



Alaska

Comprehensive System of Student Assessment

Technical Report

Spring 2008

Grades 4, 8, and 10

Science

Standards Based Assessment (SBA)



September 2008

TABLE OF CONTENTS

CHAPTER 1: BACKGROUND OF THE ALASKA SCIENCE ASSESSMENT.....	1
CHAPTER 2: TEST DESIGN AND ITEM DEVELOPMENT	2
Science Assessment Measures.....	2
Multiple-Choice Items	2
Constructed-Response Items	2
2008 Operational Plan.....	2
Grade-Level Expectations Subsumed within Reporting Categories	4
Test Development Timeline	5
Item and Test Development Process	6
Item Content Test Blueprint and Item Specifications	7
Proficiency Level Descriptor Development.....	7
Item Writer Training	8
Item Writing	9
Item Content Review.....	11
Bias and Sensitivity Review.....	12
2006 Pilot Testing	12
2007 Field-Test Plan	13
Standalone Field Test Construction.....	14
Construction of the Field Test Forms.....	15
Item Field Test Data Review.....	15
Psychometric Guidelines for Selecting Items.....	17
Proportion Correct.....	17
Average Person Logit.....	17
Item-Total Correlation	17
Fit Statistic	18
Differential Item Functioning (DIF) Analyses.....	18
Item Bank	18
Overview	18
Functionality	19
Item Cards and Reporting Options.....	19
Security	19
Quality Assurance	19
Item Bank Summary	20

Final Selection of Items and Spring 2008 SBA Operational Forms Construction	21
Steps in the Forms Construction Process	21
Construction of the Operational Forms	22
DRC Internal Review of the Items and Forms	22
CHAPTER 3: TEST ADMINISTRATION PROCEDURES	23
Overview	23
Student Population Tested.....	23
Accommodations.....	23
Test Administrator Training.....	24
Test Security.....	24
Materials.....	24
Packaging and Shipping Materials	25
Materials Return.....	25
Box Receipt.....	25
CHAPTER 4: SCORING & STUDENT PERFORMANCE REPORTING	26
Document Processing	26
Handscoring of Constructed Responses	26
Readers	26
Range-finding and Developing Training Material	26
Training the Readers	27
Imaging.....	27
Quality Control of Handscoring.....	28
Reporting.....	28
District Reports	29
State Reports	29
CHAPTER 5: FORM ANALYSIS & ITEM CALIBRATION	30
Rasch Measurement Models.....	30
Item Statistics.....	31
Form Statistics	33
Frequency Distributions	36
Items.....	36
Persons	37

Cautions for Score Use	37
CHAPTER 6: SCALING & EQUATING	38
Introduction	38
Operational Item Calibration	38
Item Bank Maintenance	39
CHAPTER 7: FIELD-TEST ITEM DATA SUMMARY	40
Field-Test Items	40
CHAPTER 8: SCALE SCORES & PERFORMANCE LEVELS	41
Overview	41
Description of Scores	41
Raw Score	41
Scale Score	41
Comparability of Scale Scores Across Grades.....	41
Science SBA Standard Setting	42
Modified Bookmark Procedure.....	42
Standard Setting Panels—Alaska Teachers and Stakeholders.....	42
Recommended Cut Scores	43
Transformations	43
Scale Score Summary Statistics	44
Proficiency Levels	46
CHAPTER 9: TEST VALIDITY & RELIABILITY	49
Introduction	49
Validity	49
Content/Curricular.....	49
Construct Validity	50
Validity Evidence for Different Student Populations	51
Reliability	52
Standard Error of Measurement.....	53
REFERENCES	55

APPENDIX 1: SPRING 2008 TEST BLUEPRINTS	1-1
APPENDIX 2: PROFICIENCY LEVEL DESCRIPTORS.....	2-1
Grade 4 Science.....	2-1
Grade 8 Science.....	2-2
Grade 10 Science.....	2-3
APPENDIX 3: ITEM WRITER ORIENTATION MANUAL.....	3-1
APPENDIX 4: DEPTH OF KNOWLEDGE LEVELS	4-1
Level 1.....	4-1
Level 2.....	4-1
Level 3.....	4-1
Level 4.....	4-2
APPENDIX 5: UNIVERSALLY DESIGNED ASSESSMENTS.....	5-1
Elements of Universally Designed Assessments.....	5-1
Guidelines for Universally Designed Items	5-3
APPENDIX 6: ITEM REVIEW TRACKING FORMS.....	6-1
Content Review Form	6-1
Data Review Form.....	6-2
APPENDIX 7: CONFIDENTIALITY AGREEMENTS.....	7-1
APPENDIX 8: FAIRNESS IN TESTING MANUAL.....	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 4	13-1
Grade 8	13-2
Grade 10	13-4
APPENDIX 14: OPERATIONAL TEST ITEM AND THRESHOLD DIFFICULTY MAPS	14-1
Grade 4	14-1
Grade 8	14-2
Grade 10	14-3
APPENDIX 15: RAW-TO-SCALE SCORE TABLES.....	15-1
APPENDIX 16: STANDARD SETTING AGENDA	16-1
APPENDIX 17: STANDARD SETTING ROUND BY ROUND SUMMARY.....	17-1
Grade 4	17-1
Grade 8	17-2
Grade 10	17-3
APPENDIX 18: SUBSCALE SCORE SUMMARY STATISTICS	18-1
Grade 4 Subscale Reporting Categories.....	18-1
Grade 8 Subscale Reporting Categories.....	18-1
Grade 10 Subscale Reporting Categories.....	18-1
APPENDIX 19: TOTAL SCORE AND SUBSCALE SCORE INTERCORRELATIONS ..	19-1
Grade 4 Subscale Reporting Categories.....	19-1
Grade 8 Subscale Reporting Categories.....	19-1
Grade 10 Subscale Reporting Categories.....	19-1
APPENDIX 20: OPERATIONAL TEST RELIABILITY BY SUBPOPULATIONS	20-1

CHAPTER 1: BACKGROUND OF THE ALASKA SCIENCE ASSESSMENT

The Science Standards Based Assessment (SBA) is a criterion based assessment, and is aligned to the Alaska academic science content standards, which include Grade Level Expectations (GLEs) for each grade. The Science SBA was first administered operationally in April 2008 to students in grades four, eight, and ten. Assessment items were extensively reviewed by Alaska educators and subsequently field tested in three standalone field tests (i.e., grades 4, 8, and 10) administered in April 2007.

Alaska academic content standards for science were developed for each grade, grades 3–11, and are aligned with the National Science Education Standards. The assessments in grades 4, 8, and 10 focus on standards designated for those grades (e.g., the grade 4 test is not cumulative for grades 3 and 4).

