides the total number of forms, total number of operational (OP) and field test (FT) items, and maximum possible points for the Winter/Trimester 2008-2009 assessments. Table 1.5 provides the total number of forms, total number of operational (OP) and field test (FT) items, and maximum possible points for the Spring 2009 assessments.

Table 1.4. Configuration of the OSTP-ACE EOI tests for Winter/Trimester 2008-2009

OSTP- ACE/EOI	Content Area	Form(s)	Total Number of		Test Items	Maximum Possible Points on Test Items Per Form			
			OP	FT		OP		FT	
			Items	Items		MC	CR	MC	CR
Winter/ Trimester 2008-09	Algebra I	1	55	20	75	55	0	20	0
	Algebra II	1	55	20	75	55	0	20	0
	Geometry	1	55	20	75	55	0	20	0
	Biology I	1	60	20	80	60	0	20	0
	English II	1	60/1*	20	80/1*	60	6	20	0
	English III	1	62/1*	20	82/1*	62	10	20	0
	U.S. History	1	60	20	80	60	0	20	0

Note: OP = Operational; FT = Field Test; MC = Multiple Choice; CR = Constructed Response; *=multiple choice/constructed response.

Table 1.5. Configuration of the OSTP-ACE/EOI tests for Spring 2009

OSTP- ACE/EOI	Content Area	Form(s)	Total Number of		Test Items	Maximum Possible Points on Test Items Per Form			
			OP Items	FT Items		OP		FT	
						MC	CR	MC	CR
Spring 2009	Algebra I	1	55	20	75	55	0	20	0
		2	55	20	75	55	0	20	0
		3	55	20	75	55	0	20	0
		4	55	20	75	55	0	20	0
		5	55	20	75	55	0	20	0
		6	55	20	75	55	0	20	0
		7	55	20	75	55	0	20	0
		8	55	20	75	55	0	20	0
		9	55	20	75	55	0	20	0
		10	55	20	75	55	0	20	0
		11	55	20	75	55	0	20	0
Spring 2009	Algebra II	1	55	20	75	55	0	20	0
		2	55	20	75	55	0	20	0
		3	55	20	75	55	0	20	0
		4	55	20	75	55	0	20	0
		5	55	20	75	55	0	20	0
		6	55	20	75	55	0	20	0
		7	55	20	75	55	0	20	0
		8	55	20	75	55	0	20	0
		9	55	20	75	55	0	20	0
		10	55	20	75	55	0	20	0
		11	55	20	75	55	0	20	0
Spring 2009	Geometry	1	55	20	75	55	0	20	0
		2	55	20	75	55	0	20	0
		3	55	20	75	55	0	20	0
		4	55	20	75	55	0	20	0
		5	55	20	75	55	0	20	0
		6	55	20	75	55	0	20	0
		7	55	20	75	55	0	20	0
		8	55	20	75	55	0	20	0
		9	55	20	75	55	0	20	0
		10	55	20	75	55	0	20	0
		11	55	20	75	55	0	20	0

Note: OP = Operational; FT = Field Test; MC = Multiple Choice; CR = Constructed Response.

Table 1.5 cont. Configuration of the OSTP-ACE/EOI tests for Spring 2009

OSTP- ACE/EOI	Content Area	Form(s)	Total Number of			Maximum Possible Points on Test Items Per Form			
			OP	FT	Test	OP		FT	
			Items	Items	Items	MC	CR	MC	CR
Spring 2009	Biology I	1	60	20	80	60	0	20	0
		2	60	20	80	60	0	20	0
		3	60	20	80	60	0	20	0
		4	60	20	80	60	0	20	0
		5	60	20	80	60	0	20	0
		6	60	20	80	60	0	20	0
		7	60	20	80	60	0	20	0
		8	60	20	80	60	0	20	0
		9	60	20	80	60	0	20	0
		10	60	20	80	60	0	20	0
		11	60	20	80	60	0	20	0
Spring 2009	English II	1	60/1*	20	80/1*	60	6	20	0
		2	60/1*	20	80/1*	60	6	20	0
		3	60/1*	20	80/1*	60	6	20	0
		4	60/1*	20	80/1*	60	6	20	0
		5	60/1*	20	80/1*	60	6	20	0
		6	60/1*	20	80/1*	60	6	20	0
		7	60/1*	20	80/1*	60	6	20	0
		8	60/1*	20	80/1*	60	6	20	0
		9	60/1*	20	80/1*	60	6	20	0
		10	60/1*	20	80/1*	60	6	20	0
		11	60/1*	20	80/1*	60	6	20	0
		12	60/1*	20	80/1*	60	6	20	0
		13	60/1*	20	80/1*	60	6	20	0
Spring 2009	English III	1	62/1*	20	82/1*	62	10	20	0
		2	62/1*	20	82/1*	62	10	20	0
		3	62/1*	20	82/1*	62	10	20	0
		4	62/1*	20	82/1*	62	10	20	0
		5	62/1*	20	82/1*	62	10	20	0
		6	62/1*	20	82/1*	62	10	20	0
		7	62/1*	20	82/1*	62	10	20	0
		8	62/1*	20	82/1*	62	10	20	0
		9	62/1*	20	82/1*	62	10	20	0
		10	62/1*	20	82/1*	62	10	20	0
		11	62/1*	20	82/1*	62	10	20	0
		12	62/1*	20	82/1*	62	10	20	0
		13	62/1*	20	82/1*	62	10	20	0
		14	62/1*	20	82/1*	62	10	20	0
		15	62/1*	20	82/1*	62	10	20	0
		16	62/1*	20	82/1*	62	10	20	0

Table 1.5 cont. Configuration of the OSTP-ACE/EOI tests for Spring 2009

		1	60	20	80	60	0	20	0
		2	60	20	80	60	0	20	0
		3	60	20	80	60	0	20	0
		4	60	20	80	60	0	20	0
		5	60	20	80	60	0	20	0
		6	60	20	80	60	0	20	0
Spring	U.S.	7	60	20	80	60	0	20	0
2009	History	8	60	20	80	60	0	20	0
		9	60	20	80	60	0	20	0
		10	60	20	80	60	0	20	0
		11	60	20	80	60	0	20	0
		12	60	20	80	60	0	20	0
		13	60	20	80	60	0	20	0
		14	60	20	80	60	0	20	0

Note: OP = Operational; FT = Field Test; MC = Multiple Choice; CR = Constructed Response; *=multiple choice/constructed response.

1.2.d Operational and Field-Test Items by Content Area

Algebra I. The Winter/Trimester 2008-2009 Algebra I administration was comprised of one form with 55 operational multiple-choice items and 20 field test MC items. There were 16 anchor items to this test, all from the Spring 2008 administration. There were eleven Algebra I test forms in the Spring 2009 administration. Each of the eleven forms contained a duplicate set of 55 operational MC items and 20 unique field test MC items, totaling 75 items per form, and 275 items across forms. The number of items and maximum points possible by content standard is shown in Table 1.6a. Note that Algebra I was reported by content standard and at the objective level. There were four or more items in each reported category. Each item was mapped to one content standard and one objective per content standard.

Table 1.6a. Number of item and points by Content Standard for Algebra I

		Total Number of Items/Points Within a Content Standard						Total	
		1		2		3			
		Its	Pts	Its	Pts	Its	Pts	Its	Pts
Winter/Trimester 2008-09	Operational	15	15	31	31	9	9	55	55
	FT-Form 1	5	5	11	11	4	4	20	20
Spring 2009	Operational	15	15	31	31	9	9	55	55
	FT-Form 1	4	4	11	11	5	5	20	20
	FT-Form 2	6	6	10	10	4	4	20	20
	FT-Form 3	6	6	11	11	3	3	20	20
	FT-Form 4	7	7	11	11	2	2	20	20
	FT-Form 5	5	5	13	13	2	2	20	20
	FT-Form 6	6	6	12	12	2	2	20	20
	FT-Form 7	5	5	13	13	2	2	20	20
	FT-Form 8	6	6	11	11	3	3	20	20
	FT-Form 9	5	5	11	11	4	4	20	20
	FT-Form 10	4	4	10	10	6	6	20	20
FT-Form 11	4	4	11	11	5	5	20	20	

Note: Its = Number of Items; Pts = Number of Points; FT = Field Test.

Algebra II. The Winter/Trimester 2008-2009 Algebra II administration was comprised of one form with 55 operational MC items and 20 field test MC items. There were 15 anchor items to this test, all from Spring 2008 operational administration. There were eleven Algebra II test forms in the Spring 2009 administration. Each of the eleven forms contained a duplicate set of 55 operational MC items and 20 unique field test MC items, totaling 75 items per form, and 275 items across forms. The number of items and maximum points possible by content standard is shown in Table 1.6b. Note that Algebra II was reported by content standard and at the objective level. There were four or more items in each reported category. Each item was mapped to one content standard and one objective per content standard.

Table 1.6b. Number of item and points by Content Standard for Algebra II

		Total Number of Items/Points Within a Content Standard						Total	
		1		2		3			
		Its	Pts	Its	Pts	Its	Pts	Its	Pts
Winter/Trimester 2008-09	Operational	15	15	31	31	9	9	55	55
	FT-Form 1	5	5	13	13	2	2	20	20
Spring 2009	Operational	15	15	31	31	9	9	55	55
	FT-Form 1	7	7	12	12	1	1	20	20
	FT-Form 2	7	7	12	12	1	1	20	20
	FT-Form 3	7	7	10	10	3	3	20	20
	FT-Form 4	7	7	9	9	4	4	20	20
	FT-Form 5	6	6	10	10	4	4	20	20
	FT-Form 6	6	6	11	11	3	3	20	20
	FT-Form 7	6	6	9	9	5	5	20	20
	FT-Form 8	6	6	10	10	4	4	20	20
	FT-Form 9	6	6	7	7	7	7	20	20
	FT-Form 10	6	6	10	10	4	4	20	20
FT-Form 11	7	7	9	9	4	4	20	20	

Note: Its = Number of Items; Pts = Number of Points; FT = Field Test.

Geometry. The Winter/Trimester2007-2008 Geometry administration was comprised of one form with 55 operational MC items and 20 field test MC items. There were 15 anchor items to this test, all from the Spring 2008 operational administration. There were eleven Geometry test forms in the Spring 2009 administration. Each of the eleven forms contained a duplicate set of 55 operational MC items and 20 unique field test MC items, totaling 75 items per form, and 275 items across forms. The number of items and maximum points possible by content standard is shown in Table 1.6c. Note that Geometry was reported by content standard and at the objective level. There were four or more items in each reported category. Each item was mapped to one content standard and one objective per content standard.

Table 1.6c. Number of item and points by Content Standard for Geometry

		Total Number of Items/Points Within a Content Standard										Total	
		1		2		3		4		5		Total	
		Its	Pts	Its	Pts	Its	Pts	Its	Pts	Its	Pts	Its	Pts
Winter/Trimester 2008-09	Operational	6	6	20	20	12	12	10	10	7	7	55	55
	FT-Form 1	3	3	5	5	4	4	4	4	4	4	20	20
Spring 2009	Operational	6	6	20	20	12	12	10	10	7	7	55	55
	FT-Form 1	3	3	7	7	5	5	5	5	.	.	20	20
	FT-Form 2	2	2	8	8	4	4	4	4	2	2	20	20
	FT-Form 3	2	2	5	5	5	5	3	3	5	5	20	20
	FT-Form 4	2	2	6	6	6	6	4	4	2	2	20	20
	FT-Form 5	2	2	6	6	4	4	6	6	2	2	20	20
	FT-Form 6	2	2	5	5	6	6	4	4	3	3	20	20
	FT-Form 7	2	2	5	5	5	5	6	6	2	2	20	20
	FT-Form 8	2	2	7	7	4	4	3	3	4	4	20	20
	FT-Form 9	2	2	7	7	5	5	3	3	3	3	20	20
	FT-Form 10	2	2	5	5	4	4	6	6	3	3	20	20
FT-Form 11	2	2	8	8	4	4	3	3	3	3	20	20	

Note: Its = Number of Items; Pts = Number of Points; FT = Field Test.

Biology I. The Winter/Trimester 2008-2009 Biology I administration was comprised of one form with 55 operational MC items and 20 field test MC items. There were 18 anchor items to this test, all from the Spring 2008 administration. There were eleven Biology I test forms in the Spring 2009 administration. Each of the eleven forms contained a duplicate set of 60 operational MC items and 20 unique field test MC items, totaling 80 items per form, and 280 items across forms. The number of items and the maximum number points possible by content standard in Biology I are shown in Table 1.6d. Note that Biology I was reported for content and process standards at the standard level. Each reported standard has four or more items. Unlike other content areas, all items in Biology I were primarily mapped to process standards. All items (except safety items) were also mapped to content standards.

Table 1.6d. Number of item and points by Content Standard for Biology I

		Total Number of Items/Points Within a Content Standard												Total*	
		1		2		3		4		5		6			
		Its	Pts	Its	Pts	Its	Pts	Its	Pts	Its	Pts	Its	Pts	Its	Pts
WI08	Operational	9	9	8	8	9	9	12	12	9	9	9	9	56	56
	FT-Form 1	2	2	4	4	3	3	5	5	4	4	.	.	18	18
SP09	Operational	10	10	9	9	9	9	12	12	9	9	7	7	56	56
	FT-Form 1	3	3	2	2	2	2	4	4	4	4	3	3	18	18
	FT-Form 2	4	4	2	2	3	3	5	5	3	3	1	1	18	18
	FT-Form 3	2	2	2	2	3	3	6	6	4	4	2	2	19	19
	FT-Form 4	3	3	4	4	1	1	4	4	4	4	3	3	19	19
	FT-Form 5	3	3	4	4	2	2	3	3	6	6	1	1	19	19
	FT-Form 6	4	4	5	5	2	2	2	2	3	3	3	3	19	19
	FT-Form 7	3	3	3	3	5	5	5	5	4	4	.	.	20	20
	FT-Form 8	3	3	4	4	3	3	4	4	4	4	1	1	19	19
	FT-Form 9	4	4	5	5	2	2	4	4	3	3	1	1	19	19
	FT-Form 10	1	1	6	6	5	5	4	4	4	4	.	.	20	20
FT-Form 11	2	2	3	3	4	4	3	3	4	4	3	3	19	19	

Note: WI08 = Winter/Trimester 2008-2009; SP09 = Spring 2009; Its = Number of Items; Pts = Number of Points; FT = Field Test; Some totals for OP forms and FT forms are less than 60 (for OP) and 20 (for FT) due to dual item alignment – an item does not map to a content standard, but maps to a process

English II. The Winter/Trimester 2008-2009 English II administration was comprised of one form with 60 operational MC items, 1 open-ended writing prompt, and 20 field test MC items. There were 21 anchor items to this test, all from the Spring 2008 administration. There were thirteen English II test forms in the Spring 2009 administration. Each of the thirteen forms contained a duplicate set of 60 operational MC items, 1 operational open-ended writing prompt, and 20 unique field test MC items, totaling 81 items per form, and 321 items across forms. Table 1.6e lists the number of items and the maximum possible number of points by content standard in the Winter/Trimester 2008-2009 and Spring 2009 tests. Note that English II was reported at the content standard level. Each item was mapped to one content standard and one objective. Note that the writing prompts in English II, both for Winter/Trimester and Spring, were scored analytically at five traits with a maximum of four score points for each trait. The scores in the analytic traits were reported in the Writing report. The trait scores were weighted differentially to derive a composite score that ranged from 1 to 6. The composite scores contributed to the English II total score.

Table 1.6e. Number of item and points by Content Standard for English II
Total Number of Items/Points Within a Content Standard

		R1		R2		R3		R4		W1/W2		W3		Total	
		Its	Pts	Its	Pts	Its	Pts	Its	Pts	Its	Pts	Its	Pts	Its	Pts
WI08	Operational	7	7	16	16	20	20	5	5	1	6	12	12	61	66
	FT-Form 1	2	2	2	2	7	7	2	2	.	.	7	7	20	20
SP09	Operational	7	7	16	16	19	19	6	6	1	6	12	12	61	66
	FT-Form 1	2	2	9	9	7	7	2	2	20	20
	FT-Form 2	1	1	5	5	7	7	1	1	.	.	6	6	20	20
	FT-Form 3	3	3	6	6	4	4	1	1	.	.	6	6	20	20
	FT-Form 4	3	3	9	9	5	5	3	3	20	20
	FT-Form 5	1	1	7	7	5	5	1	1	.	.	6	6	20	20
	FT-Form 6	2	2	3	3	7	7	2	2	.	.	6	6	20	20
	FT-Form 7	2	2	6	6	4	4	2	2	.	.	6	6	20	20
	FT-Form 8	2	2	3	3	10	10	1	1	.	.	4	4	20	20
	FT-Form 9	1	1	7	7	7	7	1	1	.	.	4	4	20	20
	FT-Form 10	2	2	3	3	8	8	2	2	.	.	5	5	20	20
	FT-Form 11	1	1	8	8	5	5	6	6	20	20
	FT-Form 12	.	.	9	9	1	1	2	2	.	.	8	8	20	20
FT-Form 13	1	1	2	2	3	3	1	1	.	.	13	13	20	20	

Note: WI08 = Winter/Trimester 2008-2009; SP09 = Spring 2009; Its = Number of Items; Pts = Number of Points; FT = Field Test.

English III. The Winter/Trimester 2008-2009 English III administration was comprised of one forms with 62 operational MC items, 1 open-ended writing prompt, and 20 field test MC items. There were 18 anchor items to this test, all from the Spring 2008 administration. There were sixteen English III test forms in the Spring 2009 administration. Each of the sixteen forms contained a duplicate set of 62 operational MC items, 1 operational open-ended writing prompt, and 20 unique field test MC items, totaling 83 items per form, and 383 items across forms (some field-test items were duplicated across forms). Table 1.6f lists the number of items and the maximum possible number of points by content standard in the Winter/Trimester 2008-2009 and Spring 2009 tests. Note that English III was reported at the content standard level. Each item was mapped to one content standard and one objective. Note that the writing prompts in English III, both for Winter/Trimester and Spring, were scored analytically at five traits with a maximum of four score points for each trait. The scores in the analytic traits were reported in the Writing report. The trait scores were weighted differentially to derive a

composite score that ranged from 1 to 10. The composite scores contributed to the English III total score.

Table 1.6f. Number of item and points by Content Standard for English III

		Total Number of Items/Points Within a Content Standard													
		R1		R2		R3		R4		W1/W2		W3		Total	
		Its	Pts	Its	Pts	Its	Pts	Its	Pts	Its	Pts	Its	Pts	Its	Pts
WI08	Operational	5	5	18	18	20	20	5	5	1	10	14	14	63	72
	FT-Form 1	2	2	7	7	4	4	7	7	20	20
SP09	Operational	6	6	17	17	19	19	6	6	1	10	14	14	63	72
	FT-Form 1	2	2	9	9	6	6	3	3	20	20
	FT-Form 2	3	3	9	9	7	7	1	1	20	20
	FT-Form 3	2	2	8	8	7	7	3	3	20	20
	FT-Form 4	5	5	6	6	8	8	1	1	20	20
	FT-Form 5	2	2	5	5	7	7	2	2	.	.	4	4	20	20
	FT-Form 6	3	3	7	7	5	5	5	5	20	20
	FT-Form 7	2	2	4	4	7	7	2	2	.	.	5	5	20	20
	FT-Form 8	2	2	6	6	5	5	3	3	.	.	4	4	20	20
	FT-Form 9	1	1	9	9	4	4	1	1	.	.	5	5	20	20
	FT-Form 10	2	2	10	10	4	4	4	4	20	20
	FT-Form 11	3	3	4	4	4	4	1	1	.	.	8	8	20	20
	FT-Form 12	2	2	3	3	6	6	1	1	.	.	8	8	20	20
	FT-Form 13	1	1	2	2	7	7	2	2	.	.	8	8	20	20
	FT-Form 14	2	2	3	3	4	4	3	3	.	.	8	8	20	20
	FT-Form 15	2	2	7	7	2	2	1	1	.	.	8	8	20	20
FT-Form 16	1	1	7	7	3	3	1	1	.	.	8	8	20	20	

Note: WI08 = Winter/Trimester 2008-2009; SP09 = Spring 2009; Its = Number of Items; Pts = Number of Points; FT = Field Test.

U.S. History The Winter/Trimester 2008-2009 U.S. History administration was comprised of one form with 60 operational multiple-choice items and 20 field test MC items. There were 20 anchor items to this test, all from the Spring 2008 administration. There were fourteen U.S. History test forms in the Spring 2009 administration. Each of the fourteen forms contained a duplicate set of 60 operational MC items and 20 unique field test MC items, totaling 80 items per form, and 240 items across forms. The number of items and maximum points possible by content standard in Winter/Trimester 2008-

2009 and Spring 2009 are shown in Table 1.6g. Note that U.S. History was reported only at the content standard level and each reported standard had four or more items.

