



Alaska

Comprehensive System of Student Assessment

Technical Report

Spring 2007

Grades 3–10

Standards Based Assessments (SBAs)



August 2007

TABLE OF CONTENTS

CHAPTER 1: BACKGROUND OF ALASKA ASSESSMENTS	1
CHAPTER 2: TEST DESIGN & ITEM DEVELOPMENT	2
Mathematics Assessment Measures	2
Multiple-Choice Items	2
Constructed-Response Items	2
Reading Assessment Measures	3
Multiple-Choice Items	3
Constructed-Response Items	3
Writing Assessment Measures.....	3
Multiple-Choice Items	4
Constructed-Response Items	4
2007 Operational Plan.....	4
Grade-Level Expectations Subsumed within Reporting Categories	8
Test Development Timeline	10
Item and Test Development Process	11
Item Writer Training	11
Reading Passage Selection.....	12
Passage Readability.....	13
Item Writing.....	13
Item Content Review.....	15
Bias and Sensitivity Review.....	16
Item Field Test	17
Item Field Test Data Review.....	17
Psychometric Guidelines for Selecting Items.....	19
Proportion Correct (also known as <i>p</i> -value)	19
Average Person Logit.....	20
Item-Total Correlation	20
Fit Statistic	20
Differential Item Functioning (DIF) Analyses.....	20
Item Bank.....	21
Overview	21
Functionality	21
Item Cards and Reporting Options.....	22
Security	22

Quality Assurance	22
Item Bank Summary	22
Final Selection of Items and Spring 2007 SBA Operational Forms Construction	28
Steps in the Forms Construction Process	28
Construction of the Operational Forms	29
DRC Internal Review of the Items and Forms	29
CHAPTER 3: TEST ADMINISTRATION PROCEDURES.....	30
Overview.....	30
Student Population Tested.....	30
Accommodations.....	31
Spiraling Plan.....	31
Test Administrator Training.....	31
Test Security.....	31
Materials.....	32
Packaging and Shipping Materials	32
Materials Return	33
Box Receipt	33
CHAPTER 4: SCORING & STUDENT PERFORMANCE REPORTING	34
Document Processing	34
Handscoring of Constructed Responses	34
Readers	34
Rangefinding and Developing Training Material	35
Training the Readers	35
Imaging.....	35
Quality Control of Handscoring	36
Data Processing.....	36
Reporting.....	37
District Reports	38
State Reports	38

CHAPTER 5: FORM ANALYSIS & ITEM CALIBRATION	39
Rasch Measurement Models.....	39
Item Statistics.....	40
Form Statistics	44
Frequency Distributions	46
Items.....	46
Persons	46
Cautions for Score Use.....	46
CHAPTER 6: SCALING & EQUATING.....	47
Introduction	47
Grades 3–9 Pre-Equating.....	47
Grades 3–9 Operational Item Calibration	47
Grade 10 Operational Item Calibration.....	48
Item Bank Maintenance.....	49
CHAPTER 7: FIELD TEST ITEM DATA SUMMARY.....	50
Field Test Items.....	50
Field Test Item Descriptive Statistics	50
Item Bank Maintenance.....	56
CHAPTER 8: SCALE SCORES & PROFICIENCY LEVELS.....	57
Overview.....	57
Description of Scores.....	57
Raw Score	57
Scale Score.....	57
Comparability of Scale Scores Across Grades.....	58
Transformations	58
Scale Score Summary Statistics	61
Proficiency Levels.....	75

CHAPTER 9: TEST VALIDITY & RELIABILITY	79
Introduction	79
Validity	79
Content/Curricular.....	79
Construct Validity	80
Validity Evidence for Different Student Populations	81
Reliability	82
Standard Error of Measurement	83
REFERENCES	84
APPENDIX 1: SPRING 2007 OPERATIONAL TEST BLUEPRINTS	1-1
Grade 3	1-1
Grade 4	1-4
Grade 5	1-7
Grade 6	1-10
Grade 7	1-13
Grade 8	1-16
Grade 9	1-19
Grade 10	1-22
APPENDIX 2: RUBRICS.....	2-1
6-Point Extended Constructed-Response (ECR) Scoring Rubric for Writing	2-1
6 Points.....	2-1
5 Points.....	2-1
4 Points.....	2-1
3 Points.....	2-2
2 Points.....	2-2
1 Point	2-2
4-Point Extended Constructed-Response (ECR) Scoring Rubric for Grades 3-9 Writing.....	2-3
4 Points.....	2-3
3 Points.....	2-3
2 Points.....	2-3
1 Point	2-3
APPENDIX 3: DRC ITEM WRITER TRAINING MANUAL	3-1

APPENDIX 4: FAIRNESS IN TESTING MANUAL	4-1
APPENDIX 5: DEPTH OF KNOWLEDGE LEVELS	5-1
Mathematics.....	5-1
Level 1.....	5-1
Level 2.....	5-1
Level 3.....	5-2
Level 4.....	5-2
Reading.....	5-3
Level 1.....	5-3
Level 2.....	5-3
Level 3.....	5-4
Level 4.....	5-4
Writing.....	5-5
Level 1.....	5-5
Level 2.....	5-5
Level 3.....	5-5
Level 4.....	5-6
Source of Challenge Criterion	5-6
APPENDIX 6: UNIVERSALLY DESIGNED ASSESSMENTS.....	6-1
Elements of Universally Designed Assessments.....	6-1
Guidelines for Universally Designed Items	6-3
APPENDIX 7: ITEM REVIEW TRACKING FORMS.....	7-1
Data Review Form.....	7-2
APPENDIX 8: CONFIDENTIALITY AGREEMENT.....	8-1
APPENDIX 9: BIAS & SENSITIVITY REVIEW FORM	9-1
APPENDIX 10: SAMPLES OF MANUALS	10-1
APPENDIX 11: INTER-RATER RELIABILITY ITEMS.....	11-1
APPENDIX 12: SAMPLES OF GUIDES TO TEST INTERPRETATION	12-1

APPENDIX 13: OPERATIONAL TEST ITEM ANALYSIS..... 13-1

Grade 3 Mathematics	13-1
Grade 3 Reading	13-2
Grade 3 Writing.....	13-4
Grade 4 Mathematics	13-5
Grade 4 Reading	13-7
Grade 4 Writing.....	13-8
Grade 5 Mathematics	13-10
Grade 5 Reading	13-12
Grade 5 Writing.....	13-13
Grade 6 Mathematics	13-15
Grade 6 Reading	13-16
Grade 6 Writing.....	13-18
Grade 7 Mathematics	13-19
Grade 7 Reading	13-21
Grade 7 Writing.....	13-22
Grade 8 Mathematics	13-24
Grade 8 Reading	13-25
Grade 8 Writing.....	13-27
Grade 9 Mathematics	13-28
Grade 9 Reading	13-30
Grade 9 Writing.....	13-32
Grade 10 Mathematics	13-33
Grade 10 Reading	13-34
Grade 10 Writing.....	13-36

APPENDIX 14: FORM STATISTICS 14-1

Grade 3 Mathematics	14-1
Grade 3 Reading	14-2
Grade 3 Writing.....	14-3
Grade 4 Mathematics	14-4
Grade 4 Reading	14-5
Grade 4 Writing.....	14-6
Grade 5 Mathematics	14-7

Grade 5 Reading	14-8
Grade 5 Writing	14-9
Grade 6 Mathematics	14-10
Grade 6 Reading	14-11
Grade 6 Writing	14-12
Grade 7 Mathematics	14-13
Grade 7 Reading	14-14
Grade 7 Writing	14-15
Grade 8 Mathematics	14-16
Grade 8 Reading	14-17
Grade 8 Writing	14-18
Grade 9 Mathematics	14-19
Grade 9 Reading	14-20
Grade 9 Writing	14-21
Grade 10 Mathematics	14-22
Grade 10 Reading	14-23
Grade 10 Writing	14-24

APPENDIX 15: OPERATIONAL TEST ITEM AND THRESHOLD

DIFFICULTY MAPS 15-1

Grade 3 Mathematics	15-1
Grade 3 Reading	15-2
Grade 3 Writing	15-3
Grade 4 Mathematics	15-4
Grade 4 Reading	15-5
Grade 4 Writing	15-6
Grade 5 Mathematics	15-7
Grade 5 Reading	15-8
Grade 5 Writing	15-9
Grade 6 Mathematics	15-10
Grade 6 Reading	15-11
Grade 6 Writing	15-12
Grade 7 Mathematics	15-13
Grade 7 Reading	15-14
Grade 7 Writing	15-15

Grade 8 Mathematics	15–16
Grade 8 Reading	15–17
Grade 8 Writing	15–18
Grade 9 Mathematics	15–19
Grade 9 Reading	15–20
Grade 9 Writing	15–21
Grade 10 Mathematics	15–22
Grade 10 Reading	15–23
Grade 10 Writing	15–24
APPENDIX 16: RAW-TO-SCALE SCORE TABLES	16–1
Grade 3	16–1
Grade 4	16–12
Grade 5	16–23
Grade 6	16–33
Grade 7	16–43
Grade 8	16–53
Grade 9	16–63
Grade 10	16–73
APPENDIX 17: SUBSCALE SCORE SUMMARY STATISTICS	17–1
Mathematics Subscale Reporting Categories	17–1
Reading Subscale Reporting Categories	17–1
Writing Subscale Reporting Categories	17–1
Grade 3	17–2
Grade 4	17–3
Grade 5	17–4
Grade 6	17–5
Grade 7	17–6
Grade 8	17–7
Grade 9	17–8
Grade 10	17–9

APPENDIX 18: GRADES 3–9 PROFICIENCY LEVEL DESCRIPTORS 18–1

Grade 3 Reading	18–1
Grade 3 Writing.....	18–2
Grade 3 Mathematics	18–3
Grade 4 Reading	18–5
Grade 4 Writing.....	18–6
Grade 4 Mathematics	18–7
Grade 5 Reading	18–9
Grade 5 Writing.....	18–10
Grade 5 Mathematics	18–11
Grade 6 Reading	18–13
Grade 6 Writing.....	18–14
Grade 6 Mathematics	18–15
Grade 7 Reading	18–17
Grade 7 Writing.....	18–18
Grade 7 Mathematics	18–19
Grade 8 Reading	18–21
Grade 8 Writing.....	18–22
Grade 8 Mathematics	18–24
Grade 9 Reading	18–25
Grade 9 Writing.....	18–26
Grade 9 Mathematics	18–27
Grade 10 Reading	18–29
Grade 10 Writing.....	18–30
Grade 10 Mathematics	18–31

APPENDIX 19: FIELD TEST ITEM ANALYSIS..... 19–1

Grade 3 Mathematics	19–1
Grade 3 Reading	19–1
Grade 3 Writing.....	19–2
Grade 4 Mathematics	19–2
Grade 4 Reading	19–3
Grade 4 Writing.....	19–3

Grade 5 Mathematics	19-4
Grade 5 Reading	19-4
Grade 5 Writing	19-5
Grade 6 Mathematics	19-5
Grade 6 Reading	19-6
Grade 6 Writing	19-6
Grade 7 Mathematics	19-7
Grade 7 Reading	19-7
Grade 7 Writing	19-8
Grade 8 Mathematics	19-8
Grade 8 Reading	19-9
Grade 8 Writing	19-9
Grade 9 Mathematics	19-10
Grade 9 Reading	19-10
Grade 9 Writing	19-11
Grade 10 Mathematics	19-12
Grade 10 Reading	19-15
Grade 10 Writing	19-19
APPENDIX 20: FIELD TEST DIFFERENTIAL ITEM FUNCTIONING (DIF)	
CLASSIFICATION RULES	20-1
Dichotomous (Multiple-Choice) DIF Classification	20-1
Polytomous (Constructed-Response) DIF Classification	20-1
APPENDIX 21: FIELD TEST DIFFERENTIAL ITEM FUNCTIONING (DIF)	
SUMMARY BY FORM	21-1
Mathematics	21-1
Reading	21-2
Writing	21-3
APPENDIX 22: OPERATIONAL TEST RELIABILITY BY SUBPOPULATION	22-1
Mathematics	22-1
Reading	22-5
Writing	22-9

APPENDIX 23: TOTAL SCORE AND SUBSCALE SCORE INTERCORRELATIONS 23-1

Mathematics Subscale Reporting Categories23-1

Reading Subscale Reporting Categories23-1

Writing Subscale Reporting Categories23-1

Grade 323-2

Grade 423-3

Grade 523-4

Grade 623-5

Grade 723-6

Grade 823-7

Grade 923-8

Grade 1023-9

APPENDIX 24: SUMMARY OF STUDENT DEMOGRAPHICS..... 24-1

Grade 324-1

Grade 424-2

Grade 524-3

Grade 624-4

Grade 724-5

Grade 824-6

Grade 924-7

Grade 1024-8

CHAPTER 1: BACKGROUND OF ALASKA ASSESSMENTS

The Standards Based Assessments (SBAs) in mathematics, reading, and writing are criterion based assessments that are aligned with the Alaska Grade Level Expectations (GLEs) for students in grades 3 through 10. The SBAs were first administered operationally in April 2005. Assessment items were extensively reviewed by Alaska educators and subsequently field tested in a standalone field test administered in October 2004.

Previously, assessments were administered in the benchmark grades 3, 6, and 8, and a commercial norm-referenced test was administered in grades 4, 5, 7, and 9. Provisions of the Improving America's School Act of 1994 mandated that states develop standards and related assessments for *each grade span*. In 1999, the Alaska State Board of Education and Early Development adopted performance standards in mathematics, reading, and writing. Subsequently, in 2001, provisions of the No Child Left Behind (NCLB) Act required that standards be developed *by grade* for grades 3 through 8 and that assessments be given that align with these grade-specific standards.

Spurred by PL 107-110, the NCLB and the timeline of having a fully compliant assessment system by 2006, Alaska's Commissioner of Education and Early Development (EED), in discussions with the Alaska State Board of Education, principals, teachers, superintendents, and other stakeholders determined that plans should be developed to administer a standards-based assessment in mathematics, reading, writing in grades 3 through 10 (the Standards Based Assessments).

The Alaska SBAs are a coherent set of assessments aligned with Alaska GLEs developed for students in grades 3 through 10. The core set of assessments consists of custom assessments in mathematics, reading, and writing in grades 3 through 10 that are suitable for reporting student achievement in relation to state proficiency standards, and for inclusion in state and federal school/district accountability programs.

CHAPTER 2: TEST DESIGN & ITEM DEVELOPMENT

MATHEMATICS ASSESSMENT MEASURES

The mathematics component of the SBAs is composed of items that address GLEs in grades 3 through 10. The assessable GLEs for each grade level are distributed among the six reporting categories: Numeration, Measurement, Estimation and Computation, Functions and Relationships, Geometry, and Statistics and Probability. The organization of the mathematics GLEs mirrors the categories used by the National Council of Teachers of Mathematics (NCTM) and the National Assessment of Educational Progress (NAEP). Information about the GLEs assessed in each reporting category, as well as the types and numbers of items used in each category, can be found in the test blueprints (Appendix 1).

Multiple-choice (MC), short constructed-response (SCR), and extended constructed-response (ECR) items are used to assess the mathematics GLEs. These item types are designed to measure students' knowledge at various cognitive levels and provide a variety of information about mathematics achievement.

Multiple-Choice Items

MC items require students to select a correct answer from four response choices with a single correct answer. Each MC item is scored as right or wrong and has a value of 1 point. MC items are used to assess a variety of skill levels, from short-term recall of facts to problem solving. The selection of incorrect response choices, or distractors, by the student commonly results from misunderstood concepts, incorrect logic, invalid application of an algorithm, or computation errors.

Constructed-Response Items

The mathematics constructed-response (CR) items are designed to address comprehension of mathematics at higher cognitive levels in ways that MC items cannot. They offer the opportunity to present real-life situations that require students to solve problems using mathematics skills learned in the classroom. Students must read the items carefully, identify the information needed to solve the tasks involved, devise a method of solution, perform the calculations, and, when required, offer explanations. This process provides insight into the students' mathematical knowledge, abilities, and reasoning processes.

There are two types of mathematics CR items: SCR and ECR. The student can earn 0–2 points on SCRs and 0–4 points on ECRs. Both types are scored using item-specific rubrics. The abbreviated tasks of SCRs and the more elaborate tasks of ECRs are carefully constructed to reflect the scoring rubrics. All item-specific scoring rubrics are based on generic rubrics, which are written by DRC test development specialists and approved by EED and committees of Alaska educators.

READING ASSESSMENT MEASURES

The reading component of the SBAs is composed of items that address GLEs in grades 3 through 10. The assessable GLEs for each grade level are distributed among the three major reporting categories: Word Identification Skills, Forming a General Understanding, and Analysis of General Content or Structure. Information about the GLEs assessed in each reporting category, as well as the types and numbers of items used in each category, can be found in the test blueprints (Appendix 1).

MC, SCR, and ECR items are used to measure the reading GLEs. All items in the reading assessment are derived from a selection of literary and informational passages.

Multiple-Choice Items

MC items require students to select a correct answer from four response choices with a single correct answer. Each MC item is scored as right or wrong and has a value of 1 point. The selection of incorrect response choices, or distractors, commonly results from misinterpretation, predisposition, unsound reasoning, or superficial reading.

Constructed-Response Items

The reading CR items are designed to address comprehension of text in ways that MC items cannot. Composing a written response requires the student to prepare an answer using supporting details or examples drawn from the text.

There are two types of reading CR items: SCR and ECR. The student can earn 0–2 points on SCRs and 0–4 points on ECRs. Both types of items are scored using an item-specific scoring rubric. The abbreviated tasks of SCRs and the more elaborate tasks of ECRs are designed to be passage-dependent. All item-specific scoring rubrics are based on generic scoring rubrics developed by DRC test development specialists and reviewed and approved by EED and committees of Alaska educators.

WRITING ASSESSMENT MEASURES

The writing component of the SBAs has three major reporting categories: Write Using a Variety of Forms, Structures and Conventions of Writing, and Revising. Information about the GLEs assessed in each reporting category, as well as the types and numbers of items used in each category, can be found in the test blueprints (Appendix 1).

The writing assessment employs four types of items: standalone and stimulus based MC, SCR, and ECR (i.e., writing prompts). The different items are designed to measure students' understanding of different writing forms, conventions, and types of revisions, and their ability to draft a rough-draft response to a writing prompt.

Multiple-Choice Items

MC items require students to select a correct answer from four response choices with a single correct answer. Each MC item is scored as right or wrong and has a value of 1 point. On the writing assessment there are two types of MC items: standalone and stimulus-linked. The stimulus-linked items are linked to a written passage, or stimulus. The selection of incorrect response choices, or distractors, typically represents misunderstanding, misinterpretation, or misidentification. MC writing items are preferred when attempting to determine whether students can identify various forms of writing, are familiar with different writing conventions, and can select correct revisions.

Constructed-Response Items

CR items are independent writing items requiring written student responses. There are two types: SCR and ECR. SCRs are 2-point items and are scored using an item-specific scoring rubric. ECRs are 4-point writing prompts at grade 3, and 4-point and 6-point writing prompts at grades 4 through 10. The ECRs are scored using a general rubric. The ECRs at all grade levels are presented with a writing checklist for students' use as they draft their response. In responding to the prompt, students may use scratch paper for planning, organizing, and drafting ideas. The student's first complete draft is what is scored, and this response to the writing prompt is **not** considered as a final, published piece of writing.

The abbreviated tasks of SCRs and the more elaborate tasks of ECRs are constructed to reflect the scoring rubrics. All item-specific rubrics are based on generic rubrics written by DRC test development specialists and reviewed and approved by EED and committees of Alaska educators. The generic rubrics for grades 4 through 10 were written using the Alaska 6-point instructional rubric as a guide, so that they provide a direct link to writing as it is taught and assessed in the classroom. The generic rubrics (Appendix 2) are placed on the EED Web site well in advance of the assessment so that educators are able to see how student responses will be scored.

2007 OPERATIONAL PLAN

The 2007 grades 3–9 SBAs in mathematics, reading, and writing were comprised of a single form at each grade level. The 2007 grade 10 SBAs in mathematics, reading, and writing were comprised of 14 forms. All of the forms contained identical core items for all students. In addition, each form also included a set of field test items that were embedded throughout each form at grades 3 through 9 and appended at grade 10.

Table 2–1 displays the design for the mathematics test for grades 3 through 10. The column entries for this table denote:

- the grade level
- number of core MC items
- number of field test MC items
- number of core SCR items
- number of core ECR items
- number of field test SCR and ECR
- total number of MC, CR (SCR and ECR) items
- total number of operational points

Table 2–1. Mathematics Test Plan 2007 per Operational Form

Grade	Multiple-Choice Items		Core SCR Items (2 pt.)	Core ECR Items (4 pt.)	FT CRs (2 pt. or 4 pt)	Total Items MC/CR	Total Operational Points
	Core	FT					
3	56	6	2	1	1	62/4	64
4	56	6	2	1	1	62/4	64
5	56	6	2	1	1	62/4	64
6	56	6	2	1	1	62/4	64
7	56	6	2	1	1	62/4	64
8	56	6	2	1	1	62/4	64
9	56	6	2	1	1	62/4	64
10	34	8–9*	1	1	0–1	42–43/ 2–3	40

* This is presented as a range because of the need to field test HSGQE, SBA, and dual-coded items on each form.