NCLB required science standards be developed by the 2005–2006 school year and science be assessed in one grade span of 3–5, 6–9, and 10–12, by no later than the 2007–2008 school year. In response, Alaska’s Commissioner of Education and Early Development (EED), after 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 science in grades 4, 8, and 10 (the Standards Based Assessment).

The Alaska Science SBA is a coherent set of assessments aligned with Alaska GLEs developed for students in grades four, eight, and ten. The core set of assessments consists of custom assessments in science in grades four, eight and ten which 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 AND ITEM DEVELOPMENT

SCIENCE ASSESSMENT MEASURES

The science component of the Standards Based Assessment (SBA) is composed of items that address GLEs in grades 4, 8, and 10. The assessable GLEs for each grade level are distributed among four reporting categories. Information about the reporting categories and the GLEs assessed in each reporting category, as well as the types and numbers of items used in each reporting 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 science GLEs. These item types are designed to measure students' knowledge at various cognitive levels and provide a variety of information about science 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, or invalid application of a concept.

Constructed-Response Items

The science constructed-response (CR) items are designed to link science process with content. These items address comprehension of knowledge and skills at higher cognitive levels in ways that MC items cannot. They offer the opportunity for students to create a response to meaningful situations aligned to the assessable GLEs. Students must read the items carefully, analyze information, and, when required, offer explanations. These items provide insight into the students' science knowledge, abilities, and reasoning processes.

There are two types of SBA science assessment CR items: short constructed-response (SCR) and extended constructed-response (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.

2008 OPERATIONAL PLAN

The 2008 grades 4, 8, and 10 SBAs in science were comprised of a single form at each grade level. Place holder items were also included at each grade. These items were not scored and did not count toward a student's score. Their purpose was to hold positions for future field testing without having to change the length of the test in the future.

Table 2–1 displays the design for the science test for each grade. The column entries for this table denote:

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

Table 2–1. Science Test Plan 2008 Operational Form

Grade	Multiple-Choice Items		Core SCR Items (2 pt.)	Core ECR Items (4 pt.)	Place Holder CRs (2 pt. or 4 pt.)	Total Items MC/CR	Total Operational Points
	Core	Place Holder					
4	46	4	2	0	1	50/3	50
8	52	4	3	1	1	56/5	62
10	52	4	2	2	1	56/5	64

An individual student’s score is based solely on the core items. The total number of operational points is 50 points at grade 4, 62 points at grade 8, and 64 points 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 Science = **Grade 4:** 46 MC items plus two 2-point SCR items = 50 points
Grade 8: 52 MC items plus three 2-point SCR items plus one 4-point ECR item = 62 points
Grade 10: 52 MC items plus two 2-point SCR items plus two 4-point ECR items = 64 points

GRADE-LEVEL EXPECTATIONS SUBSUMED WITHIN REPORTING CATEGORIES

The science content area standard categories (or strands) are subdivided for specificity and eligible content or limits. The Alaska Science GLEs are organized into the seven standard categories shown in Table 2–2. The GLEs identified specifically for local assessment purposes are not included.

Table 2–2. Distribution of Science GLEs by Standard Category for Grades 4, 8, and 10

Distribution of Science GLEs		Grade Level		
		4	8	10
STANDARD CATEGORIES	A - Science as Inquiry and Process	2	2	2
	B - Concepts of Physical Science	2	4	7
	C - Concepts of Life Science	6	6	6
	D - Concepts of Earth Science	6	3	8
	E - Science and Technology	3	0	1
	F - Cultural, Social, Personal Perspectives, and Science	0	0	1
	G - History and Nature of Science	2	1	4
TOTAL		21	16	29

As outlined in Table 2–3, the seven standard categories are organized into four reporting categories.

Table 2–3. Reporting Categories with Corresponding GLEs for Grades 4, 8, and 10

Grade Level Distribution by Reporting Category		Grade Level		
		4	8	10
		Number of GLEs per Reporting Category		
A, E–G	Nature of Science and Technology	7	3	8
B	Concepts of Physical Science	2	4	7
C	Concepts of Life Science	6	6	6
D	Concepts of Earth Science	6	3	8
TOTAL		21	16	29

TEST DEVELOPMENT TIMELINE

A series of major test development activities took place in 2005, 2006, 2007, and 2008 which culminated in the administration of the operational assessment in April 2008. These key activities included:

- Development of the Alaska GLEs.
- Development of target content test blueprints and item specifications for each grade.
- Development of proficiency level descriptors.
- Development of items and tasks for pilot testing.
- Review of proficiency level descriptors and sample items by external committees of educators (content review, bias/sensitivity review).
- Pilot Testing of a sample of items.
- Generate and analyze statistics for pilot test items.
- Development of items and tasks for field testing.
- Review of items by external committees of educators (content review, bias/sensitivity review).
- Field testing of new Science items in a standalone field test in April 2007.
- Rangefinding for the tasks.
- Review of items by external committees of educators (item review with data).
- Final selection of items used to construct the 2008 SBAs.

Table 2–4 provides a high-level timeline of these major activities, which are described in detail in this report.

Table 2–4. General Timeline Associated with 2006 Pilot Testing, 2007 Field Testing, and 2008 Operational Assessment of Science at Grades 4, 8, and 10

Time Frame	Activity
	Development/approval of Alaska GLEs
September 2005	Development of item content blueprint (test design and item specifications)
October 2005	Item writing workshop for Alaska Educators
October 2005	Development of preliminary proficiency level descriptors with Alaska educators
November–December 2005	Item development for the 2006 pilot test
February 2006	Pilot Test
February–August 2006	Item development for 2007 spring field test
September 2006	Content committee review of newly developed items for 2007 field test
September 2006	Bias/sensitivity committee review of newly developed items for 2007 field test
December 2006	Forms construction of spring 2007 standalone field test
April 2007	2007 field test assessment administration grades 4, 8, and 10
April 2007	Rangefinding
September 2007	Data review committee for items field tested in spring 2007
October–December 2007	Forms construction of spring 2008 operational administration
April 2008	2008 operational assessment administration grades 4, 8, and 10

ITEM AND TEST DEVELOPMENT PROCESS

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 were major considerations in the item and test development process. 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 develop a preliminary test blueprint, including the standards to be assessed per grade arranged by the relative importance of the content to be assessed.

- Develop preliminary 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 items for review by committees of Alaska educators (content and bias/sensitivity).
- Select and assemble items for pilot testing/field testing.
- Pilot test/field test items, scoring of the items, and analysis of the data.
- Review items and associated statistics after pilot/field testing, including bias statistics.
- Select and assemble items for operational forms (test construction).