Table 1.6g. Number of item and points by Content Standard for U.S. History

		Total Number of Items/Points Within a Content Standard																					
		01		02		03		04		05		06		07		08		09		10		Total	
		Its	Pts	Its	Pts	Its	Pts	Its	Pts	Its	Pts	Its	Pts	Its	Pts	Its	Pts	Its	Pts	Its	Pts	Its	Pts
WI08	Operational	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	60	60
	FT-Form 1	3	3	3	3	.	.	3	3	2	2	1	1	2	2	2	2	2	2	2	2	20	20
SP09	Operational	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	60	60
	FT-Form 1	2	2	2	2	2	2	2	2	2	2	3	3	2	2	2	2	2	2	1	1	20	20
	FT-Form 2	2	2	2	2	2	2	2	2	2	2	1	1	2	2	2	2	3	3	2	2	20	20
	FT-Form 3	1	1	2	2	2	2	2	2	1	1	2	2	2	2	2	2	4	4	2	2	20	20
	FT-Form 4	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	20	20
	FT-Form 5	2	2	2	2	2	2	2	2	2	2	2	2	3	3	2	2	1	1	2	2	20	20
	FT-Form 6	2	2	2	2	3	3	1	1	2	2	1	1	2	2	3	3	2	2	2	2	20	20
	FT-Form 7	2	2	2	2	1	1	2	2	.	.	2	2	2	2	4	4	1	1	4	4	20	20
	FT-Form 8	1	1	2	2	1	1	1	1	3	3	3	3	3	3	2	2	2	2	2	2	20	20
	FT-Form 9	1	1	.	.	2	2	1	1	2	2	1	1	3	3	2	2	5	5	3	3	20	20
	FT-Form 10	1	1	3	3	1	1	1	1	2	2	2	2	5	5	5	5	20	20
	FT-Form 11	3	3	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	20	20
	FT-Form 12	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	3	3	3	3	20	20
	FT-Form 13	4	4	1	1	2	2	2	2	1	1	2	2	5	5	.	.	2	2	1	1	20	20
FT-Form 14	3	3	2	2	3	3	3	3	1	1	3	3	2	2	1	1	1	1	1	1	20	20	

Note: WI08 = Winter/Trimester 2008-2009; SP09 = Spring 2009; Its = Number of Items; Pts = Number of Points; FT = Field Test.

Section 2

Administration of the ACE 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 accurately reflects their achievement in each subject.

The schools play a key role in administering the OSTP-ACE EOI assessments in a manner 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 district and school faculty members play is essential in the fair and equitable administration of successful ACE EOI assessments.

2.1 Packaging and Shipping

To provide OSTP-ACE EOI with secure and dependable services for the shipping of the Oklahoma 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 Pearson's SchoolHouse™ Web site, and optionally with data received directly from Oklahoma. Oklahoma educators also use the SchoolHouse™ 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 the accurate tracking of these materials through the entire packing, delivery, and return process. It also permits us 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 x, 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 all material progresses, any discrepancies in material received versus shipped are reported immediately to the SDE and scoring will begin. This system allows us 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 notified the SDE regarding unresolved material discrepancies within 24 hours after our initial attempt to contact the school principal. Within one week after materials are returned, our Service Center Team also notified 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 same materials for scanning within 72-hours of receipt. District status is available from a web-based system readily accessible by SDE. In addition, the Oklahoma Call Center is able to provide receipt status information if required. The receipt notification Web site'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 Field Test Design

3.1.a Sampling Plan

Population data was used for classical and item response theory (IRT) analyses for all Winter/Trimester 2008-2009 and Spring 2009 tests. All students who complete a course with an End-of-Instruction test associated with it must also take the test.

3.1.b Field-Test Design

New items are field tested to build-up the item bank for future high stakes administrations. The overall field test design used by Pearson was an embedded field test design where newly developed field test items were embedded throughout the test. The advantage of an embedded field test design is that test-takers do not know where the field test items are located and therefore will treat each item as a scored item. Twenty field test items per form were placed in common positions across forms and administrations (Winter/Trimester and Spring). Field test items were prioritized for inclusion on forms based on current item bank analyses.

3.1.c Data Receipt Activities

After all tests were scored, a data file was provided for item analyses and calibration. A data clean up process was completed that removed invalid cases, ineligible responses, absent students, and second time test takers. 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 research scientist 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 cleaning, 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. The demographic breakdown of the students in the Winter/Trimester 2008-2009 and Spring 2009 item analysis and calibration sample appear in Table 3.1 and 3.2, respectively.

Table 3.1. Demographic characteristics of calibration and equating sample for Winter/Trimester 2008-2009

Subject	Total	Male	Female	African		Hispanic	Asian	Pacific		White	Other
				American	Native American			Islander			
Algebra I	1499	733	766	308	197	124	23	0	823	24	
Algebra II	1915	928	987	325	216	121	40	1	1194	18	
Biology I	2073	1042	1031	405	228	132	40	3	1243	22	
English II	2628	1320	1308	449	367	181	59	2	1550	20	
English III	2783	1390	1381	400	404	199	60	27	1654	39	
Geometry	1901	926	975	331	235	182	33	1	1097	22	
U.S. History	2600	1288	1312	433	344	199	30	2	1566	26	

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

Table 3.2. Demographic characteristics of calibration and equating sample for Spring 2009

Subject	Total	Male	Female	African		Hispanic	Asian	Pacific		White	Other
				American	Native American			Islander			
Algebra I	35736	17817	17919	3696	6684	3159	728	54	21140	275	
Algebra II	29644	14355	15289	2591	5384	2146	702	38	18610	173	
Biology I	35347	17586	17761	3365	6444	3092	787	57	21381	221	
English II	34823	17137	17686	3214	6478	3032	742	55	21122	180	
English III	34842	17331	17511	3457	6497	2618	769	32	21269	200	
Geometry	34224	17132	17092	3132	6348	2819	760	41	20933	191	
U.S. History	32277	15993	16284	3038	5890	2529	802	39	19788	191	

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

Statistical Key Check. Administering students items that have only one correct key and are correctly scored is critical for accurate assessment of student performance. In order to screen for potentially problematic items, 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 below 0.20
- Distractor p-value greater than or equal to 0.40
- Distractor point-biserial greater than or equal to 0.05

Any flagged operational item was submitted for key review to the appropriate Pearson content specialist. Any flagged items that are identified by content experts as having answer key issues would be submitted to SDE for review before dropping the item from the operational scoring. There were no items identified in the Winter/Trimester 2008-2009 or Spring 2009 administration as having a key issue. Once the keys were verified, classical item analyses were conducted.

3.2 Classical Item Analyses

Once the data receipt activities and statistical key check were completed, the following classical item analyses were conducted for operational and field-test items:

- 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 correlation (point-biserial)
 - Mean item-test correlation (point-biserial)
 - Point-biserial 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)

Once the keys were verified and the item analysis results reviewed, the data were used for calibration and equating.

3.2.a Test-level summaries of classical item analyses

The test-level raw score descriptive statistics for the calibration samples is shown in Table 3.3. Note that students whose tests were invalidated and those students taking the test for a second time were excluded. The operational test results indicate that the omit rates were smaller than 1% for all subjects. The mean raw score and the mean percent of the maximum raw scores were relatively similar for both administrations. As indicated in the test configuration section, there were multiple forms with a duplicate set of operational items and a unique set of field test items in the Winter/Trimester 2008-2009 and Spring 2009 tests. A separate item analysis by test form indicated that, in both administrations, the omit rates were below 2% for all content areas. The mean percent of the maximum possible raw score across forms indicates that the forms were relatively similar in difficulty for all content areas.

Table 3.3. Test level summaries of classical item analyses for Winter/Trimester 2008-2009 and Spring 2009

Administration	Sample Size	Mean	Mean % of Max	Number of Items/Points	*Average P-value	Average Pt. Biserial	Omit Min	Omit Max
AlgebraI-W08	1499	30.06	0.55	55	0.55	0.42	0.00	0.53
AlgebraI-S09	35736	34.39	0.63	55	0.63	0.41	0.01	0.12
AlgebraII-W08	1915	32.14	0.58	55	0.58	0.44	0.00	0.26
AlgebraII-S09	29644	29.46	0.54	55	0.54	0.43	0.01	0.12
Biology I-W08	2073	39.11	0.65	60	0.65	0.40	0.00	0.39
Biology I-S09	35347	39.49	0.66	60	0.66	0.41	0.03	0.11
EnglishII-W08	2628	47.17	0.71	61/66	0.73	0.39	0.04	0.22
EnglishII-S09	34823	47.63	0.72	61/66	0.73	0.39	0.00	0.13
EnglishIII-W08	2783	42.42	0.59	63/72	0.60	0.39	0.00	0.40
EnglishIII-S09	34842	46.17	0.64	63/72	0.65	0.43	0.00	0.23
Geometry-W08	1901	34.21	0.62	55	0.62	0.43	0.05	0.32
Geometry-S09	34224	34.46	0.63	55	0.63	0.45	0.02	0.13
USHistory-W08	2600	37.68	0.63	60	0.63	0.43	0.00	0.15
USHistory-S09	32277	38.89	0.65	60	0.65	0.40	0.01	0.08

*Note: W08 = Winter/Trimester 2008-2009; S09 = Spring 2009; pt. biserial = point biserial.

3.3 Procedures for Detecting Item Bias

One of the goals of the OSTP-ACE 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. DIF procedures typically control for overall between-group differences on a criterion, usually total test scores. Between-group performance on each item is then compared within sets of examinees having the same 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 only used 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 OSTP-ACE EOI tests DIF analyses, DIF statistics were estimated for all major subgroups of students with sufficient sample size: African American, Hispanic, Asian, Native American, and Female. Field-test 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 the 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 referent groups for this DIF analysis were White for race and male for gender. The focal groups were females and minority race groups.

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. Positive values of *delta* indicate that the item is easier for the *focal* group, suggesting that the item favors the *focal* group. A negative value of *delta* indicates that the item is more difficult for the *focal* group. The item classifications are based on the Mantel-Haenszel chi-square and the MH delta (Δ) value 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 Different Item Functioning Results

The data in Table 3.4 summarizes the number of items in DIF categories for the seven subjects for the Winter/Trimester 2008-2009 and Spring 2009 administration. The results presented in Table 3.4 are for field test items only. Items flagged for DIF were placed before expert content specialist committees during Spring 2009 field test data review as described in the Section 3.4. Field test items that exhibit bias as a result of the content of the item were removed from the item bank excluding them from future use.

Table 3.4. DIF flag incidence across all OSTP-ACE EOI field test items for Winter/Trimester 2008-2009 and Spring 2009

Subject and Admin.	Total FT Items	Native American	Asian	African American	Hispanic	Female
Algebra I-FT Winter 2007-2008	20	1	0	0	2	0
Algebra II-FT Winter 2007-2008	20	0	0	0	0	0
Geometry-FT Winter 2007-2008	20	0	0	1	0	0
Biology I-FT Winter 2007-2008	20	0	0	0	0	1
English II-FT Winter 2007-2008	20	0	1	0	0	4
English III-FT Winter 2007-2008	20	0	0	0	1	0
U.S. History-FT Winter 2007-2008	20	0	0	1	0	0
Algebra I-FT Spring 2009	140	0	3	12	6	7
Algebra II-FT Spring 2009	140	0	2	8	10	9
Geometry-FT Spring 2009	140	1	3	4	5	2
Biology I-FT Spring 2009	140	0	5	8	14	12
English II-FT Spring 2009	260	2	1	13	11	8
English III-FT Spring 2009	320	6	4	22	22	16
U.S. History-FT Spring 2009	280	2	1	12	7	11

Note: Admin. = Administration; FT = Field Test.

3.4 Data Review

Data review represents a critical step in the test development cycle. At the Data Review meeting, SDE and Pearson staff had the opportunity to review actual student performance on the newly developed and field tested multiple choice items across the seven subjects based on the Winter/Trimester 2008-2009 and Spring 2009 field test administrations. The data review focused on the content validity, curricular alignment and statistical functioning of field tested items prior to selection for operational test forms. The field test results used in the data review provided evidence that the items were designed to yield valid results and were accessible for use by the widest possible range of students. The review of student performance should provide evidence regarding the fulfillment of requirement 200.2(b)(2) of NCLB. The purpose of the review meeting was to ensure that psychometrically sound, fair and aligned items are used in the construction of the ACE EOI assessments and entered into the respective item banks. Pearson provided technical and psychometric expertise to provide a clear explanation about the content of the items, the field test process, the scoring process, and the resulting field test data to ensure the success of these meetings and the defensibility of the program.

Data review meetings were a collaborative effort between SDE and Pearson. SDE administrators and content specialists attended the meeting facilitated by Pearson content specialists and research scientists who trained the SDE staff on how to interpret and review the field test data. Meeting materials included a document explaining the flagging criteria, a document containing flagged items, and the item images. Pearson discussed with SDE the analyses performed and the criteria for flagging the items. Flagged items were then reviewed and decisions were made as to whether to accept the item, accept the item with revisions, or reject the item. Review of the data included presentation of p-value, point-biserial, point-biserial by response option, response distributions, mean

overall score by response option, and indications of item DIF and IRT mis-fit. Items failing to meet the requirements of sound technical data were carefully considered for rejection by the review panel, thereby enhancing the reliability and improving the validity of the items left in the bank for future use. While the panel used the data as a tool to inform their judgments, the panel (and not the data alone) made the final determination as to the appropriateness or fairness of the assessment items. The flagging criteria for the ACE EOI assessments are as follows:

- P-value: $<.25$ or $>.90$
- Point-biserial: $<.15$
- Distracter point-biserial: $>.05$ (positive)
- Differential Item Functioning (DIF): Test item biases for subgroups
- IRT mis-fit as flagged by the Q1 index (please see section 4.2 for explanation)

Bias Review. One aspect of the data review meetings was to assess potential bias based on DIF results and item content. Although bias in the items had been avoided through writer training and review processes, there is always the potential for bias to be detected through statistical analysis. It is important to include this step in the development cycle because SDE and Pearson do not want to include an item that is biased in some way against a group, because the item may lead to inequitable test results. As described earlier, all field-test items were analyzed statistically for DIF using the field test data. A Pearson research scientist explained the meaning, in terms of level, and the direction of the DIF flags. The data review panel reviewed the item content, the percentage of students selecting each response option, and the point-biserial for each response option by gender and ethnicity for all items flagged for DIF. The data review panel was then asked if there were 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).

3.4.a Results of Data Review

The number of items inspected during data review is presented in Table 3.5 as a result of the item meeting the statistical flagging criteria for the classical item analyses, DIF, and IRT procedures.

Table 3.5. Number of items per subject flagged and rejected during Winter/Trimester 2008–2009 and Spring 2009 field test data review

Subject and Admin.	No. of FT Items	No. Flagged	Rejected	Accepted	Accepted with edits
Algebra I – Winter 2008-2009	20	15	1	10	4
Algebra II – Winter 2008-2009	20	8	0	8	0
Geometry – Winter 2008-2009	20	6	0	6	0
Biology I – Winter 2008-2009	20	9	3	6	0
English II – Winter 2008-2009	20	8	1	7	0
English III – Winter 2008-2009	20	7	1	6	0
U.S. History – Winter 2008-2009	20	7	1	4	2
Algebra I – Spring 2009	220	79	9	65	5
Algebra II – Spring 2009	220	83	18	56	9
Geometry – Spring 2009	220	61	13	40	8
Biology I – Spring 2009	220	85	16	54	15
English II – Spring 2009	260	85	29	55	0
English III – Spring 2009	320	126	34	92	0
U.S. History – Spring 2009	280	95	17	67	11

Note: No. = Number; Admin. = Administration.

3.5 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 measure of internal consistency. Cronbach's alpha is an internal consistency measure, which is derived from analysis of the consistency of the performance of individuals on items in a test administration. This is the formula for the most common index of reliability, namely, Cronbach's coefficient *alpha* (α). In this formula, the s_i^2 denotes the variances for the k individual items; s_{sum}^2 denotes the variance for the sum of all items:

$$\alpha = (k/(k-1)) * [1 - \sum(s_i^2)/s_{sum}^2] \quad (1)$$

Cronbach's alpha was estimated for each of the content areas for the operational portion of the test.

Table 3.6 presents the estimated reliability index, Cronbach's alpha, for the operational tests by subject area for the Winter/Trimester 2008-2009 and Spring 2009 ACE EOI administrations. These reliabilities indicate that the OSTP-ACE EOI assessments had strong internal consistency and that the tests produce relatively stable scores.

Table 3.6. Cronbach's alpha for Winter/Trimester 2008-2009 and Spring 2009 Administration by Subject

Administration	Cronbach's Alpha
Algebra-W08	0.92
Algebra-S08	0.91
AlgebraII-W08	0.92
AlgebraII-S09	0.92
Biology I-W08	0.91
Biology I-S09	0.91
EnglishII-W08	0.90
EnglishII-S09	0.90
EnglishIII-W08	0.91
EnglishIII-S09	0.92
Geometry-W08	0.92
Geometry-S09	0.93
USHistory-W08	0.92
USHistory-S09	0.91

Note: W08 = Winter/Trimester 2008-2009; S09 = Spring 2009

3.6 Test Reliability by Subgroup

Table 3.7 addresses the reliability analysis results by the different reporting subgroups for the OSTP-ACE EOI assessments in for the Spring 2009. Table 3.7 illustrates the subject of interest, the subgroups, the number of students used in the analyses and the associated Cronbach's Alpha for each subject and subgroup. In all instances, the reliability coefficients are well-above the accepted lower limit of .70.

Table 3.7. Test Reliability by Subgroup for Spring 2009.

Subject	Male	Female	African-American	Native American	Hispanic	Asian	White	ELL	IEP	ECDV
Algebra I	0.91	0.90	0.89	0.89	0.90	0.91	0.90	0.88	0.89	0.89
Algebra II	0.92	0.92	0.89	0.90	0.91	0.94	0.92	0.91	0.86	0.90
Biology I	0.92	0.91	0.90	0.90	0.91	0.93	0.91	0.89	0.90	0.91
English II	0.91	0.90	0.90	0.90	0.91	0.92	0.90	0.89	0.90	0.90
English III	0.93	0.93	0.92	0.92	0.92	0.93	0.93	0.86	0.89	0.92
Geometry	0.93	0.92	0.90	0.91	0.91	0.94	0.93	0.91	0.88	0.91
U.S. History	0.92	0.90	0.90	0.90	0.90	0.92	0.91	0.87	0.90	0.90

Note: ELL = English Language Learner, IEP = Individual Education Plan; ECDV = Economically Disadvantaged.

3.7 Inter-rater Reliability

Inter-rater reliability is interchangeably 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). Both the Winter/Trimester 2008-2009 and Spring 2009 English II and English III tests contained one operational writing prompt each. Raters were trained to implement the scoring rubrics, anchor papers, check sets, and resolution reading. The items were scored by two raters analytically on five strands in both administrations. The final writing score for a student in a given strand is the average of the two scores. The inter-rater reliability results for the operational prompt are presented in Table 3.8 for English II and Table 3.9 for English III. The results show that exact and adjacent rater agreement on trait scores for both the Winter/Trimester 2008-2009 and Spring 2009 operational writing prompts were reasonably high. The weighted Kappa statistic (Kraemer, 1982) is an indication of inter-rater reliability after correcting for chance. The Kappa values for the OSTP-ACE EOI Winter/Trimester 2008-2009 and Spring 2009 operational writing prompts are within the moderate range.

Table 3.8. Inter-rater reliability for English II operational writing prompts for Winter/Trimester 2008-2009 and Spring 2009.

Trait	Max Point	Valid N	Point Discrepancy Percentages							Agreement Percentages			Kappa
			-3	-2	-1	0	1	2	3	Exact	Adjacent	+/- 2 or more	
Winter/Trimester 2008-2009													
1	4	2,432	0.00	0.78	19.53	58.14	20.48	1.07	0.00	58.14	40.01	1.85	0.37
2	4	2,432	0.00	0.86	19.33	58.68	20.31	0.82	0.00	58.68	39.64	1.68	0.39
3	4	2,432	0.00	0.70	18.22	61.39	18.91	0.78	0.00	61.39	37.13	1.48	0.42
4	4	2,432	0.00	0.82	19.61	58.80	19.94	0.82	0.00	58.80	39.55	1.64	0.40
5	4	2,432	0.00	0.86	20.76	56.95	20.52	0.86	0.04	56.95	41.28	1.76	0.40
Spring 2009													
1	4	32,767	0.01	0.52	17.41	64.07	17.37	0.61	0.00	64.07	34.78	1.14	0.32
2	4	32,767	0.01	0.49	17.48	64.16	17.25	0.60	0.01	64.16	34.73	1.11	0.33
3	4	32,767	0.00	0.50	17.04	65.04	16.84	0.57	0.00	65.04	33.88	1.07	0.33
4	4	32,767	0.00	0.81	18.13	62.26	18.12	0.68	0.00	62.26	36.25	1.49	0.35
5	4	32,767	0.01	0.77	19.14	60.19	19.09	0.81	0.01	60.19	38.23	1.60	0.34

Table 3.9. Inter-rater reliability for English III operational writing prompts for Winter/Trimester 2008-2009 and Spring 2009.