Table 2–2 displays the design for the reading test for grades 3 through 10. The column entries for this table denote:

- the grade level
- number of core MC items
- number of field test MC items
- number of core SCR items
- number of core ECR items
- number of field test SCR and ECR
- total number of MC, CR (SCR and ECR) items
- total number of operational points

Table 2–2. Reading Test Plan 2007 per Operational Form

Grade	Multiple-Choice Items		Core SCR Items (2 pt. or 3 pt.)	Core ECR Items (4 pt.)	FT CRs (2 pt. or 4 pt.)	Total Items MC/CR	Total Operational Points
	Core	FT					
3	52	10	2	1	1	62/4	60
4	52	10	2	1	1	62/4	60
5	52	10	2	1	1	62/4	60
6	52	10	2	1	1	62/4	60
7	52	10	2	1	1	62/4	60
8	52	10	2	1	1	62/4	60
9	52	10	2	1	1	62/4	60
10	62	10	1	2	1	72/4	72

Table 2–3 displays the design for the writing test for grades 3 through 10. The column entries for this chart denote:

- the grade level
- number of core MC items
- number of field test MC items
- number of core SCR items
- number of core 4 point or 6 point ECR items (writing prompt)
- number of field test SCR items
- number of field test ECR items
- total number of MC, CR (SCR and ECR) items
- total number of operational points

Table 2–3. Writing Test Plan 2007 per Operational Form

Grade	Multiple-Choice Items		Core SCR Items (2 pt.)	Core ECR Items (4 pt. or 6 pt.)	FT SCRs (2 pt.)	FT ECRs (4 pt.)	Total Items MC/CR	Total Operational Points
	Core	FT						
3	46	8	5	1	1	0	54/7	60
4	46	8	2	2	1	1	54/6	60
5	46	8	2	2	1	1	54/6	60
6	46	8	2	2	1	1	54/6	60
7	46	8	2	2	1	1	54/6	60
8	46	8	2	2	1	1	54/6	60
9	46	8	2	2	1	1	54/6	60
10	28	11	2	4	1	1	39/8	50

An individual student's score is based solely on the core items. The total number of operational points is 64 points for mathematics, 60 points for reading, and 60 points for writing at grades 3 through 9, and 40 points for mathematics, 72 points for reading and 50 points for writing at grade 10. The total raw score is obtained by combining the points from the core MC and core CR (SCR and ECR) portions of the test as follows:

Student's Score in Mathematics = **Grades 3–9:** 56 MC items plus
two 2-point SCR items plus
one 4-point ECR item = 64 points

Grade 10: 34 MC items plus
one 2-point SCR items plus
one 4-point ECR items = 40 points

Student's Score in Reading = **Grades 3–9:** 52 MC items plus
two 2-point SCR items plus
one 4-point ECR item = 60 points

Grade 10: 62 MC items plus
one 2-point SCR items plus
two 4-point ECR items = 72 points

Student's Score in Writing = **Grade 3:** 46 MC items plus
five 2-point SCR items plus
one 4-point ECR item = 60 points

Grades 4–9: 46 MC items plus
two 2-point SCR items plus
one 4-point ECR item plus
one 6-point ECR item = 60 points

Grade 10: 28 MC items plus
two 2-point SCR items plus
three 4-point ECR items plus
one 6-point ECR item = 50 points

GRADE-LEVEL EXPECTATIONS SUBSUMED WITHIN REPORTING CATEGORIES

The mathematics, reading, and writing content area reporting categories (or strands) are further subdivided for specificity and eligible content or limits. For mathematics, each reporting category is subdivided into one or more areas of emphasis. For reading and writing, each reporting category is subdivided into one or more performance standards. If assessable and usable in statewide assessment, each respective area of emphasis or performance standard is subdivided into one or more GLEs. Areas of emphasis or performance standards that are not assessable and usable in statewide assessment are identified for local assessment. Test items are not written for these locally assessable areas of emphasis or performance standards.

Total mathematics, reading, and writing scores reported at the student level are based on the common items for each reporting category. The number of GLEs subsumed within each reporting category by content area and grade are listed in the Tables 2–4 through 2–6.

Table 2–4. Number of Assessable Mathematics GLEs by Reporting Category and Grade

Reporting Category	Grade							
	3	4	5	6	7	8	9	10
Numeration	7	10	10	8	7	9	5	7
Measurement	7	5	4	6	5	2	2	1
Estimation and Computation	5	4	3	3	5	4	4	2
Functions and Relationships	3	2	4	4	5	5	6	6
Geometry	6	6	5	7	7	8	4	6
Statistics/Probability	4	5	5	4	5	5	6	6
Total	32	32	31	32	34	33	27	38

Table 2–5. Number of Assessable Reading GLEs by Reporting Category and Grade

Reporting Category	Grade							
	3	4	5	6	7	8	9	10
Word Identification Skills	3	4	4	4	4	4	4	4
Forming a General Understanding	6	7	7	7	6	6	9	9
Analysis of General Content or Structure	4	6	8	10	9	9	11	11
Total	13	17	19	21	19	19	24	24

Table 2–6. Number of Assessable Writing GLEs by Reporting Category and Grade

Reporting Category	Grade							
	3	4	5	6	7	8	9	10
Write Using a Variety of Forms	2	3	4	4	5	5	5	5
Structures and Conventions of Writing	2	2	2	4	5	5	5	5
Revise	1	1	1	1	4	4	4	4
Total	5	6	7	9	14	14	14	14

TEST DEVELOPMENT TIMELINE

A series of major test development activities took place in 2006 and 2007, which culminated in the administration of the operational SBA assessments in April 2007. These key activities included the:

- Development of items, tasks, and writing prompts.
- Review of items by external committees of educators (content review, bias/sensitivity review).
- Field testing of new mathematics, reading, and writing items in April 2006.
- Review of items by external committees of educators (item review with data).
- Final selection of items used to construct the 2007 SBAs.

Table 2–7 provides a high-level timeline of these major activities, which are described in detail in this report.

Table 2–7. General Timeline Associated with 2006 Field Testing and 2007 Operational Assessment of Mathematics, Reading, and Writing at Grades 3–10.

Time Frame	Activity
April 2006	Administration of 2006 assessment with field test items
August 2006	Content committee review of field tested items; review for statistical quality
August 2006	Content committee review of newly developed items for 2007 field test
August 2006	Bias/sensitivity committee review of newly developed items for 2007 field test
December 2006	Forms construction of spring 2007 operational test
April 2–16, 2007	2007 operational assessment administration grades 3–9
April 3–5, 2007	2007 operational assessment administration grade 10

ITEM AND TEST DEVELOPMENT PROCESS

The most significant considerations in the item and test development process were: aligning the items to the GLEs; determining the grade-level appropriateness (reading level/interest level, etc.); depth of knowledge; cognitive level; item/task level of complexity; estimated difficulty level; relevancy of context for each item; providing rationales for distractors; and determining style, accuracy, and correct terminology. In addition, the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999) and *Universal Design* (Thompson, Johnstone, & Thurlow, 2002) guided the following steps in the item and test development process:

- Analyze the GLEs and test blueprints.
- Analyze item specifications and style guides.
- Select qualified item writers.
- Develop item-writing workshop training materials.
- Train test development specialists and item writers to write items.
- Write items that match the standards, are free of bias, and address fairness and sensitivity concerns.
- Conduct and monitor internal item reviews and quality processes.
- Prepare passages and items for review by committees of Alaska educators (content and bias/sensitivity).
- Select and assemble items for field testing.
- Field test items, scoring of the items, and analysis of the data.
- Review items and associated statistics after field testing, including bias statistics.
- Select and assemble items for operational forms (test construction).

Item Writer Training

The test items were written by internal DRC item writers who have experience writing items, and selected writers from across the country who are experienced writers, teachers, or former teachers who have a great deal of specialized knowledge in the subject area of their expertise. All writers met the following qualifications:

- A bachelor's degree or higher in mathematics, reading, writing, curriculum and instruction, and/or related field.
- In-depth understanding and knowledge of the special considerations involving the writing of standards-based multiple-choice items, including an understanding of cognitive levels,

estimated difficulty levels, grade-level appropriateness, depth of knowledge, readability, and bias considerations.

- In-depth understanding and knowledge of the special considerations involving the writing of standards-based constructed-response (0–2 point and 0–3 or 0–4 point) items, including the writing of scoring rubrics for each item.
- For the writing tests, in-depth understanding and knowledge of the special considerations involving the development of writing prompts (1–6 point) with scoring guidelines. General rubrics are found in Appendix 2.

All item writers were provided with an in-depth training workshop coupled with one-on-one writing sessions with DRC test development specialists and lead item writers. Prior to developing items for the SBAs the cadre of item writers was trained with regard to:

- Alaska content standards, performance standards, and GLEs.
- Cognitive levels, including depth of knowledge.
- Principles of universal design.
- Skill-specific and balanced test items for the grade level.
- Contextual relevance.
- Developmentally appropriate structure and content.
- Item-writing technical quality issues.
- Style considerations and item specifications approved by the EED.

The *DRC Item Writer Training Manual*, *Fairness in Testing Manual*, *Depth of Knowledge Levels*, and *The Principles of Universal Design* document that were used during the training are provided in Appendices 3–6.

Reading Passage Selection

All reading items in the reading assessment were derived from a selection of literary and informational passages. Passages acquired were “authentic” in that they were culled from published materials or commissioned from experienced passage writers. To be used in the SBAs, approval to reprint published materials was secured from the publisher.

Passage finders and reading content specialists who have teaching experience at specific grade levels were given formal training on the specific requirements of the Alaska assessments. Passages were submitted to DRC’s reading test development team for screening and editing internally. The team screened and edited passages for:

- Interest and accuracy of information in a passage to a particular grade level.
- Grade-level appropriateness of passage topic and vocabulary.
- Rich passage content to support the development of high-quality test questions.
- Bias, sensitivity, and fairness issues.

- Readability considerations and concerns.

Passages that survived this extensive screening process were prepared for a formal committee review by Alaska grade-level reading teachers who read and reviewed the passages for the same criteria listed above. The Alaska Bias and Sensitivity Committee also read and reviewed the same passages for issues related to bias, sensitivity, and fairness. Passages were accepted, edited, and/or rejected by both committees of Alaska educators. Comments and concerns were noted, and EED provided DRC with the final determination as to whether or not a passage was approved. The final selection of passages to be field tested was based on the specific requirements for each grade-level assessment such as the percent of fiction and nonfiction, gender and ethnicity considerations, and diversity of passage topics.

Passage Readability

The readability of a passage was a judgmental process made by Alaska grade-level classroom teachers, DRC's reading content specialists, and other individuals who understand each particular grade level and children of a particular age group. In addition, formal readability programs were also used by DRC to provide a "snapshot" of a passage's reading difficulty based on sentence structure, length of words, etc. All of this information, along with the classroom context and content appropriateness of a passage, was taken into consideration when selecting a passage for a particular grade.

Item Writing

To ensure that all test items met the requirements of the approved target content test blueprint and item specifications and were adequately distributed across subcategories and levels of difficulty, item writers were asked to document the following specific information as each item was written.

Alignment to the Alaska Grade-level Expectations: There must be a high degree of match between a particular question and the GLE it is intended to measure. Item writers were asked to clearly indicate what GLE each item was measuring.

Estimated Difficulty Level: Prior to field testing items, the item difficulties were not known, and writers could only make approximations as to how difficult an item might be. The estimated difficulty level was based upon the writer's own judgment as directly related to his or her classroom teaching and knowledge of the curriculum for a given subject area and grade level. The purpose for indicating estimated difficulty levels as items were written was to help ensure that the pool of items prepared for review by Alaska educators and EED and subsequent field testing would include a range of difficulty (easy, medium, and challenging).

Appropriate Grade Level, Item Context, and Assumed Student Knowledge: Item writers were asked to consider the conceptual and cognitive level of each item. They were asked to review each item to determine whether or not the item was measuring something that was important and could be successfully taught and learned in the classroom.

Multiple-choice (MC) Item Options and Distractor Rationale/Analysis: Writers were instructed to make sure that each item had only one clearly correct answer. Item writers submitted the answer key with the item. All distractors were plausible choices that represented

common errors and misconceptions in student reasoning. The rationale/distractor analysis for each distractor for mathematics was also provided.

Constructed-Response (CR): Each constructed-response item (SCR and ECR items) included specific scoring rubrics. Specific scoring rubrics were complete and explained why each score point would be assigned. The complete item-specific rubrics were also written to explain the strengths and weaknesses that were typically displayed for each score point.

Face Validity and Distribution of Complexity Levels: Writers were instructed to write items to reflect various levels of cognitive complexity using *Taxonomy of Educational Objectives* (Bloom et. al., 1956). As each item was written, the writer classified one of four cognition levels: recall, application, analysis, or evaluation for each item. The writers were instructed to write items so that the pool of items would represent a distribution of items across cognitive levels, as required by the test and item specifications.

Face Validity and Distribution of Items Based Upon Depth of Knowledge: Writers were asked to classify the depth of knowledge of each item, using a model based on Norman Webb's work on depth of knowledge (Webb, 2002). Items were classified as one of four depth of knowledge categories: recall, skill/concept, strategic thinking, and extended thinking.

Readability: For mathematics item development, writers were instructed to pay careful attention to the readability of each mathematics item to ensure that the focus was upon the concepts; not on reading comprehension. As a result, the goal for each mathematics writer was to write items that were, to the greatest degree possible, independent of the assessment of reading. Subject areas such as mathematics contain many content-specific vocabulary terms. These terms make it impossible to use the standard methods available for determining the reading level of test questions. Wherever it is practical and reasonable, every effort was made to keep the vocabulary one grade level below the tested grade level. Resources writers used to verify the vocabulary level were the *EDL Core Vocabularies* (Taylor et.al., 1989) and the *Children's Writer's Word Book* (Mogilner, 1992). In addition, every mathematics test question was taken before committees comprised of Alaska grade-level experts in the field of mathematics education. They reviewed each question from the perspective of the students they teach, and they determined the validity of the vocabulary used.

Curriculum-specific Issues: All items were to be curriculum independent with respect to both content and vocabulary. As items were written, writers were asked to document any specific curriculum issues.

Grammar and Structure for Item Stems and Item Options: All items were written to meet technical quality, including correct grammar, syntax, and usage in all items, as well as parallel construction and structure of text associated with each multiple-choice item.

Editorial Review of Items

After items were written, DRC test development specialists and editorial staff reviewed each item for item quality, making sure that the test items were in compliance with industry guidelines for clarity, style, accuracy, and appropriateness for Alaska students. While there are many published guidelines for reviewing assessment items, the list below serves to summarize some of the more major considerations DRC test development specialists and editors followed when

reviewing items to make sure they conformed to standard item quality for good, reliable, fair test questions.

Guidelines for Reviewing Assessment Items

A good item should

- have only one clear correct answer and contain answer choices that are reasonably parallel in length and structure.
- have a correctly assigned content code (item map).
- measure one main idea or problem.
- measure the objective or curriculum content standard it is designed to measure.
- be at the appropriate level of difficulty.
- be simple, direct, and free of ambiguity.
- make use of vocabulary and sentence structure that is appropriate to the grade level of the student being tested.
- be based on content that is accurate and current.
- when appropriate, contain stimulus material that are clear and concise and provide all information that is needed.
- when appropriate, contain graphics that are clearly labeled.
- contain answer choices that are plausible and reasonable in terms of the requirements of the question, as well as the students' level of knowledge.
- contain distractors that relate to the question and can be supported by a rationale.
- reflect current teaching and learning practices in the subject area.
- be free of gender, ethnic, cultural, socioeconomic, and regional stereotyping bias.

Item Content Review

Prior to field testing, all newly developed test items were submitted to content committees for review. The content committees consisted of Alaska teachers and subject-area supervisors from school districts throughout Alaska. The primary responsibility of the content committee was to evaluate items with regard to quality and content classification, including grade-level appropriateness, estimated difficulty, depth of knowledge, and source of challenge. They also suggested revisions, if appropriate. The committee also reviewed the items for adherence to the principles of universal design, including language demand and issues of bias, fairness, and sensitivity.

The content review was held July 31, August 1, 2, and 3, 2006. Committee members were selected by EED, and EED-approved invitations were sent to them by DRC. The committee consisted of 60 educators, 20 for each content area. EED also selected internal staff members for attendance. The meeting commenced with an overview of the test development process. Training was also provided by DRC senior staff members. Training included how to review items for technical quality and content quality, including depth of knowledge and adherence to principles

of universal design. In addition, training included providing committee members with the procedures for item review, including the use of tracking review forms to be used during the item content review.

DRC test development specialists in mathematics, reading, and writing facilitated the review of items. Committee members, grouped by grade span and content area, reviewed the items for quality and content, as well as for the following categories designated on the item review tracking form. An example of this form is found in Appendix 7.

- GLE Alignment
- Difficulty Level (classified as Low, Medium, or High)
- Depth of Knowledge (classified as Recall, Application, or Strategic Thinking)
- Correct Answer
- Quality of Graphics
- Appropriate Language Demand
- Freedom from Bias (classified as Yes or No)
- Overall Judgment (classified as Approved, Accept with Revisions, Move to another grade level, or Rewrite)

Security was addressed by adhering to a strict set of procedures. Items in binders did not leave the meeting rooms and were accounted for at the end of each day before attendees were dismissed. All attendees, with the exception of EED staff, were required to sign a Confidentiality Agreement (Appendix 8). All materials not in use at any time were kept in secure meeting rooms. During lunch and breaks, if meeting rooms were unused, they were locked or closely monitored by DRC personnel. While not in use by DRC, the meeting rooms were locked and unavailable to anyone other than one DRC person and the Chief of Security of the meeting facility. Rooms were attended to only under strict supervision by DRC personnel. Secure materials that did not need to be retained after the meeting were deposited in secure barrels, and their contents were shredded under supervision of a DRC employee.

Bias and Sensitivity Review

Prior to field testing, all newly developed test items were also submitted to a Bias and Sensitivity Committee for review. This took place on July 31 and August 1, 2006. The committee's primary responsibility was to evaluate passages and items as to acceptability with regard to bias and sensitivity issues. They also made recommendations for changes or deletion of items in order to remove the area of concern. The bias/sensitivity committee was composed of 11 men and women who represented the diversity of Alaska students. The committee was trained by a DRC test development lead to review items for bias and sensitivity issues using a Fairness in Testing Manual developed by DRC (Appendix 4). This manual was customized specifically for the Alaska program.

All mathematics, reading, and writing items were read by all of the committee members. Each member noted bias and/or sensitivity comments on review forms (Appendix 9). All comments were then compiled and the actions taken on these items were recorded by DRC. Committee members were required to sign a Confidentiality Agreement (Appendix 8) and strict security measures were in place to ensure that secure materials did not leave the meeting rooms. All secure materials were kept in a locked room while not in use. Secure materials that did not need to be retained after the meeting were deposited in secure barrels and their contents were shredded under supervision of a DRC employee.

Item Field Test

Items being field tested were embedded at grades 3–9 and appended at grade 10 to the spring 2007 administrations.

Item Field Test Data Review

Prior to the construction of operational forms, the following field test statistical analyses were completed:

- Proportion selecting correct response (p -values)
- Average person logit for all choices
- Number of persons attempting the item
- Item-total correlations
- Fit statistics
- Differential item functioning (DIF)
- Logit difficulty of item

Item analysis results were reviewed by DRC psychometricians to identify any items that were not performing as expected. These items were flagged so DRC test development specialists were made aware of potential areas of concern. For example, in the case of multiple-choice items, DRC test development specialists checked to make sure that the key for each item was correct and that none of the other response options were plausible. In the case of items where large values of DIF occur, DRC test development specialists reviewed each item flagged to consider whether or not a feature of the item may have caused a problem and/or contributed to the DIF. Under the guidance of DRC psychometricians, DRC test development specialists determined which of the flagged items were to be reviewed by a group of Alaska educators to determine whether or not the item was appropriate for use. In many cases, items with extreme DIF were removed from the pool of items available for use in forms construction. Additional guidelines concerning the review of item analysis results for the item-selection process are provided on page 29.

Items not identified for this review were those that had good statistical characteristics and, consequently, were regarded as statistically acceptable. Likewise, items of extremely poor statistical quality were regarded as unacceptable and needed no further review. However, there were some items that DRC deemed as needing further review by a committee of Alaska

educators. The intent was to capture all items that needed a closer look; thus the criteria employed tended to over-identify rather than under-identify items.

The review of the items with data was conducted on July 31 and August 1, 2006 and included content committees composed from 60 Alaska educators. EED also selected internal staff members to attend. Committee members were selected by EED, and EED-approved invitations were sent to them by DRC. In this session committee members were first trained by a DRC senior psychometrician with regard to the statistical indices used in item evaluation. This was followed by a discussion with examples concerning potential reasons why an item might be retained regardless of the statistics. The committee review process involved a brief exploration of possible reasons for the statistical profile of an item (such as possible bias, grade appropriateness, and instructional issues) and a decision regarding acceptance. DRC test development specialists facilitated the statistical review of the items.

Security was addressed by adhering to a strict set of procedures. Test items did not leave the meeting rooms and were accounted for at the end of each day before attendees were dismissed. All attendees, with the exception of EED staff, were required to sign a Confidentiality Agreement (Appendix 8). All materials not in use at any time were kept in secure meeting rooms. During lunch and breaks, if meeting rooms were unused, they were locked or closely monitored by DRC personnel. While not in use by DRC, the meeting rooms were locked and unavailable to anyone other than one DRC person and the Chief of Security of the meeting facility. Rooms were attended to only under strict supervision by DRC personnel. Secure materials that did not need to be retained after the meeting were deposited in secure barrels, and their contents were shredded under supervision of a DRC employee.

The results of the August 2006 Data Review are shown in Tables 2–8 through 2–10.

Table 2–8. Mathematics Items at Data Review

August 2006 Data Review

Grade	Accept	Accept with Revisions	Accept Total	% Accept	Reject	Total
3	9	1	10	40%	15	25
4	15	0	15	65%	8	23
5	7	0	7	47%	8	15
6	14	0	14	52%	13	27
7	25	0	25	89%	3	28
8	17	0	17	77%	5	22
9	24	2	26	87%	4	30
10	95	0	95	87%	14	109

Table 2–9. Reading Items at Data Review

August 2006 Data Review

Grade	Accept	Accept with Revisions	Accept Total	% Accept	Reject	Total
3	26	3	29	83%	6	35
4	28	1	29	85%	5	34
5	28	3	31	91%	3	34
6	23	0	23	82%	5	28
7	30	0	30	86%	5	35
8	21	0	21	75%	7	28
9	29	0	29	74%	10	39
10	77	0	77	85%	14	91

Table 2–10. Writing Items at Data Review

August 2006 Data Review

Grade	Accept	Accept with Revisions	Accept Total	% Accept	Reject	Total
3	3	0	3	75%	1	4
4	13	0	13	72%	5	18
5	19	0	19	100%	0	19
6	20	0	20	80%	5	25
7	24	0	24	92%	2	26
8	25	0	25	92%	2	27
9	18	1	19	68%	8	28
10	23	2	25	76%	8	33

PSYCHOMETRIC GUIDELINES FOR SELECTING ITEMS

Proportion Correct (also known as *p*-value)

The proportion correct, or *p*-value, is the proportion of the total group of test takers answering the question correctly. The proportion for an item will show how difficult the item was for the students who took that field test form. In general, MC items with a proportion somewhat higher than half the difference between the chance level and 1.00 should be recommended for selection first, and the range for selection should be between 0.40–0.90. When necessary to meet the test blueprint or other test specifications, items that fall outside this range may be used, albeit sparingly. The overall form was constructed to a target range of .63 to .67, with special care taken to select items that were at or near the cutpoints.