Item Content Test Blueprint and Item Specifications

The first step in the item and test development process involved the development of the content test blueprint and item specifications. Each test blueprint included a list of all the standards or GLEs to be assessed, organized by content domains or reporting categories. The preliminary test blueprint also contained the number of test items to be developed per learning target or GLE, an indication of the relative importance of the content to be assessed, and detailed information concerning the general characteristics of the items to be developed for each content standard or GLE. The preliminary content test blueprint for each grade level was provided to EED and committees of Alaska educators for review and feedback. DRC test development specialists revised each content test blueprint, with item specifications, based upon Alaska educators' and EED feedback. The content test blueprint, with item specifications, was then used to guide the item and test development process.

Proficiency Level Descriptor Development

The SBA science assessment proficiency level descriptors (Appendix 2) are statements that describe the knowledge and skills expected at four proficiency levels with respect to the content standards, performance standards, and GLEs. The proficiency level descriptors describe the expected level of performance at each of these four levels: Far Below Proficient, Below Proficient, Proficient, and Advanced.

The proficiency level descriptors were initially developed by a group of 45 Alaska science educators and DRC's science test development specialists in October 2005. The Alaska educators were trained by DRC test development specialists to develop proficiency descriptors for the levels. Training included background information about the grade-level expectations, purpose, and use of the proficiency level descriptors, and the process that would be used to develop the descriptors. Following the training, each grade level was divided into subcommittees. The subcommittees of Alaska educators developed proficiency level descriptors using the Alaska GLEs and the existing Benchmark Proficiency Level Descriptors. The

committees then met as one cohesive group to review their final draft proficiency level descriptors and revised them to clearly

- define what students in Alaska should know and be able to do based on the grade-level expectations,
- show how students would demonstrate knowledge and skill based on the grade-level expectations, and
- define the level of knowledge and skill necessary for each proficiency level.

The SBA science assessment proficiency level descriptors at grades 4, 8, and 10 can be found in Appendix 2. These were used to guide the standard-setting process and were finalized after standard-setting.

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 specialized knowledge and skill in the subject area of their expertise (e.g., biology, physical science). All writers met the following qualifications:

- A bachelor's degree or higher in science, curriculum and instruction, and/or related field.
- In-depth understanding and knowledge of the special considerations involving the writing of standards-based 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 0–4 point) items, including the writing of scoring rubrics for each item.

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 SBA, the cadre of item writers was trained using the DRC Item Writer Orientation Manual (Appendix 3) with regard to:

- Alaska content 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 approved by the EED.

Alaska Educators' Item Writer Training

In order to involve Alaska educators in the item-writing process, a group of 30 science educators attended a three-day item writer workgroup in Fairbanks in October 2005. They were trained to write test items in their subject area of expertise using the training described above. Items written by Alaska educators during the meeting were edited by DRC test development specialists and content editors and were included in the pool of items reviewed by Alaska educators during the content review meeting in September 2006. These items were also included in the pool of items reviewed by the bias/sensitivity committee.

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 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 pilot testing and field testing would include a range of difficulty (low, medium, and high).

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.

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, 2006). Items were classified as one of four depth of knowledge categories: recall, skill/concept, strategic thinking, and extended thinking (Appendix 4).

Readability: For science item development, writers were instructed to pay careful attention to the readability of each item to ensure that the focus was upon the concepts; not on reading comprehension. As a result, the goal for each writer was to write items that were, to the greatest degree possible, independent of the assessment of reading. Science contains 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 test question was taken before committees comprised of Alaska grade-level experts in the field of science education. They reviewed each question from the perspective of the students they teach, and they determined the grade-level appropriateness of the vocabulary used.

Curriculum-specific Issues: All items were to be curriculum independent with respect to both science content and vocabulary. In other words, items were not developed to align with any one particular science textbook series. 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 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 (Appendix 5). 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 is clear and concise and provides 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 field of science education.
- be free of gender, ethnic, cultural, socioeconomic, and regional stereotyping bias.

Item Content Review

Prior to pilot testing and field testing, all newly developed test items were submitted to content committees for review. The content committees consisted of Alaska educators from school districts throughout Alaska. The primary responsibility of the content committees 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 and sensitivity.

The content review was held September 28 and 29, 2006. Committee members were selected by EED, and EED-approved invitations were sent by DRC. The content review committee consisted of 45 educators. The meeting commenced with an overview of the test development process. Training was 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 (Appendix 6) to be used during the item content review.

DRC test development specialists facilitated the review of items. Committee members, grouped by grade level, reviewed the items for quality and content, as well as for the following categories designated on the item review tracking form.

- 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, or Rewrite)

Security was addressed by committee members and facilitators 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 7). 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 pilot testing and field testing, all newly developed test items were also submitted to a Bias and Sensitivity Committee for review. This took place on September 28 and 29, 2006. The committee's primary responsibility was to evaluate 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 individuals 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 8). This manual was revised specifically for the Alaska program.

All 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 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.

2006 Pilot Testing

A small pilot test of items was conducted in April of 2006. The purpose of the pilot was to try out a sample of items during the initial item development phase. This process was also used to try out the different types of multiple-choice and constructed-response items as related to the SBAs. This design allowed a representative number of Science GLEs to be tested by a small sample population. The inclusion of a pilot test for Alaska Science benefited the program in the following ways.

- Guided item development efforts for field test items
- Helped train item review committee members
- Assisted in the development of Proficiency Level Descriptors

Items available for the no-stakes pilot testing were developed in the summer and fall of 2005. These items were reviewed at the beginning of the Proficiency Level Descriptors meeting in 2005. The approved items were then assembled into forms – 1 per grade.

2007 FIELD-TEST PLAN

The 2007 grades 4, 8 and 10 standalone field tests in science were composed of four forms at each grade level. Each field test form contained unique field test items and was constructed to mirror the planned 2008 and 2009 operational test design.

Table 2–5 displays the design for the science field tests at grades 4, 8, and 10.