Trait	Max Point	Valid N	Point Discrepancy Percentages							Agreement Percentages			Kappa
			-3	-2	-1	0	1	2	3	Exact	Adjacent	+/- 2 or more	
Winter/Trimester 2008-2009													
1	4	2,588	0.00	0.19	15.69	67.00	16.81	0.31	0.00	67.00	32.50	0.50	0.53
2	4	2,588	0.00	0.35	16.23	65.77	17.23	0.43	0.00	65.77	33.46	0.78	0.49
3	4	2,588	0.00	0.19	15.49	68.28	15.80	0.23	0.00	68.28	31.29	0.42	0.51
4	4	2,588	0.00	0.35	16.65	65.15	17.31	0.54	0.00	65.15	33.96	0.89	0.51
5	4	2,588	0.00	0.46	19.59	58.96	20.32	0.66	0.00	58.96	39.91	1.12	0.48
Spring 2009													
1	4	32,456	0.01	0.74	18.89	60.89	18.7	0.75	0.01	60.89	37.59	1.51	0.42
2	4	32,456	0.01	0.85	18.91	60.50	18.85	0.87	0.01	60.50	37.76	1.74	0.42
3	4	32,456	0.00	0.51	17.97	63.29	17.66	0.56	0.00	63.29	35.63	1.07	0.40
4	4	32,456	0.00	0.77	19.30	60.37	18.71	0.83	0.01	60.37	38.01	1.61	0.43
5	4	32,456	0.02	1.19	21.05	55.65	20.90	1.19	0.01	55.65	41.95	2.41	0.41

Section 4

Calibration, Equating, and Scaling

4.1 Item Response Theory (IRT) models

Dichotomous Item Response Theory Model. The three-parameter logistic (3-PL) item response theory (IRT) model (Lord & Novick, 1968) was used for calibrating the multiple choice or dichotomously scored items. In the 3-PL model (Lord, 1980) the probability that a student with ability estimate of θ responds correctly to item i is:

$$P_i(\theta) = c_i + (1 - c_i) \frac{1}{1 + e^{-Da_i(\theta - b_i)}} \quad (2)$$

where θ is the student proficiency parameter, a_i is the item discrimination parameter, b_i is the item difficulty parameter, c_i is the lower asymptote parameter, and D is a scaling constant. The scaling constant is traditionally 1.7. With multiple-choice items it is assumed that, due to guessing, examinees with minimal proficiency have a probability greater than zero of responding correctly to an item. This probability is represented in the 3PL model by the c_i parameter.

Polytomous Item Response Theory Model. For calibrating the polytomously scored constructed response or open ended (OE; or writing prompt) items, the Generalized Partial Credit (GPC; Muraki, 1997) model was used. In the GPC model, the probability that a student with proficiency θ will have a score in the k^{th} category of the i^{th} item is

$$P_{ik}(\hat{\theta}) = \frac{\exp\left[\sum_{v=1}^k a_i(\hat{\theta} - b_{iv})\right]}{\sum_{c=0}^{m_i} \exp\left[\sum_{v=0}^c a_i(\hat{\theta} - b_{iv})\right]} \quad (3)$$

where m_i is the total score levels for item i for $k = v$ category responses, a_i is the slope parameter (or Da_i), and b_{iv} is the category intersection parameters (or $(b_i - d_{iv})$ where b_i is location/difficulty and d_{iv} is the threshold parameters representing category boundaries relative to the item location parameter).

The IRT models were implemented using MULTILOG 7.0 (Thissen, Chen, & Bock, 2003). MULTILOG estimates parameters simultaneously for dichotomous and polytomous items via marginal maximum likelihood procedures and implements the GPC model with the appropriate parameter coding. All item and student proficiency calibrations were independently conducted and verified by at least two Pearson research scientists.

4.2 Assessment of IRT Fit to the model

Item fit was assessed using the Yen's (1981, 1984) Q1 item fit index, which approximately follows a χ^2 distribution:

$$Q_{1i} = \sum_{r=1}^{10} \frac{N_r(O_{ir}-E_{ir})^2}{E_{ir}(1-E_{ir})} \quad (4)$$

where Q_{1i} is the fit of the i th item, N_r is the number of examinees per cell i , O_{ir} is the observed proportion of examinees in cell r that correctly answers item i , and E_{ir} is the predicted portion of examinees in cell r that correctly answers item i . The predictions are obtained by using trait and item parameter estimates in Equations 2 and 3 and summing over examinees in cell r :

$$E_{ir} = \frac{1}{N_r} \sum_{k \in R}^{N_r} \hat{P}_i(\hat{\theta}_k) \quad (5)$$

Since the chi-square statistics are affected by sample size and associated degrees of freedom, the following standardization of the Q1 statistics was used:

$$Z_j = \frac{Q_{1i} - df}{\sqrt{(2df)}} \quad (6)$$

The Z-statistic is an index of the degree to which obtained proportions of item scores are close to the proportions that would be expected based on the estimated thetas and item parameters. In order to assess item fit, a critical Z-value is computed and item Z-values above this critical Z-value may indicate poor item fit. Differences between expected and observed item performance may indicate poor item fit. The item characteristic curves, classical item analyses, and item content were reviewed for items flagged by Q1 for potential poor fit. An internally developed software program, Q1Static.exe, was used to compute the Q1 item fit index.

Operational items flagged by Q1 that are not flagged by the classical item analyses and have reasonable IRT parameter estimates were not further reviewed. If they are also flagged by classical item analyses and/or have poor IRT parameter estimates (e.g., low a parameter), items 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. No such incidences occurred for Winter/Trimester 2008-2009 or Spring 2009.

4.2.a Calibration and IRT Fit Results

4.2.a.i Winter/Trimester 2008-2009

Algebra I. For the Winter/Trimester 2008-2009 Algebra I assessment, based on the calibration sample, the Z-statistics for most operational items were smaller than the critical Z-statistic. Operational item 25 exhibited marginally poor fit with a Z-statistic of 5.2. The Item Characteristic Curves (ICCs) were reasonable and examination of the classical statistics for these items were also within a reasonable range (item 25: P-value=0.73 and Pbis=0.41).

Algebra II. For the Winter/Trimester 2008-2009 Algebra II assessment, based on the calibration sample, the Z-statistics for most operational items were smaller than the critical Z-statistic. Operational items 1, 39 (linking item), 47 (linking item), 64, and 68 exhibited marginally poor fit with Z-statistics of 4.4, 13.8, 8.9, 4.6, and 5.2, respectively. The ICCs were reasonable and examination of the classical statistics for these items were also within a reasonable range (item 1: P-value=0.82 and Pbis=0.48; item 39: P-value=0.85 and Pbis=0.41; item 47: P-value=0.60 and Pbis=0.41; item 64: P-value=0.25 and Pbis=0.31; item 68: P-value=0.56 and Pbis=0.45).

Geometry. For the Winter/Trimester 2008-2009 Geometry assessment, based on the calibration sample, the Z-statistics for most operational items were smaller than the critical Z-statistic. Operational items 43, 49, 56, 57, 65 (linking item), and 73 exhibited marginally poor fit with Z-statistics of 4.6, 4.0, 4.3, 5.5, 6.0, 4.4, respectively. The ICCs were reasonable and examination of the classical statistics for these items were also within a reasonable range (item 43: P-value=0.91 and Pbis=0.35; item 49: P-value=0.74 and Pbis=0.45; item 56: P-value=0.65 and Pbis=0.49; item 57: P-value=0.65 and Pbis=0.48; item 65: P-value=0.65 and Pbis=0.50; item 73: P-value=0.42 and Pbis=0.46).

Biology I. For the Winter/Trimester 2008-2009 Biology I assessment, based on the calibration sample, the Z-statistics for most operational items were smaller than the critical Z-statistic. Operational items 12 and 24 (linking item) exhibited marginally poor fit with Z-statistics of 5.1 and 6.4, respectively. The ICCs were reasonable and examination of the classical statistics for these items were also within a reasonable range (item 12: P-value=0.74 and Pbis=0.47; item 24: P-value=0.95 and Pbis=0.28).

English II. For the Winter/Trimester 2008-2009 English II assessment, based on the calibration sample, the Z-statistics for most operational items were smaller than the critical Z-statistic. Operational items 8 and 59 (linking item) exhibited marginally poor fit with Z-statistics of 11.7 and 8.6, respectively. The ICCs were reasonable and examination of the classical statistics for these items were also within a reasonable range (item 8: P-value=0.91 and Pbis=0.30; item 59: P-value=0.65 and Pbis=0.32). The writing prompt or open-ended item was also flagged for poor item fit with a Z-statistic of 107.8. Examination of the Item Category Response Function (ICRF) indicated that the poor fit can be explained, partly, by the fact that less than the expected proportion of students obtained a score in certain categories; however, overall observed and expected proportion

of maximum score curves indicated that the fit was reasonable. The classical statistics were also within reasonable range (P-value=0.53 and Pbis=0.58).

English III. For the Winter/Trimester 2008-2009 English III assessment, based on the calibration sample, the Z-statistics for all operational multiple-choice items were smaller than the critical Z-statistic. The writing prompt or open-ended item was also flagged for poor item fit with a Z-statistic of 18.7. Examination of the Item Category Response Function (ICRF) indicated that the poor fit can be explained, partly, by the fact that less than the expected proportion of students obtained a score in certain categories; however, overall observed and expected proportion of maximum score curves indicated that the fit was reasonable. The classical statistics were also within reasonable range (P-value=0.64 and Pbis=0.55).

U.S. History. For the Winter/Trimester 2008-2009 U.S. History assessment, based on the calibration sample, the Z-statistics for all operational items were smaller than the critical Z-statistic. There were no U.S. History items flagged by the Q1 index.

No items were dropped from any of the Winter/Trimester 2008-2009 ACE EOI assessments for calibration, equating, or scoring as a result of the Q1 results.

4.2.a.ii Spring 2009

Algebra I. For the Spring 2009 Algebra I assessment, based on the calibration sample, the Z-statistics for all operational items were smaller than the critical Z-statistic. There were no Algebra I items flagged by the Q1 index.

Algebra II. For the Spring 2009 Algebra II assessment, based on the calibration sample, the Z-statistics for most operational items were smaller than the critical Z-statistic. Operational items 35 (linking item), 39 (linking item), and 49 exhibited marginally poor fit with Z-statistics of 100.3, 140.7, and 115.9, respectively. The ICCs were reasonable and examination of the classical statistics for these items were also within the reasonable range (item 35: P-value=0.90 and Pbis=0.30; item 39: P-value=0.81 and Pbis=0.41; item 49: P-value=0.85 and Pbis=0.37).

Geometry. For the Spring 2009 Geometry assessment, based on the calibration sample, the Z-statistics for all operational items were smaller than the critical Z-statistic. There were no Geometry items flagged by the Q1 index.

Biology I. For the Spring 2009 Biology I assessment, based on the calibration sample, the Z-statistics for all operational items were smaller than the critical Z-statistic. There were no Biology I items flagged by the Q1 index.

English II. For the Spring 2009 English II assessment, based on the calibration sample, the Z-statistics for most operational items were smaller than the critical Z-statistic. One operational multiple choice item, item 61 (linking item), exhibited marginally poor fit with a Z-statistic of 145.2. The ICC was reasonable and examination of the classical

statistics for this item were also within the reasonable range (item 61: P-value=0.67 and Pbis=0.31). The writing prompt or open-ended item was also flagged for poor item fit with a Z-statistic of 7219.1. Examination of the Category Response Function (CRF) indicated that the poor fit can be explained, partly, by the fact that a different than the expected proportion of students obtained a score in certain categories mostly at the lower and higher ability levels; however, overall observed and expected proportion of maximum score curves indicated that the fit was reasonable. The classical statistics were also within reasonable range (P-value=0.56 and Pbis=0.61).

English III. For the Spring 2009 English III assessment, based on the calibration sample, the Z-statistics for all operational multiple-choice items were smaller than the critical Z-statistic. The writing prompt or open-ended item was also flagged for poor item fit with a Z-statistic of 8799.8. Examination of the CRF indicated that the poor fit can be explained, partly, by the fact that the expected proportion of students were different from the obtained a score in a particularly category mostly at the lower and upper theta estimates; however, overall observed and expected proportion of maximum score curves indicated that the fit was reasonable. The classical statistics were also within reasonable range (P-value=0.74 and Pbis=0.61).

U.S. History. For the Spring 2009 U.S. History assessment, based on the calibration sample, the Z-statistics for all operational items were smaller than the critical Z-statistic. There were no U.S. History items flagged by the Q1 index.

No items were dropped from any of the Spring 2009 ACE EOI assessments for calibration, equating, or scoring as a result of the Q1 index.

Field Test Items. The field test items across all subjects were evaluated using the Q1 statistic to evaluate the extent the obtained proportions of item scores are close to the proportions that would be expected based on the estimated thetas and item parameters. Any field test items flagged by Q1 were included in the data review for review by contest specialists from Pearson and SDE (for more on data review, please see Section 3.4).

4.3 Calibration and Equating

The 3-PL model was used for calibration of Algebra I, Algebra II, Geometry, Biology I, and U.S. History because all of these areas consisted of only multiple choice items. Since English II and English III have multiple choice and constructed response items, a simultaneous calibration with the 3-PL and GPC models was implemented.

A common item non-equivalent groups design was used for all content areas to link the current test forms (i.e., Winter/Trimester 2008-2009 and Spring 2009) to the base years' scale. The common, or anchor, items were selected to be representative of the test content in terms item difficulties and the test blueprint. The anchor items are critical to obtaining results that are comparable from year to year. The Stocking and Lord (1983) procedure was used to equate the tests to the base year, which estimates the equating transformation

constants by minimizing the distance between the test characteristic curves of the common items.

Equating was conducted employing the Stocking and Lord (1983) procedure using publicly available software, STUIRT (Kim & Kolen, 2004). Prior to conducting the equating, anchor item stability checks were performed to eliminate the impact of item drift on equating.

4.4 Anchor Item Stability Evaluation Methods

Despite the careful selection of anchor items, it is plausible that the anchor items may 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 constants, we evaluated changes in the item parameters from the Spring 2008 operational administration to the Spring 2009 administration (only operational items could serve as year-to-year linking items). The process used in this evaluation is called an anchor item parameter stability check. Our approach is iterative and the procedures are outlined next.

The anchor stability check performed is an iterative approach and uses a procedure that is analogous to examining differential item functioning and is called the d^2 procedure. The steps taken for Algebra I, Algebra II, English III, and Geometry were as follows:

- 1) Use a theoretically weighted posterior theta distribution with 40 quadrature points.
- 2) Place the current linking item parameters on the baseline scale by computing Stocking & Lord (SL) constants using STUIRT and all (k) linking items.
- 3) Apply the SL linking constants to the current item parameters and compute the current raw to scale table. The results based on all k linking items will comprise the “original table”.
- 4) For each linking item, calculate the weighted sum of the squared deviation between the Item Characteristic Curves (ICC; d^2):
 - a) Apply the SL constants to the thetas associated with the standard normal theta distribution used to generate the SL constants.
 - b) For each anchor item calculate a weighted sum of the squared deviation between the ICCs (d^2) based on old (x) and new (y) parameters at each point in this theta distribution.

$$d_i^2 = \sum_k [P_{ix}(\theta_k) - P_{iy}(\theta_k)]^2 \cdot g(\theta_k) \quad (4)$$

- c) Review and sort the items in a descending (largest to smallest) fashion according to the d^2 estimate.
- d) From Step c) results in an items with the largest area at the top:
 - i) Drop the largest d^2 item from the linking set.
 - ii) Repeat steps 2 through 3c) using k-1 linking items.
- e) Terminate when either the number of linking items remaining is 20% or the raw to scale tables across iterations do not differ. The raw score to scale score table before the last iteration becomes the final table.

The anchor stability check implemented for Biology I, English II, and U.S. History (for Spring 2009 only; Winter/Trimester 2008-2009 followed the procedures outlined above) was slightly modified from the anchor stability checks used for Algebra I, Algebra II, Geometry, and English III, which has stopping criteria at 4d) based on stability of the raw score to scale score table at each of the cut score points. Since the cut score points were not available until after the Standard Setting and the item parameter estimates were required to be on the baseline operational metric, the stopping criteria was modified in 4e). The stopping criteria for Biology I, English II, and U.S. History for Spring 2009 were as follows:

- 4e) Terminate when either the number of linking items is 11 or there are no “large” d^2 values remaining. “Large” is defined by a d^2 value that is an outlier based on the original distribution of d^2 . Outlier is defined as falling outside the 95% confidence interval around the mean of the original d^2 distribution. The item with the largest d^2 value will be dropped and the anchor stability check re-ran and items can only be dropped iteratively, one at a time. The Lord and Stocking equating constants computed during the final step, when there are no more “large” d^2 values will be the constants used for equating the operational items to the baseline operational scale. This will leave a minimum of 11 items in the linking set.

Before removing any anchor item, the following additional characteristics were examined: 1) prior and current year p-values and point-biserials, 2) prior and current year IRT parameters, 3) prior and current year item sequence, 4) standard and objective/skill of the item, 5) impact on blueprint representation, 6) Passage ID/Title if the item is part of stimulus, and 7) content review of the actual item. Decisions about whether to keep or remove an item were evaluated on a per item basis. If an item (note, only one item can be removed at a time) was removed from the anchor set, the process (beginning at the equating step) was repeated until there were no further items to be removed (the raw score to scale score table has stabilized or the item is judged that it should be included in the equating set; for example, a portion of the blueprint is not represented if the item is removed).

4.4.a Anchor Items for Winter/Trimester 2008-2009 and Spring 2009

Table 4.1 presents the number and proportion of anchor items by subject for the Winter/Trimester 2008-2009 and Spring 2009 administrations. The anchor set was comprised of approximately 20 items or greater than 25% of all operational items and as seen in Table 4.1 varies by subject. 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.

Table 4.1. Number of anchor items per subject

Operational Test	Number of Items on Test	Number of Anchors	Percent of Test
Algebra I	55	16	29%
Algebra II	55	15	27%
Biology I	60	18	30%
English II	61	21	34%
English III	63	18	29%
Geometry	55	15	27%
U.S. History	60	20	33%

4.4.b Results of the Anchor Item Stability Check

Once the anchor set was finalized, the equating constants obtained from the final Stocking and Lord (1983) run were applied to the non-anchor operational items for computation of raw score to scale score tables. For Winter/Trimester 2008-2009, three items were removed from Algebra I and English II, two items from Algebra II and Geometry, one item from English III, and zero items from Biology I and U.S. History as a result of the anchor item stability check. For Spring 2009, there were two anchor items removed from Algebra I, zero items from Algebra II, one item from Geometry, Biology I, English II, and U.S. History, and zero items from English III. Any item removed from the anchor set still contributed to student scores.

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, M1 (multiplicative) and M2 (additive), place the true scores associated with each raw score point onto the reporting or operational scale using a straightforward linear transformation:

$$\text{Scale Score} = (\hat{\tau} \times M1) + M2 \quad (5)$$

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

The raw score to number-correct scales scores were generated from equated parameter estimates using a publicly available software program POLYEQUATE (Kolen, 2004). For a particular scale score, it is associated with a performance level on the assessment that describes the types of behaviors, knowledge, and skill, a student in this score level is likely to be able to do. For the ACE EOI assessments there are 3 cut scores that divide scores into 4 performance levels, Unsatisfactory, Limited Knowledge, Proficient/Satisfactory, and Advanced. The cut scores for each of the tests appears in Table 4.3. In addition, a Conditional Standard Error of Measurement (CSEM; please see section 7.3 for computation of CSEM) was computed for each of the raw score points. The resulting raw score to scale scores conversions, CSEMs, as well as the performance levels for Algebra I, Algebra II, Geometry, Biology I, English II, English III, and U.S.

History are shown in Tables 4.4 and 4.5 for Winter/Trimester 2008-2009 and Spring 2009, respectively.

Table 4.2. LOSS, HOSS, and Scaling Constants by Subject.

Subject	LOSS	HOSS	M1	M2
Algebra I	490	999	58.0000	723.8000
Algebra II	440	999	77.1164	692.2381
Geometry	440	999	75.51595	721.9844
English III	440	999	74.32896	736.1256
Biology*	440	999	76.49429	716.76173
English II*	440	999	84.80517	734.90335
US History*	440	999	77.92698	722.20515

*Note: These are the scaling constants after the June 2009 Standard Setting and State Board of Education approval of the phased-in cut scores.

Table 4.3. Performance Level Cut Scores by Content Area.

Subject	Cut Scores		
	Limited Knowledge	Proficient	Advanced
Algebra I	639	684	746
Algebra II	651	696	774
Geometry	635	695	774
English III	649	695	795
Biology I*	627	691	775
English II*	588	693	797
U.S. History*	603	689	747

*Note: These are cut scores after the June 2009 Standard Setting and State Board of Education approval.