Average Person Logit

The average person logit for an item is the average measure of the persons attempting that item, which can vary from field test form to field test form. The average person logit for a response option is the average measure for the persons selecting that response. The average person logit for the correct response should be greater than the average logit for every other response. The difference between the average person logit for the correct response and the incorrect responses is an indication of the discrimination of the item. The larger the difference, the more discriminating the item. Item discrimination is also estimated by the item-total correlation.

Item-Total Correlation

The item-total correlation is the relationship between a student's performance on the item and the student's performance on the content-area test as a whole. If the item has a high item-total correlation, it generally means that the students who answered the item correctly achieved higher scores on the test than those who did not answer the item correctly. Item discrimination is an important statistic in the forms construction process, because the higher the average value for the test, the more reliable the test. Items with item-total correlations of 0.35 or greater were given primary consideration in the item selection phase of the test development process. The use of 0.35 is a rule of thumb that meets best practices. This value is higher for operational items because the item-total correlation for Alaska field-test items generally decreases from field test to operational test. However, items with item-total correlation values between 0.20 and 0.35 for a standards-based assessment were considered, if the inclusion of such items was necessary to satisfy specific content cells of the detailed test blueprint.

Fit Statistic

A goodness-of-fit statistic is computed as part of the calibration of all items in the field test. Essentially, a chi-square statistic is computed for each item that represents the sum of the squared standardized distances of the observed item performance from the expected performance for all persons, based on the Rasch model. This statistic evaluates how well each item fits the psychometric model. Poor fit could be a result of an item not functioning as expected or because the item measures a different construct than the remaining items. Typically items with values greater than +5 would be considered suspect.

Differential Item Functioning (DIF) Analyses

DIF analysis is conducted on all field test items to determine whether an item favors one group of students over another. DIF procedures examine the possibility that an item's characteristics may negatively affect the performance of select groups of students. Evidence of DIF is usually considered as a signal to test developers to examine an item more closely to consider whether or not it is defective.

DRC utilizes the Mantel-Haenszel (MH) or the Standardized Mean Difference (SMD) statistic for detecting DIF, depending on the item type. The MH statistic is the most commonly used technique for MC items in large-scale, educational assessment. It does not depend on the application or the fit of any specific measurement model. The SMD statistic is used for CR items with more than two score categories.

Essentially, these methods compute a value, which is the average amount more difficult that a member of the reference group found the studied item than did comparable members of the focal group. From this value, one of three severity classification categories is assigned (A, B, or C). The A category represents negligible DIF. The B category indicates moderate potential DIF, that is to say, that one group outperformed the other group once differences in skill levels between the two groups have been accounted for. The C category indicates that there is large potential DIF. Items assigned an A are given primary consideration. C items are considered if the inclusion of such items is necessary to satisfy specific content cells of the detailed test blueprint or other test specifications. However, these items must pass committee review before they are placed on an operational form.

ITEM BANK

Overview

The DRC item bank is a secure, searchable database. The item bank stores items along with associated graphic images, item characteristics (e.g., item ID, standard, answer key, subject, grade), administration information (e.g., form, sequence, year of administration), as well as item level statistics (e.g., p -values (proportion correct), item-total correlations, and omits (proportion leaving an item blank)). Items are maintained throughout an item's lifecycle from development through the form construction phase. Information about each item is accessible using the item bank's searching and reporting capabilities in the following situations: determining item development needs, constructing field test and operational test forms, locating released or rejected items, as well as verifying or researching information from committee review sessions.

Functionality

A unique, sequential item ID is assigned to items when they enter the bank. This ensures that each item is uniquely identified throughout its lifecycle with one item ID. Another client-specific item ID may also be assigned.

Current and historic information about item status and characteristics are easily accessible in the item bank. Item characteristics (e.g., standard, key, passage type, calculator status, etc.) are searchable and viewable in the item bank. The item image and associated graphics are also stored in the item bank. The items and graphics can be viewed and versioned based upon suggested modifications by committees and internal edits. Versioning allows changes to be made and archived for reference.

Item status information from committee review sessions is stored in the database. Items accepted by committees are available for form construction. Conversely, items rejected by committees remain in the database for reference and are flagged so they are not available for future test forms.

Item Cards and Reporting Options

Common outputs of the item bank include item cards and user-defined reports. DRC's item cards contain item text and associated graphics, unique item identifiers, as well as applicable administration and statistical information. Item cards are used for committee reviews, client reviews, and form construction purposes.

Information is queried in the item bank to generate reports. For example, a list of items with their associated statistics can be printed for a specific administration or a list of rejected or released items can be printed for reference.

Security

While the viewing options in the item bank are read-only only approved DRC employees are allowed to make modifications or changes to items and their associated item level administration information.

Quality Assurance

The item bank is the central repository of all item level information at DRC. All changes to an item, its graphic, and associated item-specific information are made in this database. This allows our test development specialists to access the most current, reliable information available at any time in the item and form development processes.

The integrity of the item bank is maintained by tracking changes to items, graphics, and associated information during all stages of development. Similarly, item status codes reflect the availability of an item so that only the most recent version of an item image is placed on a test form. Items which have been released or rejected are flagged so that they are not available for form construction purposes.

During the form construction process, information is extracted from the item bank: DRC relies on the accuracy of the information stored in the item bank. DRC strives to make updates to items and all item related information in a timely manner to ensure the accuracy and reliability of the bank.

Item Bank Summary

The number of eligible items before the spring 2007 SBA forms were built is presented in Tables 2–11 to 2–13. The grade 10 item summary table for each content area shows eligible items after the fall 2006 HSGQE Retest was built. Items doubled coded to both a HSGQE performance standard and a SBA Grade Level Expectation will appear in both Item Bank Summary tables in this document and the 2007 HSGQE Technical Report.

Table 2–11. SBA Mathematics Items

Grade 3

Standard	MC	CR
Numeration	42	8
Measurement	26	3
Estimation & Computation	42	4
Functions & Relations	27	2
Geometry	36	6
Statistics & Probability	24	5

Grade 4

Standard	MC	CR
Numeration	45	10
Measurement	26	3
Estimation & Computation	42	4
Functions & Relations	28	3
Geometry	35	3
Statistics & Probability	31	5

Grade 5

Standard	MC	CR
Numeration	38	7
Measurement	26	6
Estimation & Computation	48	4
Functions & Relations	28	0
Geometry	35	2
Statistics & Probability	28	9

Grade 6

Standard	MC	CR
Numeration	31	4
Measurement	33	5
Estimation & Computation	30	2
Functions & Relations	32	4
Geometry	41	4
Statistics & Probability	32	4

Table 2–11 (continued). SBA Mathematics Items

Grade 7

Standard	MC	CR
Numeration	29	6
Measurement	39	5
Estimation & Computation	27	4
Functions & Relations	43	2
Geometry	43	2
Statistics & Probability	29	7

Grade 8

Standard	MC	CR
Numeration	32	5
Measurement	26	4
Estimation & Computation	30	6
Functions & Relations	42	5
Geometry	42	4
Statistics & Probability	30	4

Grade 9

Standard	MC	CR
Numeration	28	0
Measurement	26	6
Estimation & Computation	23	0
Functions & Relations	47	4
Geometry	43	6
Statistics & Probability	35	10

Grade 10

Standard	MC	CR
Numeration	28	1
Measurement	13	0
Estimation & Computation	27	6
Functions & Relations	45	6
Geometry	37	6
Statistics & Probability	36	5

Table 2–12. SBA Reading Items

Grade 3

Standard	MC	CR
Word Identification	127	0
Forming a General Understanding	172	23
Analysis of Content and Structure	52	6

Grade 4

Standard	MC	CR
Word Identification	77	0
Forming a General Understanding	176	24
Analysis of Content and Structure	68	9

Grade 5

Standard	MC	CR
Word Identification	74	0
Forming a General Understanding	162	24
Analysis of Content and Structure	80	10

Grade 6

Standard	MC	CR
Word Identification	63	0
Forming a General Understanding	162	19
Analysis of Content and Structure	98	15

Grade 7

Standard	MC	CR
Word Identification	69	0
Forming a General Understanding	170	23
Analysis of Content and Structure	92	11

Grade 8

Standard	MC	CR
Word Identification	71	0
Forming a General Understanding	182	20
Analysis of Content and Structure	71	14

Table 2–12 (continued). SBA Reading Items

Grade 9

Standard	MC	CR
Word Identification	46	0
Forming a General Understanding	165	22
Analysis of Content and Structure	80	12

Grade 10

Standard	MC	CR
Word Identification	41	0
Forming a General Understanding	173	49
Analysis of Content and Structure	79	7

Table 2–13. SBA Writing Items

Grade 3

Standard	MC	CR
Write Using a Variety of Forms	47	15
Structures and Conventions of Writing	68	5
Revising	27	7

Grade 4

Standard	MC	CR
Write Using a Variety of Forms	37	24
Structures and Conventions of Writing	73	6
Revising	35	11

Grade 5

Standard	MC	CR
Write Using a Variety of Forms	60	32
Structures and Conventions of Writing	49	1
Revising	39	7

Grade 6

Standard	MC	CR
Write Using a Variety of Forms	29	29
Structures and Conventions of Writing	82	5
Revising	23	8

Table 2–13 (continued). SBA Writing Items

Grade 7

Standard	MC	CR
Write Using a Variety of Forms	42	26
Structures and Conventions of Writing	62	8
Revising	37	7

Grade 8

Standard	MC	CR
Write Using a Variety of Forms	42	25
Structures and Conventions of Writing	62	5
Revising	43	9

Grade 9

Standard	MC	CR
Write Using a Variety of Forms	48	28
Structures and Conventions of Writing	51	8
Revising	38	3

Grade 10

Standard	MC	CR
Write Using a Variety of Forms	49	24
Structures and Conventions of Writing	58	5
Revising	57	10

FINAL SELECTION OF ITEMS AND SPRING 2007 SBA OPERATIONAL FORMS CONSTRUCTION

The test forms for the spring 2007 SBAs were constructed to meet the target range of the content specifications set forth in the target test blueprints, as well as meet psychometric standards for excellence. Forms construction was accomplished with all forms reflecting a range of valid content at the appropriate level of difficulty. The following information documents the steps DRC's test development specialists took in the test forms construction process to ensure that the SBAs are of high quality, legally defensible, and meet the requirements as outlined by the Alaska testing program.

Steps in the Forms Construction Process

1. DRC test development specialists reviewed the content standards and test blueprints, including the number of items per domain or reporting category for each content-area test.
2. DRC psychometricians provided DRC test development specialists with the psychometric guidelines for operational forms construction.
3. DRC psychometricians analyzed item statistics for the field tested items and provided DRC test development specialists with characteristics for each item.
4. DRC test development specialists received all item cards and verified that each item image had its correct item characteristics and psychometric data.
5. DRC test development specialists reviewed all items in the operational pool and made an initial selection of items according to test blueprint guidelines and psychometric guidelines.
6. DRC test development specialists created item-mapping charts for the test.
7. Final recommendations for items selected for the operational forms were prepared for review by senior test development staff.
8. Based upon senior review, suggested replacements were made by DRC test development specialists, if necessary.
9. Operational forms were prepared for psychometric review and approval.
10. Based upon psychometric review, suggested replacements were made by DRC test development specialists, if necessary.
11. Operational forms were prepared for EED review and approval.

Construction of the Operational Forms

In constructing the forms, DRC test development specialists followed the guidelines provided in the list below.

Guidelines for Placing Items into Forms

- Forms will include an adequate objective coverage, as required by the detailed test blueprint.
- No item in a form will “clue” another item on that same form.
- “Clang” will be avoided (i.e., distractors should be unique from one another).
- Forms will be ethnically diverse, both in terms of artwork and in terms of names.
- Forms will target an equal representation of genders, both in terms of artwork and names.
- Forms will include a wide range of topics and a variety of questions.
- Correct answer distributions will be distributed such that approximately 25 percent of them are A, B, C, or D.
- Overall form will be within the target p -value range of 0.63–0.67 with particular care taken to select items at or near the cutpoints.

DRC INTERNAL REVIEW OF THE ITEMS AND FORMS

At every stage of the test development process the match of the item to the content standard was reviewed and verified since establishing content validity is one of the most important aspects in the legal defensibility of a test. As a result, it is essential that an item selected for a form link directly to the content curriculum standard and performance standard to which it is measuring. DRC test development specialists verified all items against their classification codes and item maps, both to evaluate the correctness of the classification and to ensure that the given task measures what it purports to measure.

CHAPTER 3: TEST ADMINISTRATION PROCEDURES

OVERVIEW

The 2007 SBA reading, writing, and mathematics assessments were administered to students in grades three through nine during the spring of 2007. A District Test Coordinator was assigned at every school district. The test administration window was April 2 through April 16, 2007. Specific statewide testing days were not designated this year. Districts followed the designated SBA test order: reading test (first), writing test (second), and mathematics test (last). Variations on this schedule were not permitted. Districts were not required to set the SBA testing schedule on consecutive days during the test window. DRC distributed the testing materials to each District Test Coordinator (DTC).

STUDENT POPULATION TESTED

Districts submitted their enrollment, accommodated materials counts and updates to district contact information via DRC's Online Enrollment System January 8-23, 2007. Districts also submitted their precode files January 8-23, 2007. Districts with 30 or more schools and 9,000 or more students were given the option to submit their enrollment files directly to DRC by January 23, 2007. Mat-Su, Anchorage, and Fairbanks took advantage of this offer and were locked out of DRC's Online Enrollment System. In addition, those districts were allowed to submit their precode files directly to DRC by February 23, 2007 with precode and district/school labels arriving in these districts by March 15, 2007.

The enrollment and documents processed counts were as follows:

Table 3–1. Project Counts

District Count	School Count
54	490
Enrollment Count	Processed Count
Grade 3: 10,244	Grade 3: 9,261
Grade 4: 10,372	Grade 4: 9,355
Grade 5: 10,327	Grade 5: 9,299
Grade 6: 10,634	Grade 6: 9,492
Grade 7: 10,811	Grade 7: 9,858
Grade 8: 10,765	Grade 8: 9,799
Grade 9: 11,853	Grade 9: 10,520
Grade 10: 11,358	Grade 10: 9,844

ACCOMMODATIONS

Students with disabilities may use appropriate accommodations when taking assessments. These accommodations must be documented in an Individualized Education Program (IEP) or in a 504 plan. Refer to the Participation Guidelines for examples of acceptable accommodations: (http://www.eed.state.ak.us/tls/assessment/participation_guidelines/ParticipationGuidelinesJune2005Final.pdf).

SPIRALING PLAN

- Forms were spiraled by district for the 47 smallest districts. DRC’s Psychometric Services Team determined which schools received which form.
- Forms were spiraled by student for the 7 largest districts – Anchorage, Fairbanks, Mat-Su, Kenai, Juneau, Lower Kuskokwim, and Galena.
- A single common form was provided for those students in the 7 largest districts who required a “read aloud” administration.
- The number of students in the 7 largest districts needing a “read aloud” administration was collected via the Online Enrollment System.

TEST ADMINISTRATOR TRAINING

DTCs were trained in March 2007 by EED and DRC. The training focused on test materials receipt, distribution and return procedures, and general testing information. DTCs scheduled training sessions with test administrators during March and April 2007.

TEST SECURITY

The SBA materials are considered secure materials. According to Alaska test security regulation 4 AAC 06.765, all test materials must be kept secure. Materials may not be photocopy or duplicated any portion of the test materials at any time. Except for the person testing, no person, including test administrators, is permitted to read test items on the SBA prior to, during (except for the student testing), or after administration. Teachers, proctors, test administrators, or any testing personnel may not read test items aloud, silently, to themselves, or to another individual, unless specifically required to provide a documented accommodation to an individual or student group. Parents/guardians may not read test items under any circumstances.

The DTC shall designate the school and district personnel who will have access to secure test materials, and who must sign the Test Security Agreements. All signed test security forms must be returned to the DTC who will keep them on file in the district.

Prior to the first test administration of the school year, DTCs must sign and send their District Test Coordinator Test Security Agreement to EED.

MATERIALS

The following materials were produced for this administration:

- *District Test Coordinator's Manual*
- *Test Administration Directions*
- Form B Reading Test Books - grades 3 and 4
- Form B Writing/Mathematics Test Books - grades 3 and 4
- Form B Reading/Writing/Mathematics Test Books - grades 5-9
- Form B Reading/Writing/Mathematics Answer Booklets - grades 5-9
- Form F Reading Test Books - 14 versions per grade for grade 10
- Form F Writing/Mathematics Test Books - 14 versions per grade for grade 10
- Large Print Test Books
- Braille Test Books
- Ancillary materials—rulers, protractors, large print and Braille rulers, large print and Braille protractors, precode labels, district/school labels, “Do Not Score” labels, return shipping labels, security checklists, school box range sheets, shipping rosters, and packing lists

Samples of the *District Test Coordinator's Manual* and *Test Administration Directions* are provided in Appendix 10.

Packaging and Shipping Materials

All materials were packaged by school and shipped to the districts in one shipment. All test materials arrived in the districts by March 5, 2007, as scheduled.

District ancillary materials were packed in the last box and labeled, “District Materials Enclosed.” Boxes were filled 75 percent full to allow for the fluff factor when districts returned their materials.

DRC overage was shrinkwrapped in groups of three. All secure materials were packaged by range sheet and shrinkwrapped. DRC barcoded and shrinkwrapped all accommodated materials.

DRC provided EED with a Point of Delivery Report on April 4, 2007. This report listed the date each district received their materials, the person who signed for the materials, and noted any special circumstances.

DRC entered, packed, and shipped requests for additional materials March 5-26, 2007. DRC processed 28 additional materials requests for this administration.

Materials Return

Districts returned all materials via Manna on April 19, 2007, and most materials arrived at DRC's warehouse on April 25, 2007. All districts used green DRC return shipping labels. DRC return shipping labels were district specific and included a line for District Test Coordinators (DTC) to indicate how many boxes they were returning to the DRC.

Box Receipt

As materials arrived, DRC's Materials Processing team checked the bill of lading to ensure that the number of boxes received matched the number signed for by the DTC and Manna. The Materials Processing team scanned each box using the OpsMMS box receipt system and notified DRC's Education Project Management (EPM) team of any districts that did not return a box as soon as box receipt was complete. DRC's automated system provided immediate information regarding materials return. DRC identified the date and time each box was checked in, where the box originated, and districts that did not return materials.

CHAPTER 4: SCORING & STUDENT PERFORMANCE REPORTING

DOCUMENT PROCESSING

All secure materials were scanned in by district through DRC's OpsMMS system to ensure accurate counts. Through an automated precount system, DRC counted the books before check-in and again at scanning to ensure counts matched. If a count didn't match, the books were reconciled to ensure accurate numbers. Customized testing materials were also barcoded and checked in securely.

The Materials Processing team produced a preliminary missing document report and performed a quality check based on this report. The report was then forwarded to EPM, who checked for the missing materials on the security checklists. If any documentation regarding the materials was found, the item was removed from the report.

DRC used its Image Scanning System to scan the SBA test books. Scanning test books and answer booklets was completed on May 4, 2007. All editing and validating rules were followed per the contract.

HANDSCORING OF CONSTRUCTED RESPONSES

For the Alaska SBAs, DRC employed a variety of score-point scales for scoring SCR and ECR.

Preliminary rubrics for field test items were written during the item development stage, and these rubrics were refined once live student responses are available for review. DRC staff used the rubrics and live student responses to build anchor sets and training materials for each item assessed. Writing constructed-response items were scored using "generic" (e.g., not item-specific) rubrics on 1–4 and 1–6 point scales (Appendix 2). DRC's performance assessment staff assisted in the crucial effort of writing and refining scoring rubrics.

READERS

The scorers for the Alaska SBAs were selected from DRC's larger pool of available professional test scorers. All of our readers for the Alaska SBAs had an undergraduate degree and background in the content areas being assessed.

DRC selects readers who are articulate, concerned with the task at hand, and, most importantly, flexible. Our readers must have strong content-specific backgrounds: they are educators, writers, editors, accountants, and other professionals. They are valued for their experience but, at the same time, are required to set aside their own biases about student performance and accept the scoring standards of the client's program. Candidates must demonstrate proficiency in the content areas they will be scoring. For example, mathematics scorer candidates must successfully solve a DRC mathematics problem and show all steps necessary to reach the correct answer. Reader candidates are asked to respond to a DRC writing topic.

Rangefinding and Developing Training Material

DRC's Scoring Directors and Content Specialists consensus scored "live" field test responses to create training materials for our scorers. During this process, student responses selected and the rubric and scoring guidelines were applied. DRC staff moved from item to item until a sufficient number of scored responses were compiled to construct training materials. Responses that were particularly relevant (in terms of the scoring concepts they illustrate) were annotated for use in the scoring guide. The scoring guide for each item served as the readers' constant reference. An anchor set and a training set were created for each field test item. For operational items, these materials were enhanced with the addition of further training sets and qualifying sets.

Training the Readers

The fundamental objective of any handscoring activity is that results be accurate and consistent. Therefore, it is important that high-quality methods of training and monitoring readers be employed.

Training for readers in each content area began with a room-wide presentation and discussion of the scoring guide by the Scoring Director and/or Team Leader. The scoring guide for each item contained the scoring rubric and anchor papers that were selected and annotated to define and articulate the score scale. Next, the readers "practiced" by scoring the responses in the training sets. The Scoring Director and/or Team Leaders then led a thorough discussion of each set.

After the scoring guide and all training sets were discussed, readers of operational items demonstrated their ability to apply the scoring criteria by qualifying (i.e., scoring with acceptable agreement with "true" scores) on at least one of the qualifying sets. Any readers who did not qualify by the end of the qualifying process were not allowed to score any Alaska "live" responses.

IMAGING

DRC used its Image Scanning and Scoring system for the handscoring of the responses to constructed-response items.

DRC's hardware environment to support the image handscoring system consists of a server-based solution, with hundreds of handscoring workstations (PCs). Each DRC scoring site has a server, a local area network (LAN), and workstations for readers, Team Leaders, and Scoring Directors. There is locally resident software to view the students' constructed-responses and to recall images of any student document upon demand. Each handscoring site is connected to the DRC main operation facility with multiple T1 transmission lines. The operation facility has multiple application and secure database servers that support the scanning, editing, scoring, and handscoring processes. The database backups and archived images are also housed on the secure servers.