Table 2–5. Science Field Test Plan 2007

GRADE 4		GLEs	Total Items per Form		Pt. Totals	Total No. of Items	
			MC	SR			
REPORTING CATEGORIES	A, E–G – Inquiry, Technology, and Nature of Science	7	18	2	22	20	
	B – Concepts of Physical Science	2	8	0	8	8	
	C – Concepts of Life Science	6	10	0	10	10	
	D – Concepts of Earth Science	6	10	0	10	10	
Total		21			50	48	
GRADE 8		GLEs	Total Items per Form			Pt. Totals	Total No. of Items
			MC	SR	ER		
REPORTING CATEGORIES	A, G – Inquiry and Nature of Science	3	16	0	1	20	17
	B – Concepts of Physical Science	4	12	1	0	14	13
	C – Concepts of Life Science	6	14	1	0	16	15
	D – Concepts of Earth Science	5	10	1	0	12	11
Total		18				62	56

Table 2–5 (continued). Science Field Test Plan 2007

GRADE 10		GLEs	Total Items per Form			Pt. Totals	Total No. of Items
			MC	SR	ER		
REPORTING CATEGORIES	A, E–G – Inquiry, Technology, Society, and Nature of Science	8	16	0	1	20	17
	B – Concepts of Physical Science	7	12	1	0	14	13
	C – Concepts of Life Science	6	12	1	0	14	13
	D – Concepts of Earth Science	8	12	0	1	16	13
Total		29				64	56

STANDALONE FIELD TEST CONSTRUCTION

In order to construct operational forms for 2008 and 2009, a standalone field test was administered on April 2–16, 2007. For the field test administration, the science test was one test book. The book contained multiple-choice items and short and extended constructed-response items with response space. The test books in grade 10 contained a periodic table found at the end of the test book.

Field test forms construction took place during the period from September 2006, through January 2007. In an effort to maximize the item pool for the first operational SBA test, DRC Test Development, under the approval from EED, developed completely unique forms.

Table 2–6 displays the total item counts across each grade on the four field test forms.

Table 2–6. Science Field Test Total Item Counts

Grade 4: Four Unique Field-Test Forms

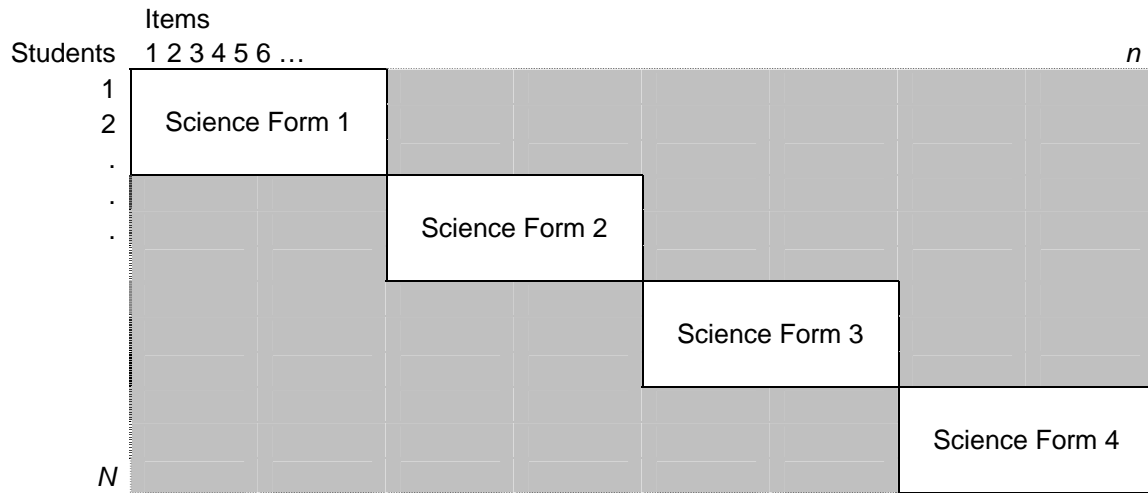
	MC	CR	Total Item Count
Number of Field Test Items per Form	46	2	48
Total Number of Field Test Items	184	8	192

Grade 8 and 10: Four Unique Field-Test Forms

	MC	CR	Total Item Count
Number of Field Test Items per Form	52	4	56
Total Number of Field Test Items	208	16	224

Figure 2–7 displays the design for the science field tests for grades 4, 8, and 10.

Figure 2–7. Science Forms Design



Construction of the Field Test Forms

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

Guidelines for Placing Items into Field Test Forms

- Forms included an adequate objective coverage, as required by the detailed test blueprint.
- No item in a form “clued” another item on that same form.
- “Clang” was avoided (i.e., distractors were unique from one another).
- Forms were ethnically diverse, both in terms of artwork and in terms of names.
- Forms targeted an equal representation of genders, both in terms of artwork and names.
- Forms included a wide range of topics and a variety of questions.
- Correct answer distributions were distributed such that approximately 25 percent of them were A, B, C, or D.

Item Field Test Data Review

A data review meeting for all items field tested during the Spring 2007 administration was held on September 27–28, 2007, in Anchorage, AK. 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 pages 17–18.

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 September 27–28, 2007 and included content committees composed of 45 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 7). 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 September 2007 Data Review are shown in Table 2–8.

Table 2–8. Science Items at Data Review

September 2007 Data Review

Grade	Accept	Accept with Revisions	Accept Total	% Accept	Reject	Total
4	15	0	15	65%	8	23
8	17	0	17	77%	5	22
10	95	0	95	87%	14	109

PSYCHOMETRIC GUIDELINES FOR SELECTING ITEMS

Proportion Correct

The proportion correct, or p -value, is the proportion of the total group of test takers answering the item 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 0.63 to 0.67, 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 answer the item incorrectly. 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 than the value 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

were included if such items were 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 that quantifies the sum of the squared standardized residual of the observed item performance from the expected performance for all persons, based on the Rasch model, is computed for each item. 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 potentially 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 in test construction. 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. Items with C DIF 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 2008 SBA forms were built is presented in Table 2–9.

Table 2–9. SBA Science Items

Grade 4

Standard	MC	CR
Science as Inquiry and Process	27	4
Concepts of Physical Science	25	0
Concepts of Life Science	32	0
Concepts of Earth Science	31	0
Science and Technology	20	0
Cultural, Social, Personal Perspectives, and Science	NA*	NA*
History and Nature of Science	11	0

Grade 8

Standard	MC	CR
Science as Inquiry and Process	38	3
Concepts of Physical Science	38	3
Concepts of Life Science	50	3
Concepts of Earth Science	34	2
Science and Technology	NA*	NA*
Cultural, Social, Personal Perspectives, and Science	NA*	NA*
History and Nature of Science	12	0

Grade 10

Standard	MC	CR
Science as Inquiry and Process	22	4
Concepts of Physical Science	34	2
Concepts of Life Science	45	2
Concepts of Earth Science	36	3
Science and Technology	4	0
Cultural, Social, Personal Perspectives, and Science	8	1
History and Nature of Science	19	0

*NA—these standards are locally assessed

FINAL SELECTION OF ITEMS AND SPRING 2008 SBA OPERATIONAL FORMS CONSTRUCTION

The test forms for the spring 2008 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 same guidelines used in the construction of the field test forms (page 15), including the following.

