OKLAHOMA STATE DEPARTMENT OF EDUCATION
Janet Barresi, State Superintendent of Public Instruction

Oklahoma Modified Alternative Assessment Program

2011 Technical Report

Grades 3-8 Tests

Submitted to
The Oklahoma State Department of Education
October 2011

PEARSON ALWAYS LEARNING

DRAFT

Executive Summary

Introduction

The Oklahoma School Testing Program (OSTP) includes the Oklahoma Core Curriculum Tests (OCCT), which are administered statewide to students in Grades 3-8. The general Oklahoma Core Curriculum Tests assess student proficiency in the subjects of mathematics, reading, science, geography, social studies, and U.S. History. Each test is a measure of a student's knowledge relative to the *Priority Academic Student Skills (PASS)*, Oklahoma's content standards. The Oklahoma Modified Alternative Assessment Program (OMAAP) tests are used to assess student proficiency in mathematics, reading, and science (grades 5 and 8 only). The OMAAP tests are intended for a population of students for whom the general OCCT exams and the Oklahoma Alternative Assessment Program (OAAP) portfolio assessments are inappropriate. The OMAAP exams are based on modified blueprints and items from the corresponding OCCT 3-8 exams. These OMAAP tests were administered in Spring 2011. The OSTP was established to improve academic achievement for all Oklahoma students and it also meets the requirements of the No Child Left Behind Act (NCLB), which was introduced by the Federal Government in 2001. In 2011, Pearson was contracted by the Oklahoma State Department of Education (SDE) to develop, administer, and maintain the OMAAP 3-8 tests. This report provides technical details of work accomplished through the end of 2011 on these tests.

Purpose

The purpose of this Technical Report is to provide objective information regarding technical aspects of the OMAAP 3-8 assessments. This volume is intended to be one source of information to Oklahoma K-12 educational stakeholders (including testing coordinators, educators, parents, and other interested citizens) about the development, implementation, scoring, and technical attributes of the OMAAP 3-8 assessments. Other sources of information regarding the OMAAP 3-8 tests include the administration manuals, interpretation manuals, student-, teacher-, and parent guides, implementation materials, and training materials.

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

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

This technical report provides accurate, complete, current, and clear documentation of the OMAAP development methods, data analysis, and results as is appropriate for use by qualified users and technical experts. Section 1 provides an overview of the test design, test content, and content standards. Section 2 provides summary information about the test administration. Section 3 details the classical item analyses and reliability results, and Section 4 details the calibration, equating, scaling analyses, and results. Section 5 provides the results of the classification accuracy and classifications studies and section 6 provides higher-level summaries of all the tests included in the OMAAP 3-8 testing program.

DRAFT

Information provided in this report presents valuable information about the OMAAP 3-8 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, and
11. Decision accuracy and classification.

Each of these facets in the OMAAP 3-8 assessment development and use cycle is critical to the validity of test scores and interpretation of the results. This technical report covers all of these topics for the 2011-11 testing year.

DRAFT

Table of Contents

Section 1 7
1.1 Overview of the OMAAP 3-8 Assessments 7
1.1.a Purpose 7
1.1.b Modified *PASS* Content Standards 7
1.2 Content Modifications 8
1.2.a Modification of Blueprint and Item Specifications 8
1.2.b Modification of Forms and Items 8
1.3 Summary of Test Development and Content Validity 9
1.3.a Aligning Test to Modified *PASS* Content Standards 9
1.3.b Item Pool Development and Selection 10
1.3.c Configuration of the Tests 11
1.3.d Operational Items by Content Area 12
Section 2 14
2.1 Packaging and Shipping 14
2.2 Materials Return 14
2.3 Materials Discrepancies Process 15
2.4 Processing Assessment Materials Returned by Schools 15
Section 3 16
3.1 Sampling Plan and Data Receipt 16
3.1.a Sampling Plan 16
3.1.b Data Receipt Activities 16
3.2 Classical Item Analyses 20
3.2.a Test-Level Summaries of Classical Item Analyses 21
3.3 Procedures for Detecting Item Bias 21
3.3.a Differential Item Functioning Results 22
3.4 Test Reliability 23
Section 4 26
4.1 Item Response Theory (IRT) Models 26
4.1.a Item Response Theory Model 26
4.2 Assessment of Fit to the Model 26
4.2.a Spring 2011 26
4.3 Calibration and Equating 27
4.4 Anchor Item Stability Evaluation Methods 27
4.4.a Anchor Items for Spring 2011 29
4.4.b Results of the Anchor Item Stability Check 30
4.5 Scaling and Scoring Results 31
Section 5 43
5.1 Classification Consistency and Accuracy 43

DRAFT

Section 6 47
6.1 Descriptive Statistics 47
6.2 Performance Level Distribution 52
6.3 Conditional Standard Error of Measurement 53
6.4 Standard Error of Measurement 53
References 55
Appendix A 56
Appendix B 71

DRAFT

List of Tables and Figures

Table 1.1. Oklahoma Content Standards by Subject 8
Table 1.2. OMAAP Item Modification Rules 9
Table 1.4. Criteria for Aligning the Test with Modified *PASS* Standards and Objectives..... 10
Table 1.5. Percentage of Items in Depth of Knowledge Levels by Grade - Math..... 11
Table 1.3. OCCT and OMAAP 3-8 Item Count Comparison 12
Table 1.9. Number of Operational and Field Test Items by Content Standard for Math 12
Table 1.10. Number of Operational and Field Test Items by Content Standard for Reading ... 12
Table 1.11. Number of Items by Content Standard for Science 13
Table 1.12. Number of Items by Process Standard for Science 13
Table 3.1. Demographic Characteristics of Calibration and Equating Sample..... 17
Table 3.2. Test-Level Summaries of Classical Item Analyses 21
Table 3.5. Cronbach's Alpha by Subject and Grade 25
Table 4.1. Number of Horizontal Linking Items Per Subject and Grade 30
Table 4.2. LOSS, HOSS, and Scaling Constants by Subject and Grade 32
Table 4.3. Performance-Level Cut Scores by Content Area 32
Table 4.6. Raw Score to Scale Score Conversion Table for Reading – Grades 3-5..... 37
Table 4.7. Raw Score to Scale Score Conversion Table for Reading – Grades 6-8..... 39
Table 4.8. Raw Score to Scale Score Conversion Table for Science – Grades 5 and 8 41
Table 5.1. Estimates of Accuracy and Consistency of Performance Classification..... 44
Table 5.2. Accuracy and Consistency Estimates by Cut Score: False Positive- and False Negative Rates 45
Table 7.1. Descriptive Statistics of the Scale Scores - Overall 47
Table 7.2. Descriptive Statistics of the Scale Scores by Gender 48
Table 7.3. Descriptive Statistics of the Scale Scores by Race/Ethnicity 49
Table 7.4. Descriptive Statistics of the Scale Scores by Free/Reduced Lunch Status 52
Table 7.5. Percentage of Students by Performance Level 53
Table 7.6. Overall Estimates of *SEM* by Subject and Grade 54

DRAFT

Section 1

Overview of the Oklahoma Modified Alternative Assessment Program (OMAAP) 3-8 Assessments

1.1 Overview of the OMAAP 3-8 Assessments

The Oklahoma Modified Alternative Assessment Program 3-8 assessments are a collection of state-mandated, criterion-referenced tests used to assess student proficiency in mathematics, reading, and science. The OMAAP is intended for a population of students for whom the Oklahoma Alternative Assessment Program (OAAP) portfolio assessment and the general Oklahoma Core Curriculum Tests (OCCT) are inappropriate.

The OMAAP 3-8 tests are used to assess student proficiency relative to a specific set of academic skills established by committees of Oklahoma educators. The OCCT exams are aligned with the *Priority Academic Student Skills* (*PASS*). The OMAAP exams are aligned with a modified version of the *PASS* standards. The mathematics and reading OMAAP exams are administered in Grades 3-8; the science OMAAP exam is administered in Grades 5 and 8. The Spring 2011 OMAAP assessments were developed by Riverside and administered by Pearson in collaboration with the Oklahoma State Department of Education (SDE) and were administered by the SDE.

Pearson scored, equated, and scaled the OMAAP 3-8 assessments. One OMAAP form was administered in Spring 2011 for each subject and grade. In addition, an equivalent form from one of the previous administrations was designated as a breach form for each Reading and Mathematic assessment and a Braille test was built for each subject. A student could receive an equivalent form for various reasons, including becoming ill during test administration or experiencing some kind of security breach. The State Department of Education Office of Accountability and Assessments determines eligibility for an equivalent form on a case-by-case basis. These students' responses were scored and reported using the scoring tables from the form's previous administration.

1.1.a Purpose

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

1.1.b Modified *PASS* Content Standards

The Oklahoma Modified content standards are shown in Table 1.1.

DRAFT

Table 1.1. Oklahoma Content Standards by Subject

Math	
Standard 1.	Algebraic Reasoning: Patterns and Relationships
Standard 2.	Number Sense and Operation
Standard 3.	Geometry
Standard 4.	Measurement
Standard 5.	Data Analysis
Reading	
Grade 3	
Standard 2.	Vocabulary
Standard 4.	Comprehension/Critical Literacy
Standard 5.	Literature
Standard 6.	Research and Information
Grades 4-8	
Standard 1.	Vocabulary
Standard 3.	Comprehension/Critical Literacy
Standard 4.	Literature
Standard 5.	Research and Information
Science	
Process Standards: Grades 5 and 8	
Standard 1.	Observe and Measure
Standard 2.	Classify
Standard 3.	Experiment
Standard 4.	Interpret and Communicate
Content Standards: Grade 5	
Standard 1.	Properties of Matter and Energy
Standard 2.	Organisms and Environments
Standard 3.	Structures of the Earth and the Solar System
Content Standards: Grade 8	
Standard 1.	Properties and Chemical Changes in Matter
Standard 2.	Motion and Forces
Standard 3.	Diversity and Adaptations of Organisms
Standard 4.	Structures/Forces of the Earth/Solar System
Standard 5.	Earth's History

1.2 Content Modifications

1.2.a Modification of Blueprint and Item Specifications

In comparison to OCCT, the OMAAP blueprints use similar proportions of items across standards. A standard must have at least six items to serve as a reporting category. Committees of teachers reviewed *PASS* item specifications and OMAAP 3-8 items. Also, the SDE reviewed and revised item specifications.

1.2.b Modification of Forms and Items

Universal- and content-specific OMAAP 3-8 exam modifications are listed in Table 1.2.

DRAFT

Table 1.2. OMAAP Item Modification Rules

Universal
• Minimize the number of questions on the page (limit to 2) • Use a larger font size • Provide only three answer options instead of four • Highlight the main points in the question or passage by underlining and using boldface • Allow for the same accommodations as in the standard assessment • Avoid questions that require students to select the better/best answer • Eliminate answer choices that give students the option of making "no change" to the item
Mathematics
• Allow for read-aloud and calculators • For lower grades, display numbers on all sides of figures for questions about perimeter • Unless required by standard, avoid items with negative and positive answer choices that use the same number (e.g., −4 and +4) • Place any items with coordinate grids on one page • Be consistent with qualifiers in the stem and answer choices (e.g., use ml throughout or milliliters throughout) • Avoid questions that use best or closest • Avoid complicated art
Reading
• Display passages in a one column format • Break passages into smaller portions, and place the questions that pertain to the smaller portion
Science
• Reduce the amount of reading • Avoid complicated art • Simplify tables and charts by removing irrelevant rows or columns • Box formulas to make them stand out

Furthermore, the number of items in the OMAAP forms was reduced to between 50-65% of the number of items on the OCCT exams. Table 1.6 shows a comparison of item counts across the Spring 2011 OCCT and OMAAP forms.

1.3 Summary of Test Development and Content Validity

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

In addition to the test Blueprints provided by the SDE, Table 1.3 describes four criteria for test alignment with the modified *PASS* Standards and Objectives.

DRAFT

Table 1.3. Criteria for Aligning the Test with Modified *PASS* Standards and Objectives

1. Categorical Concurrence	The test is constructed so that there are at least six items measuring each *PASS* standard. The number of items 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. Depth-of Knowledge Consistency	The test is constructed using items from a variety of Depth of Knowledge levels that are consistent with the processes students need in order to demonstrate proficiency for each *PASS* objective.
3. Range of Knowledge Correspondence	The test is constructed so that at least 75% of the objectives for a *PASS* standard have at least one corresponding assessment item.
4. Balance-of-Representation	The test is constructed according to the Test Blueprint which reflects the degree of representation given on the test to each *PASS* standard and/or objective in terms of the percent of total test items measuring each standard and the number of test items measuring each standard and/or objective. The test construction shall yield a balance of representation with an index of 0.7 or higher of assessed objectives related to a standard.
5. Source-of-Challenge	Each test item is constructed in such a way that the major cognitive demand comes directly from the targeted *PASS* objective or concept being assessed, not from specialized knowledge or cultural background that the test-taker may bring to the testing situation.

1.3.b Item Pool Development and Selection

The OMAAP test forms were built using modified versions of previously-used OCCT items. To equate the forms across years, a set of items from the Spring 2010 administration served as anchors for the Spring 2011 administration. Equating is necessary to account for slight year-to-year differences in form difficulty and to maintain comparability across years. Details of the equating procedures are provided in a subsequent section in this document. Content experts also targeted the percentage of items measuring various depth of knowledge (DOK) levels for assembling the tests.

DRAFT

Table 1.4 provides the DOK level percentages for the OMAAP exams. 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.