The student responses were separated for readers by item for each subject, and only qualified readers had access to student response images. The readers read each response and keyed in the correct score. After the score is entered, a new response image appeared. Images of specific sets of items (unit-specific) were sent to designated groups of readers qualified to score those items.

This process of routing and scoring sets of imaged items continued until all responses to items or prompts received the prescribed number of independent readings. Non-adjacent scores that required resolving were routed to Scoring Directors or Team Leaders for electronic review and resolution.

Quality Control of Handscoring

DRC's quality control procedures helped to ensure that constructed-response items for the Alaska assessment were scored in an objective and accurate manner using the following approach.

With the exception of grade 10 responses, ten percent of all operational (common) items were independently scored by two readers for the purposes of monitoring inter-rater reliability. The imaging system re-directed every tenth item to a second scorer for another independent reading. Likewise, the majority of the spring field test items received a random ten percent second scoring (this was done at no cost to the client). Since grade 10 items served as both SBA items and HSGQE items, the grade 10 items were scored using the HSGQE scoring rules, under which all responses are read by at least two readers. The HSGQE scoring rules are further detailed in the 2007 HSGQE Technical Report. The inter-rater reliability statistics are included in Appendix 11.

In order to monitor reader reliability and to ensure that an acceptable agreement rate was maintained, DRC monitored the daily statistics provided by the reliability reports, which documented individual reader data, including reader number and team designation, number of responses scored, individual score point distributions, and exact agreement rates. A ratio of one Team Leader for every 10–12 readers was maintained to ensure adequate monitoring of the readers. In addition to this information, Team Leaders conducted routine “read behinds” for all readers.

DATA PROCESSING

The original scanned multiple-choice data was converted into a master student file. Record counts were verified against the counts from the Document Processing staff to ensure all students were accounted for in the file.

DRC provided EED the student file so corrections and updates could be applied. After the demographic information was updated, the student file was scored against the appropriate answer key, indicating correct and incorrect responses. Correct responses were designated by converting the numeric value into an alpha value. Incorrect responses remained numeric. In addition, the original response string was stored for data verification and auditing purposes.

Scores for a student's constructed responses were systematically matched to the student's multiple-choice responses by a unique document ID. This process allowed DRC to score and create a student record for each test book returned for processing, while providing accurate and reliable data. Student scale scores and proficiency levels were determined prior to production of final data files and reports.

Once the scored master student file was deemed 100 percent accurate, DRC's Psychometric Services staff performed additional detailed analysis on the data files prior to EED's review and approval process.

REPORTING

DRC worked with EED to determine appropriate file layouts. The layouts included field names, field descriptions, field values, and starting and ending positions. DRC posted district-level data files and layouts to the DRC Online Web Reporting System and state-level data files and layouts to the FTP site.

DRC created report mockups of the production reports that were produced and delivered for this administration. The mockups comprised simulated, but realistic, data elements and were in the required report layout, displayed the approximate font and font sizes, and demonstrated paper size and printing elements.

DRC followed a review process that allowed EED to review, change, and approve all mockups prior to report development. The mockups were reviewed by DRC's Business Analysts and Software Quality Assurance Analysts for accuracy and consistency. During the review process, EED was able to evaluate the static content and layout of each report to make certain they reflected the format, verbiage, and design required. DRC worked closely with EED throughout the review process to incorporate changes or modifications.

EED identified Kenai as the sample district for quality verification. This helped DRC identify and prioritize boxes of used test books returned from that district and process those test books on a first-priority basis through check-in, scanning, scoring, and reporting.

During all phases of reporting, DRC performed a thorough quality assurance review prior to releasing of reports. A cycle of "bluedot" samples was reviewed by EED prior to producing live reports for districts and schools.

DRC provided the district and state reports outlined below. DRC also produced Parent/Student and Teacher/Staff versions of the *Guide to Test Interpretation*. Samples of these guides are provided in Appendix 12 and are also available on EED's Web site.

Final Grade 10 SBA reports were provided electronically on May 11, 2007. Paper copies of the final Grade 10 SBA reports were delivered to the districts as scheduled by May 18, 2007.

Final Grades 3–9 SBA reports were provided electronically on May 16, 2007. Paper copies of the final Grades 3–9 SBA reports were delivered to the districts as scheduled by May 24, 2007.

The erasure analysis was delivered to EED on July 2, 2007.

District Reports

- Student Reports
- School Student Rosters
- School Summary Reports
- School Subpopulation Summary Reports
- District School Rosters
- District Subpopulation Summary Reports
- Student Data File
- Abbreviated Student Data File

State Reports

- Student Data File
- Abbreviated Student Data File
- State Subpopulation Summary Reports
- DVDs

CHAPTER 5: FORM ANALYSIS & ITEM CALIBRATION

RASCH MEASUREMENT MODELS

Scale scores for the SBAs were developed using the family of Rasch (1960) measurement models for scaling and equating. The advantage of using Rasch models in scaling is that all of the items measuring performance in a particular content area can be placed on a common difficulty scale, allowing the Rasch difficulty values for the individual items to be used in computing a Rasch logit for any raw score point on any test constructed from scaled items.

Rather than percent correct, the Rasch model expresses item difficulty (and student proficiency) in units commonly referred to as logits. In the simplest case, a logit is a transformed p -value with the average p -value represented by a logit of zero. The logit metric has several mathematical advantages over p -values. It is an interval scale, meaning two items with logits of 0 and +1 are the same distance apart as items with logits of +3 and +4. Logits are independent of the ability distribution of the students taking a particular test. A specific form will have a mean logit of zero, whether the average p -value of the test is 0.8 or 0.3. The Rasch model also allows person measures and item measures to be placed on a common scale. This allows the comparison of person proficiency and item difficulty to determine the probability that a person will respond correctly to any given test item. This comparison is not possible in the percent correct metric. It is impossible to predict how well a person who answered 80% of the items correctly will perform on an item answered correctly by 80% of the persons.

The standard Rasch calibration procedure sets the mean difficulty of the items on any unanchored calibration at zero. Any item with a p -value lower than the mean receives a positive logit and any item with a p -value higher than the mean receives a negative logit. Consequently, the logits for any calibration, whether it is a third grade reading test or a high school mathematics test, relate to an arbitrary origin defined by the average of item difficulties for that form. The average third grade reading item will have a logit of zero; the average high school mathematics item will have a logit of zero in unanchored calibrations. This logit scale applies to both item difficulties and student abilities.

Because both dichotomous and polytomous items were part of the SBA assessments, DRC utilized a mixed-model item calibration approach that placed both item types onto a common scale. Multiple-choice (MC) items scored either right or wrong, were calibrated using the familiar form of the dichotomous Rasch model. Constructed-response (CR) items were calibrated using another model in the Rasch family, Master's partial-credit model (Wright and Masters, 1982). The latter model parameterizes each threshold needed to obtain the maximum score on the task. Consequently, there is one item difficulty parameter for each of the $n-1$ score transitions (0/1, 1/2, etc.), or thresholds. While the partial-credit model is a non-trivial extension of the simple logistic Rasch model, an MC item may be thought of as a partial-credit task with only one threshold.

With the partial-credit model, π_{nix} is the probability that person n scores x on item i . The conditional probability of a score of 1, given a score of 0 or 1 is:

$$\Phi_{ni1} = \frac{\pi_{ni1}}{\pi_{ni0} + \pi_{ni1}} = \frac{\exp(\beta_n - \delta_{i1})}{1 + \exp(\beta_n - \delta_{i1})},$$

where β_n is the ability of person n and δ_{i1} is the difficulty of the first threshold for item i .

The preceding equation can be expanded to obtain one general expression for the probability of person n scoring x on item i :

$$\pi_{nix} = \frac{\exp \sum_{j=0}^x (\beta_n - \delta_{ij})}{\sum_{k=0}^{m_i} \exp \sum_{j=0}^k (\beta_n - \delta_{ij})}, \quad x = 0, 1, \dots, m_i,$$

where m_i is the number of thresholds and for notational convenience,

$$\exp \sum_{j=0}^0 (\beta_n - \delta_{ij}) = 1.$$

This equation expresses the probability of person n scoring x on the m_i threshold of item i as a function of the person's measure (β_n) and the threshold difficulties of the m_i thresholds for item i . The observation x is a count of the successfully completed item thresholds.

The unconditional, joint maximum likelihood (UCON) estimation procedure estimates the person parameters (i.e., ability) simultaneously with the item parameters (i.e., difficulty). The UCON procedure was accomplished using WINSTEPS Version 3.63 (Linacre, 2006). This calibration software is commercially available and widely used in the testing industry and is considered the industry standard for Rasch calibration.

ITEM STATISTICS

Appendix 13 provides item level statistics by content area for the spring 2007 SBA operational assessments. These statistics (i.e., logit, standard error, fit, p -value, item-total correlation, and omits) represent the item characteristics most commonly used to determine whether an item functioned in an appropriate manner. Table 5-1 presents the mean or median of these statistics within each content area.

The logit column in the table and appendix provides the average ability of the persons attempting that item. The standard error (SE) column gives the asymptotic standard error associated with these values.

The Rasch fit statistics are used to determine how well items conform to the requirements of the Rasch measurement model. The items were analyzed for scale comparability by examining the residuals between observed and expected scores for the persons and items (Smith, 2000; Mead,

1978). This process investigated the underlying construct measured by a test by analyzing the patterns of item covariation within the scale. For example, when local dependence is exhibited, it may indicate violations of unidimensionality, thus introducing sources of variability that are unrelated to the construct being measured, with the caveat that, even if some minor item dependence existed in the CR item formats, they are likely to have minor influence on scores (Stout, 1987). A standardized weighted total fit (OUTFIT z-std) statistic was computed for each item. This fit statistic quantifies the sum of the squared distances of the observed item performance from the expected performance for all persons. Items may not fit the Rasch model for several reasons, all of which relate to students responding to items in an unexpected way. In many cases the reason behind why students respond in unexpected ways to a particular item is unclear. However, it is possible to determine possible causes of an item's misfit by re-examining the item and its distracters. As part of this investigative process, DRC content specialists examined all items with large fit statistics to confirm that each item exhibited the attributes of a high quality item based on best practices.

The p -value for an MC item is the percent of all students that responded to an item correctly. The p -value for a CR item represents the average score earned divided by the maximum number of points for that item. For the spring 2007 SBA forms, this score can range from 0 to 2 or 0 to 4 points in mathematics, 0 to 2, or 0 to 4 in reading, and 0 to 2, 1 to 4, or 1 to 6 in writing.

The item-total correlation (PtBis or Corr.) provides a measure of the internal consistency of the responses. It assesses how well each item measures the trait defined by the set of items as a whole. Typically, students with high proficiency (i.e., those that perform well on the SBA content area test overall) would be expected to get items correct, and students with low proficiency (i.e., those that perform poorly on the SBA content-area test overall) to get items incorrect. If these expectations are met, the item-total correlation between the item and the total test score will be high and positive, indicating that the item is a good discriminator between high ability and low ability students. An item-total correlation value above 0.30 is usually considered acceptable. An item-total correlation value below 0.30 indicates that an item may not be measuring what it was intended to measure, and should be reviewed. DRC content specialists reviewed all items with item-total correlations below .30 and verified that each item was acceptable as written and scored. As seen in Tables 5.1 through 5.8, the median item-total correlations for MC and CR items all exceeded the .30 criterion.

The omits column represents the proportion of persons leaving the item blank for MC items and the proportion of persons with blanks or other condition codes for CR items. Note, however, that the nonscorable codes are recoded as 0 points for the purposes of item calibration and scoring.

Table 5-1. Summary of Operational Item Analysis – Grade 3

Content Area	Item Type	Mean Logit	Mean SE	Mean Fit	Mean <i>p</i> -value	Median PtBis or Corr	Mean Omits
Mathematics	MC	0.009	0.027	-1.027	0.698	0.420	0.006
	CR	0.203	0.016	6.000	0.639	0.590	0.010
Reading	MC	0.137	0.025	-0.267	0.658	0.464	0.010
	CR	1.520	0.016	6.500	0.394	0.584	0.051
Writing	MC	-0.153	0.026	-1.622	0.683	0.504	0.009
	CR	0.687	0.017	9.283	0.526	0.559	0.040

Table 5-2. Summary of Operational Item Analysis – Grade 4

Content Area	Item Type	Mean Logit	Mean SE	Mean Fit	Mean <i>p</i> -value	Median PtBis or Corr	Mean Omits
Mathematics	MC	0.039	0.026	-2.114	0.691	0.402	0.006
	CR	0.457	0.015	9.800	0.578	0.528	0.011
Reading	MC	0.184	0.026	-1.375	0.670	0.475	0.007
	CR	0.861	0.015	6.967	0.499	0.557	0.032
Writing	MC	-0.100	0.026	-1.004	0.677	0.466	0.007
	CR	0.562	0.015	9.900	0.364	0.456	0.019

Table 5-3. Summary of Operational Item Analysis – Grade 5

Content Area	Item Type	Mean Logit	Mean SE	Mean Fit	Mean <i>p</i> -value	Median PtBis or Corr	Mean Omits
Mathematics	MC	0.065	0.026	-0.761	0.698	0.422	0.002
	CR	0.055	0.016	3.333	0.706	0.605	0.013
Reading	MC	0.071	0.025	-1.413	0.675	0.444	0.006
	CR	0.921	0.015	-0.733	0.456	0.622	0.032
Writing	MC	0.058	0.025	-0.835	0.673	0.435	0.003
	CR	0.300	0.016	9.900	0.485	0.470	0.017

Table 5-4. Summary of Operational Item Analysis – Grade 6

Content Area	Item Type	Mean Logit	Mean SE	Mean Fit	Mean <i>p</i> -value	Median PtBis or Corr	Mean Omits
Mathematics	MC	-0.025	0.026	-1.205	0.695	0.401	0.002
	CR	-0.030	0.016	6.367	0.679	0.599	0.013
Reading	MC	-0.003	0.026	-2.465	0.706	0.447	0.003
	CR	0.722	0.015	9.900	0.513	0.507	0.027
Writing	MC	-0.118	0.026	0.585	0.695	0.445	0.003
	CR	1.069	0.016	5.700	0.413	0.510	0.017

Table 5-5. Summary of Operational Item Analysis – Grade 7

Content Area	Item Type	Mean Logit	Mean SE	Mean Fit	Mean <i>p</i> -value	Median PtBis or Corr	Mean Omits
Mathematics	MC	-0.218	0.025	-2.171	0.684	0.414	0.002
	CR	0.231	0.014	9.900	0.589	0.488	0.019
Reading	MC	-0.016	0.024	-1.744	0.672	0.421	0.003
	CR	0.950	0.015	4.900	0.457	0.547	0.034
Writing	MC	0.083	0.024	1.630	0.641	0.399	0.003
	CR	0.671	0.015	7.675	0.451	0.528	0.023

Table 5-6. Summary of Operational Item Analysis – Grade 8

Content Area	Item Type	Mean Logit	Mean SE	Mean Fit	Mean <i>p</i> -value	Median PtBis or Corr	Mean Omits
Mathematics	MC	-0.023	0.025	-1.414	0.676	0.421	0.002
	CR	0.473	0.014	2.267	0.557	0.591	0.035
Reading	MC	0.096	0.025	-0.187	0.675	0.387	0.003
	CR	1.654	0.015	5.233	0.387	0.403	0.039
Writing	MC	-0.133	0.025	-1.291	0.676	0.442	0.003
	CR	0.973	0.015	3.350	0.397	0.511	0.036

Table 5-7. Summary of Operational Item Analysis – Grade 9

Content Area	Item Type	Mean Logit	Mean SE	Mean Fit	Mean <i>p</i> -value	Median PtBis or Corr	Mean Omits
Mathematics	MC	-0.132	0.024	-2.018	0.678	0.422	0.002
	CR	0.017	0.013	5.800	0.641	0.611	0.031
Reading	MC	0.001	0.024	-2.542	0.695	0.422	0.003
	CR	1.318	0.015	9.367	0.443	0.462	0.052
Writing	MC	0.091	0.024	-0.374	0.674	0.426	0.003
	CR	0.974	0.014	8.250	0.446	0.528	0.041

Table 5-8. Summary of Operational Item Analysis – Grade 10

Content Area	Item Type	Mean Logit	Mean SE	Mean Fit	Mean <i>p</i> -value	Median PtBis or Corr	Mean Omits
Mathematics	MC	0.027	0.027	-0.850	0.666	0.369	0.005
	CR	1.299	0.015	1.300	0.426	0.557	0.050
Reading	MC	0.661	0.026	-1.355	0.696	0.391	0.003
	CR	2.146	0.014	9.900	0.425	0.461	0.028
Writing	MC	0.391	0.026	-0.136	0.707	0.349	0.004
	CR	1.070	0.017	1.283	0.435	0.519	0.025

FORM STATISTICS

Appendix 14 contains summary descriptive statistics for student performance and item difficulty, including mean score, standard deviation, and minimum and maximum scores by content area. These statistics were generated using WINSTEPS v3.63 (Linacre, 2006) and illustrate student and item performance. The top halves of the student summary tables provide descriptive statistics for persons (i.e., students) measured. The column labeled “Measure” provides the mean and standard deviation of the estimated student proficiency measures. The “Model Error” column presents similar information for the asymptotic standard errors.

The top halves of the item summary tables provide the same descriptive statistics outlined above, with the exception that items are the unit of analysis rather than students. In this table, “Measure” refers to estimated item difficulty, so that the average measure refers to the average difficulty of the items on the test. Again, “Model Error” is the descriptive statistics for the asymptotic standard errors.

The bottom halves of the tables contain the Root Mean Square Error (RMSE). The Real RMSE corresponds to a worst case error estimate, and Model RMSE corresponds to a best case estimate. The adjusted standard deviation is an estimate of the “true” standard deviation, which adjusts for potential measurement error by removing it from the standard deviation estimate (Wright and Masters, 1982, see pages 92 and 113):

$$SA_I^2 = SD_I^2 - MSE_I ,$$

where SA_I is the adjusted standard deviation, SD_I is the observed standard deviation, and MSE_I is the mean square error, which is calculated using the following equation:

$$MSE_I = \sum_{i=1}^L s_i^2 / L ,$$

where L is the number of items and s_i is the standard error of item i .

The RMSE is computed by taking the square root of the MSE value:

$$RMSE_I = \sqrt{MSE_I} .$$

The item separation value then provides the adjusted standard deviation in RMSE units. It is calculated by finding the ratio of the adjusted standard deviation to the RMSE:

$$G_I = SA_I / RMSE_I .$$

The test reliability estimate is called the index of “item separation reliability.” This is a refined measure of internal consistency reliability, which provides the proportion of observed item variance that is not due to estimation error. The item separation reliability estimate is computed using:

$$R_I = \frac{SA_I^2}{SD_I^2} .$$

It can also be calculated using only the separation value:

$$R_I = \frac{G_I^2}{1 + G_I^2} .$$

The processes for obtaining person separation and person separation reliability values are analogous to those for calculating item separation and item separation reliability values. The previous equations should be used, substituting a “P” for each “I.”

Below the tables, the standard error of the mean for the persons and items tested, respectively, are provided. This value is an estimate of the average amount of error associated with the sample person and item means. Two additional statistics, the student raw score-to-measure correlation and Coefficient Alpha student raw score reliability, are also reported below the Student Summary tables.

FREQUENCY DISTRIBUTIONS

Items

Appendix 15 provides frequency distributions of all SBA item difficulties, including the thresholds for CR items. Each item sequence number is shown to the right of its corresponding logit, which represents the lowest possible value for that row. When more than one item falls in the logit range, the items are arranged from lowest to highest logit value. For instance, as seen in Figure 15-1 of the appendix, the logit value for Mathematics Item 38 is between 0.7 and 0.9, and it is also lower than the logit value for Item 17, which also falls between 0.7 and 0.9. In addition, each CR item sequence number is displayed to the right of its corresponding logit for each possible threshold.

Persons

Appendix 16 provides frequency distributions of raw scores and scale scores by content area for the spring 2007 SBA administration. The columns in these tables present each raw score, scale score, scale score asymptotic standard error, frequency count, frequency percent, cumulative frequency, and cumulative percent. The range of reported scale scores for the SBAs is 100 through 600.

CAUTIONS FOR SCORE USE

As with any assessment, student scores at the minimum or maximum ends of the score range will have large standard errors of measurement and should be viewed cautiously. For instance, if the maximum score for the SBA in reading is 600 and a student achieves this score, it cannot be determined whether the student would have achieved a higher scale score if that score were possible. All that is known is that the student's scale score, as revealed by this test, is at least 600. In this manner, extreme scale scores may vary from one administration to the next even if the number of items tested does not, making comparisons of students that score at the extreme ends of the score distribution difficult. To minimize confusion and the potential for misinterpretation, the maximum scale scores possible on the SBA have been fixed so they do not change across administrations.

CHAPTER 6: SCALING & EQUATING

INTRODUCTION

To assist in maintaining the same passing standard across different administrations, EED, in collaboration with DRC, constructs all tests to be of similar difficulty. This similarity is maintained from administration to administration at the total test level and, as much as possible, at the reporting standard level.

For grades 3–9, the spring 2007 operational SBA tests in mathematics, reading, and writing were constructed using items from the 2005 and 2006 field test administrations. The grade 10, spring 2007 operational SBA tests in mathematics, reading, and writing were the second forms developed to meet new NCLB requirements.

In addition to the operational items, DRC embedded field test items in order to expand the item pool for future form development. (Items are embedded for grades 3–9 only. Field test items for Grade 10 were appended.) Field test items are those being administered for the first time to gather statistical information about the item. These items do not count toward an individual student's score.

Because the grade 3–9 operational forms were new forms, pre-equating based on field test item statistics was employed. The advantages of a pre-equated test include a significant reduction in the waiting period between the initial administration of a new form and the release of score reports. Because item analysis, key finalization, and test equating are completed before new forms are administered, score reports can be issued sooner than those from post-equated assessments.

GRADES 3–9 PRE-EQUATING

In the pre-equating process, a newly developed test is linked to a previously administered test form. This allows for the new test form's difficulty level to be equated to previous administrations. This procedure utilizes common item equating, where the entire operational form serves as the common links. This produces high quality of equating from administration to administration (due to the large number of equating items).