Additional Guidelines for Placing Items into Forms

- Overall form will be within the target p -value range of 0.63–0.67.

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 2008 science SBA was administered to students in grades four, eight, and ten during the spring of 2008. A District Test Coordinator was assigned to each school district. The test administration window was March 31 through April 14, 2008. Specific statewide testing days were not designated. 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 November 12 through December 11, 2007. Districts also submitted their precode files January 7–22, 2008. 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 22, 2008. 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 22, 2008 with precode and district/school labels arriving in these districts by March 13, 2008.

The enrollment and documents processed counts were as follows:

Table 3–1. Project Counts

District Count	School Count
54	486
Enrollment Count	Processed Count
Grade 4: 9,838	Grade 4: 9,199
Grade 8: 10,428	Grade 8: 9,603
Grade 10: 10,974	Grade 10: 9,781

ACCOMMODATIONS

Appropriate accommodations were available for students with disabilities while taking the science assessment. These accommodations were 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/ParticipationGuidelinesSept2007Final.pdf).

TEST ADMINISTRATOR TRAINING

DTCs were trained in February 2008 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 2008.

TEST SECURITY

The science 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 photocopied or any portion of the test materials duplicated at any time. Except for the person testing, no person, including test administrators, was permitted to read test items on the science SBA prior to, during (except for the student testing), or after administration. Teachers, proctors, test administrators, or any testing personnel could 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 could not read test items under any circumstances.

The DTC designated the school and district personnel who had access to secure test materials, and who needed to sign the Test Security Agreements. All signed test security forms were returned to the DTC and kept on file in the district.

Prior to the first test administration of the school year, DTCs signed and sent 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 A Science Test Books—*grades 4, 8, and 10*
- Large Print Test Books
- Braille Test Books
- Ancillary materials—periodic table for grade 10 only, 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 3, 2008, 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 2, 2008. 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 3–25, 2008. DRC processed 6 additional materials requests for this administration.

Materials Return

Districts returned all materials via Assessment Distribution Services on April 17, 2008, and most materials arrived at DRC’s warehouse on April 21, 2008. All districts used pink 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 Assessment Distribution Services. 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 1, 2008. 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 (short constructed-responses) and ECR (extended constructed-response) items.

Preliminary rubrics for field test items were written during the item development stage, and these rubrics were refined once live student responses were 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. 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.

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 would be 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.

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. 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.

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 Grades 4, 8, and 10 science SBA reports were provided electronically on August 12, 2008. Paper copies of the final Grades 4, 8, and 10 science SBA reports were delivered to the districts as scheduled by August 21, 2008.

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 science 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 grade level 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 ability 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 used in the true-score model. 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 fourth-grade science test or a high-school mathematics test, relate to an arbitrary origin defined by the average of item difficulties for that form. The average fourth-grade science item will have a logit of zero; the average high-school mathematics item will have a logit of zero in unanchored calibrations. This common logit scale describes both item difficulties and student abilities.

Because both dichotomous and polytomous items were part of the science SBAs, 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_{ij} 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 (δ_{ij}) 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 3.65 (Linacre, 2008). 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 grade level for the spring 2008 science 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 grade level.

The logit column in the table and appendix provides the item difficulty. 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. Even if some minor item dependence existed in the CR item formats, they were 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 difference 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 distractors. Content specialists examined items with large fit statistics and confirmed that each item had only one correct answer and was correctly written.

The *p*-value for an MC item is the percent (or proportion) 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 2008 science SBA forms, the range of CR item scores is from 0–2 or 0–4 points.

The item-total correlation (PtBis or Corr.) provides a measure of 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 science SBA overall) would be expected to answer items correctly, and students with low proficiency (i.e., those that perform poorly on the science SBA overall) to answer items incorrectly. 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 0.30 and verified that each item was acceptable as written and scored. As seen in Table 5–1, the median item-total correlations for MC and CR items all exceeded the 0.30 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. The non-scorable codes are recoded as 0 points during item calibration.

Table 5–1. Summary of Operational Item Analysis

Grade	Item Type	Mean Logit	Mean SE	Mean Fit	Mean <i>p</i>-value	Median PtBis or Corr	Mean Omits
4	MC	0.00	0.03	-0.86	0.66	0.43	0.00
	CR	1.31	0.02	4.65	0.45	0.45	0.02
8	MC	0.00	0.03	-0.72	0.65	0.41	0.00
	CR	2.31	0.02	0.02	0.37	0.52	0.06
10	MC	0.00	0.03	-0.44	0.66	0.39	0.00
	CR	2.65	0.02	0.03	0.23	0.45	0.12

FORM STATISTICS

Tables 5–2 through 5–7 contain summary descriptive statistics for student performance and item difficulty, including mean score, standard deviation, and minimum and maximum scores by grade level. These statistics were generated using WINSTEPS v3.65 (Linacre, 2008) 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 half 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 half 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 & 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:

$$MSE_I = \sum_{i=1}^L \left[\frac{s_i^2}{L} \right] ,$$

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.

Table 5–2. Grade 4—Summary of 8,918 Measured Students

	Raw Score	Count	Measure	Model Error
Mean	32.0	48.0	0.96	0.38
SD	10.3	0.0	1.30	0.13
Max.	50	48	5.87	1.85
Min	3	48	-3.03	0.31
	RMSE	Adjusted SD	Person Separation	Person Separation Reliability
Real	0.42	1.23	2.97	0.90
Model	0.40	1.24	3.06	0.90

SE of Student Measure Mean = 0.01

Student Raw Score-to-Measure Correlation = 0.97

Coefficient Alpha Student Raw Score Reliability = 0.92

Table 5–3. Grade 4—Summary of 48 Measured Items

	Raw Score	Count	Measure	Model Error
Mean	5951.1	8918.0	0.05	0.03
SD	1241.1	0.0	0.84	0.00
Max.	8866	8918	1.65	0.04
Min	3733	8918	-1.70	0.02
	RMSE	Adjusted SD	Item Separation	Item Separation Reliability
Real	0.03	0.84	31.31	1.00
Model	0.03	0.84	31.91	1.00

SE of Item Measure Mean = 0.12

Table 5–4. Grade 8—Summary of 9,244 Measured Students

	Raw Score	Count	Measure	Model Error
Mean	37.6	56.0	0.85	0.33
SD	12.0	0.0	1.14	0.08
Max.	60	56	5.50	1.23
Min	1	56	-4.24	0.27
	RMSE	Adjusted SD	Person Separation	Person Separation Reliability
Real	0.35	1.09	3.15	0.91
Model	0.33	1.09	3.26	0.91