Table 4.4. Raw Score to Scale Score Conversion Tables for Winter/Trimester 2008-2009

Raw Score	Algebra I			Biology I			U.S. History			English II		
	Scale Score	CSEM	Perf. Level	Scale Score	CSEM	Perf. Level	Scale Score	CSEM	Perf. Level	Scale Score	CSEM	Perf. Level
0	490	55	1	440	41	1	440	54	1	440	36	1
1	490	55	1	440	41	1	440	54	1	440	36	1
2	490	55	1	440	41	1	440	54	1	440	36	1
3	490	55	1	440	41	1	440	54	1	440	36	1
4	490	55	1	440	41	1	440	54	1	440	36	1
5	490	55	1	440	41	1	440	54	1	440	36	1
6	490	55	1	440	41	1	440	54	1	440	36	1
7	490	55	1	440	41	1	440	54	1	440	36	1
8	490	55	1	440	41	1	440	54	1	440	36	1
9	490	55	1	440	41	1	440	54	1	440	36	1
10	490	55	1	440	41	1	440	54	1	440	36	1
11	507	55	1	440	41	1	440	54	1	440	36	1
12	567	59	1	440	41	1	440	54	1	440	36	1
13	593	59	1	440	41	1	440	54	1	440	36	1
14	610	57	1	461	44	1	484	58	1	440	36	1
15	623	52	1	488	48	1	519	61	1	440	36	1
16	639	46	2	509	51	1	544	62	1	443	37	1
17	643	40	2	527	51	1	562	60	1	470	42	1
18	651	35	2	542	50	1	577	56	1	490	45	1
19	658	30	2	557	48	1	590	52	1	508	46	1
20	664	26	2	570	46	1	603	47	2	523	46	1
21	670	23	2	582	43	1	611	42	2	536	45	1
22	676	21	2	593	40	1	620	38	2	548	44	1
23	684	19	3	604	38	1	629	34	2	559	41	1
24	686	18	3	614	36	1	637	31	2	569	39	1
25	690	17	3	627	34	2	644	29	2	578	36	1
26	695	16	3	632	32	2	651	27	2	588	34	2
27	699	15	3	641	31	2	658	26	2	596	33	2
28	703	15	3	649	29	2	665	24	2	604	31	2

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

Table 4.4 cont. Raw Score to Scale Score Conversion Tables for Winter/Trimester 2008-2009

Raw Score	Algebra I			Biology I			U.S. History			English II		
	Scale Score	CSEM	Perf. Level	Scale Score	CSEM	Perf. Level	Scale Score	CSEM	Perf. Level	Scale Score	CSEM	Perf. Level
29	707	15	3	657	28	2	671	23	2	612	30	2
30	711	14	3	665	27	2	677	22	2	619	28	2
31	715	14	3	672	26	2	683	22	2	627	27	2
32	719	14	3	680	26	2	689	21	3	634	27	2
33	723	14	3	691	25	3	694	20	3	641	26	2
34	727	14	3	694	24	3	700	20	3	648	25	2
35	731	13	3	700	24	3	705	19	3	654	24	2
36	735	13	3	707	23	3	711	19	3	661	24	2
37	739	13	3	714	23	3	716	19	3	667	23	2
38	746	13	4	721	23	3	721	19	3	673	23	2
39	747	13	4	727	22	3	727	19	3	680	23	2
40	751	13	4	734	22	3	732	19	3	686	22	2
41	755	14	4	740	22	3	737	19	3	693	22	3
42	760	14	4	747	22	3	747	19	4	698	22	3
43	764	14	4	754	22	3	749	19	4	704	22	3
44	769	15	4	761	22	3	754	19	4	710	22	3
45	775	15	4	768	22	3	760	19	4	717	22	3
46	780	16	4	775	22	4	767	20	4	723	22	3
47	786	17	4	783	23	4	773	20	4	729	22	3
48	793	18	4	790	23	4	780	21	4	736	22	3
49	800	21	4	798	24	4	787	22	4	743	23	3
50	809	24	4	807	25	4	795	22	4	750	23	3
51	819	30	4	816	26	4	803	24	4	757	24	3
52	833	40	4	825	27	4	812	25	4	765	25	3
53	851	53	4	836	29	4	822	27	4	773	26	3
54	884	65	4	848	31	4	833	29	4	782	27	3
55	999	34	4	861	35	4	846	33	4	797	28	4
56				877	39	4	861	37	4	802	30	4

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

Table 4.4 cont. Raw Score to Scale Score Conversion Tables for Winter/Trimester 2008-2009

Raw Score	Algebra I			Biology I			U.S. History			English II		
	Scale Score	CSEM	Perf. Level	Scale Score	CSEM	Perf. Level	Scale Score	CSEM	Perf. Level	Scale Score	CSEM	Perf. Level
57				898	42	4	881	41	4	813	32	4
58				926	42	4	907	44	4	825	34	4
59				974	34	4	950	40	4	839	37	4
60				999	28	4	999	31	4	855	41	4
61										874	44	4
62										898	47	4
63										927	46	4
64										969	39	4
65										999	32	4
66										999	32	4

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

Table 4.4 cont. Raw Score to Scale Score Conversion Tables for Winter/Trimester 2008-2009

Raw Score	Algebra II			Geometry			English III		
	Scale Score	CSEM	Perf. Level	Scale Score	CSEM	Perf. Level	Scale Score	CSEM	Perf. Level
0	440	60	1	440	61	1	440	53	1
1	440	60	1	440	61	1	440	53	1
2	440	60	1	440	61	1	440	53	1
3	440	60	1	440	61	1	440	53	1
4	440	60	1	440	61	1	440	53	1
5	440	60	1	440	61	1	440	53	1
6	440	60	1	440	61	1	440	53	1
7	440	60	1	440	61	1	440	53	1
8	440	60	1	440	61	1	440	53	1
9	440	60	1	440	61	1	440	53	1
10	440	60	1	440	61	1	440	53	1
11	471	63	1	440	61	1	440	53	1
12	523	67	1	511	67	1	440	53	1
13	555	68	1	545	69	1	440	53	1
14	577	66	1	569	68	1	487	58	1
15	595	61	1	587	63	1	522	61	1
16	609	55	1	602	58	1	546	62	1
17	622	49	1	614	51	1	564	60	1
18	633	43	1	635	45	2	578	57	1
19	643	38	1	636	39	2	590	53	1
20	651	34	2	645	35	2	601	49	1
21	660	30	2	654	31	2	611	44	1
22	667	28	2	662	29	2	619	40	1
23	674	26	2	669	26	2	628	37	1
24	681	24	2	676	25	2	636	34	1
25	688	23	2	683	23	2	649	31	2
26	696	22	3	689	22	2	650	29	2
27	700	21	3	695	21	3	657	28	2
28	706	21	3	701	20	3	663	26	2

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

Table 4.4 cont. Raw Score to Scale Score Conversion Tables for Winter/Trimester 2008-2009

Raw Score	Algebra II			Geometry			English III		
	Scale Score	CSEM	Perf. Level	Scale Score	CSEM	Perf. Level	Scale Score	CSEM	Perf. Level
29	712	20	3	707	20	3	670	25	2
30	717	19	3	712	19	3	676	24	2
31	723	19	3	718	19	3	682	23	2
32	728	19	3	723	18	3	687	23	2
33	734	19	3	728	18	3	695	22	3
34	739	18	3	734	18	3	698	21	3
35	745	18	3	739	17	3	704	21	3
36	750	18	3	744	17	3	709	21	3
37	756	18	3	749	17	3	714	20	3
38	762	18	3	754	17	3	719	20	3
39	767	18	3	760	17	3	724	20	3
40	774	19	4	765	17	3	729	19	3
41	779	19	4	774	17	4	734	19	3
42	786	19	4	777	18	4	739	19	3
43	792	19	4	783	18	4	744	19	3
44	799	20	4	789	18	4	749	19	3
45	806	20	4	796	18	4	754	18	3
46	814	21	4	803	19	4	759	18	3
47	822	22	4	810	20	4	764	18	3
48	831	24	4	818	21	4	769	18	3
49	841	27	4	827	23	4	774	18	3
50	852	30	4	837	26	4	779	18	3
51	866	34	4	848	31	4	784	19	3
52	883	39	4	863	38	4	795	19	4
53	909	43	4	883	46	4	796	19	4
54	954	39	4	918	50	4	801	19	4
55	999	30	4	999	35	4	807	20	4
56							813	20	4

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

Table 4.4 cont. Raw Score to Scale Score Conversion Tables for Winter/Trimester 2008-2009

Raw Score	Algebra II			Geometry			English III		
	Scale Score	CSEM	Perf. Level	Scale Score	CSEM	Perf. Level	Scale Score	CSEM	Perf. Level
57							820	21	4
58							826	22	4
59							834	23	4
60							841	24	4
61							849	25	4
62							858	27	4
63							868	28	4
64							879	31	4
65							891	33	4
66							905	36	4
67							921	38	4
68							940	38	4
69							965	34	4
70							999	26	4
71							999	26	4
72							999	26	4

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

Table 4.5. Raw Score to Scale Score Conversion Tables for Spring 2009

Raw Score	Algebra I			Biology I			U.S. History			English II		
	Scale Score	CSEM	Perf. Level	Scale Score	CSEM	Perf. Level	Scale Score	CSEM	Perf. Level	Scale Score	CSEM	Perf. Level
0	490	46	1	440	40	1	440	46	1	440	32	1
1	490	46	1	440	40	1	440	46	1	440	32	1
2	490	46	1	440	40	1	440	46	1	440	32	1
3	490	46	1	440	40	1	440	46	1	440	32	1
4	490	46	1	440	40	1	440	46	1	440	32	1
5	490	46	1	440	40	1	440	46	1	440	32	1
6	490	46	1	440	40	1	440	46	1	440	32	1
7	490	46	1	440	40	1	440	46	1	440	32	1
8	490	46	1	440	40	1	440	46	1	440	32	1
9	490	46	1	440	40	1	440	46	1	440	32	1
10	490	46	1	440	40	1	440	46	1	440	32	1
11	490	46	1	440	40	1	440	46	1	440	32	1
12	538	51	1	440	40	1	440	46	1	440	32	1
13	566	53	1	466	44	1	440	46	1	440	32	1
14	585	53	1	491	48	1	452	48	1	440	32	1
15	600	50	1	512	50	1	486	53	1	444	33	1
16	613	47	1	529	50	1	512	56	1	466	38	1
17	624	42	1	544	49	1	533	57	1	485	41	1
18	639	38	2	558	47	1	551	56	1	501	42	1
19	642	33	2	570	44	1	566	53	1	515	43	1
20	650	30	2	582	41	1	580	50	1	527	42	1
21	657	27	2	592	39	1	592	46	1	539	41	1
22	664	25	2	602	36	1	603	42	2	550	40	1
23	671	23	2	612	34	1	613	38	2	560	38	1
24	677	21	2	627	32	2	622	35	2	569	36	1
25	684	20	3	629	30	2	631	33	2	578	34	1
26	688	19	3	637	29	2	639	30	2	588	32	2
27	693	18	3	645	28	2	647	29	2	595	31	2
28	698	17	3	652	27	2	654	27	2	603	30	2
29	703	17	3	659	26	2	661	26	2	610	29	2

Table 4.5 cont. Raw Score to Scale Score Conversion Tables for Spring 2009

Raw Score	Algebra I			Biology I			U.S. History			English II		
	Scale Score	CSEM	Perf. Level	Scale Score	CSEM	Perf. Level	Scale Score	CSEM	Perf. Level	Scale Score	CSEM	Perf. Level
30	708	16	3	666	25	2	668	25	2	618	28	2
31	712	16	3	673	24	2	674	24	2	625	27	2
32	717	15	3	680	23	2	681	23	2	632	26	2
33	721	15	3	691	23	3	689	22	3	639	25	2
34	726	14	3	692	22	3	693	22	3	645	25	2
35	730	14	3	699	22	3	699	21	3	652	24	2
36	734	14	3	705	21	3	705	21	3	659	24	2
37	738	14	3	711	21	3	711	20	3	665	23	2
38	746	14	4	717	21	3	716	20	3	671	23	2
39	747	14	4	723	20	3	722	20	3	678	23	2
40	751	14	4	729	20	3	728	20	3	684	23	2
41	755	14	4	735	20	3	734	20	3	693	22	3
42	760	14	4	741	20	3	740	20	3	696	22	3
43	765	14	4	747	20	3	747	20	4	703	22	3
44	770	15	4	753	20	3	752	20	4	709	22	3
45	775	15	4	759	20	3	758	21	4	716	22	3
46	780	16	4	766	20	3	765	21	4	722	23	3
47	787	17	4	775	20	4	772	22	4	729	23	3
48	793	19	4	779	20	4	779	22	4	736	23	3
49	801	21	4	786	20	4	787	23	4	743	23	3
50	810	25	4	793	21	4	795	24	4	750	24	3
51	820	31	4	800	22	4	804	25	4	758	24	3
52	834	40	4	809	23	4	814	27	4	766	25	3
53	852	53	4	817	25	4	824	29	4	774	26	3
54	885	64	4	827	27	4	836	31	4	783	27	3
55	999	38	4	839	30	4	850	35	4	797	29	4
56				852	35	4	866	40	4	803	31	4
57				869	40	4	886	44	4	815	33	4
58				893	46	4	914	46	4	828	36	4
59				933	46	4	967	38	4	843	39	4

Table 4.5 cont. Raw Score to Scale Score Conversion Tables for Spring 2009

Raw Score	Algebra I			Biology I			U.S. History			English II		
	Scale Score	CSEM	Perf. Level	Scale Score	CSEM	Perf. Level	Scale Score	CSEM	Perf. Level	Scale Score	CSEM	Perf. Level
60				999	34	4	999	31	4	860	43	4
61										881	47	4
62										906	49	4
63										940	46	4
64										986	36	4
65										999	33	4
66										999	33	4

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

Table 4.5 cont. Raw Score to Scale Score Conversion Tables for Spring 2009

Raw Score	Algebra II			Geometry			English III		
	Scale Score	CSEM	Perf. Level	Scale Score	CSEM	Perf. Level	Scale Score	CSEM	Perf. Level
0	440	69	1	440	61	1	440	44	1
1	440	69	1	440	61	1	440	44	1
2	440	69	1	440	61	1	440	44	1
3	440	69	1	440	61	1	440	44	1
4	440	69	1	440	61	1	440	44	1
5	440	69	1	440	61	1	440	44	1
6	440	69	1	440	61	1	440	44	1
7	440	69	1	440	61	1	440	44	1
8	440	69	1	440	61	1	440	44	1
9	440	69	1	440	61	1	440	44	1
10	440	69	1	440	61	1	440	44	1
11	440	69	1	440	61	1	440	44	1
12	475	71	1	508	66	1	451	45	1
13	535	76	1	544	69	1	492	50	1
14	570	77	1	567	68	1	517	53	1
15	595	74	1	585	64	1	535	53	1
16	614	68	1	601	58	1	550	52	1
17	630	60	1	614	52	1	562	49	1
18	651	53	2	635	45	2	573	45	1
19	653	46	2	636	40	2	583	42	1
20	663	39	2	646	35	2	592	38	1
21	672	34	2	655	32	2	600	35	1
22	679	30	2	663	29	2	608	32	1
23	686	27	2	671	26	2	615	30	1
24	696	25	3	678	25	2	623	28	1
25	699	23	3	685	23	2	629	27	1
26	705	22	3	695	22	3	636	26	1
27	711	20	3	697	21	3	642	25	1
28	716	20	3	703	20	3	649	24	2

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

Table 4.5 cont. Raw Score to Scale Score Conversion Tables for Spring 2009

Raw Score	Algebra II			Geometry			English III		
	Scale Score	CSEM	Perf. Level	Scale Score	CSEM	Perf. Level	Scale Score	CSEM	Perf. Level
29	722	19	3	709	19	3	654	23	2
30	727	18	3	714	18	3	660	22	2
31	732	18	3	719	17	3	665	22	2
32	737	17	3	724	17	3	671	21	2
33	742	17	3	729	16	3	676	21	2
34	747	17	3	734	16	3	682	21	2
35	752	17	3	739	16	3	687	20	2
36	757	16	3	743	16	3	695	20	3
37	762	16	3	748	15	3	698	20	3
38	767	16	3	753	15	3	703	19	3
39	774	16	4	758	15	3	708	19	3
40	777	16	4	762	15	3	713	19	3
41	782	16	4	767	15	3	718	18	3
42	788	17	4	774	15	4	723	18	3
43	793	17	4	778	16	4	728	18	3
44	799	18	4	783	16	4	733	18	3
45	806	18	4	789	16	4	737	17	3
46	812	19	4	795	17	4	742	17	3
47	819	20	4	801	18	4	747	17	3
48	827	22	4	808	19	4	752	17	3
49	836	24	4	816	21	4	756	17	3
50	846	27	4	825	24	4	761	17	3
51	858	31	4	835	29	4	766	17	3
52	873	37	4	848	37	4	771	17	3
53	893	43	4	866	49	4	776	18	3
54	929	46	4	896	58	4	781	18	3
55	999	32	4	999	30	4	787	18	3
56							795	19	4

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

Table 4.5 cont. Raw Score to Scale Score Conversion Tables for Spring 2009

Raw Score	Algebra II			Geometry			English III		
	Scale Score	CSEM	Perf. Level	Scale Score	CSEM	Perf. Level	Scale Score	CSEM	Perf. Level
57							798	19	4
58							805	20	4
59							811	21	4
60							818	22	4
61							826	23	4
62							834	24	4
63							843	26	4
64							853	28	4
65							864	31	4
66							876	34	4
67							891	37	4
68							909	40	4
69							933	40	4
70							968	35	4
71							999	28	4
72							999	28	4

Note: CSEM = Conditional Standard Error of Measure; Perf. Level = Performance Level; 1 = Unsatisfactory, 2 = Limited Knowledge, 3 = Proficient, 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 an impact on how to explain the cut scores used to classify students into different performance levels. For example, some students may have a true performance level greater than a cut-score. However, due to random variations (measurement error), the observed test score may be below the cut score. As a result, the students may be classified as having a lower performance level. As discussed in Section 7.4 on SEM, a student's observed score is most likely to fall into 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, also known as 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 in order 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 3.7. 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, proficient students who were classified as proficient) and the proportion of all the students classified into that level (total proportion of students who were considered

proficient). Accuracy conditional on performance level is computed in a similar manner. The only difference is that in the consistency table where both row and column marginal sums are the same, in the accuracy table the sum based on estimated status is used as a 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 Proficient and Advanced to assess decisions at the Proficient 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 for Winter/Trimester 2008-2009 and Table 5.2 for Spring 2009 present the overall accuracy and consistency indices and accuracy and consistency conditioned on performance level for all of the ACE EOI tests. There are four performance levels on the ACE EOI tests: Unsatisfactory, Limited Knowledge, Proficient, and Advanced. Table 5.3 for Winter/Trimester 2008-2009 and Table 5.4 for Spring 2009 provide the accuracy and consistency estimates by cut-score for all subjects.

Table 5.1. Estimates of Accuracy and Consistency of Performance Classification for Winter/Trimester 2008-2009.

Winter/Trimester 2008-2009	Accuracy	Consistency	Kappa	False Positives	False Negatives
Algebra I	0.77	0.72	0.60	0.13	0.10
Algebra II	0.79	0.72	0.61	0.13	0.08
Biology I	0.76	0.71	0.59	0.08	0.15
English II	0.78	0.72	0.57	0.14	0.08
English III	0.80	0.74	0.60	0.13	0.07
Geometry	0.79	0.74	0.62	0.08	0.13
U.S. History	0.80	0.73	0.61	0.12	0.08

Table 5.2. Estimates of Accuracy and Consistency of Performance Classification for Spring 2009.

Spring 2009	Accuracy	Consistency	Kappa	False Positives	False Negatives
Algebra I	0.79	0.73	0.60	0.09	0.12
Algebra II	0.77	0.69	0.58	0.15	0.08
Biology I	0.78	0.72	0.59	0.11	0.11
English II	0.80	0.74	0.59	0.07	0.14
English III	0.81	0.75	0.62	0.13	0.06
Geometry	0.79	0.75	0.63	0.14	0.08
U.S. History	0.79	0.72	0.60	0.11	0.09

As shown in Tables 5.1 and 5.2 overall accuracy indices range between 76 and 80 percent for Winter/Trimester 2008-2009 and 77 and 81 percent for Spring 2009 and overall

consistency ranging between 71 percent and 74 percent for Winter/Trimester 2008-2009 and 69 and 75 percent for Spring 2009. Kappa coefficients range from 0.57 and 0.62 for Winter/Trimester 2008-2009 and 0.58 and 0.63 for Spring 2009. The false positive and negatives rates also appearing in Table 5.1 and 5.2 for Winter/Trimester 2008-2009 and Spring 2009, respectively. The rate of false positives for the Winter/Trimester range from 8 percent to 13 percent, which were similar to the Spring 2009 that ranged from 7 percent to 15 percent. The false negative rates were also similar across administration ranging from 7 to 15 percent for Winter/Trimester and 6 to 14 percent for Spring 2009.

Tables 5.3 and 5.4 provide the accuracy and consistency and false positive and false negative rates by cut-score for Winter/Trimester 2008-2009 and Spring 2009, respectively. The data in these tables reveals that the level of agreement for both accuracy and consistency is above 85 percent in all cases, with most above 90 percent. In general, the high rates of accuracy and consistency support the cut decisions made using these assessments. Similar to Tables 5.1 and 5.2, the false positive and false negative rates were comparable for the Winter/Trimester 2008-2009 and Spring 2009 administrations and are quite low.