GRADES 3–9 OPERATIONAL ITEM CALIBRATION

The stability (invariance) of the item difficulties for the spring 2007 administration was determined by anchoring the operational item difficulty values to those obtained from the spring 2005 and 2006 field test administrations. This anchored calibration method produced results such that the items and thresholds were on the same scale as the original (spring 2005) operational test form. The WINSTEPS (Linacre, 2006) program was used to anchor (hold the item difficulty constant) the Rasch item difficulty estimates and the constructed-response (CR) threshold estimates for the items from the 2005 and 2006 field test administrations, as well as estimate the change in item difficulty (displacement) between the field test and operational administrations.

The calibrated item and threshold difficulties were used in conjunction with actual student performance to obtain Rasch ability estimates for each possible raw score value for the overall test, as well as each subscale/reporting standard. The generation of this raw score-to-Rasch

ability was accomplished through fundamental formulas in the Rasch measurement model (Wright and Masters, 1982).

The combination of both dichotomously scored MC items as well as polytomously scored CR tasks required the use of a partial-credit model. The Newton-Raphson iterative procedure was used to obtain precise ability estimates:

$$b_r^{(t+1)} = b_r^t - \frac{r - \sum_i^L \sum_{k=1}^m k P_{rik}^{(t)}}{- \sum_i^L \left[\sum_{k=1}^m k^2 P_{rik}^{(t)} - \left(\sum_{k=1}^m k P_{rik}^{(t)} \right)^2 \right]}, \quad r=1, M-1,$$

where b_r^t is the estimated ability of the student with score r after t iterations, r is the number of thresholds, L is the number of items, $M=mL$, and $P_{rik}^{(t)}$ is the probability π_{nix} defined earlier in Chapter 5:

$$\pi_{nix} = \frac{\exp \sum_{j=0}^x (\beta_n - \delta_{ij})}{\sum_{k=0}^{m_i} \exp \sum_{j=0}^x (\beta_n - \delta_{ij})}, \quad x = 0, 1, \dots, m_i.$$

The asymptotic standard error was estimated from the denominator of the final iteration:

$$SE(b_r) = \left[\sum_i^L \left[\sum_{k=1}^m k^2 P_{rik}^{(t)} - \left(\sum_{k=1}^m k P_{rik}^{(t)} \right)^2 \right] \right]^{-1/2}.$$

The iteration was terminated using the WINSTEPS convergence criteria of 0.01 maximum logit chance.

GRADE 10 OPERATIONAL ITEM CALIBRATION

In the post-equating process, previously administered field test items are included on an operational test with items that have been previously equated onto the common scale. Although the initial intent was to pre-equate the spring 2007 grade 10 SBA, the decision to post-equate the test was made following the analysis of the fall 2006 High School Graduation Qualifying Examination (HSGQE). This analysis indicated that a small number of items had large changes in item difficulty from the field test administration to the operational administration. Similar to the fall 2006 HSGQE, items on the operational spring 2007 grade 10 SBA came from the pool of items that were field tested in spring 2005 and spring 2006. The fall 2006 analyses also indicated that there were systematic differences in the average displacement values for items from the spring 2005 and spring 2006 field test administrations. Because the SBA standards validation was based on the spring 2006 operational administration and the fact that the large majority of the items on the three forms originated from the 2006 field tests, it was determined that post-equating based on the spring 2006 field test administration should be used to equate the fall 2006

administration. It was determined that this provided the strongest link to the 2006 standards validation. The same reasoning for post-equating was applied to the spring 2007 grade 10 SBA. The spring 2007 grade 10 SBA analyses also showed systematic differences in the average displacement values for items from the spring 2005 and spring 2006 field test administrations, indicating that post-equating produced results more in line with the spring 2006 standards validation.

ITEM BANK MAINTENANCE

The item bank was then updated with the operational item statistics from this administration.

CHAPTER 7: FIELD TEST ITEM DATA SUMMARY

FIELD TEST ITEMS

Once a newly constructed item had passed committee review, it was ready for field testing. For instance, for the 2007 SBA reading test, grades 3–9, this was accomplished by embedding 10 MC items and 1 CR item within the 55 operational test items. Unlike previous administrations, there was only one field test form per grade/content areas that appeared on all tests for that grade/content area combination. Note that the field test items do not count towards an individual student's score. Only the operational test items counted towards the individual score. For the grade 10 spring 2007 SBA, 14 unique field test forms were administered. Each form contained the same 47 operational mathematics test items, 58 operational reading test items, and 32 operational writing test items. In addition, each form appended 6 to 11 MC items and 1 or 2 CR items, depending on the content area and form.

As discussed previously, the operational items were used as anchors for transforming the field test item parameters to the same logit scale that was previously established. The full sample was used to estimate the difficulty of the MC items for all grades. A random sample of approximately 1000 (of a population of approximately 9000) students was used to estimate the difficulty of the CR items for grades 3–9 given that there was only one field test form. All student responses were used to estimate the difficulty of the CR items for grade 10.

FIELD TEST ITEM DESCRIPTIVE STATISTICS

Appendix 19 provides field test item statistics by content area for the spring 2007 SBAs. These statistics represent the item characteristics most commonly used to determine whether an item functioned in an appropriate manner and are the same as those defined in Chapter 5 for operational items.

Tables 7-1 through 7-3 report the mean raw score summary statistics for the three grade 10 content areas that had multiple field test forms. Estimation of item difficulty in the Rasch Model is independent of the mean person raw scores. But the similarity of the field test raw scores supports the sampling plan described in the Spiraling Plan section of Chapter 3: Test Administration Procedures.

Table 7–1. Mathematics Raw Score Summary Statistics by Field Test Form – Grade 10

Form	N	Mean	SE of the Mean	Standard Deviation	Minimum	Maximum
01	766	24.1214	0.2883	7.9782	3	40
02	686	25.7536	0.2771	7.2564	1	40
03	677	24.8951	0.2978	7.7496	4	40
04	547	23.6764	0.3366	7.8723	4	40
05	687	24.7875	0.2998	7.8583	3	40
06	694	25.0706	0.2732	7.1976	6	40
07	630	25.1825	0.3032	7.6093	6	40
08	577	25.5494	0.3049	7.3247	5	40
09	669	24.5725	0.2948	7.6244	3	40
10	697	25.6571	0.2702	7.1325	4	40
11	755	25.7192	0.2651	7.2842	5	40
12	670	25.1388	0.2757	7.1357	6	40
13	599	25.3573	0.2952	7.2254	3	40
14	743	25.8156	0.2686	7.3205	3	40

Table 7–2. Reading Raw Score Summary Statistics by Field Test Form – Grade 10

Form	N	Mean	SE of the Mean	Standard Deviation	Minimum	Maximum
01	784	47.9209	0.4201	11.7634	7	71
02	706	48.9533	0.4426	11.7590	15	70
03	644	46.6071	0.5076	12.8806	5	70
04	633	45.1485	0.5425	13.6490	10	71
05	655	46.8489	0.5006	12.8110	3	70
06	671	47.6349	0.4816	12.4753	9	70
07	603	46.5987	0.5162	12.6761	15	71
08	684	48.3494	0.5000	13.0759	9	70
09	679	46.6834	0.5024	13.0911	9	70
10	670	48.5522	0.4724	12.2269	4	68
11	733	48.3492	0.4327	11.7161	12	71
12	625	47.6336	0.5017	12.5414	9	71
13	610	48.3344	0.5027	12.4165	8	69
14	714	49.0238	0.4542	12.1369	12	71

Table 7-3. Writing Raw Score Summary Statistics by Field Test Form – Grade 10

Form	N	Mean	SE of the Mean	Standard Deviation	Minimum	Maximum
01	773	31.9004	0.3009	8.3656	2	48
02	690	33.6333	0.2857	7.5060	9	48
03	678	32.3732	0.2976	7.7489	5	49
04	549	31.4645	0.3623	8.4896	7	47
05	682	32.5117	0.3040	7.9383	6	48
06	688	33.0858	0.2858	7.4965	5	50
07	626	33.0751	0.3075	7.6947	7	50
08	580	33.6897	0.3040	7.3224	7	48
09	675	33.1230	0.2860	7.4301	2	49
10	695	33.5482	0.2791	7.3575	5	49
11	771	33.8392	0.2591	7.1933	3	50
12	672	32.9539	0.2845	7.3741	7	48
13	595	33.3429	0.2955	7.2088	8	48
14	739	33.7564	0.2595	7.0552	8	48

DRC utilized the Mantel-Haenszel (MH) or the Standardized Mean Difference (SMD) statistic for detecting differential item functioning (DIF) depending on the item type. The MH statistic is the most commonly used technique for MC items in large-scale, educational assessment. It does not depend on the application or the fit of any specific measurement model.

The MH procedure, as implemented by DRC, compared the observed and expected totals of a two-by-two-by-four contingency table (Holland & Thayer, 1986) shown in Table 7-4. The contingency table contrasts a focal group with a reference group by item response (correct/incorrect) by four performance levels (quartiles of the total test score). Males and Caucasians were considered the reference groups for the gender and ethnicity comparisons and the focal group was females or Alaska Natives and American Indians.

Table 7-4. Mantel-Haenszel Contingency Table

Group	Correct (1)	Incorrect (0)	Total
Reference	A_j	B_j	n_{Rj}
Focal	C_j	D_j	n_{Fj}
Total	m_{1j}	m_{0j}	T_j

An odds-ratio,

$$\hat{\alpha}_{MH} = \frac{\sum \left(\frac{A_j D_j}{T_j} \right)}{\sum \left(\frac{B_j C_j}{T_j} \right)},$$

was summed across each of the j -levels and then converted into the Educational Testing Service (ETS) “delta scale”

$$\hat{\Delta}_{MH} = -2.35(\ln(\hat{\alpha}_{MH}))$$

The value $\hat{\Delta}_{MH}$ is the average amount more difficult that a member of the reference group found the studied item than did comparable members of the focal group.

The variance approximation for $\hat{\alpha}_{MH}$ was determined via the equation:

$$\text{Var}(\hat{\alpha}_{MH}) = \frac{1}{2U^2} \sum_j [T_j^{-2} (A_j D_j + \hat{\alpha}_{MH} B_j C_j)(A_j + D_j + \hat{\alpha}_{MH} (B_j + C_j))]$$

where
$$U = \sum_j \frac{A_j D_j}{T_j}$$

From the $\hat{\Delta}_{MH}$ value, one of three severity classification categories was assigned (i.e., A, B, C). Rules for the classification are found in Appendix 20. The A category represents negligible DIF. The B category indicates moderate potential DIF, that is to say, that one group outperformed the other group once the effects of differences in skill levels between the two groups have been removed. The C category indicates that there is large potential DIF. The plus (+) and minus (-) signs that follow the DIF category indicate which group is favored by the item. The minus sign indicates that the reference group outperformed the focal group once the skill level differences between the groups have been accounted for. The plus sign indicates that the focal group outperformed the reference group once the skill level differences between the groups have been removed.

The analysis on CR items was based on the SMD procedure (Zwick & Thayer, 1996). SMD takes into account the natural ordering of the response levels of the item. In contrast to the MH procedure, this summary statistic compares the means of the reference and focal groups, adjusting for differences in the distribution of each group’s members across the four ability stratifications. Data were organized into a two-by- T -by-four contingency table shown in Table 7-5, where T is the number of score categories and the plus (+) signs denote summation over a particular index.

Table 7–5. SMD Contingency Table

Group	y₁	y₂	y₃	...	y_T	Total
Reference	n_{R1k}	n_{R2k}	n_{R3k}	...	n_{RTk}	n_{R+k}
Focal	n_{F1k}	n_{F2k}	n_{F3k}	...	n_{FTk}	n_{F+k}
Total	n_{+1k}	n_{+2k}	n_{+3k}	...	n_{+Tk}	n_{++k}

The SMD statistic was calculated using the equation:

$$SMD = \sum_k p_{Fk} m_{Fk} - \sum_k p_{Fk} m_{Rk} ,$$

where the proportion of focal group members who were at the k^{th} ability stratification was found by:

$$p_{Fk} = \frac{n_{F+k}}{p_{F++}} ,$$

the mean item score for the focal group at the k^{th} stratification was calculated using:

$$m_{Fk} = \frac{\sum_T y_T n_{FTk}}{n_{RTk}} ,$$

and the mean item score for the reference group was determined from:

$$m_{Rk} = \frac{\sum_T y_T n_{RTk}}{n_{RTk}} .$$

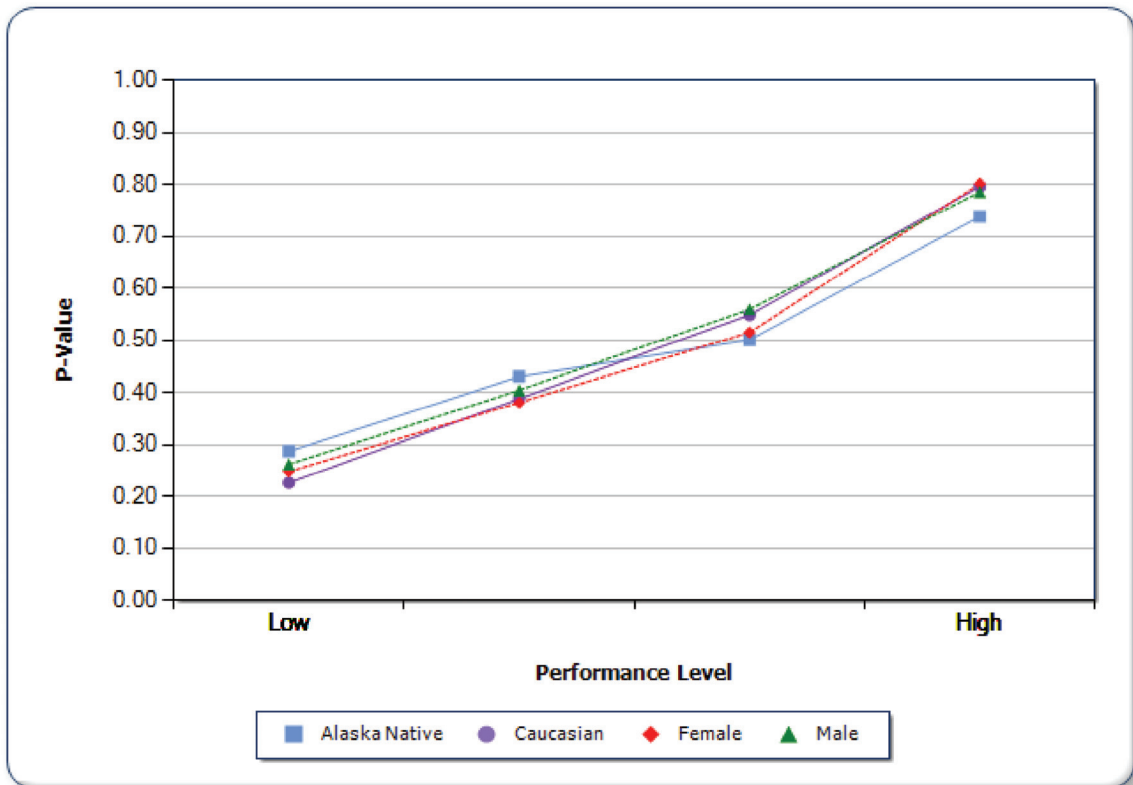
One of three severity classification categories was then assigned (A, B, or C). Appendix 20 provides rules for classification.

A summary of DIF results for field test items across forms is presented in Appendix 21. There was only one form per content area for grades 3–9 and 14 forms per content area for grade 10.

Figure 7-1 provides an example of a DIF plot used in data review. This plot shows the average proportion correct for two focal groups (female and Alaska Native) and two reference groups (male and Caucasian).

Figure 7–1. Example of DIF Plot

Item 6M-0907 Performance By Group



Performance Level Group Counts

Group	Low	Medium Low	Medium High	High
Alaska Native	1088	780	527	283
Caucasian	957	1315	1578	1883
Female	1199	1376	1256	1203
Male	1340	1200	1292	1300

Group Proportion Choosing Option

	Alaska Native	Caucasian	Female	Male
A	0.176	0.157	0.158	0.161
B	0.225	0.145	0.177	0.175
C	0.172	0.155	0.175	0.156
D*	0.419	0.540	0.484	0.503

ITEM BANK MAINTENANCE

Following field test item calibration and analysis, the item bank was then updated with the new item information. If the same field test item appeared on more than one form, then these items had multiple Rasch item difficulties. The Rasch item difficulty value corresponding to the field test form with the greater number of students tested was taken as the one to represent the item. Selected field test items were then made available for Data Review Committee final appraisal. Once approved, the operational portion of subsequent forms could be constructed from the calibrated item bank.

Item data review for the field test items administered in spring 2007 will be conducted in summer 2007.

CHAPTER 8: SCALE SCORES & PROFICIENCY LEVELS

OVERVIEW

To ensure that student proficiency results for each grade are reported on a common standard score scale, EED provides a unique scale score conversion system for each SBA assessment. In this system, raw scores are converted to a logistic metric. Logit measures are then transformed into scale scores. Scale scores are intended to make scores more meaningful by defining a scale of measurement that is not tied to a particular test form. The scales across all content areas and grades have a maximum of 600, a minimum of 100, with the proficient cut score fixed to 300.

DESCRIPTION OF SCORES

Raw Score

The basic summary statistic on all SBA assessments is the raw score. A raw score is reported for each examinee in mathematics, reading, and writing. The raw score is the number of multiple-choice (MC) items answered correctly plus the number of points earned on constructed-response (CR) items on a content-area assessment. By itself, the raw score has limited utility; it can only be interpreted in reference to the total number of items on a content-area assessment, and raw scores should not be compared across tests or administrations.

Scale Score

Given that a given raw score may not represent the same skill level on every test form, all statewide assessment score reports include scale scores. Scale scores are statistical conversions of raw scores that adjust for slight shifts in item difficulties and permit valid comparisons across all test administrations within a particular content area and grade.

When new test forms are developed, the new set of items will require slightly different levels of content-area skill to answer correctly. This depends on the difficulty of the specific questions used on each form. To be fair to students and to permit valid comparison of test scores across administrations, the skills represented by each score point must remain consistent from year to year.

As noted previously, scale scores adjust for slight shifts in underlying difficulty levels at each score point and provide valid points of comparison across all test administrations within a particular grade and content area. With scale scores, schools can reasonably compare the demonstrated knowledge and performance of groups of students across years.

Comparability of Scale Scores Across Grades

Through the process described in the previous section, the standards for Proficient were established to have consistent interpretation from grade to grade. The logit measures that defined the Proficient cut score for each content area and grade (obtained from the standard setting process described above) was thus defined to be a scale score of 300. As a result, a student who receives a scale score of 300 at each grade is making progress from grade to grade that is the same as the difference in the standards for Proficient across those two grades.

Further, the relationship between the logit measures and the scale scores was established so that the standard deviation of scale scores would be 75 on average across all the grades in the baseline year. In subsequent years, the standard deviation of the logit measures varies from grade to grade. Therefore the standard deviation of student scale scores is higher than 75 at some grades and less than that amount at others. In this administration, across all grades and content areas, the standard deviation of the scale scores ranged from a low of 68.6 for grade 8 reading to a high of 89.6 for grade 4 writing.

As a result, the interpretation of scale scores is the same for all grades and content areas in the following context: a scale score of 225, for example, means that the student scored approximately one standard deviation below the standard for Proficient. If that same student had a scale score of 250 in that subject at the next grade (meaning the student now is approximately .67 standard deviations below the standard for Proficient), the student is now closer to the standard of Proficient at this grade than he/she was the year previously to the standard for Proficient at the lower grade. Restated, a higher scale score at one grade than another means that the student is achieving better relative to the standard for Proficient at the higher grade.

TRANSFORMATIONS

Student measures were transformed mathematically to a more convenient metric. The minimum scale scores necessary for each proficiency level are provided in Tables 8–1 through 8–3. Tables 8–4 through 8–6 provide the equations used for each transformation. These equations were applied to the overall test as well as to each subscale reporting category.