DRAFT

Table 1.4. Percentage of Items in Depth of Knowledge Levels by Grade - Math

		DOK Level					
		1		2		3	
Subject	**Grade**	**Target**	**Actual**	**Target**	**Actual**	**Target**	**Actual**
Math	3	20-25	47	65-70	44	5-15	9
	4	20-25	37	65-70	58	5-15	5
	5	20-25	28	65-70	67	5-15	5
	6	10-15	28	65-70	60	15-25	12
	7	10-15	19	65-70	72	15-25	9
	8	10-15	28	65-70	63	15-25	9
Reading	3	20-25	28	65-70	63	5-15	9
	4	20-25	14	65-70	79	5-15	9
	5	20-25	7	65-70	86	5-15	9
	6	10-15	14	65-70	63	15-25	23
	7	10-15	10	65-70	72	15-25	18
	8	10-15	20	65-70	68	15-25	12
Science	5	20-25	19	65-70	67	5-15	14
	8	10-15	14	65-70	63	15-25	23

1.3.c Configuration of the Tests

Table 1.5. provides an overview of the number of operational and field test items for the Spring 2011 OSTP-OMAAP 3-8 assessments. The Spring 2011 test is comprised of a core of operational items on each form and set of field test items. The field test items are uniquely developed items that meet the modification requirements (i.e., they are developed from the ground up for OMAAP rather than the result of modifying an OCCT item). These items are embedded in the operational test forms with the intent of building the item bank for future use. Each form of the assessment was spiraled within classrooms to obtain randomly-equivalent samples of examinees for the field test items. The Spring 2011 administration is the first year that field-test items have been included in the OMAAP.

New items are field-tested to build up the item bank with items that have known psychometric properties. In past administrations, newly modified OCCT items are eligible for inclusion in the operational form, but do not have an administrative history in their modified format. Thus, these items must be evaluated for use as scored operational items after their first operational use, and given poor psychometric properties, could be removed from scoring. Adding items that have a field test history in addition to anchor items reducing this risk. 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.

DRAFT

Table 1.6. OCCT and OMAAP 3-8 Item Count Comparison

	OCCT Items		OMAAP Items	
Subject	**OP**	**FT**	**OP**	**FT**
Math	50	10	43	5
Reading	50	10	43	8
Science	45	10	43	5

*Note OP=Operational, FT=Field-test.

1.3.d Operational Items by Content Area

Math. The forms were comprised of 43 items each. There were 15 anchor items on each form, all from previous administrations. The number of items and maximum points possible by content standard is shown in Table 1.7. Math scores were reported by content and there were 6 or more operational items from each reported category on each form. Each item was mapped to one content standard and one objective per content standard.

Table 1.7. Number of Operational and Field Test Items by Content Standard for Math

	Grade					
Standard	**3**	**4**	**5**	**6**	**7**	**8**
1	7(2)	7(0)	11(0)	11(2)	12(2)	14(1)
2	16(0)	15(1)	13(2)	13(2)	9(0)	9(2)
3	6(0)	7(1)	6(2)	6(1)	7(0)	7(1)
4	8(3)	8(2)	6(0)	6(0)	8(2)	6(1)
5	6(0)	6(1)	7(1)	7(0)	7(1)	7(0)
Total	**43**	**43**	**43**	**43**	**43**	**43**

Reading. The forms were comprised of 43 items each. There were 12-15 anchor items on each form, all from previous administrations. The number of items and maximum points possible by content standard is shown in Table 1.8. Reading scores were reported by content standard and there were 7 or more operational items in each reported category. Each item was mapped to one content standard and one objective per content standard.

Table 1.8. Number of Operational and Field Test Items by Content Standard for Reading

	Grade					
Standard	**3**	**4**	**5**	**6**	**7**	**8**
1	-	9(2)	10(2)	7(1)	8(2)	6(1)
2	10 (2)	-	-	-	-	-
3	-	19(4)	16(3)	17(4)	17(4)	17(3)
4	19 (4)	8(1)	10(2)	13(2)	11(1)	13(2)
5	7 (0)	7(1)	7(1)	7(0)	7(1)	7(2)
6	9 (2)	-	-	-	-	-
Total	**43**	**43**	**43**	**43**	**43**	**43**

Science. Each Science form was comprised of 43 items, of which 39 Grade 5 items and 38 Grade 8 items were mapped to a content standard (all items were linked to process standards). The forms were comprised of 43 items each. There were 15 anchor items on each form, all from previous administrations. The number of items and the maximum number of points possible by content standard is shown in Table 1.9 and the number of points possible by process standard is shown in Table 1.12. Scores were reported for content and process standards at the standard level. Each reported standard has 7 or more items. Unlike other

DRAFT

subjects, all Science items were primarily mapped to process standards. With the exception of safety items, all items were also mapped to content standards.

Table 1.9. Number of Items by Content Standard for Science

	Grade	
Standard	**5**	**8**
1	17 (2)	8 (0)
2	12 (1)	8 (1)
3	11 (2)	8 (2)
4	-	8 (2)
5	-	7 (0)
Total	**39**	**38**

Note: Items classified by content standard do not sum to the total number of items on the forms (43) because several items on the forms map to process standards, but not content standards.

Table 1.10. Number of Items by Process Standard for Science

	Grade	
Process	**5**	**8**
1	9 (0)	7 (1)
2	10 (2)	8 (1)
3	10 (2)	15 (0)
4	14 (1)	13 (3)
Total	**43**	**43**

DRAFT

Section 2

Administration of the OMAAP 3-8 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 OMAAP 3-8 assessments in a manner that is consistent with established procedures, monitoring the fair administration of the assessment, and working with the SDE office to address deviations from established assessment administration procedures. The role that district and school faculty members play is essential in the fair and equitable administration of successful OMAAP 3-8 assessments.

2.1 Packaging and Shipping

To provide secure and dependable services for the shipping of the OMAAP 3-8 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 Access™ website, and optionally with data received directly from Oklahoma. Oklahoma educators also use the Access™ site to provide Pearson with the Pre-Identification information needed to print the student identification section on answer documents. Bar-coding of all secure materials during the pre-packaging effort allows for accurate tracking of these materials through the entire packing, delivery, and return process. It also permits Pearson to inventory all materials throughout the packaging and delivery process along with the ability to provide the customer with status updates at any time. Use of handheld radio-frequency scanners in the packaging process help to eliminate the possibility of packing the wrong materials. The proprietary "pick-and-pack" process prompts packaging personnel as to what materials are to go in which shipping box. If the packer tries to pack the wrong item (or number of items into a shipping carton), the system signals an alert.

2.2 Materials Return

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

DRAFT

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

2.3 Materials Discrepancies Process

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

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

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

2.4 Processing Assessment Materials Returned by Schools

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

DRAFT

Section 3

Classical Item Analysis and Results

3.1 Sampling Plan and Data Receipt

3.1.a Sampling Plan

Population data were used for classical and item response theory (IRT) analyses for all tests. Students who meet the criteria may take the modified (OMAAP) exam and were included in the student data file for equating.

3.1.b Data Receipt Activities

After all tests were scored, a data file was provided for item analyses and calibration. A data clean-up process that removed invalid cases, ineligible responses, absent students, and repeat test-takers was completed. A statistical key check was also performed at this time. This 'cleaned' sample was used for classical item analyses, calibration, and equating. Upon receipt of data, a 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 clean-up, exclusionary rules were applied to form the final sample that was used for classical item analyses, calibration, and equating. Any student who had attempted at least five responses was included in the data analyses; however, students who took the Braille form, were a second time test-taker, had invalidated codes, or attended a private school was not included in the equating and scaling processes. The demographic breakdown of the students in the item analysis and calibration sample appears in Table 3.1.

DRAFT

Table 3.1. Demographic Characteristics of Calibration and Equating Sample

Subject/Grade		Total	Female	Male
MATH	3	3240	1199	2041
	4	3612	1344	2268
	5	4217	1615	2602
	6	4143	1597	2546
	7	4214	1568	2646
	8	4001	1470	2531
READING	3	3861	1355	2506
	4	4266	1513	2753
	5	4593	1666	2927
	6	4373	1599	2774
	7	4231	1532	2699
	8	4043	1411	2632
SCIENCE	5	3572	1271	2301
	8	3163	1114	2049

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

DRAFT

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

Subject/Grade		African America	Native American	Hispanic	Asian	Pacific Islander	White	Other
MATH	**3**	470	610	416	16	3	1574	151
	4	523	703	425	39	2	1752	168
	5	622	811	541	25	5	2038	175
	6	602	829	458	31	5	2058	160
	7	604	816	474	18	7	2140	155
	8	568	860	371	15	3	2033	151
READING	**3**	533	712	497	24	4	1902	189
	4	588	825	537	44	3	2060	209
	5	649	867	634	29	5	2216	193
	6	626	863	519	35	5	2154	171
	7	607	825	501	14	7	2124	153
	8	574	863	396	18	4	2039	149
SCIENCE	**5**	508	698	470	21	4	1727	144
	8	439	705	279	16	3	1605	116
	All	947	1403	749	37	7	3332	260

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

Statistical Key Check. Administering items that have only one correct key and are correctly scored is critical for accurate assessment of student performance. 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 correlation less than 0.20
- Distractor *p*-value greater than or equal to 0.40
- Distractor point-biserial correlation greater than or equal to 0.05

Any flagged operational items are submitted for key review by the appropriate Pearson content specialist. Any flagged items that are identified by content experts as having key issues are submitted to the SDE for review before dropping the item from the operational scoring. There were no items identified as having a key issue. Once the keys were verified, a secondary statistical key check and evaluation of items was conducted for the potential of removing items from scoring.

Secondary Statistical Key Check – removal of operational items

As testing materials were returned for scanning and scoring following the Spring 2011 operational administration, a preliminary data file (N must be larger than 200 per item) was created for a statistical key check (a key check using empirical data) and key verification (a

DRAFT

review of test items and keys by Content Specialists). The purpose of these analyses was to confirm the answer keys by scoring and analyzing the multiple-choice items. Classical item analyses were used for this process. For the empirical checks, items meeting the statistical criteria defined in Table 3.2 were flagged and the content of the item was reviewed by a Content Specialist. If there was a possible mis-key, Pearson contacted SDE for evaluation and review.

Table 3.2. Secondary Statistical Key Check Criteria.

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

Dropping Operational Items

Once the statistical key check was complete, a thorough review of the items and their statistics was conducted by the Pearson Research Scientist and Content Specialists reviewing items with poor statistics for potential removal from scoring. After a thorough review of the item statistics, Pearson's review identified the potential need for the removal of items based on poor statistics for SDE evaluation. A final recommendation was made by the SDE to exclude any items from students' scores and are listed by subject in Table 3.3. These items removed from operational scoring for the current operational administration and will also be removed from the OMAAP 3-8 item bank. Once the final set of operationally scored items were agreed upon, classical item analyses were conducted.

DRAFT

Table 3.3. Items Removed from Scoring by Subject.

Subject	Grade	Item Numbers
Math	3	
	4	21,30,45,46
	5	45
	6	15,48
	7	
	8	25,37
Reading	3	6, 41
	4	42,48,49
	5	8,46
	6	
	7	33
	8	5
Science	5	
	8	9,15

*Note: HL = Horizontal Linking or anchor item

Table 3.4 provides the final number of points possible on the 2011 OMAAP 3-8 after the removal of items with poor statistics.

Table 3.4. Total Possible Points on the 2011 OMAAP 3-8 after Item Review

Subject	Grade	Score Points
Math	3	43
	4	39
	5	42
	6	41
	7	43
	8	41
Reading	3	41
	4	40
	5	41
	6	43
	7	42
	8	42
Science	5	43
	8	41

3.2 Classical Item Analyses

Following completion of the data receipt activities and statistical key check, the following classical item analyses were conducted for operational and field test items:

- Percentage of students endorsing each multiple choice option (overall and broken down by gender and ethnicity)
- Overall p-value for each item
- Point-biserial correlation (overall and broken down by gender and ethnicity)
- Point-biserial for non-key response options (overall and broken down by gender and ethnicity)
- Omit percentage per item
- Mean score by response option (overall and broken down by gender and ethnicity)

DRAFT

Once the keys were verified, the item analysis results reviewed, and the dropped items removed, 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 are shown in Table 3.2. Note that students whose tests were invalidated and those students taking the test for a second time were excluded. The test results indicate that the omit rates were smaller than 3% for all subjects, with the exception of Reading Grade 4 where the omit rate was3.39%.

Table 3.2. Test-Level Summaries of Classical Item Analyses

Subject	Grade	Sample Size	Mean	Mean % of Max	Items Points	Mean P	Mean rpb	Omit Min	Omit Max
MATH	03	3240	27.89	0.65	43	0.65	0.39	0.04	0.78
	04	3612	23.48	0.60	39	0.57	0.34	0.10	0.59
	05	4217	22.12	0.53	42	0.52	0.32	0.06	0.68
	06	4143	21.40	0.52	41	0.53	0.30	0.03	1.22
	07	4214	21.25	0.49	43	0.49	0.26	0.06	1.44
	08	4001	19.69	0.48	41	0.46	0.29	0.03	1.24
READING	03	3861	21.06	0.51	41	0.50	0.35	0.26	2.09
	04	4266	21.57	0.54	40	0.52	0.33	0.23	3.39
	05	4593	23.63	0.58	41	0.56	0.36	0.11	1.82
	06	4373	22.70	0.53	43	0.54	0.35	0.23	2.10
	07	4231	23.21	0.55	42	0.56	0.32	0.15	2.75
	08	4043	26.12	0.62	42	0.63	0.34	0.22	1.53
SCIENCE	05	3572	24.23	0.56	43	0.58	0.32	0.04	0.84
	08	3163	23.54	0.57	41	0.58	0.28	0.04	2.61

Note: r_{pb} = point biserial correlation; *item counts after dropping items.

3.3 Procedures for Detecting Item Bias

One of the goals of the OMAAP 3-8 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

DRAFT

I error. As a result, DIF statistics are used only to identify potential sources of item bias. Subsequent review by content experts and bias committees are required to determine the source and meaning of performance differences. For the OMAAP 3-8 test DIF analyses, DIF statistics were estimated for race (African American and white) and gender (male and female). Items with statistically-significant differences in performance were flagged so that items could be carefully examined for possible biased or unfair content that was undetected in earlier fairness and bias content review meetings held prior to form construction.