The importance of the dichotomous categorization is particularly notable when they map onto pass/fail decisions for the assessments. For the EOI tests, the U+L/P+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) students' observed performance is accurately reflective of their true ability (i.e., passed and should have passed); 2) students' true ability is below the standard, but they score above the standard (false positives); and 3) students' true ability is above the standard, but they score below the standard (false negatives). In examining Tables 5.3 and 5.4, in Winter/Trimester 2008-2009 Algebra I, for example, 92 percent of students are correctly classified as pass or fail based on their performance (scenario 1), 6 percent passed but their true performance is below the standard (scenario 2), and 3 percent failed although their true performance is above the standard (scenario 3). Overall, the accuracy rates for accurate classification are above 90% for the Winter/Trimester and Spring administrations for all subjects – students are appropriately (more than 90% of the time) categorized into pass/fail classifications based on their true ability using their observed score (raw score) as their classification score.

Table 5.3. Accuracy and Consistency estimates by cut-score: False positives and false negatives rates for Winter/Trimester 2008-2009.

Winter/ Trimester 2008-2009	Accuracy			Consistency			False Positives			False Negatives		
	U	U+L	U+L+P	U	U+L	U+L+P	U	U+L	U+L+P	U	U+L	U+L+P
	/	/	/	/	/	/	/	/	/	/	/	/
	L+P+A	P+A	A	L+P+A	P+A	A	L+P+A	P+A	A	L+P+A	P+A	A
Algebra I	0.93	0.92	0.92	0.92	0.89	0.90	0.01	0.06	0.06	0.06	0.02	0.01
Algebra II	0.94	0.92	0.92	0.92	0.89	0.90	0.04	0.04	0.05	0.02	0.04	0.03
Biology I	0.95	0.92	0.89	0.93	0.89	0.89	0.04	0.03	0.01	0.01	0.05	0.09
English II	0.98	0.91	0.89	0.97	0.90	0.85	0.01	0.07	0.05	0.01	0.02	0.06
English III	0.95	0.93	0.92	0.94	0.90	0.89	0.04	0.04	0.05	0.01	0.03	0.03
Geometry	0.95	0.91	0.92	0.94	0.90	0.89	0.04	0.01	0.03	0.01	0.07	0.05
U.S. History	0.96	0.93	0.92	0.94	0.90	0.89	0.02	0.04	0.06	0.02	0.04	0.02

Note: U=Unsatisfactory; L = Limited Knowledge; P = Proficient; and A = Advanced.

Note: U / L+P+A = Unsatisfactory divided by Limited Knowledge plus Proficient plus Advanced; U+L / P+A = Unsatisfactory plus Limited Knowledge divided by Proficient plus Advanced; U+L+P / A = Unsatisfactory plus Limited Knowledge plus Proficient divided by Advanced.

Table 5.4. Accuracy and Consistency estimates by cut-score: False positives and false negatives rates for Spring 2009.

Spring 2009	Accuracy			Consistency			False Positives			False Negatives		
	U	U+L	U+L+P	U	U+L	U+L+P	U	U+L	U+L+P	U	U+L	U+L+P
	/	/	/	/	/	/	/	/	/	/	/	/
	L+P+A	P+A	A	L+P+A	P+A	A	L+P+A	P+A	A	L+P+A	P+A	A
Algebra I	0.96	0.92	0.91	0.94	0.90	0.88	0.01	0.01	0.07	0.03	0.07	0.02
Algebra II	0.91	0.91	0.94	0.88	0.88	0.92	0.07	0.05	0.04	0.03	0.04	0.02
Biology I	0.95	0.91	0.91	0.94	0.89	0.89	0.03	0.02	0.07	0.02	0.07	0.02
English II	0.98	0.93	0.88	0.98	0.90	0.86	0.01	0.03	0.03	0.01	0.04	0.09
English III	0.96	0.94	0.91	0.94	0.91	0.89	0.03	0.03	0.07	0.02	0.03	0.01
Geometry	0.95	0.92	0.92	0.93	0.90	0.91	0.01	0.06	0.07	0.05	0.02	0.01
U.S. History	0.96	0.92	0.91	0.95	0.89	0.88	0.02	0.04	0.06	0.02	0.04	0.03

Note: U=Unsatisfactory; L = Limited Knowledge; P = Proficient; and A = Advanced.

Note: U / L+P+A = Unsatisfactory divided by Limited Knowledge plus Proficient plus Advanced; U+L / P+A = Unsatisfactory plus Limited Knowledge divided by Proficient plus Advanced; U+L+P / A = Unsatisfactory plus Limited Knowledge plus Proficient divided by Advanced.

Section 6

Standard Setting

6.1 Overview and Standard Setting Process

Committees of Oklahoma educators convened June 1 through June 4, 2009, in Oklahoma City, Oklahoma, to set standards for the Achieving Classroom Excellence (ACE) Biology I, English II, and U.S. History assessments. A total of 75 educators participated for three or four days to recommend cut scores. The item mapping procedure was applied to set the standards. The outcomes of the committee meetings are described in this summary and more detailed information is provided in a Standard Setting Technical Report (please see “Oklahoma State Testing Program Standard Setting for ACE Biology I, English II, and U.S. History”).

On the afternoon of Monday, June 1, prior to the standard setting conference, training was held for all table leaders. During this training the table leaders were introduced to the standard setting facilitators, briefed on their role in the standard setting process, and received advance instruction on the item mapping process.

The standard setting conference began on Tuesday, June 2. The morning of Tuesday, June 2, was devoted to introductions of the SDE and Pearson staff, a description of standard setting process, a description of the ACE Biology I, English II, and U.S. History tests, and a general overview of the agenda for the meeting. For this stage of the conference, all panelists met together in one large room.

Following the midmorning break, the committees were dispersed into subject-specific conference rooms and took the appropriate ACE test in order to gain familiarity with the content represented on the test. Once the committee members had completed the test they were asked to review the Performance Level Descriptors (PLDs) for their subject in order to obtain a clear and concrete understanding of the performance levels and the differences between adjacent levels. Committee members were asked to identify general themes of the performance levels and behavioral anchors that describe “threshold students” – those students who could be described as minimally competent at a particular performance level.

The item mapping procedure was the judgmental process used in this standard setting. In this procedure, panelists were instructed to identify the last item in an ordered item book that a threshold student at a given level would have a response probability of answering correctly more often than not. After additional small-group training on the item mapping procedure the three committees began the standard setting process in the late afternoon of Tuesday, June 2. These committees were to set Limited Knowledge, Proficient, and Advanced performance level cuts. The standard setting process consisted of three rounds of judgments.

Round 1. In Round 1, panelists were asked to move through the item ordered booklet and indicate their independent recommendation for three cuts based on judgments about the performance of threshold students at each of the three levels. Panelists were provided with feedback between each round. The feedback was intended to inform the panelist's decisions, but not to dictate their ratings. Following Round 1, panelists met in small groups of 5 or 6 and were provided the cut-scores for each panelist and the mean and median cut-score at each level for that table. In reviewing the cut-score report panelists were asked to think about the following:

- How similar are their cut-scores are to that of the group (i.e., is a given panelist more lenient or stringent than the other panelists)?
- If so, why is this the case?
- Do panelists have different conceptualization of these borderline students?

Panelists were informed that it was not necessary for them to come to consensus on their cut-score judgments, but they should discuss differences to get a feel for why differences exist. Next, panelists were given the mean and median cut-scores for the committee (across tables). The facilitator lead the discussion with all tables combined. The facilitator noted the differences and similarities across tables but reminded the panelists that consensus was not required.

Round 2. In Round 2, based on the discussion at the end of Round 1 and judgments about the items, panelists were asked reevaluate their recommended cut from Round 1, and move if desired. Following Round 2, panelists received the same feedback for each table and for the full committee that was provided following Round 1. Additionally, panelists were provided a graphical display of the impact (distribution of students at each performance level) if using the committee's median cut-score. The impact data graphic representation provided panelists with information on what percentages of students are at each performance level for the populations of interest (all students, African-American/White/Hispanic/Native American, female/male). Panelists were given time to discuss, within the big group, the appropriateness of the committee level cut-scores given the proportion of students that would fall in each level.

Round 3. In Round 3, based on further discussion and review of the impact data, panelists were asked to make a final review of their cut-score recommendation and make adjustments if desired. Following Round 3, panelists were shown the cut-scores they were recommending based on this round of ratings, given the mean and median cut-scores for the committee (across tables), and provided a graphical display of the impact of using the median cut score for all students. Panelists were informed that this was their final cut-score judgments that would be sent to the State Board of Education for approval.

6.2 Results – Biology I, English II, and U.S. History Cut Scores

The Biology I, English II, and U.S. History ordered item books were comprised of 82, 88 and 87 ordered score points, respectively. Table 6.1 summarizes the cut scores after the Final Round of ratings for Biology I, English II, and U.S. History. These are the recommendations from the committees based on item location in the ordered item book.

The scale score cuts (and raw scores) associated with these recommendations and the percentage of students in the Unsatisfactory, Limited Knowledge, Proficient, and Advanced performance levels based upon these cuts are presented in Table 6.2. The impact resulting from the final cut-score recommendations appears in Figure 6.1.

Table 6.1. OIB Cut Scores after the Final Round of Rating by Subject.

Subject	Score	Limited Knowledge		
		Proficient	Advanced	
Biology I	Mean	15.41	40.52	72.96
	Median	15.00	40.00	74.00
English II	Mean	14.00	49.22	77.43
	Median	14.00	50.00	78.00
U.S. History	Mean	13.16	36.84	69.72
	Median	13.00	36.00	70.00

Table 6.2. Raw Score and Scale Score Cut Scores After the Final Round of Rating

Subject	% Unsatisfactory	Limited Knowledge		% Limited Knowledge	Proficient		% Proficient	Advanced		% Advanced
		RS*	SS		RS*	SS		RS*	SS	
Biology I	9	24	594	19	33	659	42	47	745	30
English II	3	26	543	19	41	646	49	55	749	29
U.S. History	7	22	579	23	33	663	29	43	722	41

Note: Biology I and U.S. History have a total possible score of 60 points and English II has a total possible score of 66 points; RS = Raw Score; SS = Scale Score.

* These are the scale scores associated with cut scores from the committee median recommendation (prior to approval by the State Board of Education and to final scaling).

Figure 6.1 shows the percentage of students in each performance level using the cut scores after the Final Round of rating for Biology I, English II, and U.S. History.

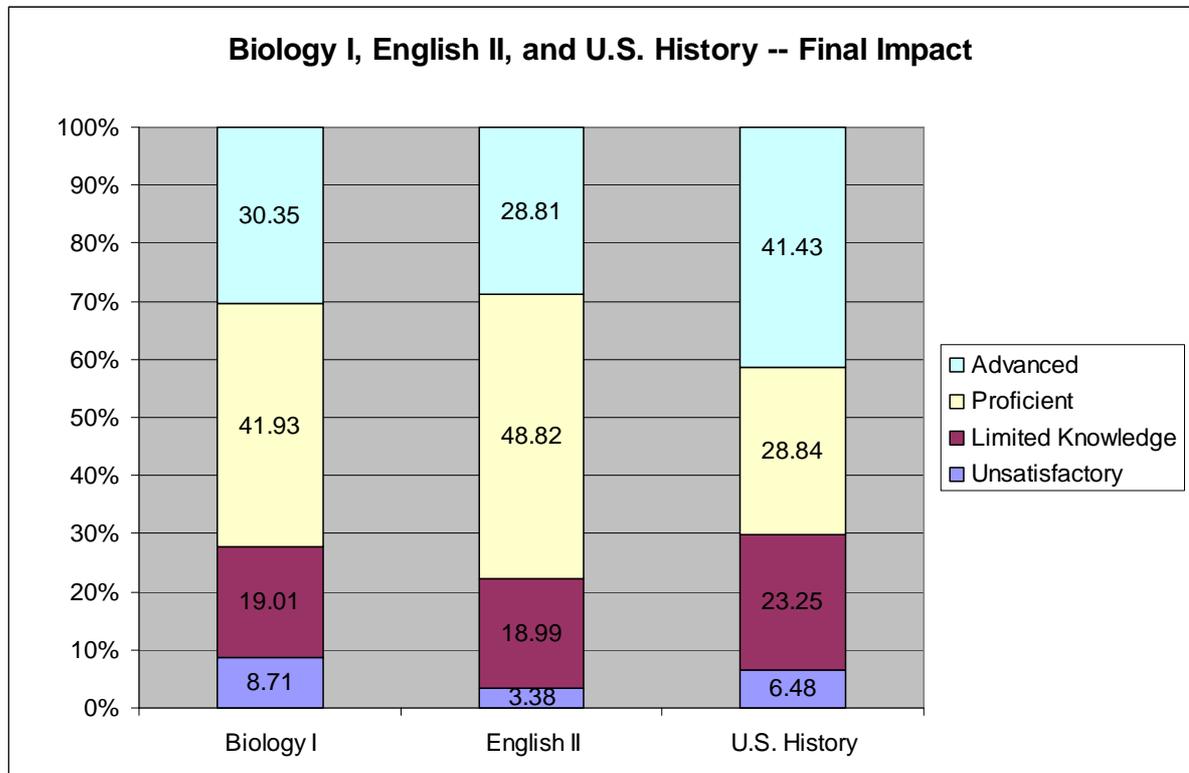


Figure 6.1. The percentage of students in each performance level using the cut scores after the Final Round of rating for Biology I, English II, and U.S. History.

Section 7

Summary Statistics

7.1 Means and Standard Deviations

The summary descriptive statistics (mean, median, and standard deviation) of the scale scores for Winter/Trimester 2008-2009 and Spring 2009 appears in Table 7.1 and 7.2, respectively. The scales scores presented exclude invalid student cases and second time testers.

Table 7.1. Descriptive Statistics of the Scale Scores for Winter/Trimester 2008-2009

Winter/Trimester 2008-2009	Total				Female				Male			
	N	Mean	SD	Med.	N	Mean	SD	Med.	N	Mean	SD	Med.
Algebra I	1,499	704.8	64.6	711	766	709.3	61.0	715	733	700.0	68.0	707
Algebra II	1,915	724.1	87.2	734	987	721.6	83.5	728	928	726.7	90.8	737
Biology I	2,073	728.8	86.1	734	1,031	722.8	79.9	727	1,042	734.6	91.5	740
English II	2,628	740.0	84.6	743	1,308	747.2	82.2	750	1,320	733.0	86.4	736
English III	2,783	740.0	76.0	744	1,381	745.9	71.3	749	1,390	735.0	79.4	739
Geometry	1,901	732.8	76.1	739	975	732.2	74.5	734	926	733.4	77.8	739
U.S. History	2,600	719.1	84.6	724	1,312	704.2	82.9	711	1,288	734.3	83.6	737

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

Table 7.2. Descriptive Statistics of the Scale Scores for Spring 2009

Spring 2009	Total				Female				Male			
	N	Mean	SD	Med.	N	Mean	SD	Med.	N	Mean	SD	Med.
Algebra I	35,736	725.5	58.6	730	17,919	725.2	57.4	730	17,817	725.7	59.7	730
Algebra II	29,644	710.5	89.9	716	15,289	711.5	86.9	716	14,355	709.4	92.9	716
Biology I	35,347	726.6	80.8	729	17,761	722.7	77.1	723	17,586	730.5	84.2	735
English II	34,823	743.9	82.8	750	17,686	751.9	80.1	750	17,137	735.8	84.7	743
English III	34,842	745.0	77.0	752	17,511	754.7	74.1	756	17,331	735.2	78.6	742
Geometry	34,224	733.1	77.7	739	17,092	733.4	75.6	739	17,132	732.8	79.8	739
U.S. History	32,277	722.9	82.9	728	16,284	710.6	78.8	711	15,993	735.5	85.2	740

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

7.2 Performance Level Distribution

The percentage distributions of students in the four performance levels based on student performance in the Winter/Trimester 2008-2009 and Spring 2009 administration and the cut-scores (please see Table 4.3 in section 4.6 for cut scores) are presented in Table 7.3 (please see Appendix B and C for distribution by scale score for Winter/Trimester 2008-2009 and Spring 2009, respectively). As above, these percentages exclude invalid student cases and second time test takers. The percentage distributions for each of the content areas are comparable to previous administrations (e.g., Winter/Trimester 2007-2008 and Spring 2008).

Table 7.3. Percentage of Students by Performance Level for Winter/Trimester 2008-2009 and Spring 2009

Subject	N	Unsatisfactory	Limited	Proficient	Advanced
			Knowledge		
Winter 2008-09					
Algebra I	1,499	10.0%	20.7%	39.9%	29.4%
Algebra II	1,915	17.7%	13.7%	38.8%	29.8%
Biology I	2,073	9.6%	18.7%	39.6%	32.2%
English II	2,628	3.8%	20.4%	49.1%	26.7%
English III	2,783	8.9%	14.4%	49.7%	27.0%
Geometry	1,901	7.0%	19.6%	40.9%	32.5%
U.S. History	2,600	7.3%	24.3%	26.1%	42.3%
Spring 2009					
Algebra I	35,736	5.1%	14.7%	38.9%	41.3%
Algebra II	29,644	17.0%	19.4%	39.1%	24.6%
Biology I	35,347	8.7%	19.0%	41.9%	30.4%
English II	34,823	3.4%	19.0%	48.8%	28.8%
English III	34,842	10.0%	12.3%	49.5%	28.2%
Geometry	34,224	6.9%	18.6%	43.2%	31.3%
U.S. History	32,277	6.5%	23.2%	28.8%	41.4%

7.3 Conditional Standard Error of Measurement

The conditional standard error of measurement (CSEM) was computed for each reported scale score. The CSEMs were computed using an IRT-based approach based on the following formula:

$$\text{CSEM}(O_x | \theta) = \sqrt{\left[\sum_{X=0}^{\text{Max}X} O_x^2 p(X | \theta) \right] - \left[\sum_{X=0}^{\text{Max}X} O_x \cdot p(X | \theta) \right]^2} \quad (6)$$

where O_x is the observed (scaled) score for a particular number right score X , θ is the IRT ability scale value conditioned on, and $p(\bullet)$ is the probability function. Pearson has implemented a computational approach for estimating $\text{CSEM}(O_x | \theta)$ in which $p(X | \theta)$ is computed using a recursive algorithm given by Thissen, Pommerich, Billeaud, and Williams (1995). Their algorithm is a polytomous generalization of the algorithm for dichotomous items given by Lord and Wingersky (1984). The values of θ used with the algorithm are obtained through the true score equating process (i.e., by solving for θ through the test characteristic curve for each number right score x). There is one CSEM per number correct or raw score and the CSEMs by subject appear in Section 4.6 in Tables 4.4 and 4.5 for Winter/Trimester 2008-2009 and Spring 2009, respectively.

7.4 Standard Error of Measurement

Measurement error is associated with every test score. A student's true score is the hypothetical average score that would result if the student took the test repeatedly under similar conditions. The Standard Error of Measurement (SEM), as an overall test-level

measure of error, provides a range around any given observed test score that likely includes the student's true score. This SEM is computed by taking the square root of the average value of the variances of the error of measurement associated with each of the raw score or scales scores:

$$SEM = \sqrt{\frac{\sum_j (CSEM_j^2 \cdot N_j)}{N_T}} \quad (7)$$

where,

SEM = Standard Error of Measurement

CSEM = Conditional Standard of Measurement

N_j = number of examinees obtaining score j in the population

N_T = total number of students in test sample

SEM was computed for each of the content areas. Table 7.4 presents the overall estimates of SEM for each of the content areas for the Winter/Trimester 2008-2009 and Spring 2009 administrations.

Table 7.4. Overall Estimates of SEM by Subject

Subject	SEM
Winter/Trimester 2008-2009	
Algebra I	4.51
Algebra II	5.66
Biology I	5.02
Geometry	5.31
English II	4.68
English III	4.98
U.S. History	5.17
Spring 2009	
Algebra I	4.81
Algebra II	5.27
Biology I	5.22
Geometry	5.27
English II	4.86
English III	5.00
U.S. History	5.16

Note: SEM = Standard Error of Measurement.