Table 8–1. Mathematics Raw and Scale Score Cutpoints for Each Proficiency Level

Grade	Raw Score Cut Point			Below Proficient		Proficient		Advanced	
	Below Proficient	Proficient	Advanced	SS Cut	SSSE	SS Cut	SSSE	SS Cut	SSSE
3	26	34	52	263	17	300	17	390	21
4	27	35	51	260	18	300	18	383	22
5	24	35	50	252	17	300	16	373	19
6	27	37	52	258	17	300	17	376	21
7	26	38	53	248	17	300	17	383	22
8	28	37	52	258	17	300	17	379	21
9	30	40	53	258	16	300	17	370	21
10	16	22	33	252	21	300	21	392	26

Table 8–2. Reading Raw and Scale Score Cutpoints for Each Proficiency Level

Grade	Raw Score Cut Point			Below Proficient		Proficient		Advanced	
	Below Proficient	Proficient	Advanced	SS Cut	SSSE	SS Cut	SSSE	SS Cut	SSSE
3	16	24	44	261	18	300	16	392	18
4	17	25	47	260	19	300	18	415	21
5	15	24	48	251	19	300	17	418	20
6	16	30	48	234	19	300	17	394	21
7	17	27	47	246	20	300	19	406	22
8	16	26	44	243	20	300	18	402	21
9	15	28	44	229	20	300	18	382	20
10	18	34	54	222	20	300	18	400	20

Table 8–3. Writing Raw and Scale Score Cutpoints for Each Proficiency Level

Grade	Raw Score Cut Point			Below Proficient		Proficient		Advanced	
	Below Proficient	Proficient	Advanced	SS Cut	SSSE	SS Cut	SSSE	SS Cut	SSSE
3	13	28	48	218	20	300	17	402	21
4	13	28	47	204	24	300	21	420	26
5	12	31	48	187	23	300	19	406	24
6	18	33	47	215	22	300	21	396	26
7	19	31	51	234	19	300	18	423	26
8	20	32	53	232	20	300	19	460	30
9	20	32	54	238	20	300	19	470	32
10	18	28	45	233	21	300	23	485	40

Table 8–4. Equations Used for Each Transformation in Mathematics

Grade	Conversion Equation	Logit Cuts		
		BP	P	A
3	Scale Score = (61.9835 x Logit) + 291.0799	-0.4508	0.1358	1.6023
4	Scale Score = (66.9643 x Logit) + 280.6843	-0.3091	0.2810	1.5209
5	Scale Score = (61.9835 x Logit) + 283.6666	-0.5175	0.2554	1.4354
6	Scale Score = (63.0252 x Logit) + 282.7344	-0.3923	0.2660	1.4826
7	Scale Score = (64.1026 x Logit) + 288.3632	-0.6263	0.1737	1.4709
8	Scale Score = (63.5593 x Logit) + 277.1580	-0.3074	0.3515	1.6095
9	Scale Score = (63.0252 x Logit) + 275.7762	-0.2816	0.3764	1.5009
10	Scale Score = (58.1395 x Logit) + 274.5000	-0.3861	0.4300	2.0123

Table 8–5. Equations Used for Each Transformation in Reading

Grade	Conversion Equation	Logit Cuts		
		BP	P	A
3	Scale Score = (57.6923 x Logit) + 311.5058	-0.8672	-0.2081	1.3915
4	Scale Score = (64.1026 x Logit) + 308.1343	-0.7534	-0.1347	1.6742
5	Scale Score = (61.9835 x Logit) + 314.7427	-1.0230	-0.2459	1.6635
6	Scale Score = (61.9835 x Logit) + 298.1758	-1.0297	0.0214	1.5498
7	Scale Score = (68.8073 x Logit) + 307.7720	-0.8924	-0.1202	1.4297
8	Scale Score = (65.7895 x Logit) + 306.6157	-0.9669	-0.1082	1.4532
9	Scale Score = (65.2174 x Logit) + 302.9500	-1.1337	-0.0529	1.2131
10	Scale Score = (70.0935 x Logit) + 251.6402	-0.4299	0.6828	2.1200

Table 8–6. Equations Used for Each Transformation in Writing

Grade	Conversion Equation	Logit Cuts		
		BP	P	A
3	Scale Score = (61.4754 x Logit) + 309.6745	-1.4902	-0.1655	1.4982
4	Scale Score = (73.5294 x Logit) + 309.6277	-1.4319	-0.1377	1.4958
5	Scale Score = (67.5676 x Logit) + 289.1747	-1.5169	0.1528	1.7239
6	Scale Score = (73.5294 x Logit) + 281.3051	-0.9032	0.2475	1.5653
7	Scale Score = (65.2174 x Logit) + 284.0147	-0.7680	0.2374	2.1252
8	Scale Score = (67.5676 x Logit) + 288.0692	-0.8368	0.1692	2.5389
9	Scale Score = (67.5676 x Logit) + 274.6686	-0.5415	0.3675	2.8945
10	Scale Score = (68.1818 x Logit) + 252.9795	-0.2972	0.6823	3.3982

Complete raw-to-scale score tables are provided in Appendix 16.

SCALE SCORE SUMMARY STATISTICS

Table 8–7 includes scale score descriptive information for each overall content area test by grade. Subscale descriptive statistics can be found in Appendix 17. Histograms of the overall test scale scores are also provided in Figures 8–1 to 8–24. It should be noted that the spikes in the histograms occur when two raw score points fall in the same scale score range represented by the bar.

Table 8–7. Content Area Scale Score Information

Grade	Statistic	Spring 2007 SBA		
		Mathematics	Reading	Writing
3	Mean	364.08	371.31	370.09
	Standard Error of Mean	0.81	0.77	0.90
	Median	365	375	369
	Mode	404	436	461
	Standard Deviation	77.32	73.50	86.08
4	Mean	355.54	378.26	378.27
	Standard Error of Mean	0.82	0.82	0.93
	Median	356	385	381
	Mode	420	424	448
	Standard Deviation	78.59	79.16	89.62
5	Mean	358.66	376.51	355.82
	Standard Error of Mean	0.79	0.78	0.83
	Median	356	379	358
	Mode	418	432	410
	Standard Deviation	75.48	74.53	79.67
6	Mean	349.48	367.74	357.66
	Standard Error of Mean	0.76	0.75	0.91
	Median	350	373	363
	Mode	412	432	413
	Standard Deviation	73.25	72.90	87.71
7	Mean	340.37	368.16	341.83
	Standard Error of Mean	0.78	0.75	0.76
	Median	340	373	339
	Mode	363	427	345
	Standard Deviation	76.42	74.16	74.95
8	Mean	338.09	376.50	347.58
	Standard Error of Mean	0.74	0.70	0.79
	Median	340	378	349
	Mode	350	403	391
	Standard Deviation	73.13	68.56	77.46
9	Mean	330.56	370.39	350.95
	Standard Error of Mean	0.76	0.72	0.79
	Median	325	373	348
	Mode	379	404	397
	Standard Deviation	76.80	72.35	79.44
10	Mean	334.30	374.11	356.88
	Standard Error of Mean	0.74	0.72	0.73
	Median	335	379	355
	Mode	344	420	396
	Standard Deviation	71.34	69.93	71.05