SE of Student Measure Mean = 0.01

Student Raw Score-to-Measure Correlation = 0.98

Coefficient Alpha Student Raw Score Reliability = 0.92

Table 5–5. Grade 8—Summary of 56 Measured Items

	Raw Score	Count	Measure	Model Error
Mean	6204.0	9244.0	0.17	0.02
SD	1695.1	0.0	1.01	0.00
Max.	14673	9244	4.96	0.04
Min	3022	9244	-1.86	0.01
	RMSE	Adjusted SD	Item Separation	Item Separation Reliability
Real	0.03	0.01	40.00	1.00
Model	0.02	0.01	40.59	1.00

SE of Item Measure Mean = 0.14

Table 5–6. Grade 10—Summary of 8,950 Measured Students

	Raw Score	Count	Measure	Model Error
Mean	36.8	56.0	0.85	0.31
SD	11.7	0.0	1.08	0.05
Max.	62	56	5.32	1.02
Min	1	56	-4.18	0.28
	RMSE	Adjusted SD	Person Separation	Person Separation Reliability
Real	0.33	1.03	3.10	0.91
Model	0.32	1.03	3.25	0.91

SE of Student Measure Mean = 0.01

Student Raw Score-to-Measure Correlation = 0.99

Coefficient Alpha Student Raw Score Reliability = 0.92

Table 5–7. Grade 10—Summary of 56 Measured Items

	Raw Score	Count	Measure	Model Error
Mean	5881.6	8950.0	0.19	0.02
SD	1361.7	.0	0.96	0.00
Max.	11701	8950	3.68	0.04
Min	2757	8950	-1.64	0.01
	RMSE	Adjusted SD	Item Separation	Item Separation Reliability
Real	0.03	0.96	37.96	1.00
Model	0.02	0.96	38.54	1.00

SE of Item Measure Mean = 0.13

FREQUENCY DISTRIBUTIONS

Items

Appendix 14 provides frequency distributions of all science 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 14–1 of the appendix, the logit value for grade 4 Item 16 is between 0.5 and 0.7, and it is also lower than the logit value for Item 29, which is located on the same line. In addition, each CR item sequence number is displayed to the right of its corresponding logit for each possible threshold.

Persons

Appendix 15 provides frequency distributions of raw scores and scale scores by grade for the spring 2008 science 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 science 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 grade eight science SBA 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 SBAs have been fixed so they do not change across administrations.

CHAPTER 6: SCALING & EQUATING

INTRODUCTION

To maintain the same passing standard across different administrations, EED, in association with testing vendors, 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.

The spring 2008 science SBAs were new forms developed from the spring 2007 field-test administration. The 2008 administration was used to define the logit metric for future equating and the standard score scale for reporting.

In addition to the operational items, DRC embedded placeholder items in order to allow for field-test items to be placed in the same sequence in future tests. These items do not count toward an individual student's score.

The spring 2008 operational science SBAs were the first forms developed to meet new NCLB requirements. Each assessment became the baseline calibration to which future science SBA forms will be equated.

OPERATIONAL ITEM CALIBRATION

For each grade level of the spring 2008 science SBA administration, a separate calibration was performed with WINSTEPS (Linacre, 2008). Initially, all dichotomous item difficulties were estimated. These values were then 'anchored' in a second WINSTEPS run to determine the polytomous item threshold values. This anchored calibration method places all items and thresholds on the same scale. Because this was the first operational administration of these three tests and the standard had not been set or approved, the initial multiple-choice (MC) calibrations were unanchored. The two calibrations described above provided the information necessary to develop the ordered item booklets for the May 2008 Science SBA Standard Setting.

The WINSTEPS raw score logit conversion tables from the combined MC/CR calibration and the distribution of raw scores for the total test were used to provide impact data at the May 2008 science SBA standard setting.

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, k 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 change.

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

There were no field-test items for this test administration; therefore, no field-test item data summary is provided.

CHAPTER 8: SCALE SCORES & PERFORMANCE 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 grades have a theoretical 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 taking the science SBA. The raw score is the number of multiple-choice (MC) items answered correctly plus the number of points earned on constructed-response (CR) items. By itself, the raw score has limited utility; it can only be interpreted in reference to the total number of items on an assessment, and raw scores should not be compared across reporting categories or administrations.

Scale Score

Since 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 comparison across all test administrations within a particular grade and content area.

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 grade 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. As seen in the science SBAs, 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.

As a result, the interpretation of scale scores is the same for all grades 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 0.67 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.

SCIENCE SBA STANDARD SETTING

Traditionally, standard-setting methods have fallen into two types: test-centered methods and examinee-centered methods (Jaeger, 1989). With test-centered methods, the standard-setting judgment is made primarily by referencing the test itself, most often based upon an inspection of the actual test items. Examinee-centered methods call for judgments to be made about the performance of examinees, such as the performance of established mastery and non-mastery groups. As Kane (1995) points out, “all standard setting is based on judgments.” Proficiency standard setting uses a process from which the best judgments are obtained from the people in the best position to make those judgments. Typically, these are content experts, people familiar with the skills and knowledge to be learned.

EED, based on recommendation of the Technical Advisory Committee, used a test-centered method for all standard settings required by the NCLB legislation. Of the many test-centered methods, EED has employed the Bookmark Method (Lewis, Mitzel, Green, & Patz, 1999) for previous NCLB standard settings.

Modified Bookmark Procedure

Because there was a need to maintain consistency with the mathematics, reading, and writing SBAs, the same Modified Bookmark Procedure (Lewis, Mitzel, Green, & Patz, 1999) was utilized to set the cut scores. A schedule for the May 2008 Standard Setting is provided in Appendix 16.

Standard Setting Panels—Alaska Teachers and Stakeholders

Representatives of two broad groups typically comprise the standard setting panel. These groups include teachers and non-teacher educators. Examples of non-teacher educators are curriculum directors or administrators. The panel member selection process also considered size and location of districts and schools, socioeconomic conditions and other demographics.

While there is no established number of panelists required to establish reliable cut scores, the larger the number involved, the more confidence there can be that the result will generalize to another selection of panelists. The current plan sought 15 panelists for each grade level for a total

of 45. EED successfully recruited 45 participants (15 panelists per grade); however only 42 panelists were able to participate in the standard setting meeting. Separate panels were used for grades 4, 8, and 10.

Recommended Cut Scores

Appendix 17 shows the round by round recommended cut score summaries for each grade.