References

- American Educational Research Association (AERA), American Psychological Association (APA), and 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., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading Massachusetts: Addison-Wesley Publishing Company.
- 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.
- Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, 32, 179-197.
- 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>
- Muraki, E. (1997). The generalized partial credit model. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 153-164). New York: Springer Verlag.
- Stocking, M.L., & Lord, F.M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7, 201-210.
- Thissen, D., Chen, W-H., & Bock, R. D. (2003). *MUTILOG for Windows, Version 7* [Computer Software]. Lincolnwood, IL: Scientific Software International.
- 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 Process 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

Algebra II	
Standard 1: Number Sense and Algebraic Operations	
Standard 1.1	Rational Exponents
	1.1a Convert expressions from radical notations to rational exponents and vice versa.
	1.1b Add, subtract, multiply, divide, and simplify radical expressions and expressions containing rational exponents.
Standard 1.2	Polynomial and Rational Expressions
	1.2a Divide polynomial expressions by lower degree polynomials.
	1.2b Add, subtract, multiply, divide, and simplify rational expressions, including complex fractions.
Standard 1.3	Complex Numbers
	1.3b Add, subtract, multiply, divide, and simplify expressions involving complex numbers.
Standard 2: Relations and Functions	
Standard 2.1	Functions and Function Notation
	2.1a Recognize the parent graphs of polynomial, exponential, and logarithmic functions and predict the effects of transformations on the parent graphs, using various methods and tools which may include graphing calculators.
	2.1b Use function notation to add, subtract, multiply, and divide functions.
	2.1c Combine functions by composition.
	2.1d Use algebraic, interval, and set notations to specify the domain and range of functions of various types.
	2.1e Find and graph the inverse of a function, if it exists.
Standard 2.2	Systems of Equations
	2.2a Model a situation that can be described by a system of equations and inequalities and use the model to answer questions about the situation.
	2.2b Solve systems of linear equations and inequalities using various methods and tools which may include substitution, elimination, matrices, graphing, and graphing calculators.
	2.2c Use either one quadratic equation and one linear equation or two quadratic equations to solve problems.
Standard 2.3	Quadratic Equations and Functions
	2.3a Solve quadratic equations by graphing, factoring, completing the square and quadratic formula.
	2.3b Graph a quadratic function and identify the x- and y-intercepts and maximum or minimum value, using various methods and tools which may include a graphing calculator.
	2.3c Model a situation that can be described by a quadratic function and use the model to answer questions about the situation.

Algebra II continued	
Standard 2.4	Identify, graph, and write the equations of the conic sections (circle, ellipse, parabola, and hyperbola).
Standard 2.5	Exponential and Logarithmic Functions
	2.5a Graph exponential and logarithmic functions.
	2.5b Apply the inverse relationship between exponential and logarithmic functions to convert from one form to another.
	2.5c Model a situation that can be described by an exponential or logarithmic function and use the model to answer questions about the situation.
Standard 2.6	Polynomial Equations and Functions
	2.6a Solve polynomial equations using various methods and tools which may include factoring and synthetic division.
	2.6b Sketch the graph of a polynomial function.
	2.6c Given the graph of a polynomial function, identify the x- and y-intercepts, relative maximums and relative minimums, using various methods and tools which may include a graphing calculator.
	2.6d Model a situation that can be described by a polynomial function and use the model to answer questions about the situation.
Standard 2.7	Rational Equations and Functions
	2.7a Solve rational equations.
	2.7b Sketch the graph of a rational function.
	2.7c Given the graph of a rational function, identify the x- and y-intercepts, asymptotes, using various methods and tools which may include a graphing calculator.
	2.7d Model a situation that can be described by a rational function and use the model to answer questions about the situation.
Standard 3: Data Analysis, Probability, & Statistics	
Standard 3.1	Analysis of Collected Data ...
	3.1a Display data on a scatter plot.
	3.1b Interpret results using a linear, exponential or quadratic model/equation.
	3.1c Identify whether the model/equation is a curve of best fit for the data, using various methods and tools which may include a graphing calculator.
Standard 3.3	Identify and use arithmetic and geometric sequences

Geometry	
Standard 1: Logical Reasoning	
Standard 1.1	Identify and use logical reasoning skills (inductive and deductive) to make and test conjectures, formulate counter examples, and follow logical arguments.
Standard 1.2	State, use, and examine the validity of the converse, inverse, and contrapositive of “if-then” statements.
Standard 2: Properties of 2-Dimensional Figures	
Standard 2.2	Line and Angle Relationships
	2.2a Use the angle relationships formed by parallel lines cut by a transversal to solve problems.
	2.2b Use the angle relationships formed by two lines cut by a transversal to determine if the two lines are parallel and verify, using algebraic and deductive proofs.
	2.2c Use relationships between pairs of angles (for example, adjacent, complementary, vertical) to solve problems.
Standard 2.3	Polygons and Other Plane Figures
	2.3a Identify, describe, and analyze polygons (for example, convex, concave, regular, pentagonal, hexagonal, n-gonal).
	2.3b Apply the interior and exterior angle sum of convex polygons to solve problems, and verify using algebraic and deductive proofs.
	2.3c Develop and apply the properties of quadrilaterals to solve problems (for example, rectangles, parallelograms, rhombi, trapezoids, kites).
	2.3d Use properties of 2-dimensional figures and side length, perimeter or circumference, and area to determine unknown values and correctly identify the appropriate unit of measure of each.
Standard 2.4	Similarity
	2.4a Determine and verify the relationships of similarity of triangles, using algebraic and deductive proofs.
	2.4b Use ratios of similar 2-dimensional figures to determine unknown values, such as angles, side lengths, perimeter or circumference, and area.
Standard 2.5	Congruence
	2.5a Determine and verify the relationships of congruency of triangles, using algebraic and deductive proofs.
	2.5b Use the relationships of congruency of 2-dimensional figures to determine unknown values, such as angles, side lengths, perimeter or circumference, and area.
Standard 2.6	Circles
	2.6a Find angle measures and arc measures related to circles.
	2.6b Find angle measures and segment lengths using the relationships among radii, chords, secants, and tangents of a circle.

Geometry continued	
Standard 3: Triangles and Trigonometric Ratios	
Standard 3.1	Use the Pythagorean Theorem and its converse to find missing side lengths and to determine acute, right, and obtuse triangles, and verify using algebraic and deductive proofs.
Standard 3.2	Apply the 45-45-90 and 30-60-90 right triangle relationships to solve problems, and verify using algebraic and deductive proofs.
Standard 3.3	Express the trigonometric functions as ratios and use sine, cosine, and tangent ratios to solve real-world problems.
Standard 4: Properties of 3-Dimensional Figures	
Standard 4.1	Polyhedra and Other Solids
	4.1a Identify, describe, and analyze polyhedra (for example, regular, decahedral).
	4.1b Use properties of 3-dimensional figures; side lengths, perimeter or circumference, and area of a face; and volume, lateral area, and surface area to determine unknown values and correctly identify the appropriate unit of measure of each.
Standard 4.2	Similarity and Congruence
	4.2a Use ratios of similar 3-dimensional figures to determine unknown values, such as angles, side lengths, perimeter or circumference of a face, area of a face, and volume.
	4.2b Use the relationships of congruency of 3-dimensional figures to determine unknown values, such as angles, side lengths, perimeter or circumference of a face, area of a face, and volume.
4.3	Create a model of a 3-dimensional figure from a 2-dimensional drawing and make a 2-dimensional representation of a 3-dimensional object (for example, nets, blueprints, perspective drawings).
Standard 5: Coordinate Geometry	
Standard 5.1	Use coordinate geometry to find the distance between two points; the midpoint of a segment; and to calculate the slopes of a parallel, perpendicular, horizontal, and vertical lines.
Standard 5.2	Properties of Figures
	5.2a Given a set of points determine the type of figure formed based on its properties.
	5.2b Use transformations (reflection, rotation, translation) on geometric figures to solve problems within coordinate geometry.

Biology I	
PASS Process/Inquiry Standards and Objectives	
Process 1 Observe and Measure	
P1.1	Qualitative/quantitative observations and changes
P1.2	Use appropriate System International (SI) units and tools
P1.3	
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	Identify a testable hypothesis, variables, and control in an experiment
P3.4	
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

English III	
Reading/Literature	
Standard 1 Vocabulary	
Standard 2 Comprehension	
2.1	Literal Understanding
2.2	Inference 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 English Usage
3.2	Mechanics and Spelling
3.3	Sentence Structure
3.4	Manuscript Conventions

U.S. History	
Standard 1	Social Studies Process Skills
Standard 2	Civil War/Reconstruction Era
Standard 3	Immigration/Westward Movement
Standard 4	Industrial Revolution
Standard 5	Imperialism/Isolationism
Standard 6	Twenties Culture/Change
Standard 7	Great Depression
Standard 8	World War II
Standard 9	Post-War Foreign Policy
Standard 10	Post-War Domestic Policy

Appendix B

Scale Score Distributions for Winter/Trimester 2008-2009

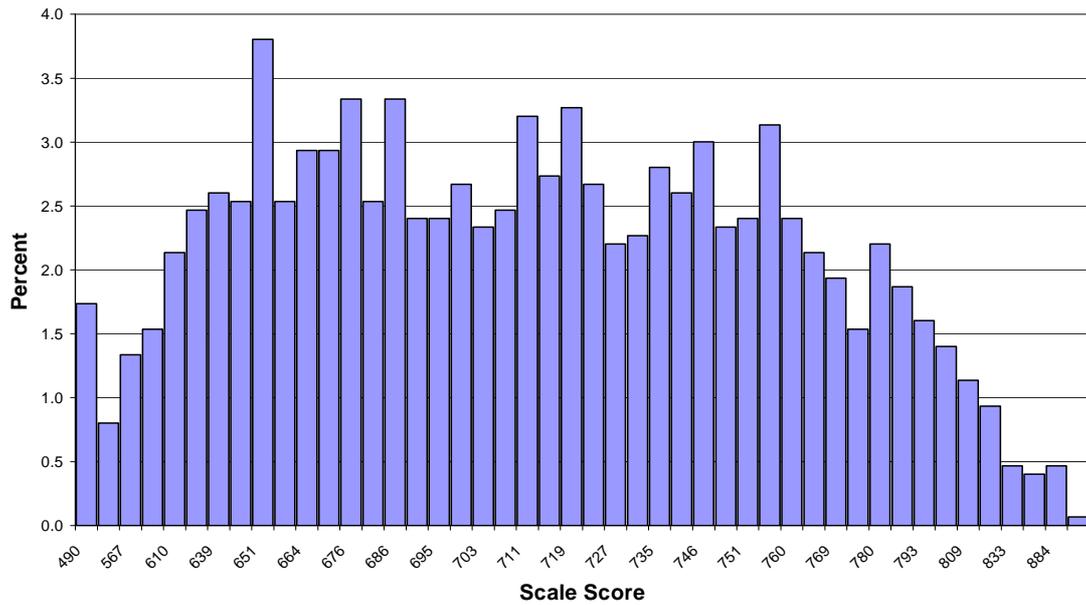
Algebra I Scale Score Distribution for Winter/Trimester 2008-2009

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
490	26	1.7	26	1.7
507	12	0.8	38	2.5
567	20	1.3	58	3.9
593	23	1.5	81	5.4
610	32	2.1	113	7.5
623	37	2.5	150	10.0
639	39	2.6	189	12.6
643	38	2.5	227	15.1
651	57	3.8	284	18.9
658	38	2.5	322	21.5
664	44	2.9	366	24.4
670	44	2.9	410	27.4
676	50	3.3	460	30.7
684	38	2.5	498	33.2
686	50	3.3	548	36.6
690	36	2.4	584	39.0
695	36	2.4	620	41.4
699	40	2.7	660	44.0
703	35	2.3	695	46.4
707	37	2.5	732	48.8
711	48	3.2	780	52.0
715	41	2.7	821	54.8
719	49	3.3	870	58.0
723	40	2.7	910	60.7
727	33	2.2	943	62.9
731	34	2.3	977	65.2
735	42	2.8	1,019	68.0
739	39	2.6	1,058	70.6
746	45	3.0	1,103	73.6
747	35	2.3	1,138	75.9
751	36	2.4	1,174	78.3
755	47	3.1	1,221	81.5
760	36	2.4	1,257	83.9
764	32	2.1	1,289	86.0
769	29	1.9	1,318	87.9
775	23	1.5	1,341	89.5
780	33	2.2	1,374	91.7
786	28	1.9	1,402	93.5
793	24	1.6	1,426	95.1
800	21	1.4	1,447	96.5
809	17	1.1	1,464	97.7
819	14	0.9	1,478	98.6
833	7	0.5	1,485	99.1

Algebra I Scale Score Distribution for Winter/Trimester 2008-2009 continued

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
851	6	0.4	1,491	99.5
884	7	0.5	1,498	99.9
999	1	0.1	1,499	100.0

Winter 2008-09 Algebra I Scale Score Distribution



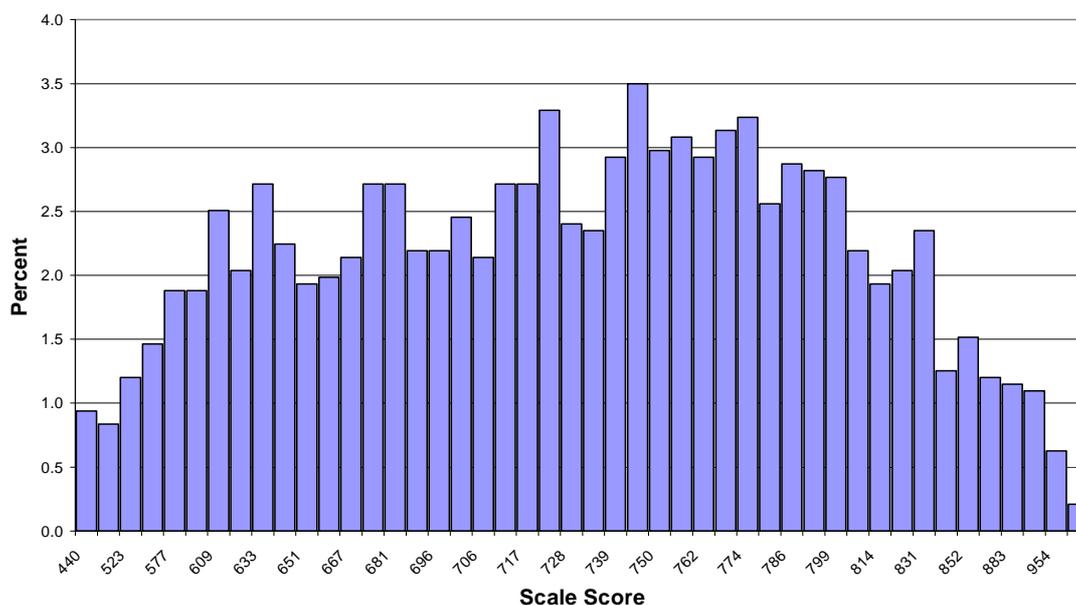
Algebra II Scale Score Distribution for Winter/Trimester 2008-2009

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
440	18	0.9	18	0.9
471	16	0.8	34	1.8
523	23	1.2	57	3.0
555	28	1.5	85	4.4
577	36	1.9	121	6.3
595	36	1.9	157	8.2
609	48	2.5	205	10.7
622	39	2.0	244	12.7
633	52	2.7	296	15.5
643	43	2.2	339	17.7
651	37	1.9	376	19.6
660	38	2.0	414	21.6
667	41	2.1	455	23.8
674	52	2.7	507	26.5
681	52	2.7	559	29.2
688	42	2.2	601	31.4
696	42	2.2	643	33.6
700	47	2.5	690	36.0
706	41	2.1	731	38.2
712	52	2.7	783	40.9
717	52	2.7	835	43.6
723	63	3.3	898	46.9
728	46	2.4	944	49.3
734	45	2.3	989	51.6
739	56	2.9	1,045	54.6
745	67	3.5	1,112	58.1
750	57	3.0	1,169	61.0
756	59	3.1	1,228	64.1
762	56	2.9	1,284	67.0
767	60	3.1	1,344	70.2
774	62	3.2	1,406	73.4
779	49	2.6	1,455	76.0
786	55	2.9	1,510	78.9
792	54	2.8	1,564	81.7
799	53	2.8	1,617	84.4
806	42	2.2	1,659	86.6
814	37	1.9	1,696	88.6
822	39	2.0	1,735	90.6
831	45	2.3	1,780	93.0
841	24	1.3	1,804	94.2
852	29	1.5	1,833	95.7
866	23	1.2	1,856	96.9

Algebra II Scale Score Distribution for Winter/Trimester 2008-2009 continued

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
883	22	1.1	1,878	98.1
909	21	1.1	1,899	99.2
954	12	0.6	1,911	99.8
999	4	0.2	1,915	100.0

Winter 2008-09 Algebra II Scale Score Distribution



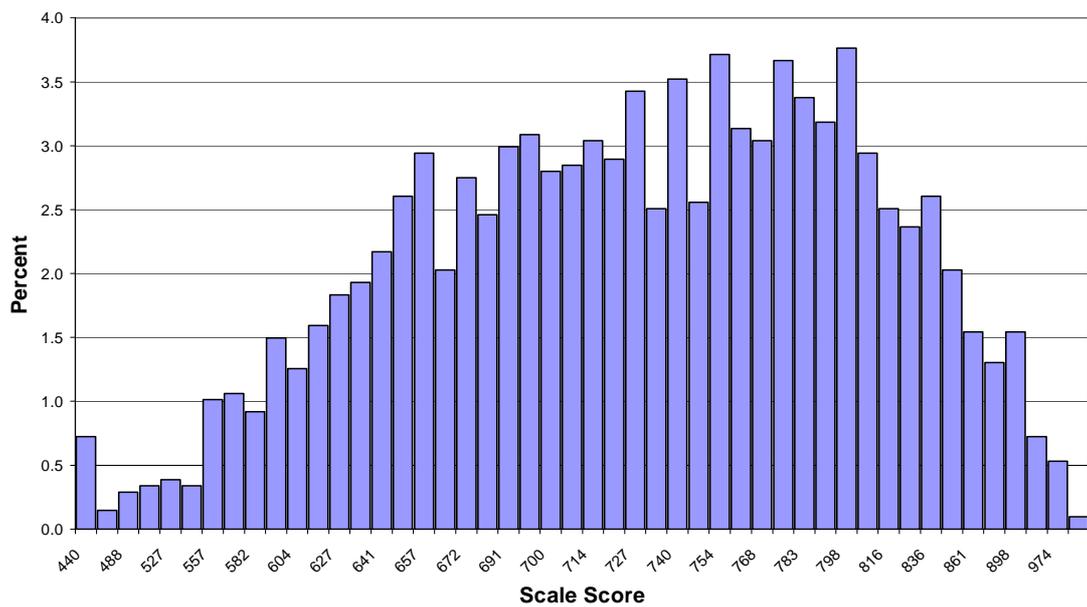
Biology I Scale Score Distribution for Winter/Trimester 2008-2009

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
440	15	0.7	15	0.7
461	3	0.1	18	0.9
488	6	0.3	24	1.2
509	7	0.3	31	1.5
527	8	0.4	39	1.9
542	7	0.3	46	2.2
557	21	1.0	67	3.2
570	22	1.1	89	4.3
582	19	0.9	108	5.2
593	31	1.5	139	6.7
604	26	1.3	165	8.0
614	33	1.6	198	9.6
627	38	1.8	236	11.4
632	40	1.9	276	13.3
641	45	2.2	321	15.5
649	54	2.6	375	18.1
657	61	2.9	436	21.0
665	42	2.0	478	23.1
672	57	2.7	535	25.8
680	51	2.5	586	28.3
691	62	3.0	648	31.3
694	64	3.1	712	34.3
700	58	2.8	770	37.1
707	59	2.8	829	40.0
714	63	3.0	892	43.0
721	60	2.9	952	45.9
727	71	3.4	1,023	49.3
734	52	2.5	1,075	51.9
740	73	3.5	1,148	55.4
747	53	2.6	1,201	57.9
754	77	3.7	1,278	61.6
761	65	3.1	1,343	64.8
768	63	3.0	1,406	67.8
775	76	3.7	1,482	71.5
783	70	3.4	1,552	74.9
790	66	3.2	1,618	78.1
798	78	3.8	1,696	81.8
807	61	2.9	1,757	84.8
816	52	2.5	1,809	87.3
825	49	2.4	1,858	89.6
836	54	2.6	1,912	92.2
848	42	2.0	1,954	94.3
861	32	1.5	1,986	95.8

Biology I Scale Score Distribution for Winter/Trimester 2008-2009 continued

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
877	27	1.3	2,013	97.1
898	32	1.5	2,045	98.6
926	15	0.7	2,060	99.4
974	11	0.5	2,071	99.9
999	2	0.1	2,073	100.0

Winter 2008-09 Biology I Scale Score Distribution



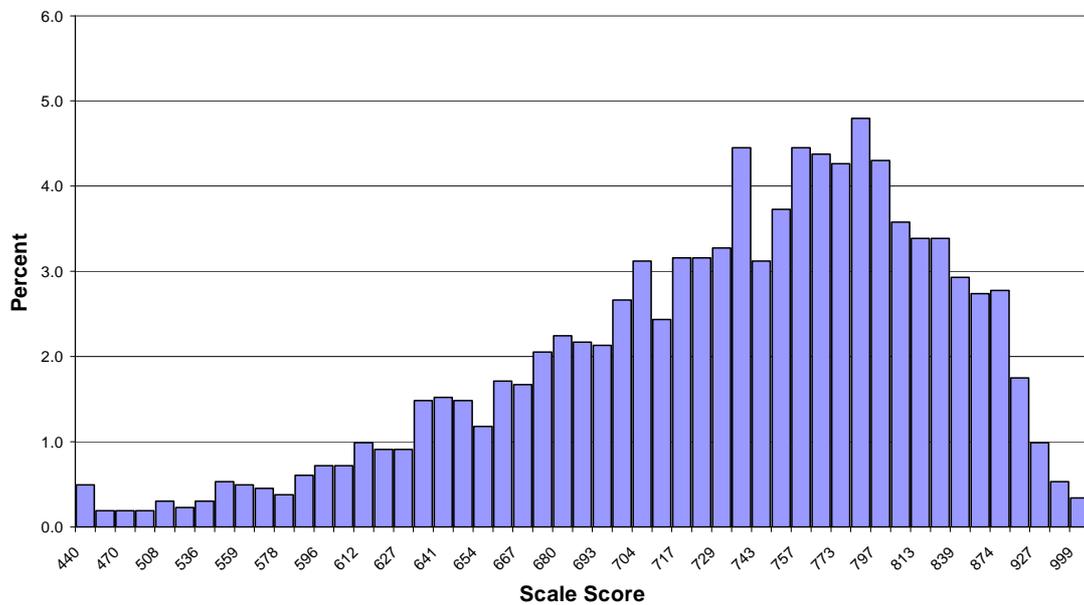
English II Scale Score Distribution for Winter/Trimester 2008-2009