Figure 8-1

Mathematics Scale Score Frequencies

GRADE: 03

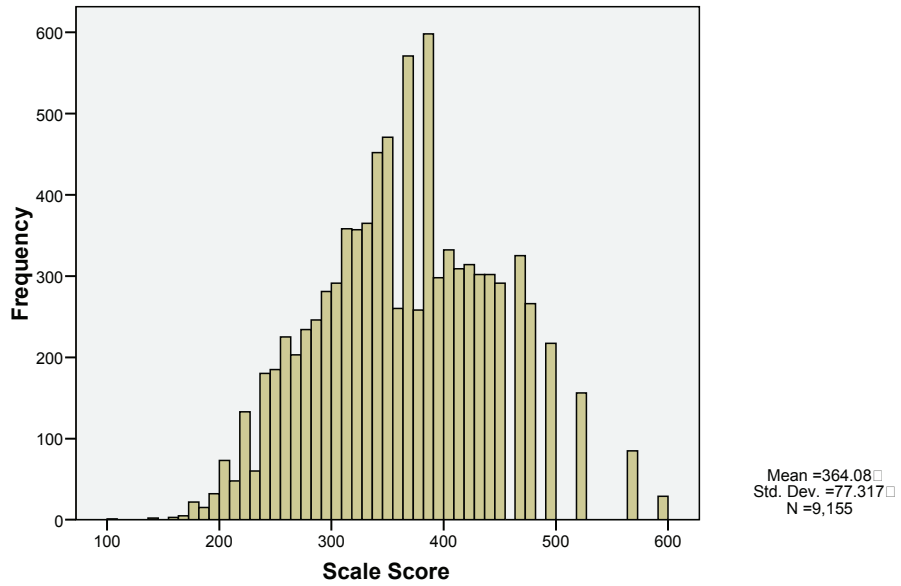


Figure 8-2

Mathematics Scale Score Frequencies

GRADE: 04

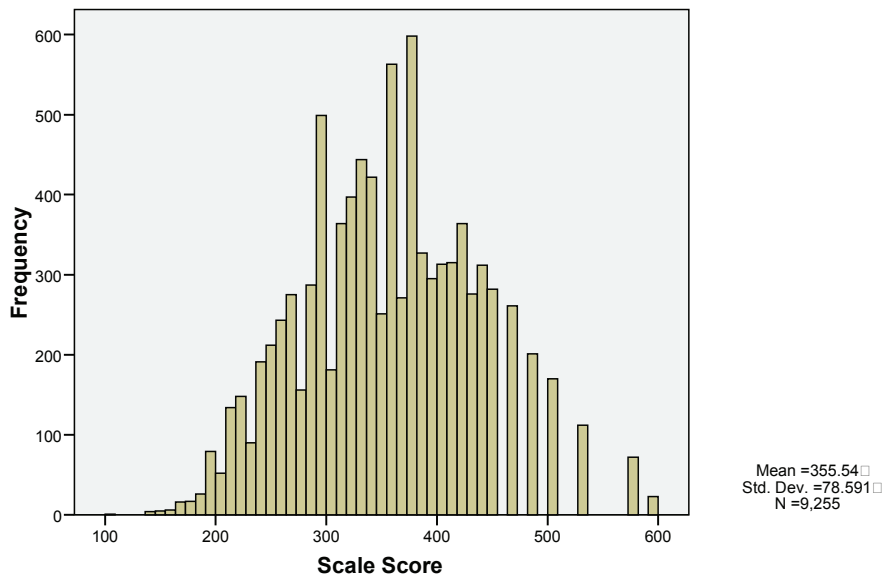


Figure 8-3

Mathematics Scale Score Frequencies

GRADE: 05

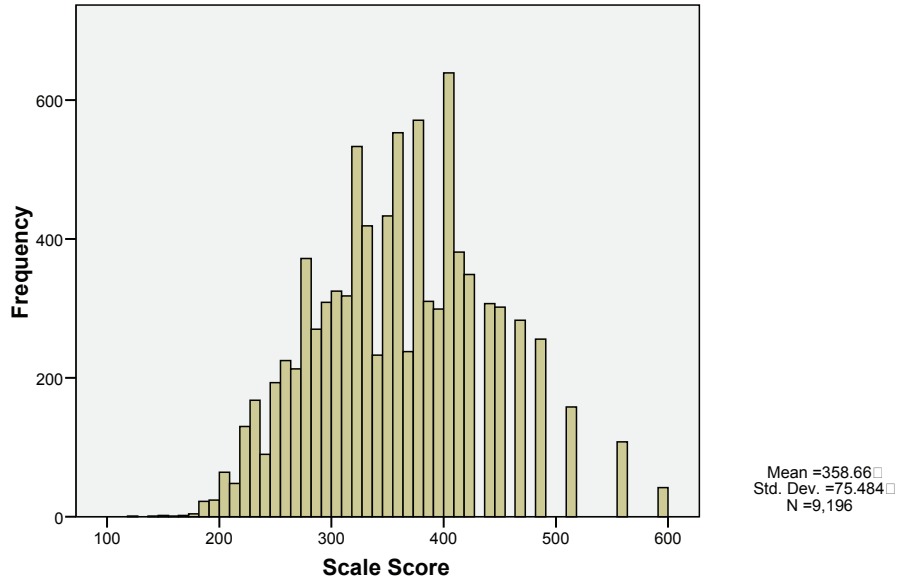


Figure 8-4

Mathematics Scale Score Frequencies

GRADE: 06

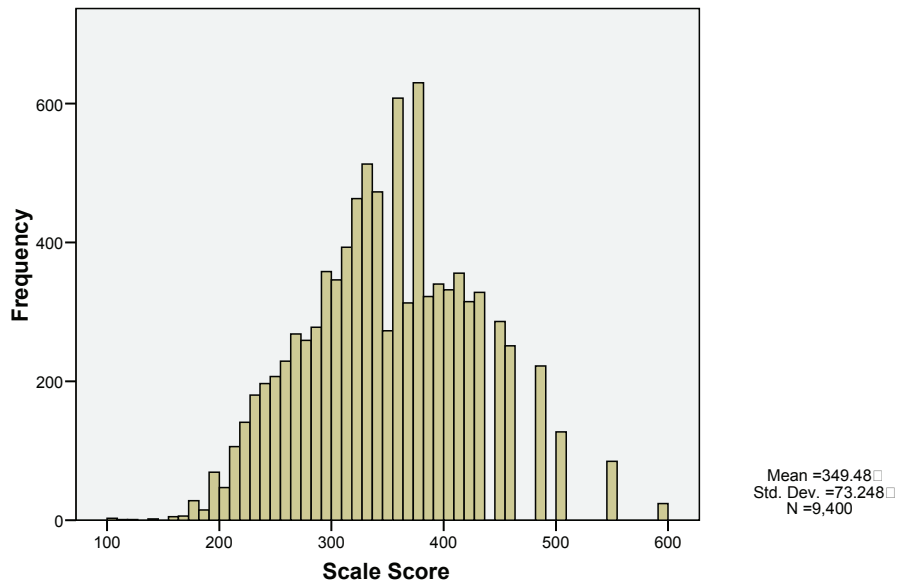


Figure 8-5

Mathematics Scale Score Frequencies

GRADE: 07

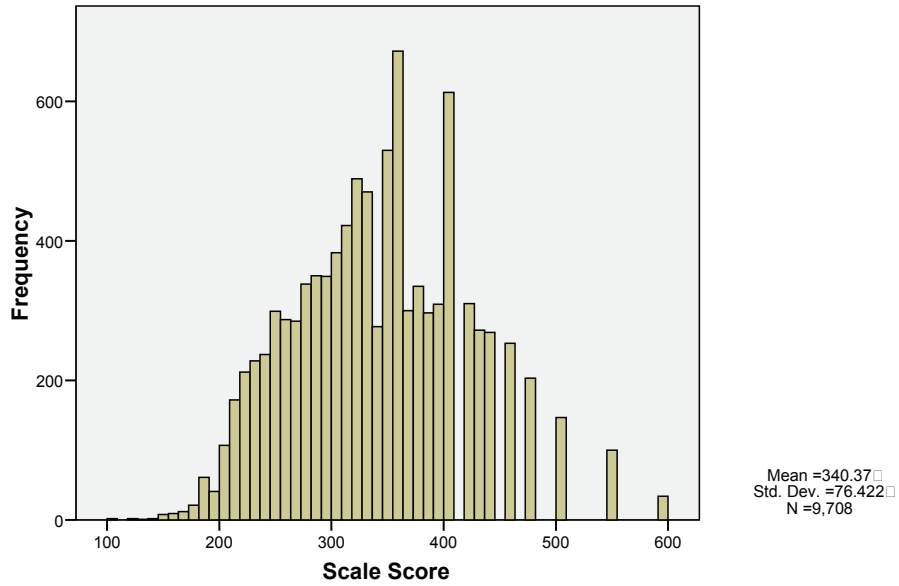


Figure 8-6

Mathematics Scale Score Frequencies

GRADE: 08

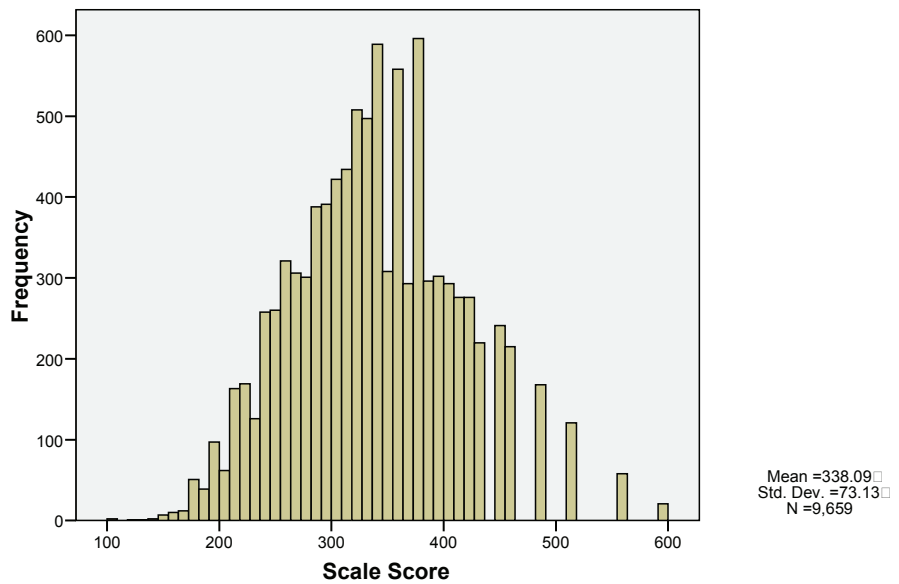


Figure 8-7

Mathematics Scale Score Frequencies

GRADE: 09

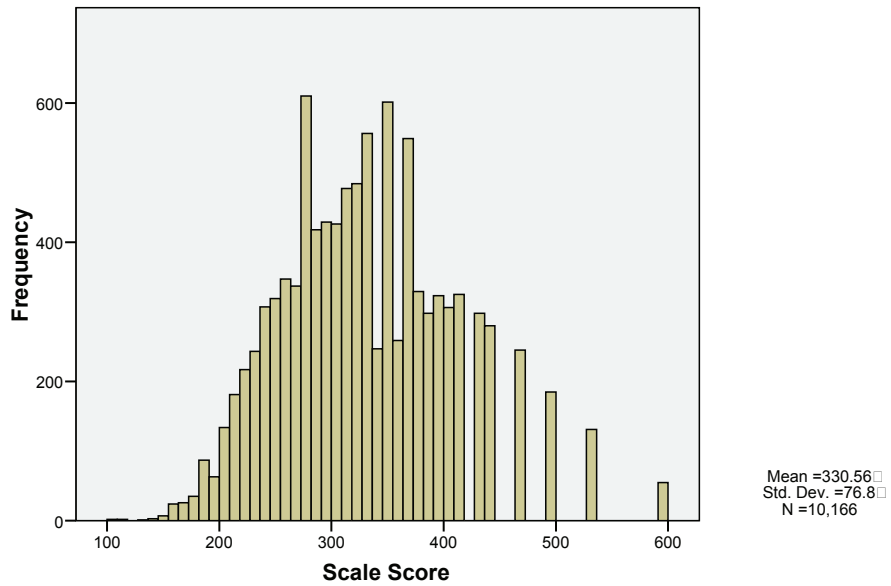


Figure 8-8

Mathematics Scale Score Frequencies

GRADE: 10

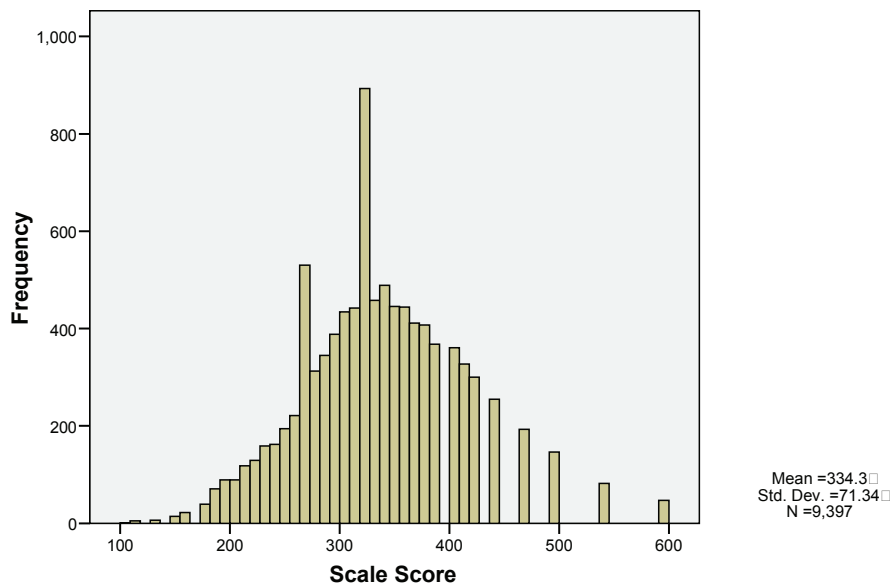


Figure 8–9

Reading Scale Score Frequencies

GRADE: 03

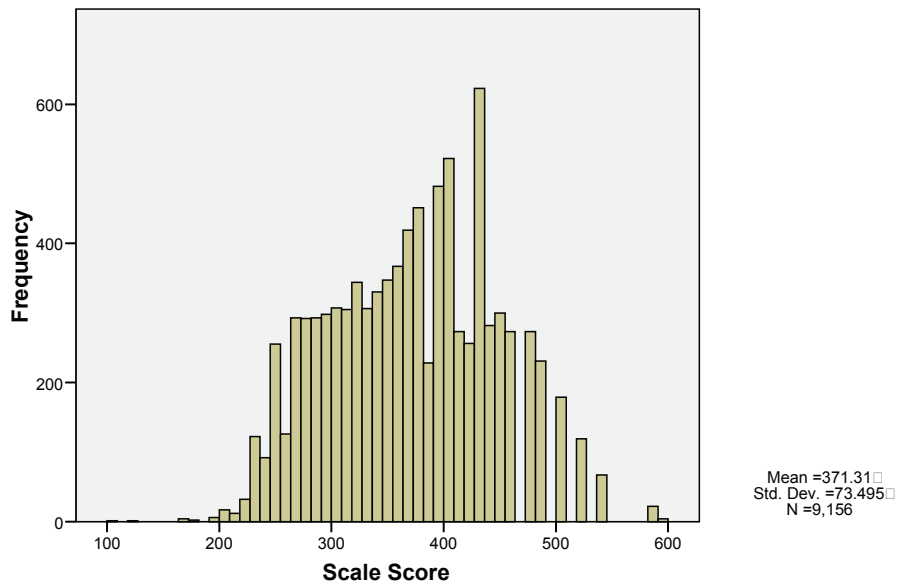


Figure 8–10

Reading Scale Score Frequencies

GRADE: 04

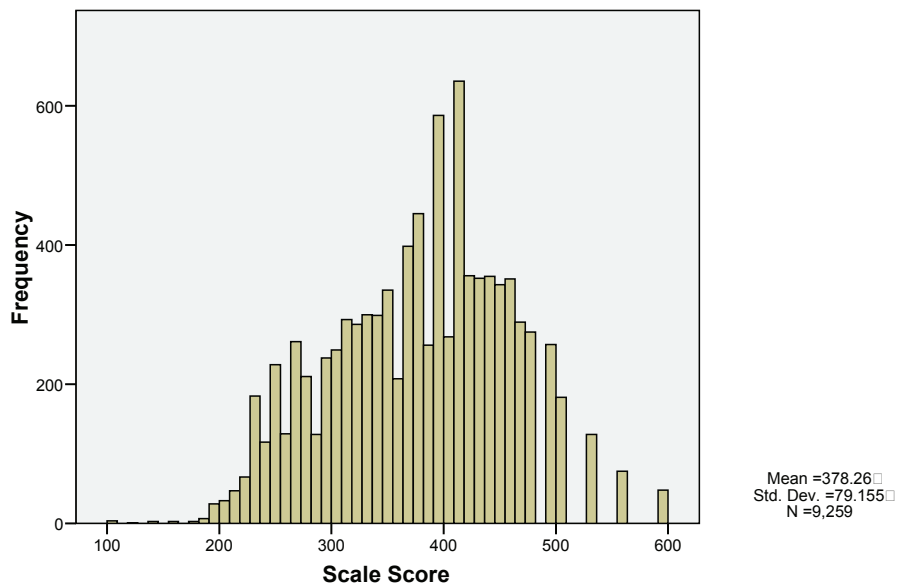


Figure 8–11

Reading Scale Score Frequencies

GRADE: 05

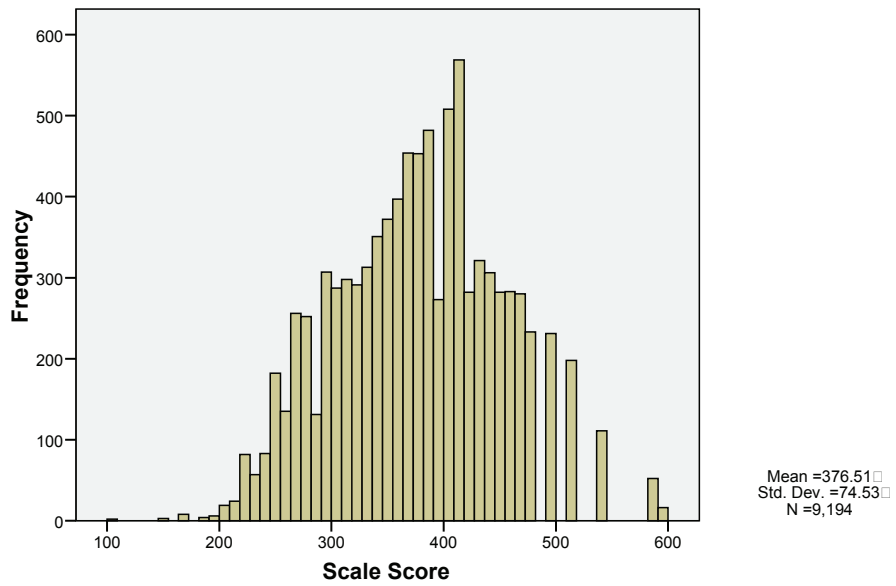


Figure 8–12

Reading Scale Score Frequencies

GRADE: 06

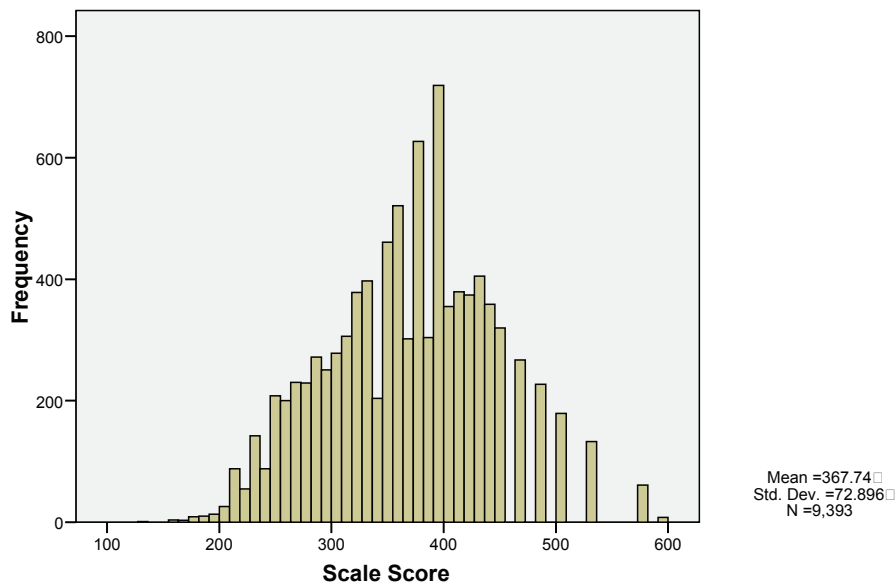


Figure 8–13

Reading Scale Score Frequencies

GRADE: 07

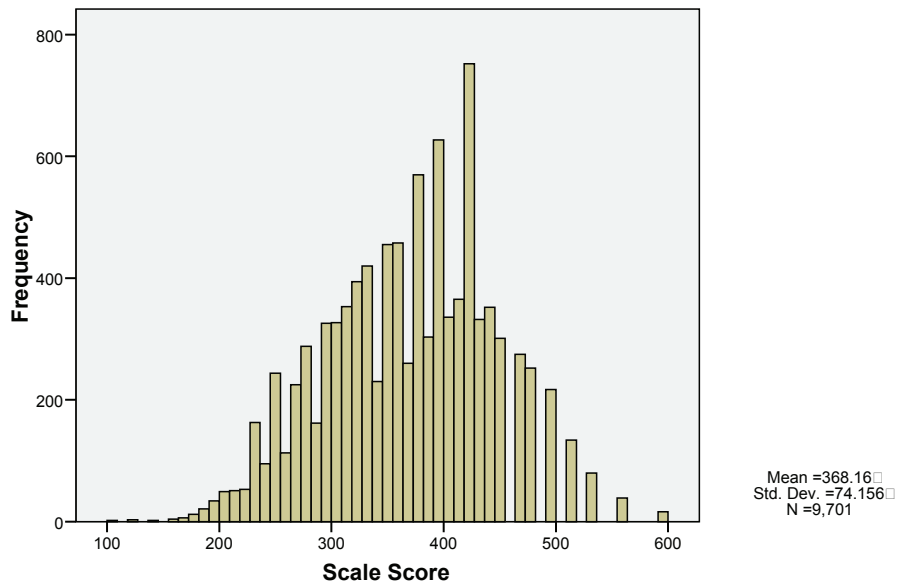


Figure 8–14

Reading Scale Score Frequencies

GRADE: 08

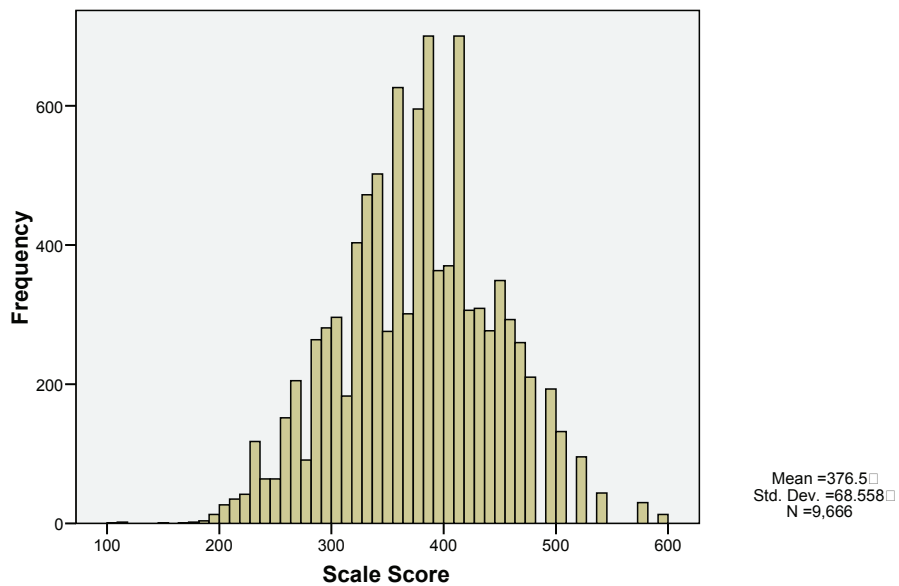


Figure 8-15

Reading Scale Score Frequencies

GRADE: 09

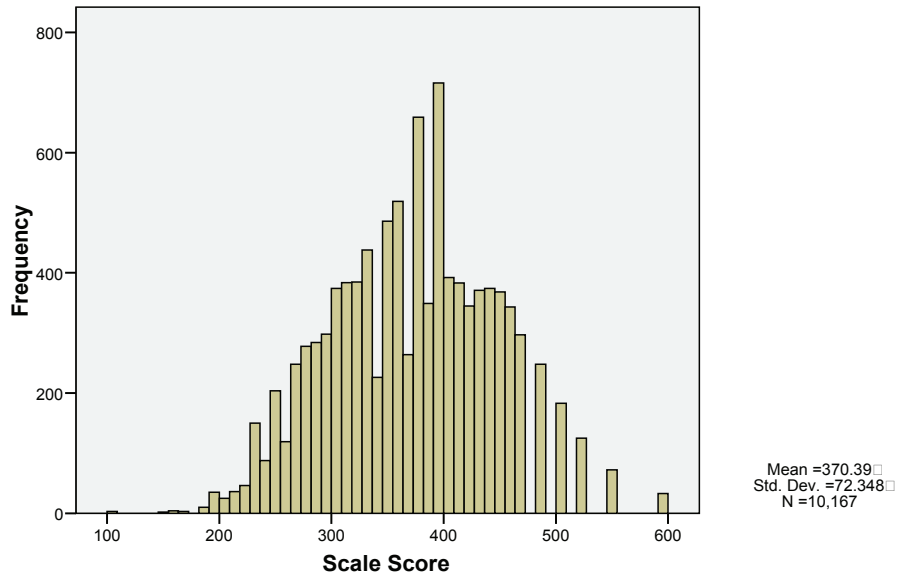


Figure 8-16

Reading Scale Score Frequencies

GRADE: 10

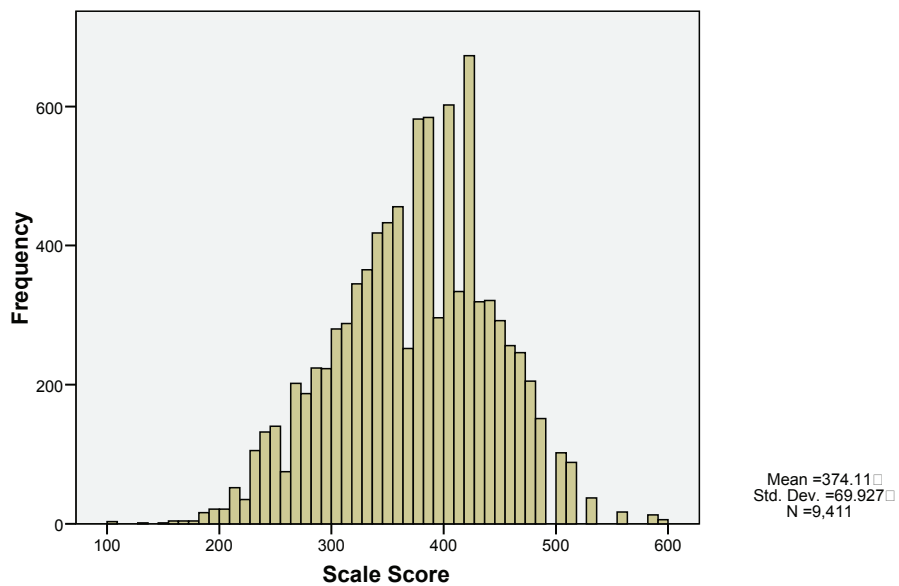


Figure 8-17

Writing Scale Score Frequencies

GRADE: 03

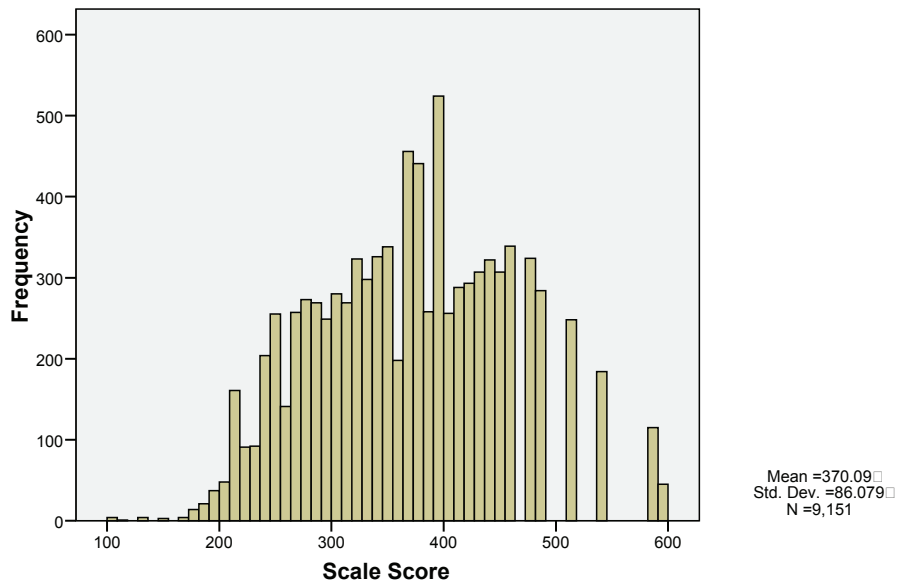


Figure 8-18

Writing Scale Score Frequencies

GRADE: 04

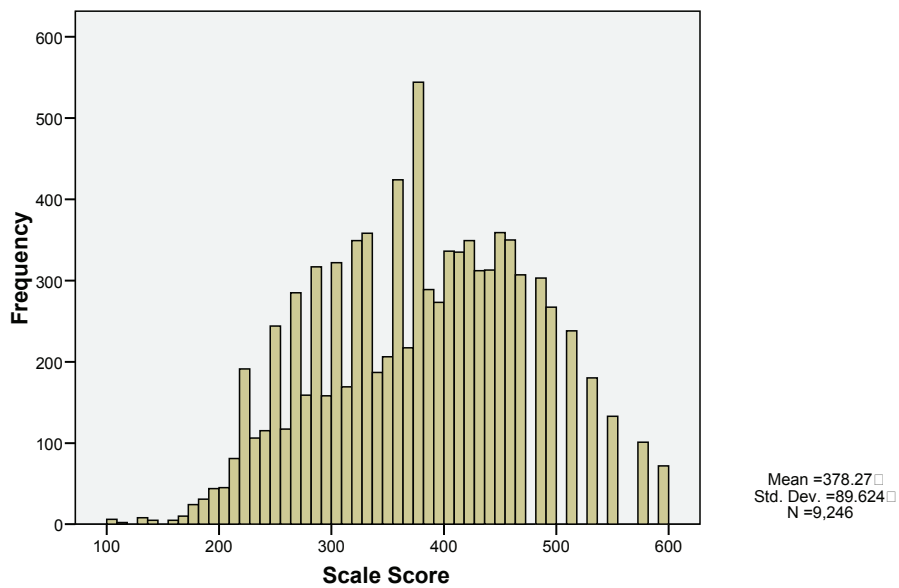


Figure 8–19

Writing Scale Score Frequencies

GRADE: 05

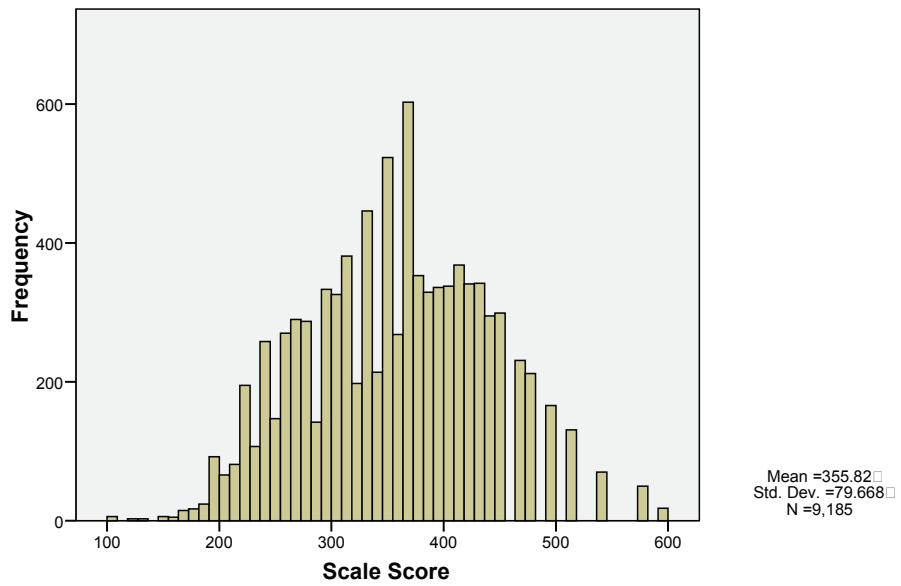


Figure 8–20

Writing Scale Score Frequencies

GRADE: 06

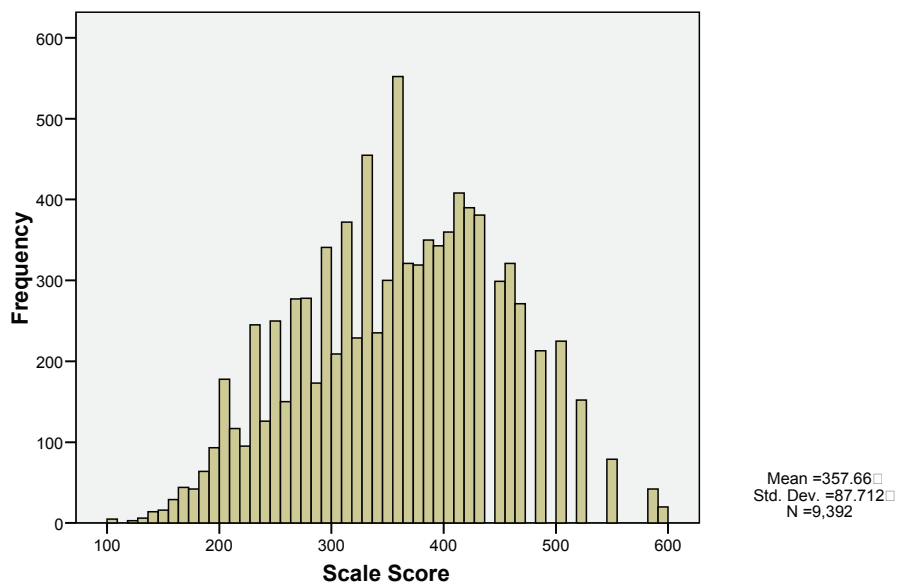


Figure 8–21

Writing Scale Score Frequencies

GRADE: 07

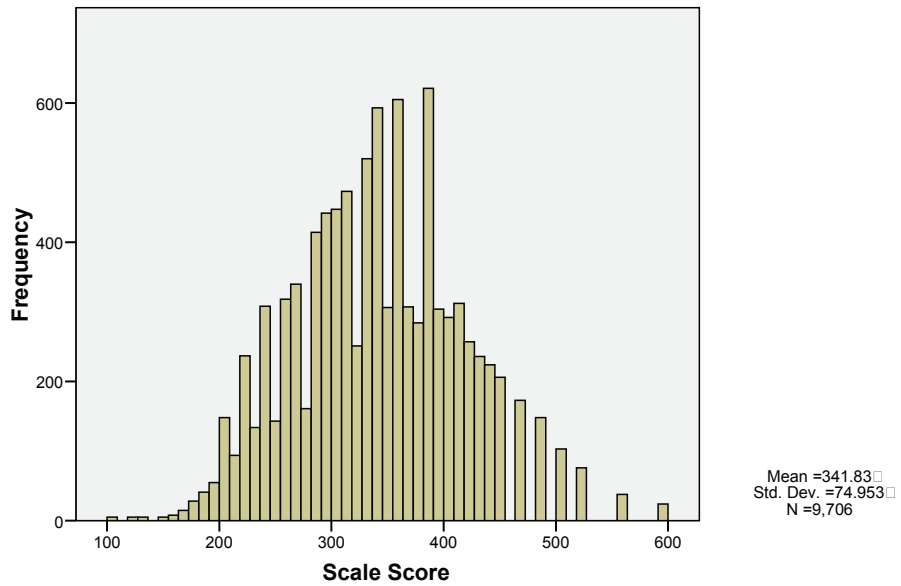


Figure 8–22

Writing Scale Score Frequencies

GRADE: 08

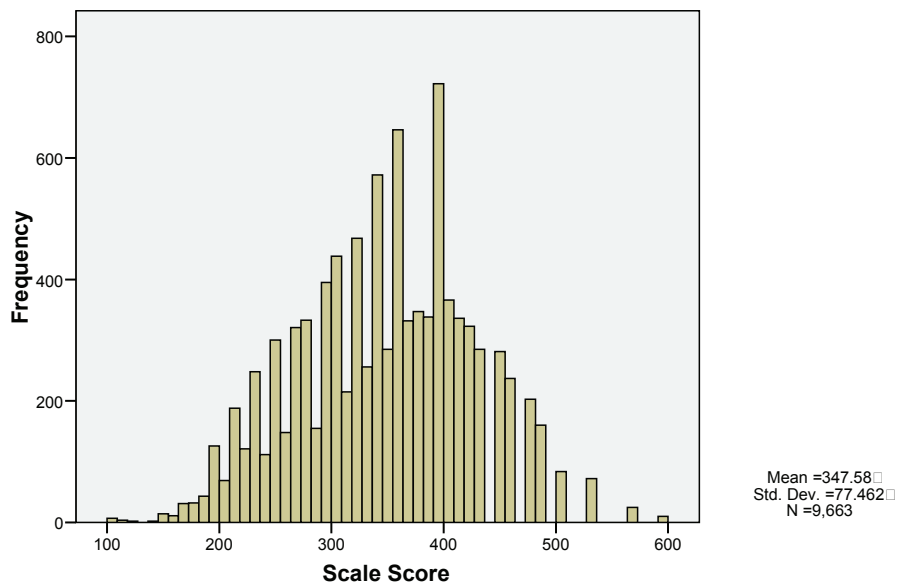


Figure 8-23

Writing Scale Score Frequencies

GRADE: 09

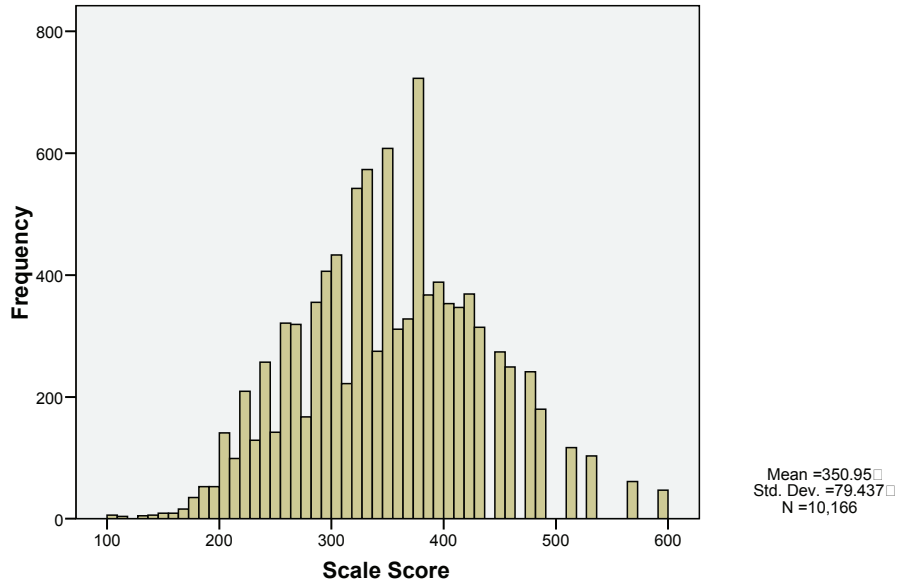
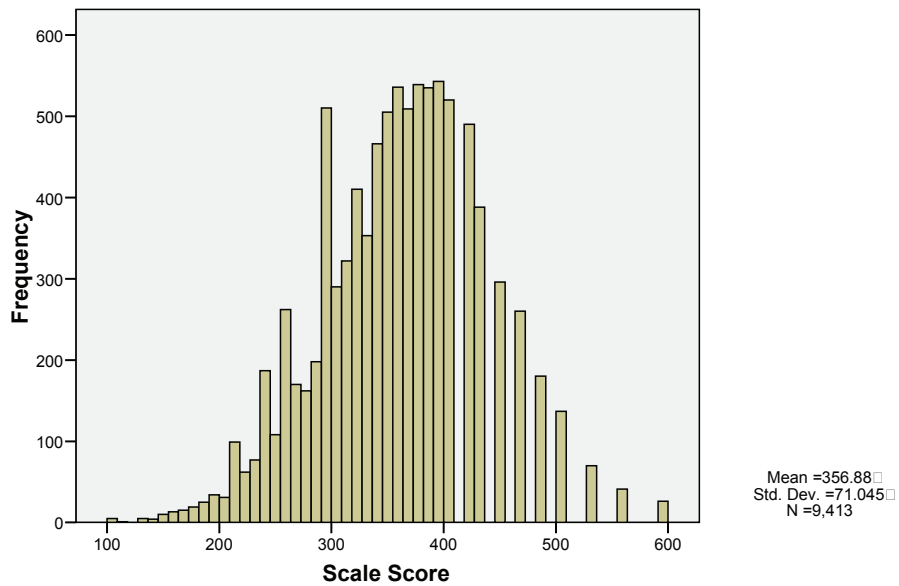


Figure 8-24

Writing Scale Score Frequencies

GRADE: 10



PROFICIENCY LEVELS

Information from the SBA is used to determine whether adequate yearly progress has been met in each school and district. Alaska has four levels of achievement on the SBA tests: Far Below Proficient (FB), Below Proficient (BP), Proficient (P), and Advanced (A).