TRANSFORMATIONS

As previously mentioned, raw scores were transformed into logits in the initial calibration. Logits in turn were mathematically transformed into scale scores to provide a more convenient metric for reporting. To maintain consistency from administration to administration, the minimum scale score necessary for proficiency was set at 300 for each grade. The minimum scale scores necessary for each proficiency level are provided in Table 8–1. Table 8–2 provides the equations and minimum logits used for each transformation. These equations were applied to the overall test as well as to each reporting subscale. Refer to Appendix 14 to locate the logit cut scores compared to item difficulties for each grade.

Table 8–1. Science 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
4	23	35	42	233	18	300	19	357	24
8	30	39	48	258	18	300	19	359	22
10	25	35	47	245	20	300	20	369	22

Table 8–2. Transformation Equation

Grade	Conversion Equation	Logit Cut Scores		
		FBP/BP	BP/P	P/A
4	Scale Score = (57.6923 x Logit) + 241.0692	-0.1375	1.0128	2.0086
8	Scale Score = (65.7895 x Logit) + 249.5526	0.1338	0.7592	1.6675
10	Scale Score = (69.4444 x Logit) + 257.9306	-0.1807	0.5986	1.5922

Complete raw-to-scale score tables are provided in Appendix 15.

SCALE SCORE SUMMARY STATISTICS

Table 8–3 includes scale score descriptive information for each overall grade level. Subscale descriptive statistics can be found in Appendix 18. Histograms of the overall test scale scores are also provided in Figures 8–1 to 8–3.

Table 8–3. Content Area Scale Score Information

	Grade 4 (n=8918)	Grade 8 (n=9244)	Grade 10 (n=8950)
Mean	296.42	305.75	316.96
Standard Error of Mean	0.79	0.78	0.79
Median	292	305	318
Mode	371	369	342; 355
Standard Deviation	74.97	74.81	75.03

Figure 8–1. Grade 4 Scale Score Frequencies

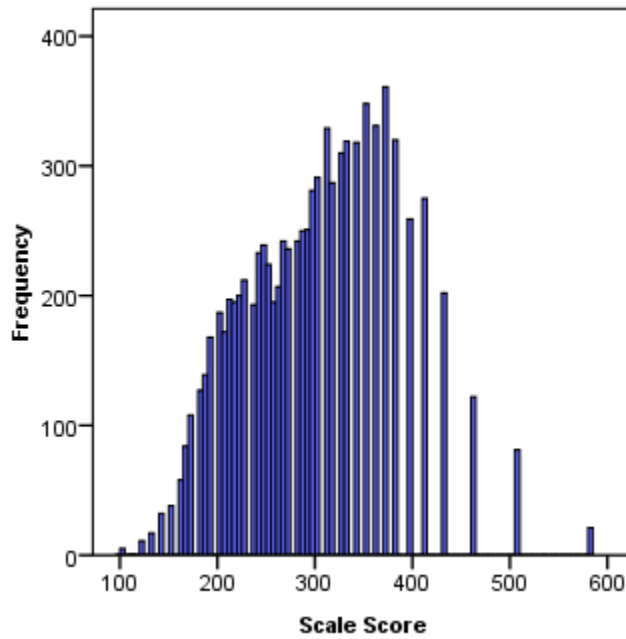


Figure 8–2. Grade 8 Scale Score Frequencies

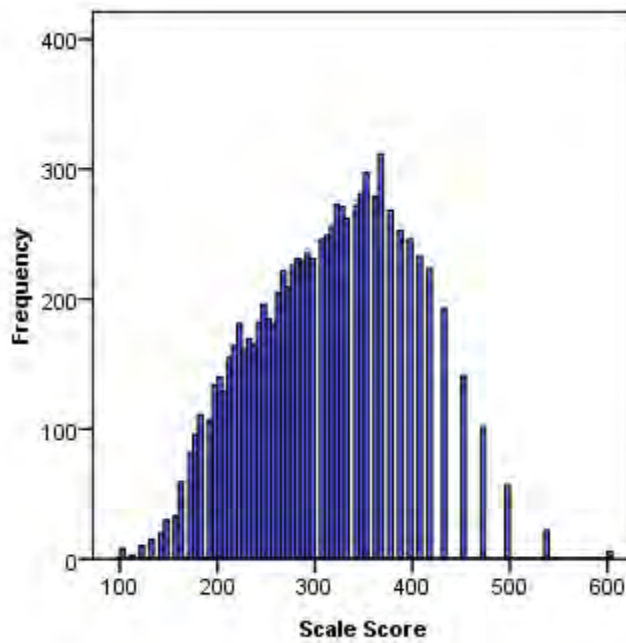
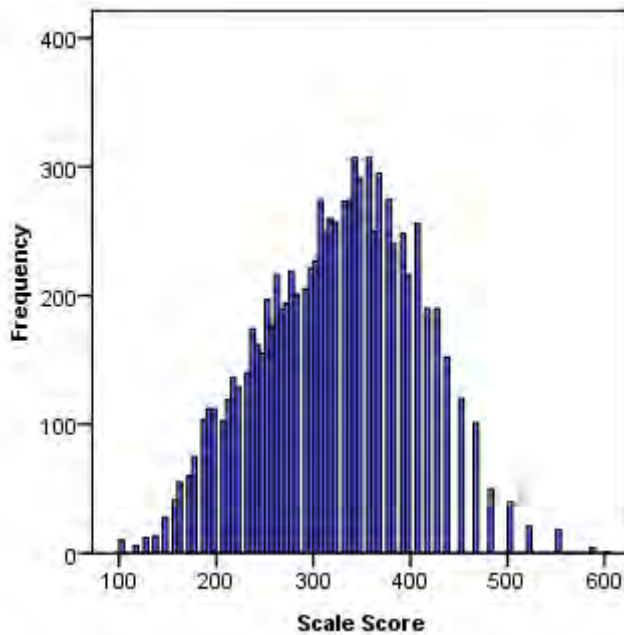


Figure 8–3. Grade 10 Scale Score Frequencies



PROFICIENCY LEVELS

Information from the science SBAs 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 (FBP), Below Proficient (BP), Proficient (P), and Advanced (A).

Scale score cutpoints at each level of proficiency are the same each year. Appendix 2 provides detailed information about the proficiency level as well as the Proficiency Level Definitions and Descriptors in each grade and content area tested.

Table 8–4 provides the distribution of students in each of the proficiency levels for all grades.

Table 8–4. Student Distribution of the Four Proficiency Levels

	Grade 4		Grade 8		Grade 10	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
Far Below Proficient	1951	21.88	2535	27.42	1591	17.78
Below Proficient	2793	31.32	1969	21.30	1973	22.04
Proficient	2202	24.69	2406	26.03	3264	36.47
Advanced	1972	22.11	2334	25.25	2122	23.71

Indicators of Consistency

Criterion-referenced tests are often used to place the examinees into two or more performance classifications. It is then 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 8–5 depicts the general framework of multiple decisions.