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

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

- The item is classified as C category if the MH D-DIF is significantly greater than 1.0 in absolute value, and its absolute value is at least 1.5.
- The item is classified as B category if the MH D-DIF is significantly different from zero, its absolute value is at least 1.0, and its absolute value is either less than 1.5 or not significantly greater than 1.0.
- The item is classified as A category if the MH D-DIF is not significantly different from zero ($p \geq 0.05$), or if its absolute value is less than 1.0.

3.3.a Differential Item Functioning Results

The data in Table 3.3 summarize the number of items flagged for possible DIF in each subject Items flagged for DIF were placed before expert content specialists from SDE and Pearson during Spring 2011 prior to inclusion as part of the operational scored set. The review panel reviewed the item content, the percentage of students selecting each response option and the point-biserial correlation for each response option by gender and race for all items flagged for DIF. The data review panel was then asked if there was context (for example, cultural barriers) or language in an item that might result in bias (i.e., an explanation for the existence of the statistical DIF flag). Operational test items that were found to exhibit bias as a result of the content of the item would be removed from scoring and the item bank excluding them from future use. No items were removed from scoring due to DIF.

DRAFT

Table 3.3. DIF Flag Incidence Across All OMAAP 3-8 Items

Subject/Grade		Total items	Female	African American	Hispanic	Asian
MATH	03	43	0	0	0	0
	04	39	0	1	2	0
	05	42	1	0	0	0
	06	41	0	1	0	0
	07	43	0	0	0	0
	08	41	0	0	0	0
READING	03	41	0	0	0	0
	04	40	0	0	0	1
	05	41	0	0	0	0
	06	43	0	1	0	0
	07	42	0	0	0	0
	08	42	0	1	1	0
SCIENCE	05	43	0	0	0	0
	08	41	0	1	1	0

3.4 Test Reliability

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

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

Cronbach's alpha was estimated for each subject and grade.

DRAFT

Table 3.5 presents Cronbach's alpha for each of the OMAAP 3-8 tests. These reliability coefficients indicate that the OMAAP 3-8 assessments had adequate internal consistency and that the tests produce relatively stable scores. Table 3.4 also addresses the reliability analysis results by the different reporting subgroups for the OMAAP 3-8 assessments. This table illustrates the subject, the subgroups, the number of students used in the analyses and the associated Cronbach's Alpha reliability coefficient for each subject and subgroup. With few exceptions, the reliability coefficients are well above the accepted lower limit of .70.

DRAFT

Table 3.5. Cronbach's Alpha by Subject and Grade

Subject	Grade	N	All	Female	Male	African American	Native American	Hispanic	Asian	White	Other
MATH	3	43	0.87	0.85	0.88	0.85	0.86	0.89	0.87	0.88	0.87
	4	43	0.83	0.82	0.83	0.80	0.84	0.72	0.80	0.84	0.82
	5	43	0.80	0.78	0.81	0.78	0.78	0.71	0.90	0.81	0.78
	6	43	0.76	0.75	0.77	0.70	0.78	0.65	0.81	0.77	0.74
	7	43	0.71	0.69	0.73	0.67	0.67	0.61	0.75	0.73	0.72
	8	43	0.78	0.76	0.79	0.74	0.79	0.73	0.93	0.79	0.77
READING	3	43	0.85	0.83	0.86	0.83	0.84	0.86	0.89	0.86	0.85
	4	43	0.82	0.81	0.83	0.80	0.82	0.76	0.85	0.83	0.82
	5	43	0.85	0.84	0.86	0.84	0.85	0.84	0.92	0.86	0.84
	6	43	0.83	0.82	0.84	0.80	0.83	0.80	0.90	0.84	0.83
	7	43	0.79	0.79	0.79	0.75	0.79	0.82		0.80	0.79
	8	43	0.82	0.80	0.82	0.83	0.83	0.72	0.79	0.81	0.81
SCIENCE	5	43	0.81	0.78	0.82	0.78	0.79	0.88	0.86	0.82	0.79
	8	43	0.72	0.66	0.74	0.66	0.71	0.00	0.85	0.73	0.71

DRAFT

Section 4

Calibration, Equating, and Scaling

4.1 Item Response Theory (IRT) Models

4.1.a Item Response Theory Model

Rasch Model. The Rasch model (Rasch, 1960)—also commonly referred to as the one-parameter logistic (1-PL) item response theory (IRT) model—was used for calibrating all OMAAP 3-8 Items. In the Rasch model, the probability that a student with an ability level of θ responds correctly to item i is

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

where b_i is the item difficulty parameter.

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

4.2 Assessment of Fit to the Model

Item fit was assessed using the Winsteps fit function. Winsteps provides two statistics for indicating how well the Rasch model fits test data. Infit is sensitive to aberrations in item response patterns and indicates overfit for "perfect" response patterns and underfit for items targeting different curricula. The infit statistic is expressed as a mean square (MS) statistic and was considered for assessing fit. The expected value is 1.0, with values greater than 1.0 indicating mis-fitting items. For the OMAAP 3-8 equating, infit values greater 1.3 were flagged and examined further.

Operational items flagged for IRT fit that were not flagged by the classical item analyses and had reasonable estimated IRT parameters were not reviewed. Items that were also flagged by classical item analyses and/or had poor IRT parameter estimates (e.g., *b* parameter estimate with absolute value greater than 3) were reviewed by Pearson content specialists. Any item that was potentially mis-keyed was presented to SDE to make a decision regarding whether to keep or remove the item.

In addition to the fit statistics provided by Winsteps, an item fit plot program was run. In this program, the expected and observed item characteristics curves for each item were output to a PDF document for all multiple choice and constructed response items. All item fit plots were examined by the Lead Research Scientist. Items with poor fit were examined for potential removal from scoring. Any item proposed for removal from scoring was approved by the Oklahoma SDE before being removed.

4.2.a Spring 2011

Math. For the Spring 2011 mathematics assessments, based on the calibration sample, zero items were flagged for Grades 3, 4, 5, 6, 7, and 8.

DRAFT

Reading. For the Spring 2011 reading assessments, based on the calibration sample, zero items were flagged for Grades 3, 4, 5, 6, 7, and 8.

Science. For the Spring 2011 science assessments, based on the calibration sample, zero items were flagged for Grades 5 and 8.

No items were dropped from any of the Spring 2011 OMAAP 3-8 assessments for calibration, equating, or scoring as a result of IRT model fit analyses.

4.3 Calibration and Equating

The Rasch model was used for calibration for all OMAAP 3-8 assessments because all of these tests consist of only multiple choice items. A common item, non-equivalent groups (CINEG) design was used for all content areas to link the current test forms (i.e., Spring 2011) to the base year's scale. The horizontal linking items were selected to be representative of the test content in terms of difficulty and the test blueprint. Including horizontal linking items is critical for obtaining results that are comparable from year to year. With this equating design, the mean Rasch difficulty of the horizontal linking items on the current form is compared to their difficulty on the baseline form to derive an equating constant. The equating constant is then added to the *b*-parameter estimates for all items to put them on the same scale as the baseline form.

4.4 Anchor Item Stability Evaluation Methods

Despite the careful selection of anchor items, it is possible for the anchor items to perform differentially across administrations. Dramatic changes in anchor item parameter values can result in systematic errors in equating results (Kolen & Brennan, 2004). As a result, prior to finalizing the equating constant, Pearson evaluated changes in the item parameters from the Spring 2010 operational administration to the Spring 2011 administration. The process used in this evaluation is referred to as an anchor item parameter stability check.

The OMAAP 3-8 anchor stability check approach used an iterative post-equating stability check procedure to eliminate test items whose Rasch item difficulty calibration differed more than expected from the pre-equated value. This iterative approach used procedures that are relatively common in testing programs employing the Rasch model; it combined the displacement and Robust Z approaches.

Each anchor item has two parameters, the $b_{baseline}$ and $b_{2011calibrated}$ and these are both used to compute the displacement value. Once an equating constant is computed and the $b_{2011calibrated}$ is brought to the base scale, b_{scaled} the displacement value can be computed. The displacement value is the absolute value of the difference (*D*) between the $b_{baseline}$ and b_{scaled}. An anchor item is flagged when the *D* is larger than 0.30 logits. If, upon further review, it is decided that an item should be eliminated from the anchor item set then the equating constant is re-calculated, reapplied, and the stability check process is repeated until all *D* values are within desirable range or when 20% of anchor items are dropped.

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

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

$$\textit{Robust } z_j = \frac{\left(bdif_i - MDdif_j\right)}{\left(IQ_j \times 0.74\right)}$$

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

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

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

1. Compute the equating constant, displacement, and robust z.
2. Flag items with displacement > 0.30.
3. Sort items in this order:
   - Displacement Flag (descending – assumes flag = 1, else = blank), then
   - Absolute value of Robust Z (descending), and then
   - Item Sequence (ascending – for a tie-breaker)
4. Drop flagged item with largest absolute value of Robust Z.
5. Recompute scaling constant and displacement based on the new anchor set. <u>Do not</u> recompute robust z.
6. Flag items with displacement > 0.30.
7. Sort items and drop one (if necessary) based on criteria outlined above.
8. Stop criteria – Stop if either of the following occur:
   - Anchor set is 20% of the total test, OR
   - No additional items are flagged based on displacement.

The order for dropping items from the anchor set will occur based on collective rank ordering of the items change from the two approaches. Decisions about whether to keep or remove an item will be evaluated on a per item basis. If an item (note, only one item can be removed at a time) is removed from the set, then this process (beginning at the equating step) will be repeated until there are no further items to be removed. Even though an item may be

DRAFT

removed for equating purpose, it will still contribute to student scores. Items flagged for removal during the anchor stability check, will also be evaluated using the following factors:

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

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

4.4.a Anchor Items for Spring 2011

Table 4.1 presents the number and percentages of horizontal linking items by subject for the Spring 2011 administration. For all subjects and grades, the anchor set was comprised of 37% of all operational items. 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.

DRAFT

Table 4.1. Number of Horizontal Linking Items Per Subject and Grade

Subject	Grade	Number of Items on Test	Number of Linking Items	Percent of Test
Math	3	43	15	35%
	4	39	14	36%
	5	42	13	31%
	6	41	13	32%
	7	43	12	28%
	8	41	13	32%
Reading	3	41	12	29%
	4	40	14	35%
	5	41	14	34%
	6	43	14	33%
	7	42	11	26%
	8	42	13	31%
Science	5	43	14	33%
	8	41	14	34%

*Note: one linking item was dropped from scoring

4.4.b Results of the Anchor Item Stability Check

Once the anchor set was finalized, the equating constant was applied to the non-anchor items for computation of raw score to scale score tables. Table 4.1 displays the number of items dropped during the anchor stability check. Any item removed from the anchor set still contributed to student scores.

DRAFT

Table 4.2. Number of Horizontal Linking Items Dropped during the Anchor Stability Check

Subject	Grade	Number of Items Dropped from Linking Study
Math	3	0
	4	1
	5	2
	6	2
	7	3
	8	2
Reading	3	2
	4	1
	5	1
	6	1
	7	1
	8	2
Science	5	1
	8	1

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 4.3. The scaling constants, *M*1 (multiplicative) and *M*2 (additive), place the true scores associated with each raw score point onto the reporting or operational scale using a straightforward linear transformation:

$$\text{Scale Score} = \left(\hat{\tau} \times M1\right) + M2 \tag{3}$$

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

The raw score to number-correct scales scores were generated using Winsteps using the equated Rasch Item Difficulties (RIDs). 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 OMAAP 3-8 assessments, there are 3 cut scores that divide scores into 4 performance levels: Unsatisfactory, Limited Knowledge, Satisfactory, and Advanced. The cut scores for each of the tests appear in Table 4.4. In addition, a conditional standard error of measurement (*CSEM*; please see Section 6.3 for computation of *CSEM*) was computed for each of the raw score points. The resulting raw score to scale score conversions, *CSEM*s, and performance levels are shown in Table 4.5 through

DRAFT

Table 4.9.

