



Alaska

Comprehensive System of Student Assessment

Technical Report

**Spring 2006
High School Graduation
Qualifying Examination (HSGQE)
and Retest**



August 2006

TABLE OF CONTENTS

CHAPTER 1: BACKGROUND OF ALASKA ASSESSMENTS.....	1
CHAPTER 2: TEST DESIGN & ITEM DEVELOPMENT	2
Item Development for Current Administration	2
2006 HSGQE Operational Plan, with Appended Field Test Items.....	2
Test Development Timeline	6
Item and Test Development Process	7
Item Writer Training	7
Reading Passage Selection.....	8
Passage Readability.....	9
Item Writing.....	9
Item Content Review.....	12
Bias and Sensitivity Review.....	13
Item Field Test	13
Item Field Test Data Review.....	13
Item Bank.....	15
Overview	15
Functionality	15
Item Cards and Reporting Options.....	16
Security	16
Quality Assurance	16
Item Bank Summary	16
Psychometric Guidelines for Selecting Items.....	17
Proportion Correct (also known as <i>p</i> -value)	17
Average Person Logit.....	17
Item-Total Correlation	18
Fit Statistic	18
Differential Item Functioning (DIF Analyses).....	18
Final Selection of Items and Spring 2006 HSGQE Operational Forms Construction.....	19
Steps in the Forms Construction Process.....	19
Construction of the Operational Forms.....	20
DRC Internal Review of the Items and Forms	20

CHAPTER 3: TEST ADMINISTRATION PROCEDURES	21
Student Population Tested.....	21
Spiraling Plan.....	21
Testing Schedule	21
Materials.....	22
Packaging and Shipping Materials.....	22
Materials Feedback	23
Materials Return.....	23
Box Receipt.....	23
CHAPTER 4: SCORING & STUDENT PERFORMANCE REPORTING.....	24
Document Processing	24
Handscoring of Constructed-Response Items.....	24
Readers.....	24
Rangefinding and Developing Training Material	25
Training the Readers	25
Imaging.....	25
Quality Control of Handscoring.....	26
Data Processing.....	27
Report Mockups.....	27
Reporting.....	28
District Reports	28
State Reports	28
CHAPTER 5: FORM ANALYSIS & ITEM CALIBRATION.....	29
Introduction	29
Rasch Measurement Models.....	29
Item Statistics.....	31
Form Statistics	32
Frequency Distributions	33
Items.....	33
Persons	33
Cautions for Score Use.....	33

CHAPTER 6: SCALING & EQUATING.....	34
Introduction	34
Pre-Equating.....	34
Operational Item Calibration.....	34
CHAPTER 7: FIELD TEST ITEM DATA SUMMARY.....	36
Field Test Items.....	36
Field Test Item Descriptive Statistics	36
Item Bank Maintenance.....	39
CHAPTER 8: SCALE SCORES & PROFICIENCY LEVELS.....	40
Rationale.....	40
Description of Scores	40
Raw Score	40
Scale Score	40
Transformations	41
Scale Score Summary Statistics	42
Proficiency Levels.....	44
CHAPTER 9: TEST VALIDITY & RELIABILITY.....	47
Introduction	47
Validity	47
Content/Curricular.....	47
Construct Validity	48
Criterion-Related Validity.....	49
Validity Evidence for Different Student Populations	49
Reliability	50
Standard Error of Measurement	51
REFERENCES	52

APPENDIX 1: SPRING 2006 HSGQE TEST BLUEPRINT.....1

APPENDIX 2: RUBRICS1

6-Point Extended Constructed-Response (ECR) Scoring Rubric for Writing1

6 Points.....1

5 Points.....1

4 Points.....1

3 Points.....2

2 Points.....2

1 Point2

4-Point Extended Constructed-Response (ECR) Scoring Rubric for Grades 10/10+ Writing.....3

4 Points.....3

3 Points.....3

2 Points.....3

1 Point3

APPENDIX 3: DRC ITEM WRITER ORIENTATION MANUAL1

APPENDIX 4: FAIRNESS IN TESTING MANUAL.....1

APPENDIX 5: DEPTH OF KNOWLEDGE LEVELS.....1

Mathematics.....1

Level 1.....1

Level 2.....1

Level 3.....2

Level 4.....2

Reading.....3

Level 1.....3

Level 2.....3

Level 3.....3

Level 4.....4

Writing.....5

Level 1.....5

Level 2.....5

Level 3.....5

Level 4.....6

Source of Challenge Criterion6

APPENDIX 6: UNIVERSALLY DESIGNED ASSESSMENTS.....	1
Elements of Universally Designed Assessments.....	1
Guidelines for Universally Designed Items	3
APPENDIX 7: ITEM REVIEW TRACKING FORMS	1
Content Review Form	1
Data Review Form.....	2
APPENDIX 8: CONFIDENTIALITY AGREEMENT	1
APPENDIX 9: BIAS & SENSITIVITY REVIEW FORM.....	1
APPENDIX 10: SAMPLES OF MANUALS	1
APPENDIX 11: INTER-RATER RELIABILITY	1
APPENDIX 12: SAMPLES OF GUIDES TO TEST INTERPRETATION	1
APPENDIX 13: OPERATIONAL TEST ITEM ANALYSIS.....	1
Mathematics.....	1
Reading.....	3
Writing.....	5
APPENDIX 14: FORM STATISTICS	1
Mathematics.....	1
Reading.....	2
Writing.....	3
APPENDIX 15: OPERATIONAL TEST ITEM AND THRESHOLD DIFFICULTY MAPS ...	1
Mathematics.....	1
Reading.....	2
Writing.....	3