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
440	13	0.5	13	0.5
443	5	0.2	18	0.7
470	5	0.2	23	0.9
490	5	0.2	28	1.1
508	8	0.3	36	1.4
523	6	0.2	42	1.6
536	8	0.3	50	1.9
548	14	0.5	64	2.4
559	13	0.5	77	2.9
569	12	0.5	89	3.4
578	10	0.4	99	3.8
588	16	0.6	115	4.4
596	19	0.7	134	5.1
604	19	0.7	153	5.8
612	26	1.0	179	6.8
619	24	0.9	203	7.7
627	24	0.9	227	8.6
634	39	1.5	266	10.1
641	40	1.5	306	11.6
648	39	1.5	345	13.1
654	31	1.2	376	14.3
661	45	1.7	421	16.0
667	44	1.7	465	17.7
673	54	2.1	519	19.7
680	59	2.2	578	22.0
686	57	2.2	635	24.2
693	56	2.1	691	26.3
698	70	2.7	761	29.0
704	82	3.1	843	32.1
710	64	2.4	907	34.5
717	83	3.2	990	37.7
723	83	3.2	1,073	40.8
729	86	3.3	1,159	44.1
736	117	4.5	1,276	48.6
743	82	3.1	1,358	51.7
750	98	3.7	1,456	55.4
757	117	4.5	1,573	59.9
765	115	4.4	1,688	64.2
773	112	4.3	1,800	68.5
782	126	4.8	1,926	73.3
797	113	4.3	2,039	77.6
802	94	3.6	2,133	81.2
813	89	3.4	2,222	84.6

English II Scale Score Distribution for Winter/Trimester 2008-2009 continued

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
825	89	3.4	2,311	87.9
839	77	2.9	2,388	90.9
855	72	2.7	2,460	93.6
874	73	2.8	2,533	96.4
898	46	1.8	2,579	98.1
927	26	1.0	2,605	99.1
969	14	0.5	2,619	99.7
999	9	0.3	2,628	100.0

Winter 2008-09 English II Scale Score Distribution

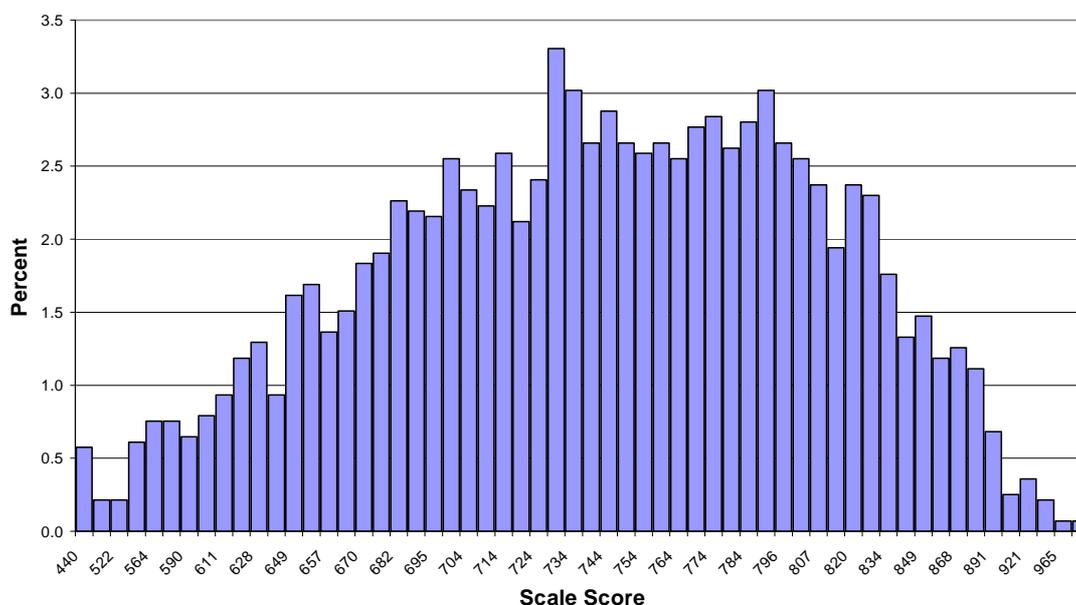


English III Scale Score Distribution for Winter/Trimester 2008-2009

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
440	16	0.6	16	0.6
487	6	0.2	22	0.8
522	6	0.2	28	1.0
546	17	0.6	45	1.6
564	21	0.8	66	2.4
578	21	0.8	87	3.1
590	18	0.6	105	3.8
601	22	0.8	127	4.6
611	26	0.9	153	5.5
619	33	1.2	186	6.7
628	36	1.3	222	8.0
636	26	0.9	248	8.9
649	45	1.6	293	10.5
650	47	1.7	340	12.2
657	38	1.4	378	13.6
663	42	1.5	420	15.1
670	51	1.8	471	16.9
676	53	1.9	524	18.8
682	63	2.3	587	21.1
687	61	2.2	648	23.3
695	60	2.2	708	25.4
698	71	2.6	779	28.0
704	65	2.3	844	30.3
709	62	2.2	906	32.6
714	72	2.6	978	35.1
719	59	2.1	1,037	37.3
724	67	2.4	1,104	39.7
729	92	3.3	1,196	43.0
734	84	3.0	1,280	46.0
739	74	2.7	1,354	48.7
744	80	2.9	1,434	51.5
749	74	2.7	1,508	54.2
754	72	2.6	1,580	56.8
759	74	2.7	1,654	59.4
764	71	2.6	1,725	62.0
769	77	2.8	1,802	64.8
774	79	2.8	1,881	67.6
779	73	2.6	1,954	70.2
784	78	2.8	2,032	73.0
795	84	3.0	2,116	76.0
796	74	2.7	2,190	78.7
801	71	2.6	2,261	81.2
807	66	2.4	2,327	83.6
813	54	1.9	2,381	85.6
820	66	2.4	2,447	87.9

826	64	2.3	2,511	90.2
834	49	1.8	2,560	92.0
841	37	1.3	2,597	93.3
849	41	1.5	2,638	94.8
858	33	1.2	2,671	96.0
868	35	1.3	2,706	97.2
879	31	1.1	2,737	98.3
891	19	0.7	2,756	99.0
905	7	0.3	2,763	99.3
921	10	0.4	2,773	99.6
940	6	0.2	2,779	99.9
965	2	0.1	2,781	99.9
999	2	0.1	2,783	100.0

Winter 2008-09 English III Scale Score Distribution



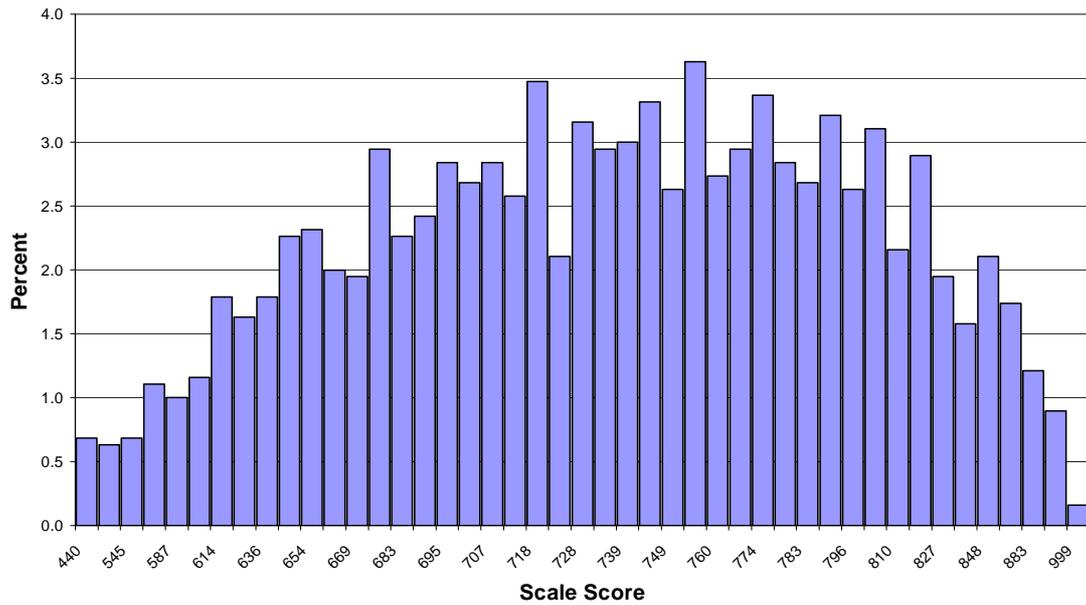
Geometry Scale Score Distribution for Winter/Trimester 2008-2009

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
440	13	0.7	13	0.7
511	12	0.6	25	1.3
545	13	0.7	38	2.0
569	21	1.1	59	3.1
587	19	1.0	78	4.1
602	22	1.2	100	5.3
614	34	1.8	134	7.0
635	31	1.6	165	8.7
636	34	1.8	199	10.5
645	43	2.3	242	12.7
654	44	2.3	286	15.0
662	38	2.0	324	17.0
669	37	1.9	361	19.0
676	56	2.9	417	21.9
683	43	2.3	460	24.2
689	46	2.4	506	26.6
695	54	2.8	560	29.5
701	51	2.7	611	32.1
707	54	2.8	665	35.0
712	49	2.6	714	37.6
718	66	3.5	780	41.0
723	40	2.1	820	43.1
728	60	3.2	880	46.3
734	56	2.9	936	49.2
739	57	3.0	993	52.2
744	63	3.3	1,056	55.5
749	50	2.6	1,106	58.2
754	69	3.6	1,175	61.8
760	52	2.7	1,227	64.5
765	56	2.9	1,283	67.5
774	64	3.4	1,347	70.9
777	54	2.8	1,401	73.7
783	51	2.7	1,452	76.4
789	61	3.2	1,513	79.6
796	50	2.6	1,563	82.2
803	59	3.1	1,622	85.3
810	41	2.2	1,663	87.5
818	55	2.9	1,718	90.4
827	37	1.9	1,755	92.3
837	30	1.6	1,785	93.9
848	40	2.1	1,825	96.0
863	33	1.7	1,858	97.7

Geometry Scale Score Distribution for Winter/Trimester 2008-2009 continued

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
883	23	1.2	1,881	98.9
918	17	0.9	1,898	99.8
999	3	0.2	1,901	100.0

Winter 2008-09 Geometry Scale Score Distribution



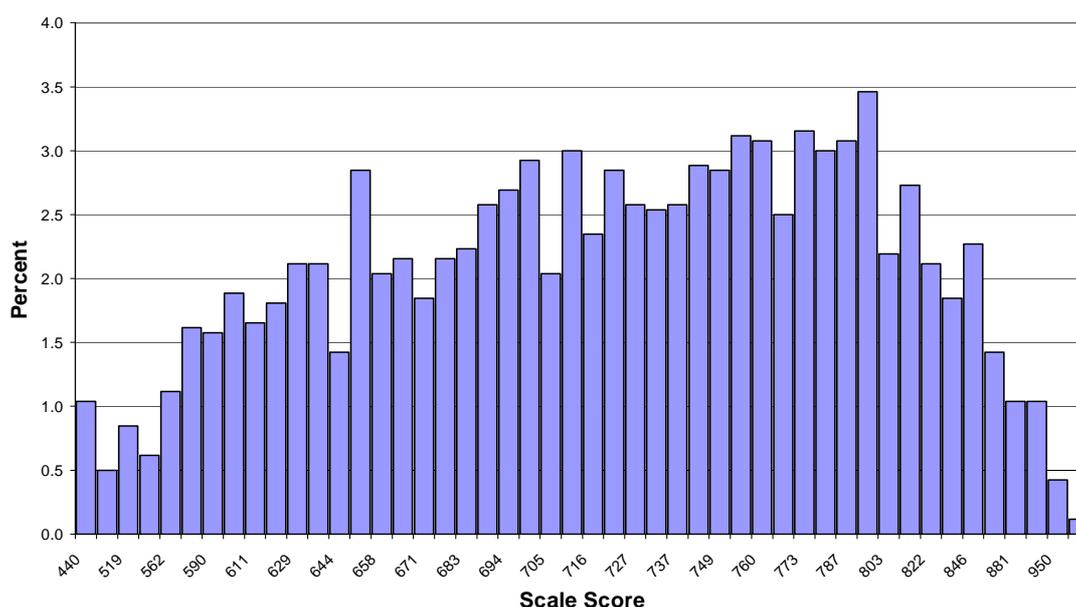
U.S. History Scale Score Distribution for Winter/Trimester 2008-2009

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
440	27	1.0	27	1.0
484	13	0.5	40	1.5
519	22	0.8	62	2.4
544	16	0.6	78	3.0
562	29	1.1	107	4.1
577	42	1.6	149	5.7
590	41	1.6	190	7.3
603	49	1.9	239	9.2
611	43	1.7	282	10.8
620	47	1.8	329	12.7
629	55	2.1	384	14.8
637	55	2.1	439	16.9
644	37	1.4	476	18.3
651	74	2.8	550	21.2
658	53	2.0	603	23.2
665	56	2.2	659	25.3
671	48	1.8	707	27.2
677	56	2.2	763	29.3
683	58	2.2	821	31.6
689	67	2.6	888	34.2
694	70	2.7	958	36.8
700	76	2.9	1,034	39.8
705	53	2.0	1,087	41.8
711	78	3.0	1,165	44.8
716	61	2.3	1,226	47.2
721	74	2.8	1,300	50.0
727	67	2.6	1,367	52.6
732	66	2.5	1,433	55.1
737	67	2.6	1,500	57.7
747	75	2.9	1,575	60.6
749	74	2.8	1,649	63.4
754	81	3.1	1,730	66.5
760	80	3.1	1,810	69.6
767	65	2.5	1,875	72.1
773	82	3.2	1,957	75.3
780	78	3.0	2,035	78.3
787	80	3.1	2,115	81.3
795	90	3.5	2,205	84.8
803	57	2.2	2,262	87.0
812	71	2.7	2,333	89.7
822	55	2.1	2,388	91.8
833	48	1.8	2,436	93.7
846	59	2.3	2,495	96.0

U.S. History Scale Score Distribution for Winter/Trimester 2008-2009 continued

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
861	37	1.4	2,532	97.4
881	27	1.0	2,559	98.4
907	27	1.0	2,586	99.5
950	11	0.4	2,597	99.9
999	3	0.1	2,600	100.0

Winter 2008-09 US History Scale Score Distribution



Appendix C

Scale Score Distributions for Spring 2009

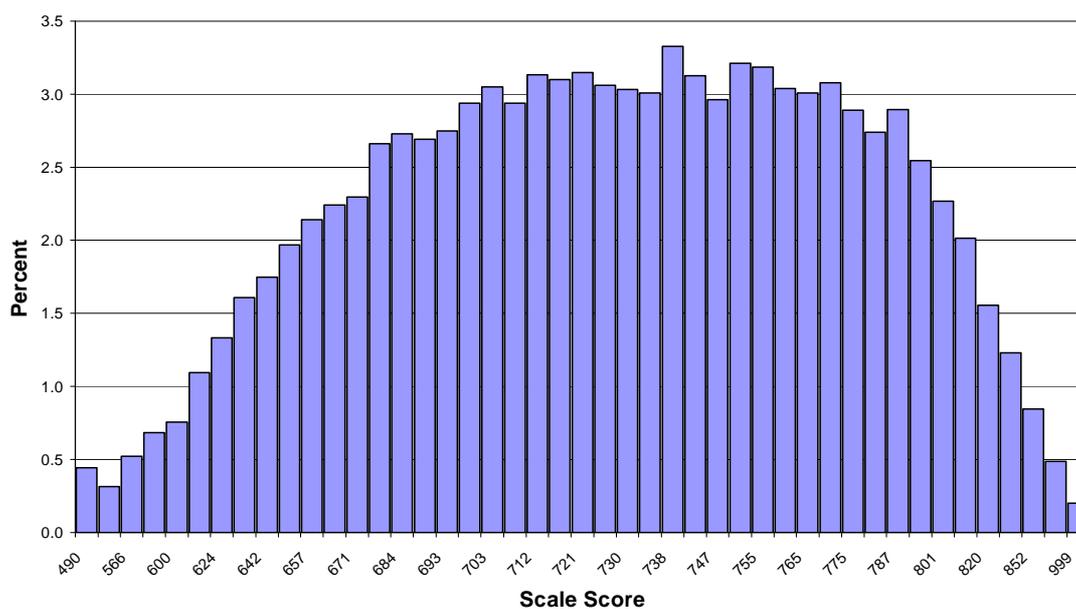
Algebra I Score Distribution for Spring 2009

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
490	158	0.4	158	0.4
538	112	0.3	270	0.8
566	186	0.5	456	1.3
585	244	0.7	700	2.0
600	270	0.8	970	2.7
613	391	1.1	1,361	3.8
624	476	1.3	1,837	5.1
639	575	1.6	2,412	6.7
642	625	1.7	3,037	8.5
650	703	2.0	3,740	10.5
657	765	2.1	4,505	12.6
664	801	2.2	5,306	14.8
671	821	2.3	6,127	17.1
677	951	2.7	7,078	19.8
684	975	2.7	8,053	22.5
688	962	2.7	9,015	25.2
693	982	2.7	9,997	28.0
698	1050	2.9	11,047	30.9
703	1090	3.1	12,137	34.0
708	1050	2.9	13,187	36.9
712	1120	3.1	14,307	40.0
717	1108	3.1	15,415	43.1
721	1125	3.1	16,540	46.3
726	1094	3.1	17,634	49.3
730	1084	3.0	18,718	52.4
734	1075	3.0	19,793	55.4
738	1189	3.3	20,982	58.7
746	1118	3.1	22,100	61.8
747	1059	3.0	23,159	64.8
751	1148	3.2	24,307	68.0
755	1138	3.2	25,445	71.2
760	1086	3.0	26,531	74.2
765	1075	3.0	27,606	77.2
770	1100	3.1	28,706	80.3
775	1033	2.9	29,739	83.2
780	979	2.7	30,718	86.0
787	1035	2.9	31,753	88.9
793	910	2.5	32,663	91.4
801	810	2.3	33,473	93.7
810	720	2.0	34,193	95.7
820	556	1.6	34,749	97.2
834	439	1.2	35,188	98.5

Algebra I Score Distribution for Spring 2009 continued

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
852	302	0.8	35,490	99.3
885	174	0.5	35,664	99.8
999	72	0.2	35,736	100.0

Spring 2009 Algebra I Scale Score Distribution



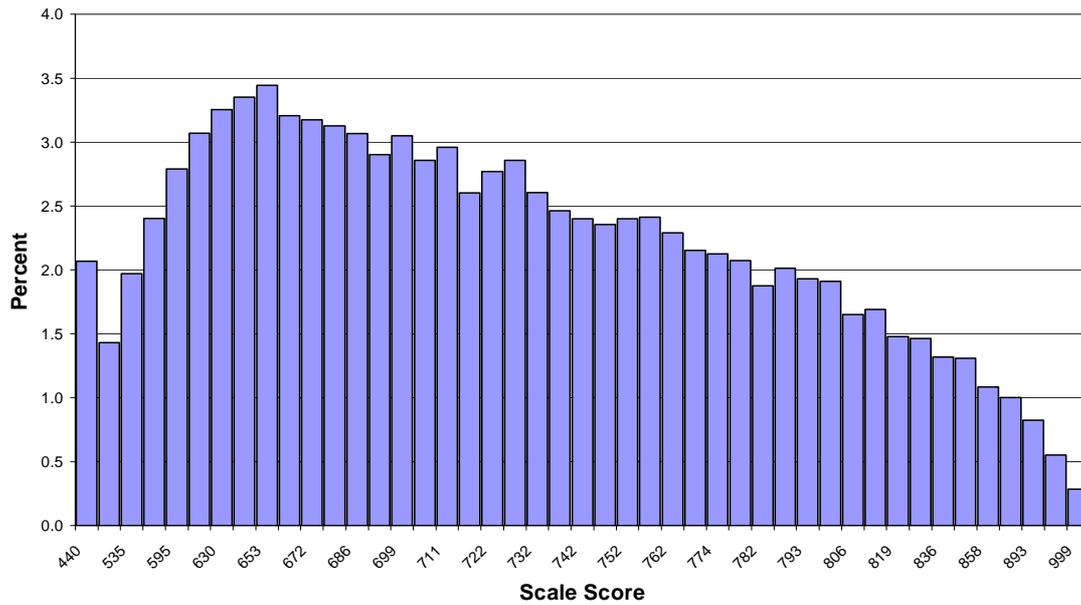
Algebra II Score Distribution for Spring 2009