DRAFT

Table 4.3. LOSS, HOSS, and Scaling Constants by Subject and Grade

Subject	Grade	LOSS	HOSS	*M*1	*M*2
Math	3	100	350	19.94	244.59
	4	100	350	20.09	247.68
	5	100	350	20.31	243.66
	6	100	350	19.36	245.86
	7	100	350	20.53	240.05
	8	100	350	20.15	244.33
Reading	3	100	350	20.89	247.50
	4	100	350	20.75	247.42
	5	100	350	20.73	248.66
	6	100	350	21.77	242.02
	7	100	350	21.82	242.05
	8	100	350	21.31	241.48
Science	5	100	350	20.92	245.91
	8	100	350	20.13	251.39

Table 4.4. Performance-Level Cut Scores by Content Area

		Cut Scores		
Subject	Grade	Limited Knowledge	Satisfactory	Advanced
Math	3	233	250	277
	4	238	250	277
	5	240	250	271
	6	237	250	272
	7	232	250	265
	8	235	250	271
Reading	3	238	250	269
	4	237	250	266
	5	231	250	269
	6	229	250	261
	7	229	250	271
	8	236	250	276
Science	5	238	250	277
	8	241	250	288

DRAFT

Table 4.5. Raw Score to Scale Score Conversion Table for Math – Grades 3-5

	Grade 3			Grade 4			Grade 5		
Raw Score	OPI Scale	Perf. Level	CSEM	OPI Scale	Perf. Level	CSEM	OPI Scale	Perf. Level	CSEM
0	139	1	37	146	1	37	146	1	37
1	164	1	20	171	1	21	172	1	21
2	178	1	15	186	1	15	187	1	15
3	187	1	12	195	1	13	196	1	12
4	194	1	11	202	1	11	202	1	11
5	199	1	10	208	1	10	208	1	10
6	203	1	9	212	1	9	212	1	9
7	207	1	9	216	1	9	216	1	9
8	211	1	8	220	1	8	220	1	8
9	214	1	8	223	1	8	224	1	8
10	217	1	8	227	1	8	227	1	8
11	220	1	7	230	1	8	230	1	8
12	223	1	7	232	1	7	232	1	7
13	225	1	7	235	1	7	235	1	7
14	228	1	7	238	2	7	238	1	7
15	230	1	7	240	2	7	240	2	7
16	233	2	7	242	2	7	242	2	7
17	235	2	7	245	2	7	245	2	7
18	237	2	7	247	2	7	247	2	7
19	239	2	7	250	3	7	250	3	7
20	242	2	7	252	3	7	252	3	7
21	244	2	7	254	3	7	254	3	7
22	246	2	7	256	3	7	256	3	7
23	250	3	7	259	3	7	259	3	7
24	251	3	7	261	3	7	261	3	7
25	253	3	7	264	3	7	263	3	7
26	255	3	7	266	3	7	265	3	7
27	257	3	7	269	3	7	268	3	7
28	260	3	7	271	3	7	271	4	7
29	262	3	7	274	3	8	273	4	7
30	265	3	7	277	4	8	275	4	7
31	267	3	7	281	4	8	278	4	8
32	270	3	8	284	4	9	281	4	8
33	273	3	8	288	4	9	284	4	8
34	277	4	8	293	4	10	288	4	8

DRAFT

35	280	4	8	298	4	11	291	4	9
36	283	4	9	304	4	12	295	4	9
37	287	4	9	313	4	15	300	4	10
38	292	4	10	328	4	20	305	4	11
39	297	4	11	350	4	37	312	4	12
40	304	4	12	-	-	-	321	4	15
41	313	4	15	-	-	-	336	4	21
42	328	4	20	-	-	-	350	4	37
43	350	4	37	-	-	-	-	-	-

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

DRAFT

Table 4.6. Raw Score to Scale Score Conversion Table for Math – Grades 6-8

	Grade 6			Grade 7			Grade 8		
Raw Score	OPI Scale	Perf. Level	CSEM	OPI Scale	Perf. Level	CSEM	OPI Scale	Perf. Level	CSEM
0	152	1	36	138	1	38	149	1	37
1	176	1	20	163	1	21	174	1	21
2	191	1	14	179	1	15	189	1	15
3	199	1	12	188	1	13	198	1	12
4	206	1	11	195	1	11	205	1	11
5	211	1	10	201	1	10	210	1	10
6	215	1	9	205	1	10	214	1	9
7	219	1	8	210	1	9	218	1	9
8	223	1	8	213	1	9	222	1	8
9	226	1	8	217	1	8	225	1	8
10	229	1	8	220	1	8	228	1	8
11	232	1	7	223	1	8	231	1	7
12	235	1	7	226	1	8	235	2	7
13	237	2	7	229	1	7	236	2	7
14	240	2	7	232	2	7	239	2	7
15	242	2	7	234	2	7	241	2	7
16	244	2	7	236	2	7	244	2	7
17	247	2	7	239	2	7	246	2	7
18	250	3	7	241	2	7	248	2	7
19	251	3	7	243	2	7	250	3	7
20	253	3	7	246	2	7	253	3	7
21	256	3	7	248	2	7	255	3	7
22	258	3	7	250	3	7	257	3	7
23	260	3	7	252	3	7	259	3	7
24	262	3	7	255	3	7	261	3	7
25	265	3	7	257	3	7	264	3	7
26	267	3	7	259	3	7	266	3	7
27	269	3	7	262	3	7	268	3	7
28	272	4	7	265	4	7	271	4	7
29	274	4	7	266	4	7	273	4	7
30	277	4	7	269	4	7	276	4	7
31	280	4	7	272	4	7	279	4	8
32	283	4	8	274	4	8	282	4	8
33	286	4	8	277	4	8	285	4	8
34	290	4	8	280	4	8	289	4	9

DRAFT

35	294	4	9	284	4	8	292	4	9
36	298	4	10	287	4	9	297	4	10
37	303	4	11	291	4	9	302	4	11
38	310	4	12	296	4	10	309	4	12
39	318	4	14	301	4	11	318	4	15
40	333	4	20	308	4	13	333	4	21
41	350	4	36	317	4	15	350	4	37
42	-	-	-	332	4	21	-	-	-
43	-	-	-	350	4	38	-	-	-

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

DRAFT

Table 4.7. Raw Score to Scale Score Conversion Table for Reading – Grades 3-5

	Grade 3			Grade 4			Grade 5		
Raw Score	OPI Scale	Perf. Level	CSEM	OPI Scale	Perf. Level	CSEM	OPI Scale	Perf. Level	CSEM
0	150	1	38	153	1	38	146	1	38
1	176	1	21	178	1	21	171	1	21
2	191	1	15	193	1	15	186	1	15
3	200	1	13	202	1	13	195	1	13
4	206	1	11	209	1	11	202	1	11
5	212	1	10	214	1	10	207	1	10
6	216	1	9	219	1	9	212	1	9
7	220	1	9	223	1	9	216	1	9
8	224	1	8	226	1	8	219	1	8
9	227	1	8	230	1	8	222	1	8
10	230	1	8	233	1	8	225	1	8
11	233	1	8	237	2	8	228	1	8
12	235	1	7	238	2	7	231	2	7
13	238	2	7	241	2	7	233	2	7
14	240	2	7	243	2	7	236	2	7
15	243	2	7	245	2	7	238	2	7
16	245	2	7	248	2	7	240	2	7
17	247	2	7	250	3	7	243	2	7
18	250	3	7	252	3	7	245	2	7
19	251	3	7	255	3	7	247	2	7
20	254	3	7	257	3	7	250	3	7
21	256	3	7	259	3	7	251	3	7
22	258	3	7	261	3	7	253	3	7
23	260	3	7	263	3	7	256	3	7
24	262	3	7	266	4	7	258	3	7
25	264	3	7	268	4	7	260	3	7
26	267	3	7	270	4	7	262	3	7
27	269	4	7	273	4	7	265	3	7
28	271	4	7	275	4	7	269	4	7
29	274	4	7	278	4	8	270	4	7
30	276	4	7	281	4	8	272	4	8
31	279	4	8	284	4	8	275	4	8
32	282	4	8	287	4	8	278	4	8
33	285	4	8	291	4	9	281	4	8
34	289	4	9	294	4	9	285	4	9

DRAFT

35	293	4	9	299	4	10	289	4	9
36	297	4	10	304	4	11	294	4	10
37	303	4	11	311	4	13	299	4	11
38	309	4	13	320	4	15	306	4	13
39	318	4	15	335	4	21	315	4	15
40	333	4	21	350	4	38	330	4	21
41	350	4	38	-	-	-	350	4	38
42	-	-	-	-	-	-	-	-	-
43	-	-	-	-	-	-	-	-	-

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

DRAFT

Table 4.8. Raw Score to Scale Score Conversion Table for Reading – Grades 6-8

	Grade 6			Grade 7			Grade 8		
Raw Score	OPI Scale	Perf. Level	CSEM	OPI Scale	Perf. Level	CSEM	OPI Scale	Perf. Level	CSEM
0	142	1	40	134	1	40	134	1	39
1	169	1	22	161	1	22	160	1	22
2	185	1	16	178	1	16	176	1	16
3	194	1	13	188	1	14	186	1	13
4	201	1	12	195	1	12	193	1	12
5	207	1	11	201	1	11	198	1	11
6	212	1	10	206	1	10	203	1	10
7	216	1	9	211	1	10	208	1	9
8	219	1	9	215	1	9	211	1	9
9	223	1	8	218	1	9	215	1	8
10	226	1	8	222	1	8	218	1	8
11	229	2	8	225	1	8	221	1	8
12	232	2	8	229	2	8	224	1	8
13	234	2	7	230	2	8	227	1	8
14	237	2	7	233	2	8	230	1	7
15	239	2	7	236	2	7	232	1	7
16	242	2	7	238	2	7	236	2	7
17	244	2	7	241	2	7	237	2	7
18	246	2	7	243	2	7	240	2	7
19	250	3	7	245	2	7	242	2	7
20	251	3	7	248	2	7	244	2	7
21	253	3	7	250	3	7	247	2	7
22	255	3	7	252	3	7	250	3	7
23	257	3	7	255	3	7	251	3	7
24	261	4	7	257	3	7	253	3	7
25	262	4	7	259	3	7	256	3	7
26	264	4	7	262	3	7	258	3	7
27	266	4	7	264	3	7	261	3	7
28	268	4	7	267	3	7	263	3	7
29	271	4	7	271	4	8	266	3	7
30	273	4	7	272	4	8	268	3	8
31	276	4	8	275	4	8	271	3	8
32	278	4	8	278	4	8	276	4	8
33	281	4	8	281	4	8	277	4	8
34	284	4	8	284	4	9	281	4	9

DRAFT

35	288	4	9	288	4	9	284	4	9
36	291	4	9	292	4	10	289	4	10
37	295	4	10	297	4	11	293	4	10
38	300	4	10	302	4	12	299	4	11
39	306	4	12	309	4	13	306	4	13
40	313	4	13	319	4	16	315	4	16
41	322	4	16	335	4	22	331	4	22
42	338	4	22	350	4	40	350	4	39
43	350	4	40	-	-	-	-	-	-

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

DRAFT

Table 4.9. Raw Score to Scale Score Conversion Table for Science – Grades 5 and 8

	Grade 5			Grade 8		
Raw Score	OPI Scale	Perf. Level	CSEM	OPI Scale	Perf. Level	CSEM
0	144	1	38	151	1	37
1	170	1	21	176	1	21
2	186	1	15	192	1	15
3	195	1	13	201	1	13
4	202	1	11	208	1	11
5	207	1	10	214	1	10
6	212	1	10	219	1	10
7	216	1	9	224	1	9
8	220	1	9	228	1	9
9	224	1	8	231	1	8
10	227	1	8	235	1	8
11	230	1	8	238	1	8
12	232	1	8	241	2	8
13	235	1	7	244	2	8
14	238	2	7	246	2	7
15	240	2	7	250	3	7
16	242	2	7	252	3	7
17	245	2	7	254	3	7
18	247	2	7	257	3	7
19	250	3	7	259	3	7
20	252	3	7	262	3	7
21	254	3	7	264	3	7
22	256	3	7	266	3	7
23	258	3	7	269	3	7
24	260	3	7	271	3	7
25	263	3	7	274	3	7
26	265	3	7	276	3	7
27	267	3	7	279	3	7
28	269	3	7	281	3	7
29	272	3	7	284	3	7
30	274	3	7	288	4	8
31	277	4	7	290	4	8
32	280	4	8	293	4	8
33	283	4	8	296	4	8
34	286	4	8	300	4	9

DRAFT

35	289	4	8	304	4	9
36	292	4	9	308	4	10
37	296	4	9	314	4	11
38	301	4	10	320	4	12
39	307	4	11	329	4	15
40	313	4	13	344	4	21
41	323	4	15	350	4	37
42	338	4	21	-	-	-
43	350	4	38	-	-	-

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

DRAFT

Section 5

Classification Consistency and Accuracy Studies

5.1 Classification Consistency and Accuracy

Every test administration will result in some error in classifying examinees. The concept of the standard error of measurement (SEM) has implications for the interpretation of cut scores used to classify students into different performance levels. For example, a given student may have a true performance level greater than a cut score. However, due to random variations (measurement error), the student's observed test score may be below the cut score. As a result, the student may be classified as having a lower performance level. As discussed in Section 6.4, a student's observed score is most likely to fall 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—the true score—is not available, Livingston and Lewis provide a method to estimate the true score distribution of a test and create the cross-tabulation of the true score and observed variable (raw score) classifications. Consistency is "the agreement between classifications based on two non-overlapping, equally-difficult forms of the test" (p. 180). Consistency is estimated using actual response data from a test and the test's reliability to statistically model two parallel forms of the test and compare the classifications on those alternate forms. There are three types of accuracy and consistency indices that can be generated using Livingston and Lewis' approach: overall, conditional on level, and by cut score.

The overall accuracy of performance level classifications is computed as a sum of the proportions on the diagonal of the joint distribution of true score- and observed score levels. Essentially, overall accuracy is a proportion (or percentage) of correct classifications across all levels. The overall consistency index is computed as the sum of the diagonal cells in a consistency table. Another way to express overall consistency is to use the kappa coefficient, which is commonly used to assess inter-rater reliability. 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, students with satisfactory true ability levels who were classified as satisfactory by the test) and the proportion of all the students classified into that level (total proportion of students who were classified as satisfactory). Accuracy conditional on performance level is computed in a similar manner except that in the consistency table where both row and column marginal sums are the same, the accuracy table uses the sum based on estimated status as the total for computing accuracy conditional on performance level.

DRAFT

To evaluate decisions at specific cut scores, the joint distribution of all the performance levels are collapsed into dichotomized distributions around that specific cut score (for example collapsing Unsatisfactory and Limited Knowledge and then Satisfactory and Advanced to assess decisions at the Satisfactory cut score). The accuracy index at cut score is computed as the sum of the proportions of correct classifications around this selected cut score. The consistency at a specific cut score is obtained in a similar way, but by dichotomizing the distributions at the cut score performance level and between all other performance levels combined. Table 5.1 presents the overall accuracy and consistency indices for the OMAAP 3-8 tests.