Table 8–5. 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 8–6.

Table 8–6. Decision Consistency Indices

	Huynh (1976)				Livingston and Lewis (1995)	
	4 categories (FBP, BP, P, A)		2 categories (Not Proficient, Proficient)		4 categories (FBP, BP, P, A)	2 categories (Not Proficient, Proficient)
Grade	Consistency Index	κ	Consistency Index	κ	Consistency Index	Consistency Index
4	0.69	0.58	0.87	0.74	0.75	0.89
8	0.68	0.58	0.87	0.74	0.69	0.89
10	0.68	0.57	0.87	0.72	0.71	0.89

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 2008 science SBAs 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 a necessary, but not sufficient, condition for making valid interpretations of the results.

VALIDITY

Content/Curricular

The science SBA is a criterion-referenced assessment. This assessment is based on an extensive definition of the content it assesses. Therefore, the science SBA is 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 science SBA, 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 science SBA. These review committees participate in this process to ensure test content validity of the science SBA.

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 science 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 can later be embedded in operational science 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 science SBAs.

Educator Input

For the spring 2008 science 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 science SBA. 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 science SBA. See Chapter 2 for details regarding the content review process.

Developer Input

For the items included in the spring 2008 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 science SBAs.

Item to Content Area Match

Expert judgments from educators, test developers, and assessment specialists provide support for the alignment of the science SBAs with the statewide content standards. In addition, because expert teachers in the science content area 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 science 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 science SBA test blueprint provides this documentation. The blueprints for science 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, & NCME, 1999) provides the following list of possible sources:

- High intercorrelations 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 2008 science 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 2008 science SBAs can be found in Appendix 13. The majority of items have item-total correlations of at least 0.30 (90% of items). 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 2008 SBAs can be found in Appendix 13. In this administration, 79% of item fit statistics are below +5.00, indicating good construct validity.

Intercorrelations

A third indicator of construct validity is the intercorrelations between the grade-level total scale scores and the subscale reporting category scale scores. This information is contained in Appendix 19.

Validity Evidence for Different Student Populations

The primary evidence for the validity of the science SBAs lies in the content and construct being measured. Because the test assesses the statewide content standards required to be taught to all students, the test is not more or less valid for use with one sub-population of students over another sub-population. In other words, because the science SBA is measuring what is required to be taught to all students and is given under the same standardized conditions to all students, the validity of score interpretations should apply to all students. Table 9–1 presents the student demographic information for the science SBAs.

Table 9–1. Summary of Student Demographics

Demographics	Grade 4		Grade 8		Grade 10	
	<i>N</i>	%	<i>N</i>	%	<i>N</i>	%
ALL STUDENTS	8918	100.00	9244	100.00	8950	100.00
ETHNICITY						
White (Caucasian)	4623	51.84	5138	55.58	5039	56.30
African American	326	3.66	318	3.44	299	3.34
Hispanic	559	6.27	493	5.33	451	5.04
Asian/Pacific Islander/Native Hawaiian	632	7.09	683	7.39	640	7.15
Alaska Native and American Indian	2134	23.93	2075	22.45	2000	22.35
Two or more races	624	7.00	517	5.59	469	5.24
Unknown	20	0.22	20	0.22	52	0.58
SOCIOECONOMIC STATUS						
Not Low Income	4924	55.21	5765	62.36	6148	68.69
Low Income	3994	44.79	3479	37.64	2802	31.31
ENGLISH PROFICIENCY STATUS						
English Proficient	7757	86.98	8117	87.81	7935	88.66
Limited English Proficient	1161	13.02	1127	12.19	1015	11.34
MIGRANT STATUS						
Non-Migrant	8340	93.52	8653	93.61	8364	93.45
Migrant	578	6.48	591	6.39	586	6.55
SPECIAL EDUCATION STATUS						
Regular Education	7568	84.86	8177	88.46	8116	90.68
Individualized Education Plan	1350	15.14	1067	11.54	834	9.32
GENDER						
Female	4377	49.08	4513	48.82	4454	49.77
Male	4541	50.92	4731	51.18	4496	50.23
ACCOMMODATIONS						
Total	6	100.00	1	100.00	4	100.00
Braille	1	16.67	0	0.00	0	0.00
Large Print	5	83.33	1	100.00	4	100.00

Great care has been taken to ensure that the items comprising the science 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

True-score theory 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 2008 administration of the science SBAs was 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. Tables 5–2 through 5–7 provide Coefficient Alpha for each grade. As can be seen in the tables, the grades 4, 8, and 10 forms have Coefficient Alphas of 0.92, 0.92, and 0.92 respectively. These high α values provide evidence for good reliability. Appendix 20 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:

$$SEM = \sigma_X \sqrt{1 - \rho_{xx}}, \tag{5}$$

where, σ_x is the standard deviation of the total test (observed measure scores), and ρ_{xx} is the reliability estimate (Coefficient Alpha) 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., raw 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 science SBA scale to obtain the final SEM for each raw score. These values for the spring 2008 science SBAs are provided in the raw-to-scale score tables in Appendix 15. In addition, person separation reliability and item separation reliability values, which use these asymptotic standard errors, are provided in Tables 5–2 through 5–7. Person separation reliability is the Rasch equivalence of reliability described in Equation 2.

REFERENCES

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (2nd ed.). Washington, DC: American Educational Research Association.
- Bloom, B. S., Englehart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of Educational Objectives: The classification of educational goals: Handbook 1: Cognitive Domain*. New York: Longmans, Green, and Co.
- Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.
- Holland, P., & Thayer, D. (1986, April). *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.
- Lewis, L., Mitzel, H., Green, D. & Patz, R. (1999). *The Bookmark Standard Setting Procedure*. Monterey, CA: McGraw-Hill Companies.
- Linacre, J. M. (2008). WINSTEPS Rasch measurement (Version 3.65). [Computer program]. Chicago: WINSTEPS.com.
- Linn, R., & Gronlund, N. (1995). *Measurement in assessment and teaching* (7th ed.). Upper Saddle River, NJ: Prentice-Hall.
- Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, *32*, 179–197.
- Mead, R. J. (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., & Brisner, E. P. (1989). *EDL Core Vocabularies in Reading, Mathematics, Science, and Social Studies*. Orlando, FL: Steck-Vaughn Company.

Thompson, S., Johnston, C. J., & 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, DC: Council of Chief State School Officers.

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

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