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
440	613	2.1	613	2.1
475	424	1.4	1,037	3.5
535	584	2.0	1,621	5.5
570	712	2.4	2,333	7.9
595	827	2.8	3,160	10.7
614	910	3.1	4,070	13.7
630	965	3.3	5,035	17.0
651	994	3.4	6,029	20.3
653	1021	3.4	7,050	23.8
663	951	3.2	8,001	27.0
672	941	3.2	8,942	30.2
679	927	3.1	9,869	33.3
686	909	3.1	10,778	36.4
696	860	2.9	11,638	39.3
699	904	3.0	12,542	42.3
705	847	2.9	13,389	45.2
711	877	3.0	14,266	48.1
716	771	2.6	15,037	50.7
722	821	2.8	15,858	53.5
727	847	2.9	16,705	56.4
732	772	2.6	17,477	59.0
737	730	2.5	18,207	61.4
742	711	2.4	18,918	63.8
747	698	2.4	19,616	66.2
752	711	2.4	20,327	68.6
757	715	2.4	21,042	71.0
762	679	2.3	21,721	73.3
767	638	2.2	22,359	75.4
774	630	2.1	22,989	77.6
777	614	2.1	23,603	79.6
782	556	1.9	24,159	81.5
788	597	2.0	24,756	83.5
793	572	1.9	25,328	85.4
799	566	1.9	25,894	87.3
806	489	1.6	26,383	89.0
812	501	1.7	26,884	90.7
819	438	1.5	27,322	92.2
827	434	1.5	27,756	93.6
836	391	1.3	28,147	95.0
846	388	1.3	28,535	96.3
858	321	1.1	28,856	97.3
873	297	1.0	29,153	98.3

Algebra II Score Distribution for Spring 2009 continued

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
893	244	0.8	29,397	99.2
929	163	0.5	29,560	99.7
999	84	0.3	29,644	100.0

Spring 2009 Algebra II Scale Score Distribution



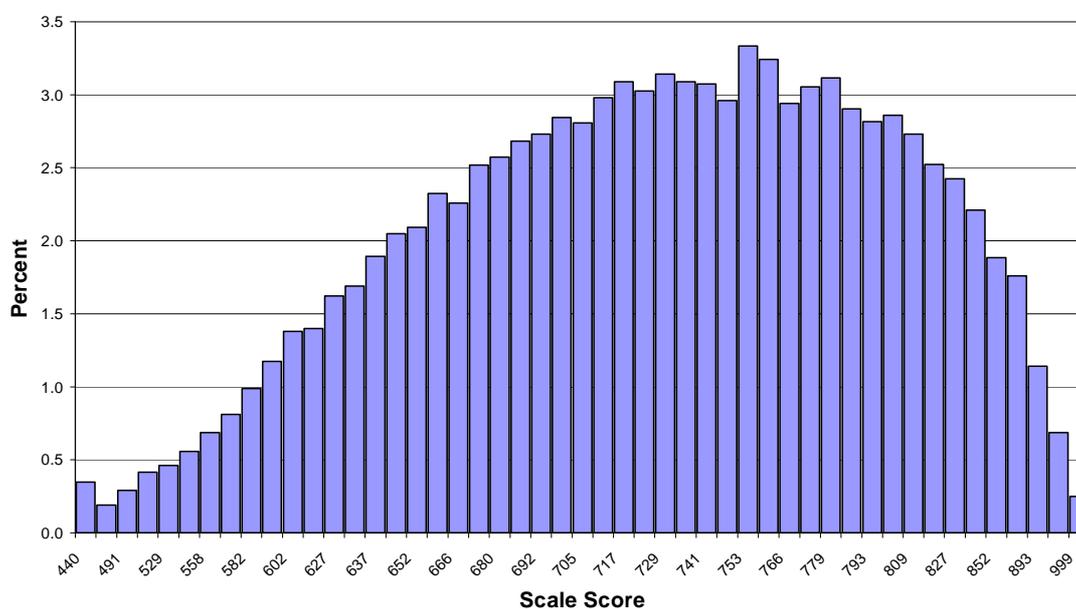
Biology I Score Distribution for Spring 2009

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
440	123	0.3	123	0.3
466	68	0.2	191	0.5
491	103	0.3	294	0.8
512	147	0.4	441	1.2
529	163	0.5	604	1.7
544	197	0.6	801	2.3
558	243	0.7	1,044	3.0
570	287	0.8	1,331	3.8
582	349	1.0	1,680	4.8
592	415	1.2	2,095	5.9
602	488	1.4	2,583	7.3
612	495	1.4	3,078	8.7
627	573	1.6	3,651	10.3
629	597	1.7	4,248	12.0
637	669	1.9	4,917	13.9
645	724	2.0	5,641	16.0
652	739	2.1	6,380	18.0
659	821	2.3	7,201	20.4
666	798	2.3	7,999	22.6
673	890	2.5	8,889	25.1
680	909	2.6	9,798	27.7
691	948	2.7	10,746	30.4
692	965	2.7	11,711	33.1
699	1005	2.8	12,716	36.0
705	992	2.8	13,708	38.8
711	1053	3.0	14,761	41.8
717	1092	3.1	15,853	44.8
723	1069	3.0	16,922	47.9
729	1110	3.1	18,032	51.0
735	1092	3.1	19,124	54.1
741	1086	3.1	20,210	57.2
747	1046	3.0	21,256	60.1
753	1178	3.3	22,434	63.5
759	1146	3.2	23,580	66.7
766	1039	2.9	24,619	69.6
775	1079	3.1	25,698	72.7
779	1101	3.1	26,799	75.8
786	1026	2.9	27,825	78.7
793	995	2.8	28,820	81.5
800	1011	2.9	29,831	84.4
809	965	2.7	30,796	87.1
817	891	2.5	31,687	89.6

Biology I Score Distribution for Spring 2009 continued

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
827	857	2.4	32,544	92.1
839	781	2.2	33,325	94.3
852	666	1.9	33,991	96.2
869	622	1.8	34,613	97.9
893	403	1.1	35,016	99.1
933	243	0.7	35,259	99.8
999	88	0.2	35,347	100.0

Spring 2009 Biology I Scale Score Distribution



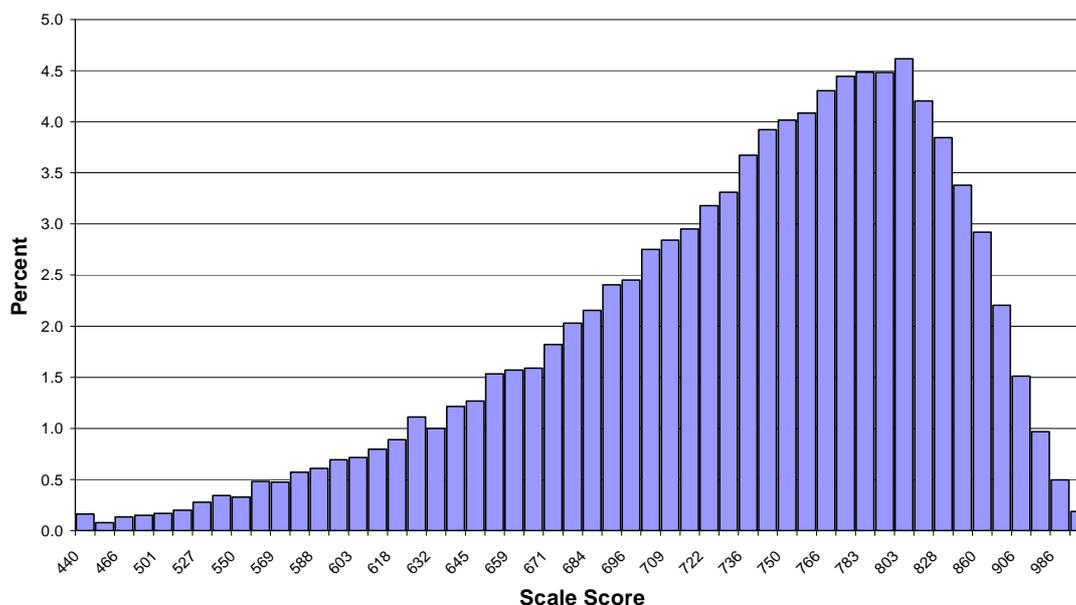
English II Score Distribution for Spring 2009

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
440	57	0.2	57	0.2
444	28	0.1	85	0.2
466	47	0.1	132	0.4
485	53	0.2	185	0.5
501	59	0.2	244	0.7
515	70	0.2	314	0.9
527	97	0.3	411	1.2
539	120	0.3	531	1.5
550	114	0.3	645	1.9
560	168	0.5	813	2.3
569	166	0.5	979	2.8
578	199	0.6	1,178	3.4
588	212	0.6	1,390	4.0
595	241	0.7	1,631	4.7
603	249	0.7	1,880	5.4
610	277	0.8	2,157	6.2
618	310	0.9	2,467	7.1
625	387	1.1	2,854	8.2
632	348	1.0	3,202	9.2
639	423	1.2	3,625	10.4
645	442	1.3	4,067	11.7
652	534	1.5	4,601	13.2
659	547	1.6	5,148	14.8
665	553	1.6	5,701	16.4
671	634	1.8	6,335	18.2
678	707	2.0	7,042	20.2
684	750	2.2	7,792	22.4
693	837	2.4	8,629	24.8
696	854	2.5	9,483	27.2
703	958	2.8	10,441	30.0
709	990	2.8	11,431	32.8
716	1028	3.0	12,459	35.8
722	1107	3.2	13,566	39.0
729	1153	3.3	14,719	42.3
736	1279	3.7	15,998	45.9
743	1366	3.9	17,364	49.9
750	1398	4.0	18,762	53.9
758	1422	4.1	20,184	58.0
766	1499	4.3	21,683	62.3
774	1547	4.4	23,230	66.7
783	1561	4.5	24,791	71.2
797	1560	4.5	26,351	75.7

English II Score Distribution for Spring 2009 continued

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
803	1607	4.6	27,958	80.3
815	1463	4.2	29,421	84.5
828	1339	3.8	30,760	88.3
843	1177	3.4	31,937	91.7
860	1016	2.9	32,953	94.6
881	768	2.2	33,721	96.8
906	527	1.5	34,248	98.3
940	337	1.0	34,585	99.3
986	173	0.5	34,758	99.8
999	65	0.2	34,823	100.0

Spring 2009 English II Scale Score Distribution



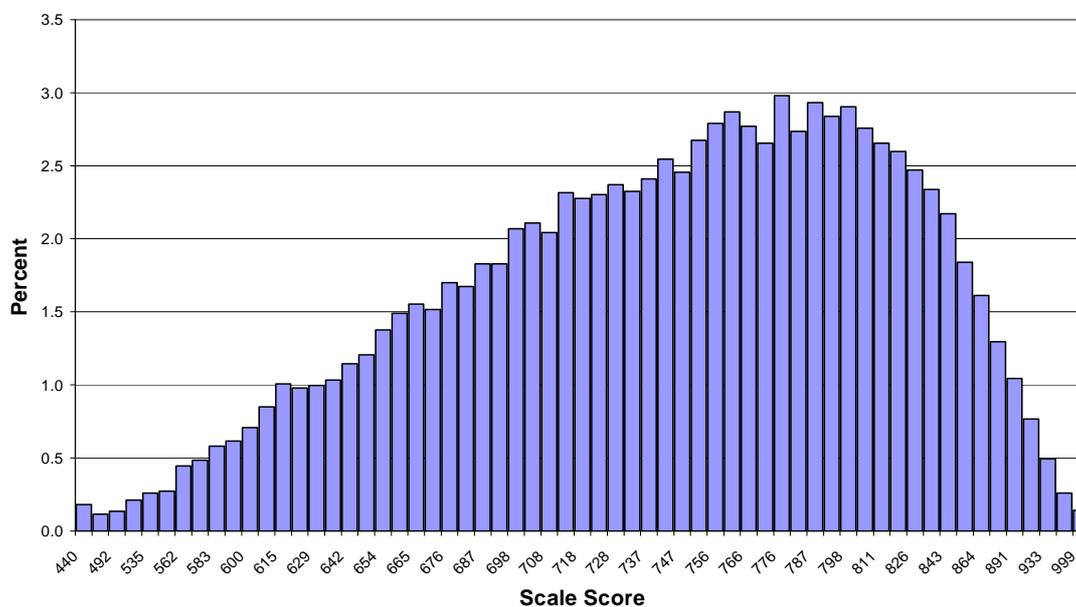
English III Score Distribution for Spring 2009

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
440	63	0.2	63	0.2
451	40	0.1	103	0.3
492	47	0.1	150	0.4
517	73	0.2	223	0.6
535	90	0.3	313	0.9
550	95	0.3	408	1.2
562	155	0.4	563	1.6
573	169	0.5	732	2.1
583	202	0.6	934	2.7
592	214	0.6	1,148	3.3
600	246	0.7	1,394	4.0
608	296	0.8	1,690	4.9
615	351	1.0	2,041	5.9
623	341	1.0	2,382	6.8
629	347	1.0	2,729	7.8
636	360	1.0	3,089	8.9
642	399	1.1	3,488	10.0
649	420	1.2	3,908	11.2
654	479	1.4	4,387	12.6
660	519	1.5	4,906	14.1
665	541	1.6	5,447	15.6
671	528	1.5	5,975	17.1
676	592	1.7	6,567	18.8
682	583	1.7	7,150	20.5
687	637	1.8	7,787	22.3
695	637	1.8	8,424	24.2
698	721	2.1	9,145	26.2
703	735	2.1	9,880	28.4
708	712	2.0	10,592	30.4
713	807	2.3	11,399	32.7
718	793	2.3	12,192	35.0
723	802	2.3	12,994	37.3
728	826	2.4	13,820	39.7
733	810	2.3	14,630	42.0
737	840	2.4	15,470	44.4
742	887	2.5	16,357	46.9
747	856	2.5	17,213	49.4
752	932	2.7	18,145	52.1
756	972	2.8	19,117	54.9
761	1000	2.9	20,117	57.7
766	965	2.8	21,082	60.5

English III Score Distribution for Spring 2009 continued

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
771	925	2.7	22,007	63.2
776	1039	3.0	23,046	66.1
781	953	2.7	23,999	68.9
787	1022	2.9	25,021	71.8
795	989	2.8	26,010	74.7
798	1012	2.9	27,022	77.6
805	961	2.8	27,983	80.3
811	925	2.7	28,908	83.0
818	905	2.6	29,813	85.6
826	861	2.5	30,674	88.0
834	815	2.3	31,489	90.4
843	757	2.2	32,246	92.5
853	641	1.8	32,887	94.4
864	562	1.6	33,449	96.0
876	451	1.3	33,900	97.3
891	364	1.0	34,264	98.3
909	267	0.8	34,531	99.1
933	172	0.5	34,703	99.6
968	90	0.3	34,793	99.9
999	49	0.1	34,842	100.0

Spring 2009 English III Scale Score Distribution



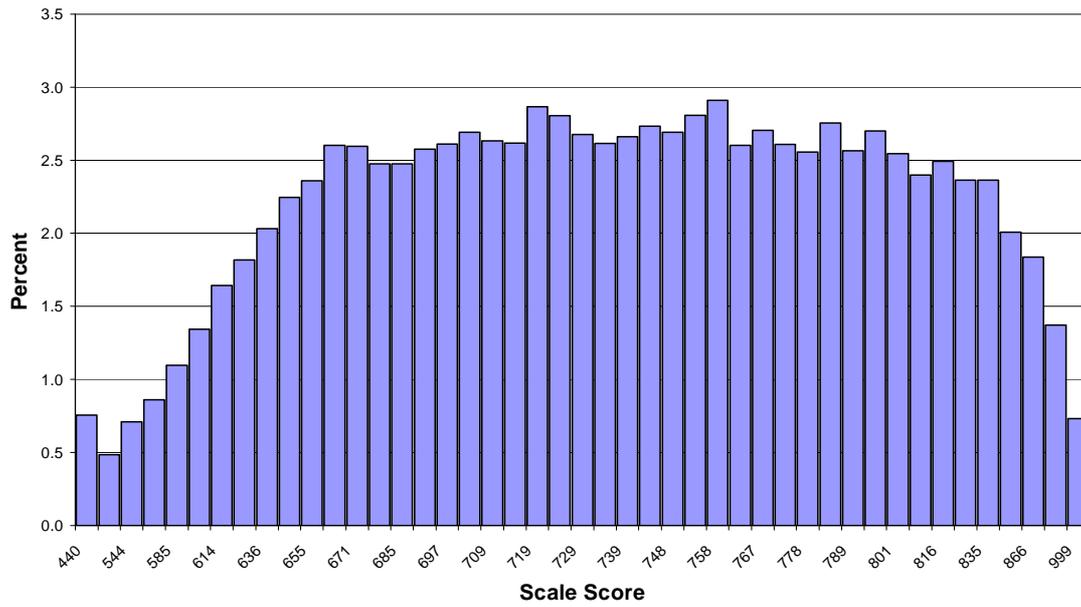
Geometry Score Distribution for Spring 2009

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
440	258	0.8	258	0.8
508	166	0.5	424	1.2
544	243	0.7	667	1.9
567	294	0.9	961	2.8
585	375	1.1	1,336	3.9
601	460	1.3	1,796	5.2
614	562	1.6	2,358	6.9
635	622	1.8	2,980	8.7
636	695	2.0	3,675	10.7
646	769	2.2	4,444	13.0
655	808	2.4	5,252	15.3
663	891	2.6	6,143	17.9
671	888	2.6	7,031	20.5
678	847	2.5	7,878	23.0
685	847	2.5	8,725	25.5
695	882	2.6	9,607	28.1
697	894	2.6	10,501	30.7
703	921	2.7	11,422	33.4
709	901	2.6	12,323	36.0
714	896	2.6	13,219	38.6
719	981	2.9	14,200	41.5
724	960	2.8	15,160	44.3
729	916	2.7	16,076	47.0
734	895	2.6	16,971	49.6
739	911	2.7	17,882	52.2
743	935	2.7	18,817	55.0
748	921	2.7	19,738	57.7
753	961	2.8	20,699	60.5
758	996	2.9	21,695	63.4
762	891	2.6	22,586	66.0
767	926	2.7	23,512	68.7
774	893	2.6	24,405	71.3
778	875	2.6	25,280	73.9
783	943	2.8	26,223	76.6
789	878	2.6	27,101	79.2
795	924	2.7	28,025	81.9
801	871	2.5	28,896	84.4
808	821	2.4	29,717	86.8
816	853	2.5	30,570	89.3
825	809	2.4	31,379	91.7
835	809	2.4	32,188	94.1
848	687	2.0	32,875	96.1

Geometry Score Distribution for Spring 2009 continued

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
866	629	1.8	33,504	97.9
896	470	1.4	33,974	99.3
999	250	0.7	34,224	100.0

Spring 2009 Geometry Scale Score Distribution



U.S. History Score Distribution for Spring 2009

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
440	140	0.4	140	0.4
452	91	0.3	231	0.7
486	115	0.4	346	1.1
512	173	0.5	519	1.6
533	228	0.7	747	2.3
551	254	0.8	1,001	3.1
566	321	1.0	1,322	4.1
580	349	1.1	1,671	5.2
592	420	1.3	2,091	6.5
603	518	1.6	2,609	8.1
613	549	1.7	3,158	9.8
622	562	1.7	3,720	11.5
631	635	2.0	4,355	13.5
639	652	2.0	5,007	15.5
647	681	2.1	5,688	17.6
654	711	2.2	6,399	19.8
661	740	2.3	7,139	22.1
668	788	2.4	7,927	24.6
674	846	2.6	8,773	27.2
681	821	2.5	9,594	29.7
689	804	2.5	10,398	32.2
693	883	2.7	11,281	35.0
699	923	2.9	12,204	37.8
705	902	2.8	13,106	40.6
711	922	2.9	14,028	43.5
716	978	3.0	15,006	46.5
722	967	3.0	15,973	49.5
728	939	2.9	16,912	52.4
734	978	3.0	17,890	55.4
740	1014	3.1	18,904	58.6
747	979	3.0	19,883	61.6
752	1020	3.2	20,903	64.8
758	1035	3.2	21,938	68.0
765	1089	3.4	23,027	71.3
772	1040	3.2	24,067	74.6
779	1022	3.2	25,089	77.7
787	941	2.9	26,030	80.6
795	998	3.1	27,028	83.7
804	869	2.7	27,897	86.4
814	902	2.8	28,799	89.2
824	765	2.4	29,564	91.6
836	719	2.2	30,283	93.8

U.S. History Score Distribution for Spring 2009 continued

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
850	631	2.0	30,914	95.8
866	489	1.5	31,403	97.3
886	405	1.3	31,808	98.5
914	273	0.8	32,081	99.4
967	143	0.4	32,224	99.8
999	53	0.2	32,277	100.0

Spring 2009 US History Scale Score Distribution