Table 5.1. Estimates of Accuracy and Consistency of Performance Classification

Subject	Grade	Accuracy	Consistency	Kappa	False Positive	False Negative
MATH	3	0.77	0.68	0.51	0.12	0.12
	4	0.74	0.64	0.44	0.13	0.13
	5	0.69	0.59	0.4	0.16	0.15
	6	0.71	0.6	0.35	0.15	0.14
	7	0.64	0.52	0.29	0.19	0.16
	8	0.69	0.57	0.37	0.16	0.15
READING	3	0.69	0.59	0.44	0.16	0.15
	4	0.71	0.61	0.43	0.14	0.15
	5	0.74	0.64	0.48	0.13	0.13
	6	0.72	0.63	0.44	0.13	0.15
	7	0.71	0.6	0.41	0.15	0.14
	8	0.67	0.57	0.35	0.16	0.16
SCIENCE	5	0.75	0.65	0.41	0.13	0.12
	8	0.78	0.69	0.19	0.19	0.02

As shown in Table 5.1, the overall accuracy indices range between 64 and 77 percent and overall consistency ranges between 52 and 29 percent. Kappa coefficients range from 0.19 and 0.51. The rate of false positives range from 12 to 19 percent. The false negative rates ranged from 2 to 16 percent.

Table 5.2 provides the accuracy-, consistency-, false positive-, and false negative rates by cut-score. The data in these tables reveal that the level of agreement for both accuracy and consistency is above 75 percent in all cases, with most at or above 90 percent. In general, the high rates of accuracy and consistency support the cut decisions made using these assessments. Similar to Table 5.1, the false positive- and false negative rates are quite low.

The importance of the dichotomous categorization is particularly notable when they map onto pass/fail decisions for the assessments. For the OMAAP 3-8 tests, the U+L/S+A is the important dichotomization, because it directly translates to the pass/fail decision point. Similar to other dichotomization distinctions, there are three main scenarios at this cut point: 1) observed performance is accurately reflective of the true ability level (i.e., the examinee passed and should have passed); 2) the true ability level is below the standard, but the observed test score is above the standard (i.e., a false positive); and 3) the true ability level is above the standard, but the observed test score is below the standard (i.e., a false negative). In examining Table 5.2, Math Grade 3, for example, 92 percent of students are correctly classified as pass or fail based on their performance (scenario 1), 4 percent passed, but their true ability is below the standard (scenario 2), and 5 percent failed although their

DRAFT

true ability is above the standard (scenario 3). Overall, the accuracy rates for accurate classification are above 82% for all tests – students are appropriately (more than 72% of the time) categorized into pass/fail classifications based on their true ability using their observed score (raw score) as their classification score.

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

		Accuracy			Consistency		
Subject	Grade	U / L+P+A	U+L / P+A	U+L+P / A	U / L+P+A	U+L / P+A	U+L+P / A
MATH	3	0.96	0.92	0.89	0.95	0.88	0.85
	4	0.95	0.9	0.89	0.92	0.86	0.85
	5	0.93	0.87	0.89	0.89	0.82	0.85
	6	0.95	0.88	0.89	0.92	0.83	0.84
	7	0.95	0.82	0.88	0.91	0.76	0.84
	8	0.94	0.85	0.91	0.9	0.8	0.87
READING	3	0.91	0.88	0.91	0.87	0.83	0.88
	4	0.96	0.88	0.87	0.94	0.83	0.82
	5	0.96	0.89	0.89	0.94	0.84	0.84
	6	0.97	0.88	0.87	0.95	0.83	0.82
	7	0.97	0.86	0.88	0.95	0.81	0.83
	8	0.96	0.92	0.81	0.94	0.88	0.75
SCIENCE	5	0.96	0.91	0.89	0.94	0.86	0.85
	8	0.98	0.97	0.83	0.96	0.94	0.77

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

DRAFT

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

		False Positives			False Negatives		
Subject	Grade	U / L+S+A	U+L / S+A	U+L+S / A	U / L+S+A	U+L / S+A	U+L+S / A
Math	3	0.02	0.04	0.06	0.02	0.05	0.04
	4	0.02	0.04	0.07	0.03	0.06	0.04
	5	0.03	0.06	0.07	0.04	0.07	0.04
	6	0.02	0.05	0.08	0.03	0.07	0.04
	7	0.02	0.08	0.09	0.03	0.09	0.03
	8	0.03	0.07	0.07	0.03	0.08	0.03
Reading	3	0.05	0.06	0.05	0.05	0.06	0.03
	4	0.02	0.05	0.07	0.02	0.07	0.05
	5	0.02	0.05	0.07	0.02	0.06	0.04
	6	0.01	0.05	0.07	0.02	0.06	0.06
	7	0.01	0.06	0.08	0.01	0.07	0.04
	8	0.02	0.03	0.11	0.02	0.05	0.08
Science	5	0.02	0.04	0.07	0.02	0.05	0.04
	8	0.01	0.02	0.17	0.01	0.02	0.00

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

DRAFT

Section 6

Summary Statistics

6.1 Descriptive Statistics

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

Table 6.1. Descriptive Statistics of the Scale Scores - Overall

		Total			
Subject	**Grade**	***N***	**Mean**	***SD***	**Med.**
Math	3	3240	260.5	33.6	262
	4	3612	259.7	31.4	261
	5	4217	254.9	31.1	256
	6	4143	254.0	33.6	256
	7	4214	245.6	33.1	248
	8	4001	247.5	39.6	250
Reading	3	3861	255.0	30.4	254
	4	4266	259.8	27.2	259
	5	4593	257.0	31.2	258
	6	4373	254.6	34.8	255
	7	4231	253.8	31.4	255
	8	4043	256.4	40.1	261
Science	5	3572	259.1	34.0	260
	8	3163	265.7	44.2	271

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

DRAFT

Table 6.2. Descriptive Statistics of the Scale Scores by Gender

Subject	Grade	Female				Male			
		N	Mean	*SD*	Med.	*N*	Mean	*SD*	Med.
Math	3	1199	260.1	27.3	260	2041	260.8	36.7	262
	4	1344	258.9	29.8	261	2268	260.2	32.4	261
	5	1615	253.8	31.0	256	2602	255.5	31.2	256
	6	1597	253.7	33.0	256	2546	254.2	34.0	258
	7	1568	245.6	30.8	248	2646	245.6	34.4	248
	8	1470	248.4	37.6	253	2531	247.0	40.8	250
Reading	3	1355	256.2	26.0	254	2506	254.4	32.5	251
	4	1513	260.9	26.6	261	2753	259.3	27.5	259
	5	1666	259.3	29.3	260	2927	255.7	32.2	256
	6	1599	257.3	30.4	257	2774	253.0	37.0	255
	7	1532	256.1	31.4	257	2699	252.5	31.4	252
	8	1411	259.7	40.3	263	2632	254.7	39.9	258
Science	5	1271	257.5	33.6	260	2301	260.0	34.1	260
	8	1114	265.2	42.3	269	2049	266.0	45.2	271

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

DRAFT

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

		African-American				Native American			
Subject	**Grade**	***N***	**Mean**	***SD***	**Med.**	***N***	**Mean**	***SD***	**Med.**
Math	3	470	255.8	27.9	257	610	259.3	38.2	260
	4	523	256.1	28.9	256	703	260.1	30.8	261
	5	622	254.5	21.2	254	811	254.7	29.9	256
	6	602	248.9	36.4	253	829	254.4	34.3	258
	7	604	241.4	37.9	246	816	246.2	28.9	248
	8	568	244.3	38.6	247	860	250.0	34.8	253
Reading	3	533	252.9	23.9	251	712	252.9	35.9	251
	4	588	255.4	29.2	255	825	260.4	24.9	259
	5	649	255.9	24.1	256	867	256.3	33.0	258
	6	626	251.4	33.1	252	863	255.4	34.0	255
	7	607	248.4	35.5	250	825	256.0	26.6	257
	8	574	252.5	39.8	257	863	258.6	35.7	263
Science	5	508	254.7	31.9	256	698	257.6	39.0	260
	8	439	261.4	42.0	266	705	268.1	40.9	274

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

DRAFT

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

		Hispanic				Asian			
Subject	**Grade**	***N***	**Mean**	***SD***	**Med.**	***N***	**Mean**	***SD***	**Med.**
Math	3	416	261.6	22.6	260	16	266.4	29.2	267
	4	425	257.8	30.8	259	39	258.5	19.2	256
	5	541	254.7	29.4	256	25	262.3	23.8	256
	6	458	254.3	27.9	256	31	244.9	47.2	250
	7	474	246.4	30.9	248	18	237.0	60.8	247
	8	371	242.7	46.8	248	15	245.1	70.8	261
Reading	3	497	253.6	21.2	251	24	257.9	23.0	254
	4	537	256.8	28.3	257	44	261.2	19.4	254
	5	634	256.1	20.6	256	29	260.7	21.8	258
	6	519	252.4	31.8	253	35	255.6	20.0	253
	7	501	252.7	22.4	252	14	257.4	18.8	252
	8	396	253.6	38.2	258	18	245.4	62.5	260
Science	5	470	257.0	30.8	258	21	260.6	19.8	260
	8	279	263.2	38.4	266	16	272.1	19.1	271

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

DRAFT

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

		White				Other			
Subject	**Grade**	***N***	**Mean**	***SD***	**Med.**	***N***	**Mean**	***SD***	**Med.**
Math	3	1574	262.5	35.1	265	151	256.8	37.2	255
	4	1752	261.0	32.9	261	168	261.1	29.0	261
	5	2038	255.1	34.5	256	175	253.6	32.2	254
	6	2058	255.4	33.4	258	160	256.4	25.2	256
	7	2140	246.2	33.9	248	155	248.0	23.6	248
	8	2033	248.1	40.9	253	151	250.4	25.0	250
Reading	3	1902	257.0	31.0	256	189	253.3	38.1	251
	4	2060	261.7	27.3	261	209	259.8	25.6	259
	5	2216	257.4	35.4	258	193	261.9	22.3	262
	6	2154	255.7	36.5	257	171	254.1	32.8	257
	7	2124	254.9	33.4	257	153	252.0	34.2	255
	8	2039	257.4	42.0	263	149	255.1	40.7	258
Science	5	1727	261.4	33.6	263	144	261.6	28.3	263
	8	1605	266.1	47.6	271	116	266.5	37.8	271

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

DRAFT

Table 6.4. Descriptive Statistics of the Scale Scores by Free/Reduced Lunch Status

		Free/Reduced Lunch = Yes				Free/Reduced Lunch = No			
Subject	**Grade**	***N***	**Mean**	***SD***	**Med.**	***N***	**Mean**	***SD***	**Med.**
Math	3	782	262.8	34.9	265	2458	259.8	33.1	260
	4	917	261.1	33.7	264	2695	259.2	30.6	259
	5	1888	256.0	31.3	256	2329	253.9	31.0	254
	6	1123	256.3	32.2	260	3020	253.2	34.1	256
	7	1234	244.1	41.5	250	2980	246.2	28.9	248
	8	1906	247.9	41.2	253	2095	247.2	38.1	250
Reading	3	1048	258.7	30.3	258	2813	253.7	30.3	251
	4	1158	263.0	27.9	263	3108	258.7	26.8	259
	5	2008	259.1	29.9	258	2585	255.4	32.2	256
	6	1220	258.4	32.9	261	3153	253.1	35.4	253
	7	1206	255.3	36.3	257	3025	253.2	29.2	255
	8	1922	257.2	41.7	263	2121	255.7	38.6	261
Science	5	1553	260.5	33.5	263	2019	258.0	34.2	260
	8	1630	265.6	46.6	271	1533	265.8	41.5	271

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

6.2 Performance Level Distribution

The distributions of students in the four performance levels are presented in Table 6.5 (please see Appendix B for distributions by scale score). As above, these percentages exclude invalid student data and second-time test-takers. The percentage distributions for each of the content areas are comparable to previous administrations although students scoring satisfactory or above decreased slightly from 2010 for all grades and subjects except for Math Grade 3 and Science Grade 5, which saw slight increases.

DRAFT

Table 6.5. Percentage of Students by Performance Level

Subject	Grade	*N*	Unsatisfactory	Limited Knowledge	Satisfactory	Advanced
Math	3	3209	6.4%	18.1%	47.8%	27.7%
	4	3579	7.6%	17.7%	51.9%	22.8%
	5	4172	10.9%	19.9%	47.0%	22.1%
	6	4085	5.4%	20.4%	56.9%	17.3%
	7	4151	5.8%	46.5%	33.2%	14.5%
	8	3916	7.3%	37.2%	41.1%	14.5%
Reading	3	3829	12.0%	26.6%	34.4%	27.0%
	4	4240	3.5%	24.5%	31.0%	40.9%
	5	4553	3.8%	28.7%	33.3%	34.2%
	6	4314	2.2%	28.3%	25.0%	44.5%
	7	4186	1.9%	33.4%	40.5%	24.2%
	8	3966	6.5%	17.1%	49.7%	26.6%
Science	5	3526	3.7%	16.1%	60.0%	20.1%
	8	3087	9%	2.4%	79.1%	17.6%

6.3 Conditional Standard Error of Measurement

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

$$CSEM\,(O_X \mid \theta) = \sqrt{\left[\sum_{X=0}^{MaxX} O_X^2\, p(X \mid \theta)\right] - \left[\sum_{X=0}^{MaxX} O_X \cdot p(X \mid \theta)\right]^2} \qquad (4)$$

where O_X is the observed scaled score for a particular number-correct 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 $CSEM(O_x \mid \theta)$ in which $p(X \mid \theta)$ is computed using a recursive algorithm given by Thissen, Pommerich, Billeaud, and Williams (1995). This algorithm is a polytomous generalization of the algorithm for dichotomous items given by Lord and Wingersky (1984). The values of 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-correct score, X). There is one *CSEM* per number-correct score. The *CSEM*s by subject appear in Table 4.5 through

DRAFT

Table 4.9.

6.4 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, can be used to construct a range around any given observed test score that likely includes the student's true score. *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:

$$SEM = \sqrt{\frac{\sum_j (CSEM_j^{\,2} \cdot N_j)}{N_T}} \tag{5}$$

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 6.6 presents the overall estimates of *SEM* for each of the content areas.