Scale score cutpoints at each level of proficiency are the same each year. Appendix 18 provides detailed information about the proficiency level as well as the Proficiency Level Definitions and Descriptors in each grade and content area tested.

Table 8–8 provides the distribution of students in each of the proficiency levels for all grades and content areas.

Table 8–8. Student Distribution of the Four Proficiency Levels

Grade	Level	Mathematics		Reading		Writing	
		Count	Percent	Count	Percent	Count	Percent
3	Far Below Proficient	984	10.7	670	7.3	297	3.2
	Below Proficient	964	10.5	1176	12.8	1831	20.0
	Proficient	3981	43.5	3640	39.8	3711	40.6
	Advanced	3226	35.2	3670	40.1	3312	36.2
4	Far Below Proficient	1098	11.9	724	7.8	135	1.5
	Below Proficient	1162	12.6	967	10.4	1818	19.7
	Proficient	3672	39.7	4216	45.5	4358	47.1
	Advanced	3323	35.9	3352	36.2	2935	31.7
5	Far Below Proficient	750	8.2	371	4.0	79	0.9
	Below Proficient	1389	15.1	1180	12.8	2268	24.7
	Proficient	3052	33.2	4771	51.9	4315	47.0
	Advanced	4005	43.6	2872	31.2	2523	27.5
6	Far Below Proficient	1008	10.7	286	3.0	494	5.3
	Below Proficient	1392	14.8	1543	16.4	2052	21.8
	Proficient	3679	39.1	4142	44.1	3685	39.2
	Advanced	3321	35.3	3422	36.4	3161	33.7
7	Far Below Proficient	1115	11.5	495	5.1	780	8.0
	Below Proficient	1908	19.7	1358	14.0	2126	21.9
	Proficient	3878	39.9	4733	48.8	5572	57.4
	Advanced	2807	28.9	3115	32.1	1228	12.7
8	Far Below Proficient	1408	14.6	310	3.2	783	8.1
	Below Proficient	1559	16.1	1057	10.9	1879	19.4
	Proficient	3907	40.4	4717	48.8	6210	64.3
	Advanced	2785	28.8	3582	37.1	791	8.2

9	Far Below Proficient	1813	17.8	231	2.3	774	7.6
	Below Proficient	1979	19.5	1602	15.8	1967	19.3
	Proficient	3325	32.7	3735	36.7	6676	65.7
	Advanced	3049	30.0	4599	45.2	749	7.4
10	Far Below Proficient	1098	11.7	127	1.3	400	4.2
	Below Proficient	1797	19.1	1323	14.1	1597	17.0
	Proficient	4791	51.0	4299	45.7	7142	75.9
	Advanced	1711	18.2	3662	38.9	274	2.9

Indicators of Consistency

Criterion-referenced tests are often used to place the examinees into two or more performance classifications. It is thus useful to have some indication of how consistent such classifications are.

Decision Consistency Index

Method I

In a personal communication to DRC from Dr. Huynh Huynh on the DRC South Carolina project, an extension of the two parameter beta-binomial model (Huynh, 1976) to polytomous constructed-response items was detailed. This extension was used in these computations. Table 8–9 depicts the general framework of multiple decisions.

Table 8–9. Multiple Decisions—General Framework

	Category 1	Category 2	Category 3	Category 4	Total
Category 1	p_{11}				$p_{1.}$
Category 2		p_{22}			$p_{2.}$
Category 3			p_{33}		$p_{3.}$
Category 4				p_{44}	$p_{4.}$
Total	$p_{.1}$	$p_{.2}$	$p_{.3}$	$p_{.4}$	$p_{..}$

From this general framework the reliability index can be computed:

$$\kappa = \frac{1 - p}{p - p_c},$$

where $p = p_{..}$,

$$p_c = \sum_i p_i^2,$$

$$p_{11} = \sum_{x,y=c_1}^n f(x,y),$$

and
$$p_1 = \sum_{x=c_1}^n f(x).$$

Method II

To solve the problem of a complex assessment, Livingston and Lewis (1995) proposed an effective test length,

$$n = \frac{(\mu_x - X_{\min})(X_{\max} - \mu_x) - r\sigma_x^2}{\sigma_x^2(1-r)},$$

which transforms the original raw score random variable from $X = 0, \dots, K$ into a new random variable $X' = 0, \dots, n$, where n is the number of dichotomous, locally independent, equally difficult items required to produce a raw score of the same reliability. Then, using the transformed observed distribution X' , parameters are estimated for a four parameter beta-binomial model where the conditional error distribution is assumed to be binomial. The X' distribution is then converted back onto the original X scale using interpolation. This method is designed only to estimate a contingency table, not a full bivariate distribution which means the probability of a consistent decision by chance, and subsequently kappa, cannot be estimated.

The results of both consistency analyses are presented in Table 8-10.

Table 8–10. Decision Consistency Indices

	Huynh (1976)				Livingston and Lewis (1995)	
	Mathematics					
	4 categories (FB, BP, P, A)		2 categories (Not Proficient, Proficient)		4 categories (FB, BP, P, A)	2 categories (Not Proficient, Proficient)
Grade	Consistency Index	κ	Consistency Index	κ	Consistency Index	Consistency Index
3	.7691	.6522	.9188	.7568	.7666	.9238
4	.7499	.6338	.9092	.7533	.7559	.9177
5	.7694	.6562	.9165	.7650	.7679	.9214
6	.7487	.6350	.9067	.7545	.7538	.9158
7	.7462	.6397	.8999	.7663	.7454	.9083
8	.7319	.6199	.8959	.7552	.7423	.9058
9	.7307	.6324	.8950	.7752	.7295	.9008
10	.6684	.4949	.8539	.6577	.6822	.8599

Reading						
	4 categories (FB, BP, P, A)		2 categories (Not Proficient, Proficient)		4 categories (FB, BP, P, A)	2 categories (Not Proficient, Proficient)
Grade	Consistency Index	κ	Consistency Index	κ	Consistency Index	Consistency Index
3	.7847	.6721	.9257	.7666	.7853	.9301
4	.7943	.6793	.9299	.7625	.7893	.9386
5	.8009	.6736	.9324	.7558	.7840	.9295
6	.7861	.6674	.9202	.7443	.7865	.9306
7	.7765	.6467	.9183	.7334	.7784	.9291
8	.7698	.6227	.9230	.6807	.7785	.9309
9	.7818	.6562	.9205	.7289	.7992	.9312
10	.7882	.6572	.9242	.7072	.7948	.9343
Writing						
	4 categories (FB, BP, P, A)		2 categories (Not Proficient, Proficient)		4 categories (FB, BP, P, A)	2 categories (Not Proficient, Proficient)
Grade	Consistency Index	κ	Consistency Index	κ	Consistency Index	Consistency Index
3	.8069	.7083	.9258	.7908	.7918	.9284
4	.7874	.6652	.9137	.7387	.7755	.9204
5	.7809	.6585	.9004	.7376	.7724	.9140
6	.7502	.6334	.8963	.7380	.7467	.9121
7	.7679	.6135	.8892	.7355	.7634	.9001
8	.7861	.6062	.8915	.7285	.7768	.9081
9	.7958	.6117	.8949	.7332	.7872	.9068
10	.7901	.4868	.8625	.5954	.8313	.8934

CHAPTER 9: TEST VALIDITY & RELIABILITY

INTRODUCTION

Validity is the process of collecting evidence to support inferences from the use of the scores derived from the assessment process. Evidence on content validity of the spring 2007 SBA is presented in terms of how the assessments were assembled to reflect the EED-prescribed blueprints that in turn reflect state content standards in each grade and content area.

Reliability is defined as the consistency of measures. The ability to measure consistently is necessary, but not sufficient, condition for making valid interpretations of the results.

VALIDITY

Content/Curricular

The SBAs is a criterion-referenced assessment. This assessment is based on an extensive definition of the content it assesses. Therefore, the SBAs are content-based and aligned directly to the Alaska statewide content standards and should demonstrate good content validity. Content validity addresses whether the test adequately samples the relevant material it purports to cover.

Relation to Statewide Content Standards

From the inception of the SBAs, a committee of educators, item development experts, assessment experts, and EED staff have met to review new and field tested items. A sequential review process has been put in place by EED. This provides many opportunities for these professionals to offer suggestions for improving or eliminating items as well as offer insights into the interpretation of the statewide content standards for the SBAs. These review committees participate in this process to ensure test content validity of the SBAs.

In addition to providing information on the difficulty, appropriateness, and fairness of these items, committee members provide a needed check on the alignment between the items and the content standards they are intended to measure. When items are judged relevant, that is, representative of the content defined by the standards, this judgment provides evidence to support the validity of inferences made (regarding knowledge of this content) with SBA results. When items are judged to be unacceptable for any reason, the committee can either suggest revisions (e.g., reclassification, rewording) or elect to eliminate the item from the field test item pool. Items that are approved by the review committee are later embedded in operational SBA forms to allow for the collection of performance data. In essence, these committees review and verify the alignment of the test items with the objectives and measurement specifications to ensure that the items measure appropriate content. The nature and specificity of these review procedures provide strong evidence for the content validity of the SBAs.

Educator Input

For the Spring 2007 SBAs, Alaska educators provided valuable input on the alignment of the items and the statewide content standards during item development. Items were written specifically to measure the objectives and specifications of the content standards for the SBAs. Because many different people with different backgrounds wrote the items, the process included a built-in system of checks-and-balances for item development and review that reduced single

source bias. This direct input from educators offers evidence regarding the content validity of the SBAs. See Chapter 2 for details regarding the content review process.

Developer Input

For the items included in the spring 2007 forms, EED and DRC staff provided a history of test building experience, including content-related expertise. The input and review by these assessment professionals provided further support of the item being an accurate measure of the intended objective. Thus, these reviews offer additional evidence for the content validity of the SBAs.

Item to Content Area Match

Expert judgments from educators, test developers, and assessment specialists provide support for the alignment of the SBAs with the statewide content standards. In addition, because expert teachers in the content areas were involved in establishing the content standards, the judgments of these same expert teachers in the review process provide a measure of content validity. A match between the content standards and the components of the SBAs provides evidence that the assessment measures the content standards. A table showing the number of assessment components, tasks, or items matching each content-standard is often used to provide documentation of the content validity of an assessment. The SBA test blueprint provides this documentation. The blueprints for mathematics, reading, and writing are presented in Appendix 1.

Construct Validity

The term construct validity refers to the degree to which the test score is a measure of the educational domain (i.e., construct) of interest. A construct is an individual characteristic that is assumed to exist in order to explain some aspect of behavior (Linn & Gronlund, 1995). When a particular individual characteristic from the assessment results is inferred, a generalization or interpretation of some construct is made. For example, problem solving is a construct. An inference that students who master the mathematical reasoning portion of an assessment are “good problem-solvers” implies an interpretation of the results of the assessment in terms of a construct. To make such an inference, it is important to demonstrate that this is a reasonable and valid use of the results.

Construct-related validity evidence can come from many sources. *The Standards for Educational and Psychological Testing* (AERA, APA, and NCME, 1999) provides the following list of possible sources:

- High inter-correlations among assessment items or tasks attest that the items are measuring the same trait, such as a content objective, sub-domain, or construct.
- Substantial relationships between the assessment results and other measures of the same defined construct.
- Little or no relationship between the assessment results and other measures that are clearly not of the defined construct.
- Substantial relationships between different methods of measurement regarding the same defined construct.

- Relationships to non-assessment measures of the same defined construct.

Evidence of Construct Validity

The collection of construct-related evidence is a continuous and ongoing process. Three indicators of construct validity for the spring 2007 SBAs are item-total correlations, Rasch item fit statistics, and intercorrelations.

Item-Total Correlations

An item-total correlation is the correlation between an item and the total test score, excluding that item score. Conceptually, if an item has a high item-total correlation (i.e., 0.40 or above), it indicates that students who performed well on the test overall usually answered the item correctly and students who performed poorly on the test overall usually answered the item incorrectly. That is, the item did a good job discriminating between high performing and low performing students. Assuming that the total test score represents the extent to which a student possesses the construct being measured by the test, high item-total correlations indicate that the items on the test require knowledge of this construct in order to be answered correctly. Item-total correlations for items on the spring 2007 SBA can be found in Appendix 13. The majority of items have item-total correlations over .30 (91% overall and 90% for math, 90% for reading, and 94% for writing for grades 3–9). These high item-total correlations provide evidence for construct validity.

Fit Statistics

In addition to item-total correlations, Rasch fit statistics also provide good evidence of construct validity. The Rasch model requires unidimensional data. Therefore, statistics showing that the items fit the measurement model also provide evidence of construct validity. Fit statistics for the spring 2007 SBA can be found in Appendix 13. Note that 36% of mathematics item fit statistics, 32% of the reading item fit statistics, and 39% writing item fit statistics for grades 3–9 are between -5.00 and +5.00, indicating good construct validity.

Intercorrelations

A third indicator of construct validity is the intercorrelations between the content area total scale scores and the subscale reporting category scale scores. This information is contained in Appendix 23 and is reported by grade. In addition, intercorrelations between the scale scores for the three content area total scale scores are presented.

Validity Evidence for Different Student Populations

The primary evidence for the validity of the SBAs lies in the content and constructs being measured. Because the test assesses the statewide content standards required to be taught to all students, the test should not be more or less valid for use with one sub-population of students over another sub-population. In other words, because the SBAs are measuring what is required to be taught to all students and are given under the same standardized conditions to all students, the validity of score interpretations should apply to all students. A summary of student demographic information is presented in Appendix 24 for grades 3 to 10.

Great care has been taken to ensure that the items comprising the SBAs are fair and representative of the content domain expressed in the content standards. Much scrutiny is applied to the items and their possible impact on minority or other sub-populations making up the population in the state of Alaska. Every effort is made to eliminate items that may have gender, ethnic, or cultural biases. See Chapter 2 for the discussion of how potential item bias is identified.

RELIABILITY

The classical view of measurement considers all measures as having a “true” component and an error component. Errors occur as a natural part of the measurement process and can never be eliminated entirely. For example, uncontrollable factors such as differences in the physical world and changes in examinee disposition may work to increase error and decrease reliability. This is the fundamental premise of true-score reliability analysis and measurement theory. Stated explicitly, this relationship can be seen as the following:

$$X = T + E, \tag{1}$$

where X represents the observed test score, T , the student’s true score, and E , random error.

If the variance of the observed measures is denoted by σ_X^2 and the variance of error by σ_E^2 then the reliability (ρ_{XX}) is given by:

$$\rho_{XX} = \frac{\sigma_X^2 - \sigma_E^2}{\sigma_X^2}. \tag{2}$$

The variance of the observed measures can be estimated from the variance of the raw scores using the usual variance formula and the error variance can be estimated by:

$$\Sigma p(1-p), \tag{3}$$

where p is the proportion correct for each item.

The reliability index used for the 2007 administration of the SBAs was the Coefficient Alpha (Cronbach, 1951):

$$\alpha = \left(\frac{k}{k-1} \right) \left(1 - \frac{\sum \sigma_i^2}{\sigma_X^2} \right), \tag{4}$$

where k is the number of items, σ_i^2 is the variance of the set of scores associated with item i , and σ_X^2 is the variance of the set of observed total scores.

Acceptable α values generally range in the high 0.80s to low 0.90s. When there is no error, the reliability index is the true score variance divided by the true score variance, which is one. Appendix 14 provides Coefficient Alpha for each grade/content area combination. As can be seen in the tables, all mathematics, reading, and writing tests for grades 3–9 have Coefficient Alpha's over 0.90. Only two tests, mathematics and reading for grade 10, have reliability lower than 0.90, but the values are in the high 80s. These high α values provide evidence for good reliability. Appendix 22 provides the reliability of the assessments for each subpopulation required by NCLB.

Standard Error of Measurement

The standard error of measurement uses the information from the test along with an estimate of reliability to make statements about the degree to which error is impacting individual scores. The standard error of measurement is based on the premise that underlying traits, such as academic achievement, cannot be measured exactly. The standard error expresses unreliability in terms of the raw score metric. Using the standard error of measurement, an error band can be placed around an individual score indicating the degree to which error might be affecting that score. In true-score test theory, the standard error of measurement can be calculated by:

$$SEM = \sigma_x \sqrt{1 - \rho_{XX}} , \quad (5)$$

where, σ_x is the standard deviation of the total test (observed measure scores), and ρ_{XX} is the Coefficient Alpha reliability estimate for the test.

The true-score test theory approach to judging a test's consistency can be useful for making overall comparisons between alternate forms. However, it is not very useful for judging the precision with which a specific student's score is known. The Rasch measurement model provides asymptotic standard errors that pertain to each unique ability estimate (i.e., scale score).

Ability estimates from scores near the center of the test are known with greater precision than are abilities associated with extremely high or low scores. The expression for computing the asymptotic standard error via WINSTEPS was provided in Chapter 6. This value is then transformed to the SBA scale to obtain the final SEM for each raw score. These values for the spring 2007 SBAs are provided in the raw-to-scale score tables in Appendix 16. In addition, person separation reliability and item separation reliability values, which use these asymptotic standard errors are provided in Appendix 14. Person separation reliability is the Rasch equivalence of reliability described in Equation 2.

REFERENCES

- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. 1999. *Standards for educational and psychological testing*. 2nd ed. Washington, D.C.: American Educational Research Association.
- Bloom, B. S., Englehart, M. D., Furst, E. J., Hill, W. H., and Krathwohl, D. R. 1956. *Taxonomy of Educational Objectives: The classification of educational goals: Handbook 1: Cognitive Domain*. New York: Longman, Green, and Co.
- Cronbach, L. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika* 16: 297–334.
- Holland, P., and Thayer, D. 1986. *Differential item performance and the Mantel-Haenszel procedure*. Paper presented at the American Educational Research Association annual meeting, San Francisco, CA.
- Huynh, H. 1976. On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement* 13: 253–64.
- Jaeger, R. 1989. Certification of student competence. In R. Linn (Ed), *Educational Measurement*, 3rd edition, pp.485–514. New York: American Council on Education/Macmillan.
- Kane, M. 1995. Examinee-centered vs. task-centered standard setting. In the *Proceedings of the Joint Conference on Standard Setting for Large Scale Assessments*, Washington D.C.: National Assessment Governing Board and the National Center for Education Statistics.
- Lewis, L., Mitzel, H., Green, D., and Patz, R. 1999. *The bookmark standard setting procedure*. Monterey, CA: McGraw-Hill Companies.
- Linacre, J. M. 2006. *WINSTEPS Rasch measurement (Version 3.60.1)*. Chicago: WINSTEPS.com. Computer program.
- Linn, R., and N. Gronlund. 1995. *Measurement in assessment and teaching*. 7th ed. New Jersey: Prentice-Hill.
- Mead, R. 1978. Examining residuals from the Rasch model. *Proceedings of the 1978 conference on adaptive testing*. Minneapolis, MN: University of Minnesota.
- Mogilner, A. 1992. *Children's Writer's Word Book*. Cincinnati, OH: Writer's Digest Books.
- Rasch, G. 1960. *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research (Expanded Edition, 1980. Chicago: University of Chicago Press).
- Smith, R. M. 2000. Fit analysis in latent trait measurement models. *Journal of Applied Measurement* 1: 199–218.
- Stout, W. 1987. A non-parametric approach to assessing latent trait unidimensionality. *Psychometrika* 52: 589–617.

Taylor, S. E., Frackenpohl, H., White, C. E., Nieroroda, B. W., Browning, C. L., and Brisner, E. P. 1989. *EDL Core Vocabularies in Reading, Mathematics, Science, and Social Studies*. Orlando, FL: Steck-Vaughn Company.

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

Webb, N. L. 2002. *Alignment Study in Language Arts, Mathematics, Science, and Social Studies of State Standards and Assessment for Four States*. Washington, D.C.: Council of Chief State School Officers.

Wright, B. D., and G. N. Masters. 1982. *Rating scale analysis*. Chicago: MESA Press.

Zwick, R., and Thayer, D. 1996. Evaluating the magnitude of differential item functioning in polytomous items. *Journal of Educational and Behavioral Statistics*, 21, 187–201.