Table 6.6. Overall Estimates of *SEM* by Subject and Grade

Subject	Grade	SEM
Math	3	2.81
	4	2.74
	5	2.70
	6	2.56
	7	2.68
	8	2.69
Reading	3	2.69
	4	2.71
	5	2.71
	6	2.70
	7	2.74
	8	2.80
Science	5	2.82
	8	2.73

Note: *SEM* = Standard Error of Measurement.

DRAFT

References

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

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

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

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

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

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

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

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

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

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

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

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

South Carolina Department of Education. (2001). Technical documentation for the 2000 Palmetto achievement challenge tests of English language arts and mathematics (Technical Report). Columbia: South Carolina Department of Education.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.

DRAFT

Appendix A

Standards, Objectives/Skills, and Processes Assessed by Subject

DRAFT

Grade 3 Mathematics
Algebraic Reasoning: Patterns and Relationships
Algebra Patterns (1.1)
Equations (1.2)
Number Properties (1.3)
Number Sense and Operation
Number Sense (2.1)
Number Operations (2.2)
Geometry
Properties of shapes (3.1)
Spatial Reasoning (3.2)
Coordinate Geometry (3.3)
Measurement
Measurement (4.1)
Time and Temperature (4.2)
Money (4.3)
Data Analysis
Data Analysis (5.1)
Probability (5.2)

DRAFT

Grade 4 Mathematics
Algebraic Reasoning: Patterns and Relationships
Algebra Patterns (1.1)
Equations (1.2)
Number Properties (1.3)
Number Sense and Operation
Number Sense (2.1)
Number Operations (2.2)
Geometry
Lines (3.1)
Angles (3.2)
Polygons (3.3)
Transformations (3.4)
Measurement
Measurement (4.1)
Time and Temperature (4.2)
Money (4.3)
Data Analysis
Data Analysis (5.1)
Probability (5.2)
Central Tendency (5.3)

DRAFT

Grade 5 Mathematics
Algebraic Reasoning: Patterns and Relationships
Algebra Patterns (1.1)
Equations (1.2)
Number Properties (1.3)
Number Sense and Operation
Number Sense (2.1)
Number Operations (2.2)
Geometry
Circles and Polygons (3.1)
Angles (3.2)
Measurement
Measurement (4.1)
Money (4.2)
Data Analysis
Data Analysis (5.1)
Probability (5.2)
Central Tendency (5.3)

DRAFT

Grade 6 Mathematics
Algebraic Reasoning: Patterns and Relationships
Algebra Patterns (1.1)
Expressions and Equations (1.2)
Number Properties (1.3)
Solving Equations (1.4)
Number Sense and Operation
Number Sense (2.1)
Number Operations (2.2)
Geometry
Three Dimensional Figures (3.1)
Congruent and Similar Figures (3.2)
Coordinate Geometry (3.3)
Measurement
Circles (4.1)
Conversions (4.2)
Data Analysis
Data Analysis (5.1)
Probability (5.2)

DRAFT

Grade 7 Mathematics
Algebraic Reasoning: Patterns and Relationships
Linear Relationships (1.1)
Solving Equations (1.2)
Solving and Graphing Inequalities (1.3)
Number Sense and Operation
Number Sense (2.1)
Number Operations (2.2)
Geometry
Classifying Figures (3.1)
Lines and Angles (3.2)
Transformations (3.3)
Measurement
Perimeter and Area (4.1)
Circles (4.2)
Composite Figures (4.3)
Data Analysis
Data Analysis (5.1)
Probability (5.2)

DRAFT

Grade 8 Mathematics
Algebraic Reasoning: Patterns and Relationships
Equations (1.1)
Inequalities (1.2)
Number Sense and Operation
Number Sense (2.1)
Number Operations (2.2)
Geometry
Three Dimensional Figures (3.1)
Pythagorean Theorem (3.2)
Measurement
Surface Area and Volume (4.1)
Ratio and Proportions (4.2)
Composite Figures (4.3)
Data Analysis
Data Analysis (5.1)
Central Tendency (5.3)

DRAFT

Grade 3 Reading
Standard 2: Vocabulary
Words in Context (2.1)
Affixes (2.2)
Synonyms, Antonyms, and Homonyms/Homophones (2.3)
Using Resource Materials (2.4)
Standard 4: Comprehension/Critical Literacy
Literal Understanding (4.1)
Inferences and Interpretation (4.2)
Summary and Generalization (4.3)
Analysis and Evaluation (4.4)
Standard 5: Literature
Literary Elements (5.2) & Figurative Language/Sound Devices (5.3)
Standard 6: Research and Information
Accessing Information (6.1)

DRAFT

Grade 4 Reading
Standard 1: Vocabulary
Words in Context (1.1)
Affixes, Roots, and Derivatives (1.2)
Synonyms, Antonyms, and Homonyms/Homophones (1.3)
Standard 3: Comprehension/Critical Literacy
Literal Understanding (3.1)
Inferences and Interpretation (3.2)
Summary and Generalization (3.3)
Analysis and Evaluation (3.4)
Standard 4: Literature
Literary Elements (4.2)
Figurative Language and Sound Devices (4.3)
Standard 5: Research and Information
Accessing Information (5.1)

DRAFT

Grade 5 Reading
Standard 1: Vocabulary
Words in Context (1.1)
Affixes, Roots, and Stems (1.2)
Synonyms, Antonyms, and Homonyms/Homophones (1.3)
Standard 3: Comprehension/Critical Literacy
Literal Understanding (3.1)
Inferences and Interpretation (3.2)
Summary and Generalization (3.3)
Analysis and Evaluation (3.4)
Standard 4: Literature
Literary Genres (4.1)
Literary Elements (4.2)
Figurative Language and Sound Devices (4.3)
Standard 5: Research and Information
Accessing Information (5.1)
Interpreting Information (5.2)

DRAFT

Grade 6 Reading
Standard 1: Vocabulary
Words in Context (1.1)
Word Origins (1.2)
Standard 3: Comprehension/Critical Literacy
Literal Understanding (3.1)
Inferences and Interpretation (3.2)
Summary and Generalization (3.3)
Analysis and Evaluation (3.4)
Standard 4: Literature
Literary Genres (4.1)
Literary Elements (4.2)
Figurative Language and Sound Devices (4.3)
Standard 5: Research and Information
Accessing Information (5.1)
Interpreting Information (5.2)

DRAFT

Grade 7 Reading
Standard 1: Vocabulary
Words in Context (1.1)
Word Origins (1.2)
Idioms and Comparisons (1.3)
Standard 3: Comprehension
Literal Understanding (3.1)
Inference and Interpretation (3.2)
Summary and Generalization (3.3)
Analysis and Evaluation (3.4)
Standard 4: Literature
Literary Genres (4.1)
Literary Elements (4.2)
Figurative Language and Sound Devices (4.3)
Standard 5: Research and Information
Accessing Information (5.1)
Interpreting Information (5.2)

DRAFT

Grade 8 Reading
Standard 1: Vocabulary
Words in Context (1.1)
Word Origins (1.2)
Idioms and Comparisons (1.3)
Standard 3: Comprehension
Literal Understanding (3.1)
Inferences and Interpretation (3.2)
Summary and Generalization (3.3)
Analysis and Evaluation (3.4)
Standard 4: Literature
Literary Genres (4.1)
Literary Elements (4.2)
Figurative Language and Sound Devices (4.3)
Standard 5: Research and Information
Accessing Information (5.1)
Interpreting Information (5.2)

DRAFT

Grade 5 Science: Process Standards
Standard 1: Observe and Measure
SI Metric (P1.1)
Similar/different characteristics (P1.2)
Standard 2: Classify
Observable properties (P2.1)
Serial order (P2.2)
Standard 3: Experiment
Experimental design (P3.2)
Hazards/practice safety (P3.4)
Standard 4: Interpret and Communicate
Data tables/line/bar/trend and circle graphs (P4.2)
Prediction based on data (P4.3)
Explanations based on data (P4.4)

Grade 5 Science: Content Standards
Standard 1: Properties of Matter and Energy
Matter has physical properties (1.1)
Physical properties can be measured (1.2)
Energy can be transferred (1.3)
Standard 2: Organisms and Environments
Dependence upon community (2.1)
Individual organism and species survival (2.2)
Standard 3: Structures of the Earth and the Solar System
Weather patterns (3.2)
Earth as a planet (3.3)

DRAFT

Grade 8 Science: Process Standards
Standard 1: Observe and Measure
Qualitative/quantitative observations/changes (P1.1)
SI (metrics) units/appropriate tools (P1.2 and P1.3)
Standard 2: Classify
Classification system (P2.1)
Properties ordered (P2.2)
Standard 3: Experiment
Experimental design (P3.2)
Identify variables (P3.3)
Hazards/practice safety (P3.6)
Standard 4: Interpret and Communicate
Data tables/line/bar/trend and circle graphs (P4.2)
Explanations/prediction (P4.3)

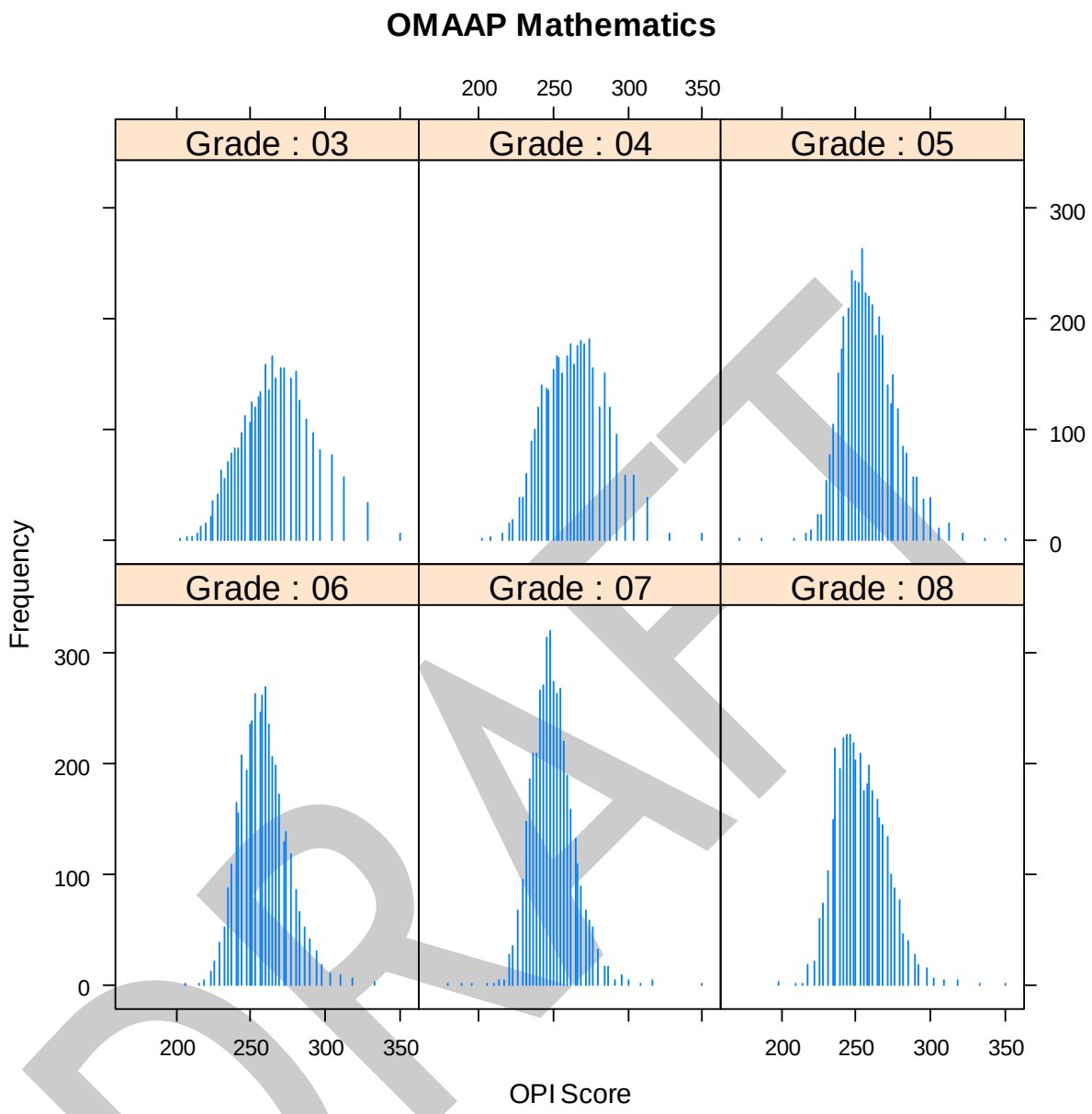
Grade 8 Science: Content Standards
Standard 1: Properties and Chemical Changes in Matter
Chemical reactions (1.1)
Conservation of matter (1.2)
Standard 2: Motion and Forces
Motion of an object (2.1)
Object subjected to a force (2.2)
Standard 3: Diversity and Adaptations of Organisms
Classification (3.1)
Internal and external structures (3.2)
Standard 4: Structures/Forces of the Earth/Solar System
Landforms result from constructive and destructive forces (4.1)
Rock cycle (4.2)
Standard 5: Earth's History
Catastrophic events (5.1)
Fossil evidence (5.2)

DRAFT

Appendix B

Scale Score Distributions

DRAFT


OMAAP Mathematics
Grade : 03
Grade : 04
Grade : 05
Grade : 06
Grade : 07
Grade : 08
Frequency
OPI Score
0
100
200
300
200
250
300
350

DRAFT

Math Grade 3 Scale Score Distribution

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
203	1	0.03	1	0.03
207	3	0.09	4	0.12
211	3	0.09	7	0.22
214	6	0.19	13	0.41
217	12	0.37	25	0.78
220	15	0.47	40	1.25
223	22	0.69	62	1.93
225	36	1.12	98	3.05
228	42	1.31	140	4.36
230	64	1.99	204	6.36
233	56	1.75	260	8.10
235	71	2.21	331	10.31
237	79	2.46	410	12.78
239	83	2.59	493	15.36
242	83	2.59	576	17.95
244	97	3.02	673	20.97
246	113	3.52	786	24.49
250	107	3.33	893	27.83
251	125	3.90	1018	31.72
253	121	3.77	1139	35.49
255	130	4.05	1269	39.55
257	134	4.18	1403	43.72
260	159	4.95	1562	48.68
262	136	4.24	1698	52.91
265	167	5.20	1865	58.12
267	146	4.55	2011	62.67
270	155	4.83	2166	67.50
273	155	4.83	2321	72.33
277	147	4.58	2468	76.91
280	152	4.74	2620	81.65
283	127	3.96	2747	85.60
287	109	3.40	2856	89.00
292	97	3.02	2953	92.02
297	82	2.56	3035	94.58
304	77	2.40	3112	96.98
313	57	1.78	3169	98.75
328	34	1.06	3203	99.81
350	6	0.19	3209	100.00

DRAFT

Math Grade 4 Scale Score Distribution

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
202	1	0.03	1	0.03
208	3	0.08	4	0.11
216	6	0.17	10	0.28
220	15	0.42	25	0.70
223	18	0.50	43	1.20
227	39	1.09	82	2.30
230	39	1.09	121	3.39
232	61	1.71	182	5.10
235	89	2.49	271	7.59
238	100	2.80	371	10.39
240	120	3.36	491	13.75
242	141	3.95	632	17.70
245	138	3.86	770	21.56
247	135	3.78	905	25.34
250	154	4.31	1059	29.66
252	166	4.65	1225	34.30
254	165	4.62	1390	38.92
256	151	4.23	1541	43.15
259	166	4.65	1707	47.80
261	178	4.98	1885	52.79
264	159	4.45	2044	57.24
266	176	4.93	2220	62.17
269	180	5.04	2400	67.21
271	177	4.96	2577	72.16
274	182	5.10	2759	77.26
277	155	4.34	2914	81.60
281	121	3.39	3035	84.99
284	151	4.23	3186	89.22
288	120	3.36	3306	92.58
293	96	2.69	3402	95.27
298	58	1.62	3460	96.89
304	59	1.65	3519	98.54
313	39	1.09	3558	99.64
328	7	0.20	3565	99.83
350	6	0.17	3571	100.00

DRAFT

Math Grade 5 Scale Score Distribution

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
172	1	0.02	1	0.02
187	1	0.02	2	0.05
208	1	0.02	3	0.07
216	6	0.14	9	0.22
220	10	0.24	19	0.46
224	24	0.58	43	1.03
227	23	0.55	66	1.59
230	54	1.30	120	2.88
232	77	1.85	197	4.73
235	105	2.52	302	7.26
238	151	3.63	453	10.89
240	173	4.16	626	15.04
242	202	4.85	828	19.90
245	210	5.05	1038	24.95
247	244	5.86	1282	30.81
250	234	5.62	1516	36.43
252	232	5.58	1748	42.01
254	263	6.32	2011	48.33
256	224	5.38	2235	53.71
259	221	5.31	2456	59.02
261	212	5.09	2668	64.12
263	185	4.45	2853	68.57
265	202	4.85	3055	73.42
268	185	4.45	3240	77.87
271	140	3.36	3380	81.23
273	124	2.98	3504	84.21
275	149	3.58	3653	87.79
278	119	2.86	3772	90.65
281	85	2.04	3857	92.69
284	79	1.90	3936	94.59
288	57	1.37	3993	95.96
291	57	1.37	4050	97.33
295	37	0.89	4087	98.22
300	38	0.91	4125	99.13
305	11	0.26	4136	99.40
312	15	0.36	4151	99.76
321	6	0.14	4157	99.90
336	2	0.05	4159	99.95
350	2	0.05	4161	100.00

DRAFT

Math Grade 6 Scale Score Distribution

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
206	1	0.02	1	0.02
215	2	0.05	3	0.07
219	5	0.12	8	0.20
223	12	0.29	20	0.49
226	22	0.54	42	1.03
229	38	0.93	80	1.96
232	53	1.30	133	3.26
235	87	2.13	220	5.39
237	109	2.67	329	8.06
240	164	4.02	493	12.07
242	155	3.80	648	15.87
244	208	5.09	856	20.96
247	194	4.75	1050	25.71
250	235	5.75	1285	31.46
251	238	5.83	1523	37.29
253	263	6.44	1786	43.73
256	246	6.02	2032	49.76
258	262	6.42	2294	56.17
260	269	6.59	2563	62.76
262	235	5.75	2798	68.51
265	207	5.07	3005	73.58
267	198	4.85	3203	78.43
269	173	4.24	3376	82.66
272	129	3.16	3505	85.82
274	138	3.38	3643	89.20
277	119	2.91	3762	92.12
280	86	2.11	3848	94.22
283	66	1.62	3914	95.84
286	52	1.27	3966	97.11
290	41	1.00	4007	98.11
294	30	0.73	4037	98.85
298	18	0.44	4055	99.29
303	11	0.27	4066	99.56
310	9	0.22	4075	99.78
318	6	0.15	4081	99.93
333	3	0.07	4084	100.00

DRAFT

Math Grade 7 Scale Score Distribution

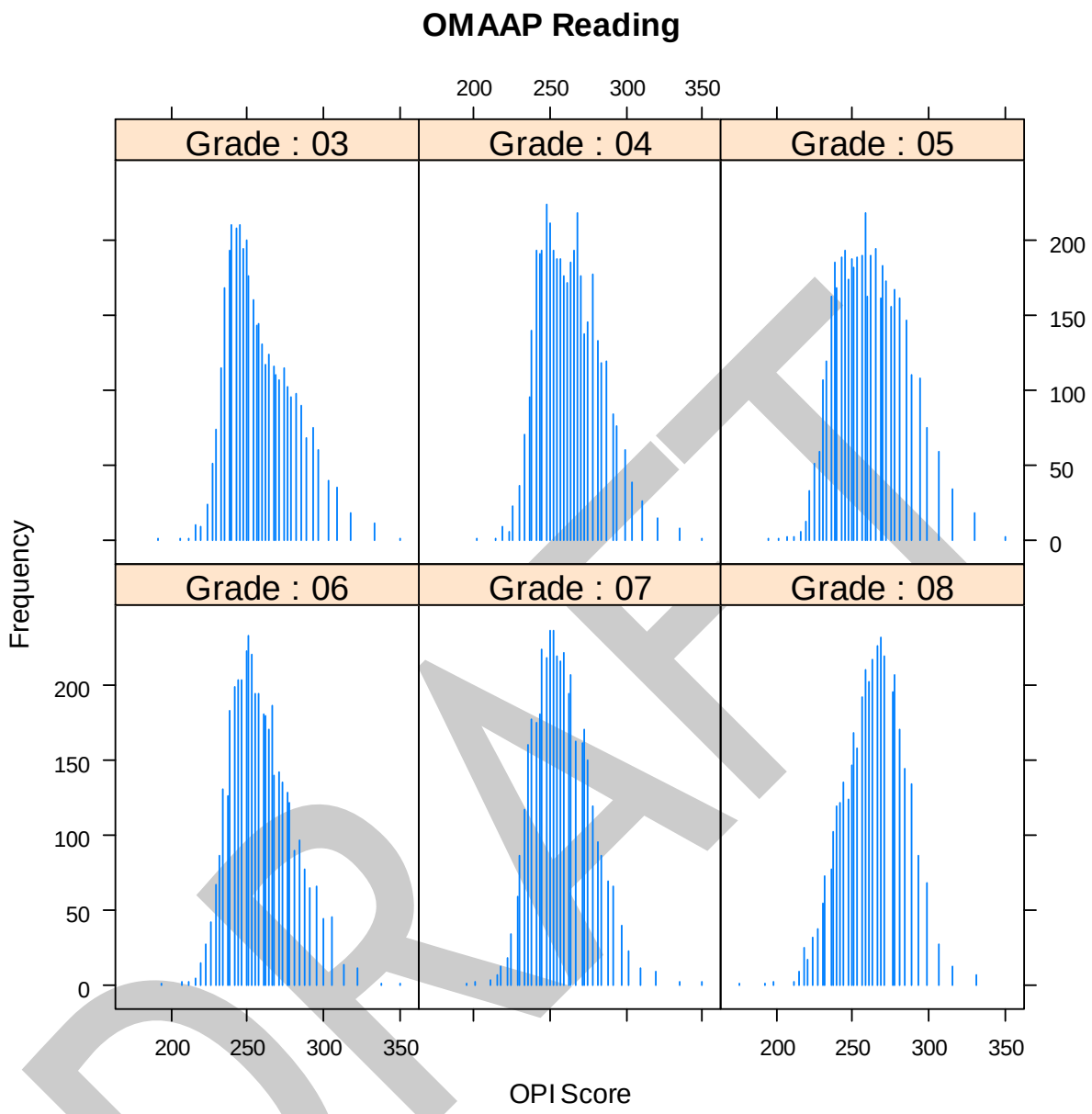
Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
179	1	0.02	1	0.02
188	1	0.02	2	0.05
195	1	0.02	3	0.07
205	1	0.02	4	0.10
210	1	0.02	5	0.12
213	4	0.10	9	0.22
217	5	0.12	14	0.34
220	28	0.68	42	1.01
223	35	0.85	77	1.86
226	68	1.64	145	3.50
229	96	2.32	241	5.82
232	148	3.57	389	9.39
234	186	4.49	575	13.89
236	209	5.05	784	18.93
239	210	5.07	994	24.00
241	266	6.42	1260	30.43
243	271	6.54	1531	36.97
246	314	7.58	1845	44.55
248	321	7.75	2166	52.31
250	274	6.62	2440	58.92
252	264	6.38	2704	65.30
255	268	6.47	2972	71.77
257	220	5.31	3192	77.08
259	189	4.56	3381	81.65
262	159	3.84	3540	85.49
265	132	3.19	3672	88.67
266	110	2.66	3782	91.33
269	90	2.17	3872	93.50
272	68	1.64	3940	95.15
274	58	1.40	3998	96.55
277	52	1.26	4050	97.80
280	32	0.77	4082	98.58
284	17	0.41	4099	98.99
287	17	0.41	4116	99.40
291	5	0.12	4121	99.52
296	9	0.22	4130	99.73
301	4	0.10	4134	99.83
308	2	0.05	4136	99.88
317	4	0.10	4140	99.98
350	1	0.02	4141	100.00

DRAFT

Math Grade 8 Scale Score Distribution

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
198	3	0.08	3	0.08
210	1	0.03	4	0.10
214	2	0.05	6	0.15
218	19	0.49	25	0.64
222	22	0.56	47	1.20
225	60	1.53	107	2.74
228	74	1.89	181	4.63
231	103	2.63	284	7.26
235	150	3.84	434	11.10
236	214	5.47	648	16.57
239	195	4.99	843	21.55
241	224	5.73	1067	27.28
244	226	5.78	1293	33.06
246	226	5.78	1519	38.84
248	218	5.57	1737	44.41
250	203	5.19	1940	49.60
253	210	5.37	2150	54.97
255	176	4.50	2326	59.47
257	182	4.65	2508	64.13
259	199	5.09	2707	69.22
261	175	4.47	2882	73.69
264	168	4.30	3050	77.99
266	151	3.86	3201	81.85
268	144	3.68	3345	85.53
271	134	3.43	3479	88.95
273	100	2.56	3579	91.51
276	88	2.25	3667	93.76
279	77	1.97	3744	95.73
282	46	1.18	3790	96.91
285	40	1.02	3830	97.93
289	28	0.72	3858	98.64
292	19	0.49	3877	99.13
297	16	0.41	3893	99.54
302	6	0.15	3899	99.69
309	4	0.10	3903	99.80
318	5	0.13	3908	99.92
333	1	0.03	3909	99.95
350	2	0.05	3911	100.00

DRAFT


OMAAP Reading
Grade : 03
Grade : 04
Grade : 05
Grade : 06
Grade : 07
Grade : 08
Frequency
OPI Score
0
50
100
150
200
250
300
350

DRAFT

Reading Grade 3 Scale Score Distribution

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
191	1	0.03	1	0.03
206	2	0.05	3	0.08
212	2	0.05	5	0.13
216	11	0.29	16	0.42
220	10	0.26	26	0.68
224	24	0.63	50	1.31
227	51	1.33	101	2.64
230	74	1.94	175	4.58
233	115	3.01	290	7.59
235	168	4.39	458	11.98
238	193	5.05	651	17.03
240	211	5.52	862	22.55
243	208	5.44	1070	27.99
245	211	5.52	1281	33.51
247	195	5.10	1476	38.61
250	200	5.23	1676	43.84
251	177	4.63	1853	48.47
254	161	4.21	2014	52.68
256	143	3.74	2157	56.42
258	145	3.79	2302	60.21
260	131	3.43	2433	63.64
262	117	3.06	2550	66.70
264	124	3.24	2674	69.95
267	116	3.03	2790	72.98
269	111	2.90	2901	75.88
271	107	2.80	3008	78.68
274	115	3.01	3123	81.69
276	103	2.69	3226	84.38
279	96	2.51	3322	86.90
282	98	2.56	3420	89.46
285	90	2.35	3510	91.81
289	69	1.80	3579	93.62
293	75	1.96	3654	95.58
297	61	1.60	3715	97.17
303	40	1.05	3755	98.22
309	36	0.94	3791	99.16
318	18	0.47	3809	99.63
333	12	0.31	3821	99.95
350	2	0.05	3823	100.00

DRAFT

Reading Grade 4 Scale Score Distribution

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
202	2	0.05	2	0.05
214	1	0.02	3	0.07
219	9	0.21	12	0.28
223	6	0.14	18	0.42
226	23	0.54	41	0.97
230	37	0.87	78	1.84
233	71	1.68	149	3.52
237	96	2.27	245	5.78
238	140	3.30	385	9.09
241	194	4.58	579	13.67
243	191	4.51	770	18.17
245	194	4.58	964	22.75
248	224	5.29	1188	28.04
250	212	5.00	1400	33.04
252	193	4.56	1593	37.60
255	188	4.44	1781	42.03
257	188	4.44	1969	46.47
259	177	4.18	2146	50.65
261	172	4.06	2318	54.71
263	185	4.37	2503	59.07
266	193	4.56	2696	63.63
268	218	5.15	2914	68.78
270	177	4.18	3091	72.95
273	138	3.26	3229	76.21
275	146	3.45	3375	79.66
278	178	4.20	3553	83.86
281	133	3.14	3686	87.00
284	119	2.81	3805	89.80
287	120	2.83	3925	92.64
291	84	1.98	4009	94.62
294	77	1.82	4086	96.44
299	61	1.44	4147	97.88
304	39	0.92	4186	98.80
311	26	0.61	4212	99.41
320	15	0.35	4227	99.76
335	8	0.19	4235	99.95
350	2	0.05	4237	100.00

DRAFT

Reading Grade 5 Scale Score Distribution

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
195	1	0.02	1	0.02
202	1	0.02	2	0.04
207	3	0.07	5	0.11
212	3	0.07	8	0.18
216	6	0.13	14	0.31
219	13	0.29	27	0.59
222	33	0.73	60	1.32
225	51	1.12	111	2.44
228	60	1.32	171	3.76
231	107	2.35	278	6.12
233	120	2.64	398	8.76
236	163	3.59	561	12.35
238	185	4.07	746	16.42
240	168	3.70	914	20.11
243	189	4.16	1103	24.27
245	194	4.27	1297	28.54
247	174	3.83	1471	32.37
250	188	4.14	1659	36.51
251	182	4.01	1841	40.51
253	189	4.16	2030	44.67
256	190	4.18	2220	48.86
258	218	4.80	2438	53.65
260	163	3.59	2601	57.24
262	190	4.18	2791	61.42
265	195	4.29	2986	65.71
269	162	3.57	3148	69.28
270	183	4.03	3331	73.31
272	173	3.81	3504	77.11
275	156	3.43	3660	80.55
278	167	3.68	3827	84.22
281	162	3.57	3989	87.79
285	147	3.24	4136	91.02
289	111	2.44	4247	93.46
294	108	2.38	4355	95.84
299	75	1.65	4430	97.49
306	59	1.30	4489	98.79
315	34	0.75	4523	99.54
330	18	0.40	4541	99.93
350	3	0.07	4544	100.00

DRAFT

Reading Grade 6 Scale Score Distribution

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
194	1	0.02	1	0.02
207	2	0.05	3	0.07
212	3	0.07	6	0.14
216	5	0.12	11	0.26
219	15	0.35	26	0.61
223	28	0.66	54	1.27
226	42	0.99	96	2.25
229	67	1.57	163	3.83
232	86	2.02	249	5.84
234	131	3.07	380	8.92
237	126	2.96	506	11.88
239	183	4.29	689	16.17
242	199	4.67	888	20.84
244	204	4.79	1092	25.63
246	203	4.76	1295	30.39
250	223	5.23	1518	35.63
251	233	5.47	1751	41.09
253	221	5.19	1972	46.28
255	194	4.55	2166	50.83
257	195	4.58	2361	55.41
261	181	4.25	2542	59.66
262	180	4.22	2722	63.88
264	171	4.01	2893	67.89
266	186	4.37	3079	72.26
268	140	3.29	3219	75.55
271	142	3.33	3361	78.88
273	135	3.17	3496	82.05
276	129	3.03	3625	85.07
278	122	2.86	3747	87.94
281	90	2.11	3837	90.05
284	97	2.28	3934	92.33
288	78	1.83	4012	94.16
291	65	1.53	4077	95.68
295	66	1.55	4143	97.23
300	44	1.03	4187	98.26
306	46	1.08	4233	99.34
313	14	0.33	4247	99.67
322	12	0.28	4259	99.95
338	1	0.02	4260	99.98
350	1	0.02	4261	100.00

DRAFT

Reading Grade 7 Scale Score Distribution

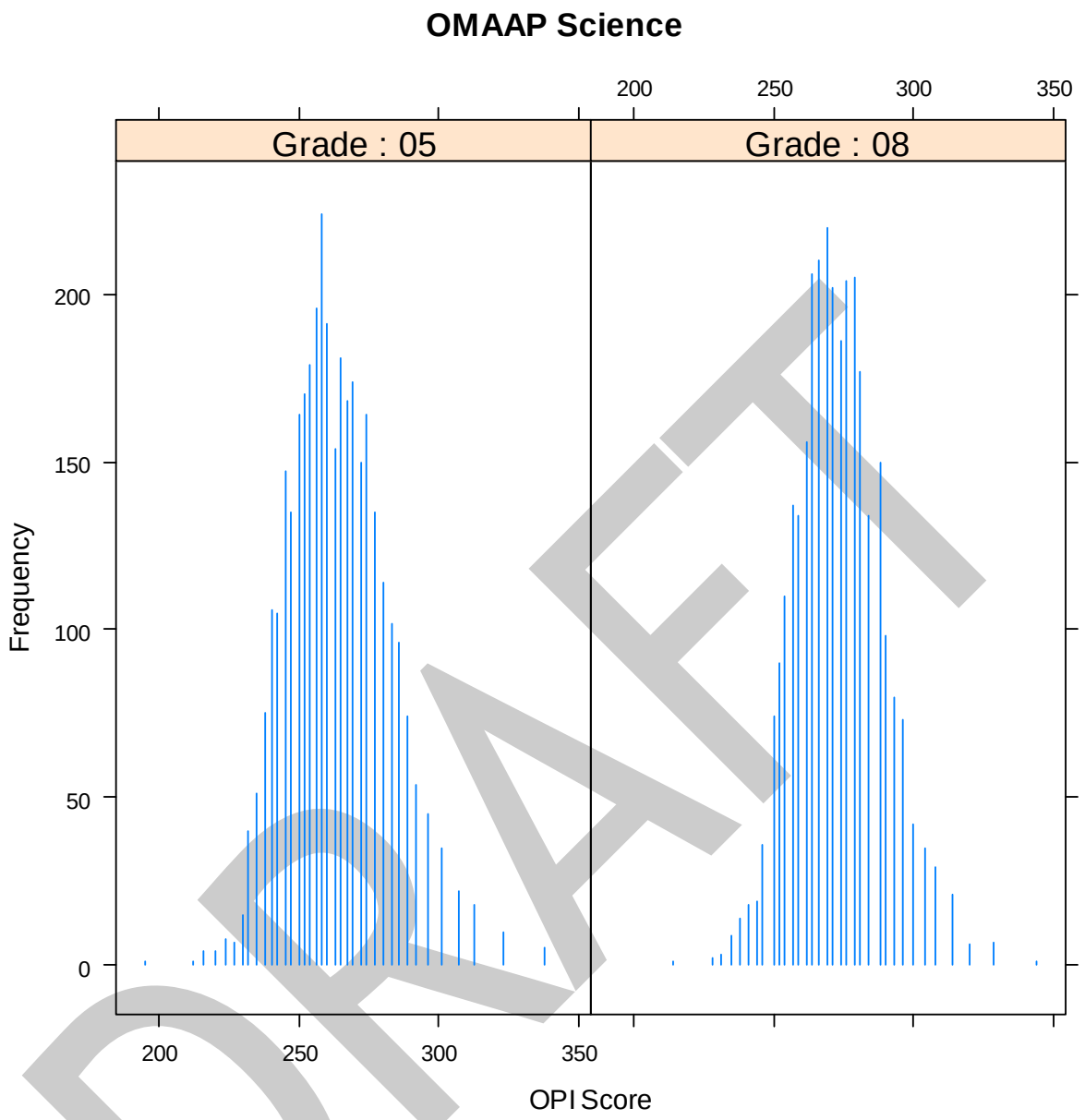
Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
195	1	0.02	1	0.02
201	3	0.07	4	0.10
211	4	0.10	8	0.19
215	7	0.17	15	0.36
218	13	0.31	28	0.67
222	18	0.43	46	1.10
225	34	0.81	80	1.91
229	59	1.41	139	3.33
230	86	2.06	225	5.38
233	117	2.80	342	8.18
236	160	3.83	502	12.01
238	177	4.23	679	16.24
241	175	4.19	854	20.43
243	181	4.33	1035	24.76
245	224	5.36	1259	30.12
248	218	5.22	1477	35.33
250	237	5.67	1714	41.00
252	236	5.65	1950	46.65
255	219	5.24	2169	51.89
257	216	5.17	2385	57.06
259	222	5.31	2607	62.37
262	195	4.67	2802	67.03
264	207	4.95	3009	71.99
267	163	3.90	3172	75.89
271	162	3.88	3334	79.76
272	171	4.09	3505	83.85
275	150	3.59	3655	87.44
278	119	2.85	3774	90.29
281	96	2.30	3870	92.58
284	86	2.06	3956	94.64
288	70	1.67	4026	96.32
292	66	1.58	4092	97.89
297	40	0.96	4132	98.85
302	23	0.55	4155	99.40
309	12	0.29	4167	99.69
319	9	0.22	4176	99.90
335	2	0.05	4178	99.95
350	2	0.05	4180	100.00

DRAFT

Reading Grade 8 Scale Score Distribution

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
176	1	0.03	1	0.03
193	1	0.03	2	0.05
198	2	0.05	4	0.10
211	3	0.08	7	0.18
215	9	0.23	16	0.40
218	25	0.63	41	1.03
221	17	0.43	58	1.46
224	32	0.81	90	2.27
227	38	0.96	128	3.23
230	55	1.39	183	4.62
232	73	1.84	256	6.46
236	78	1.97	334	8.43
237	102	2.57	436	11.00
240	119	3.00	555	14.00
242	122	3.08	677	17.08
244	135	3.41	812	20.49
247	124	3.13	936	23.62
250	147	3.71	1083	27.33
251	168	4.24	1251	31.57
253	158	3.99	1409	35.55
256	192	4.84	1601	40.40
258	210	5.30	1811	45.70
261	202	5.10	2013	50.79
263	217	5.48	2230	56.27
266	226	5.70	2456	61.97
268	232	5.85	2688	67.83
271	220	5.55	2908	73.38
276	196	4.95	3104	78.32
277	207	5.22	3311	83.55
281	171	4.31	3482	87.86
284	145	3.66	3627	91.52
289	134	3.38	3761	94.90
293	87	2.20	3848	97.10
299	68	1.72	3916	98.81
306	27	0.68	3943	99.50
315	13	0.33	3956	99.82
331	7	0.18	3963	100.00

DRAFT


OMAAP Science
Grade : 05
Grade : 08
Frequency
0
50
100
150
200
200
250
300
350
OPI Score

DRAFT

Science Grade 5 Scale Score Distribution

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
195	1	0.03	1	0.03
212	1	0.03	2	0.06
216	4	0.11	6	0.17
220	4	0.11	10	0.28
224	8	0.23	18	0.51
227	7	0.20	25	0.71
230	15	0.43	40	1.14
232	40	1.14	80	2.27
235	51	1.45	131	3.72
238	75	2.13	206	5.85
240	106	3.01	312	8.85
242	105	2.98	417	11.83
245	147	4.17	564	16.00
247	135	3.83	699	19.84
250	164	4.65	863	24.49
252	170	4.82	1033	29.31
254	179	5.08	1212	34.39
256	196	5.56	1408	39.95
258	224	6.36	1632	46.31
260	191	5.42	1823	51.73
263	154	4.37	1977	56.10
265	181	5.14	2158	61.24
267	168	4.77	2326	66.00
269	174	4.94	2500	70.94
272	150	4.26	2650	75.20
274	164	4.65	2814	79.85
277	135	3.83	2949	83.68
280	114	3.23	3063	86.92
283	102	2.89	3165	89.81
286	96	2.72	3261	92.54
289	74	2.10	3335	94.64
292	54	1.53	3389	96.17
296	45	1.28	3434	97.45
301	35	0.99	3469	98.44
307	22	0.62	3491	99.06
313	18	0.51	3509	99.57
323	10	0.28	3519	99.86
338	5	0.14	3524	100.00

DRAFT

Science Grade 8 Scale Score Distribution

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
214	1	0.03	1	0.03
228	2	0.06	3	0.10
231	3	0.10	6	0.19
235	9	0.29	15	0.49
238	14	0.45	29	0.94
241	18	0.58	47	1.52
244	19	0.62	66	2.14
246	36	1.17	102	3.30
250	74	2.40	176	5.70
252	90	2.91	266	8.61
254	110	3.56	376	12.17
257	137	4.44	513	16.61
259	134	4.34	647	20.95
262	156	5.05	803	26.00
264	206	6.67	1009	32.66
266	210	6.80	1219	39.46
269	220	7.12	1439	46.58
271	202	6.54	1641	53.12
274	186	6.02	1827	59.15
276	204	6.60	2031	65.75
279	205	6.64	2236	72.39
281	177	5.73	2413	78.12
284	134	4.34	2547	82.45
288	150	4.86	2697	87.31
290	98	3.17	2795	90.48
293	80	2.59	2875	93.07
296	73	2.36	2948	95.44
300	42	1.36	2990	96.80
304	35	1.13	3025	97.93
308	29	0.94	3054	98.87
314	21	0.68	3075	99.55
320	6	0.19	3081	99.74
329	7	0.23	3088	99.97
344	1	0.03	3089	100.00

DRAFT