APPENDIX 16: RAW-TO-SCALE SCORE TABLES.....	1
Mathematics.....	1
Reading.....	5
Writing.....	9
APPENDIX 17: FIELD TEST ITEM ANALYSIS.....	1
Mathematics.....	1
Reading.....	7
Writing.....	11
APPENDIX 18: FIELD TEST DIFFERENTIAL ITEM FUNCTIONING (DIF) CLASSIFICATION RULES	1
Dichotomous (Multiple-Choice) DIF Classification	1
Polytomous (Constructed-Response) DIF Classification	1
APPENDIX 19: FIELD TEST DIFFERENTIAL ITEM FUNCTIONING (DIF) SUMMARY BY FORM.....	1
Mathematics.....	1
Reading.....	2
Writing.....	3
APPENDIX 20: SUBSCALE SCORE SUMMARY STATISTICS.....	1
APPENDIX 21: HSGQE PROFICIENCY DESCRIPTORS OF THE MINIMUM COMPETENCIES IN ESSENTIAL SKILLS.....	1
Mathematics.....	1
Reading.....	3
Writing.....	5

CHAPTER 1: BACKGROUND OF ALASKA ASSESSMENTS

The Alaska High School Graduation Qualifying Examination (HSGQE) was developed to determine student competency in the areas of mathematics, reading, and writing. The HSGQE provides this information in the form of test scores that reflect the essential skills that students should know as a result of their public school experience. The requirement to pass the HSGQE in order to earn a high school diploma has been in effect since 2004.

CHAPTER 2: TEST DESIGN & ITEM DEVELOPMENT

ITEM DEVELOPMENT FOR CURRENT ADMINISTRATION

This section of the technical report covers the period of time from March 2005 through the field testing and subsequent data review of the HSGQE items that were operational in spring 2006. Coverage includes the item and test development process, item reviews, field testing, item data analysis, and the selection of items comprising the spring 2006 HSGQE form.

In order to construct the spring 2006 HSGQE form, Data Recognition Corporation (DRC) developed test items and conducted an appended field test of those items in spring 2005. The items were appended onto the spring 2005 CTB/McGraw-Hill-leased operational form. For field test purposes, HSGQE mathematics, reading, and writing items were written to assess the high school performance standards.

2006 HSGQE OPERATIONAL PLAN, WITH APPENDED FIELD TEST ITEMS

The 2006 HSGQE mathematics, reading, and writing test was comprised of 14 forms. (Note: A Form 15 was printed for retest use only. It did not include field test items.) The spring 2006 HSGQE Test Blueprint is in Appendix 1. All of the forms contained the core items identical for all students. In addition, each form also included a set of field test items that were appended on each of the 14 forms. The appended field test items fulfilled several purposes. These purposes included the development of

- Enough items to build a fall 2006 retest HSGQE operational form, designed to measure the performance standards.
- Enough items to build a spring 2007 HSGQE operational form, designed to measure the performance standards and the grade 10 GLEs.
- Enough additional or replacement items to build the item samplers/practice tests, should EED request.
- Other items based upon future needs as determined by EED.

Table 2–1 displays the design for the mathematics test for forms 1 through 14. The column entries for this table denote:

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

**Table 2–1. Mathematics Test Plan 2006 per Operational Form
(14 Forms)**

Grade	Multiple-Choice Items		Core SCR Items (2 pt.)	Core ECR Items (4 pt.)	FT CRs (2 pt. or 4 pt)	Total Items MC/CR	Total Operational Points
	Core	FT					
10	44	8–9*	3	0	0–1	52–53/ 3–4	50

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

Table 2–2 displays the design for the reading test for forms 1 through 14. The column entries for this table denote:

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

Table 2–2. Reading Test Plan 2006 per Operational Form (14 Forms)

Grade	Multiple-Choice Items		Core SCR Items (2 pt. or 3 pt.)	Core ECR Items (4 pt.)	FT CRs (2 pt. or 4 pt.)	Total Items MC/CR	Total Operational Points
	Core	FT					
10	56	10	4	0	1	66/5	65

Table 2–3 displays the design for the writing test for forms 1 through 14. The column entries for this chart denote:

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

**Table 2–3. Writing Test Plan 2006 per Operational Form
(14 Forms)**

Grade	Multiple-Choice Items		Core SCR Items (2 pt.)	Core ECR Items (4 pt. or 6 pt.)*	FT SCRs (2 pt.)	FT ECRs (4 pt.)	Total Items MC/CR	Total Operational Points
	Core	FT						
10	26	11	2	4	1	1	39/8	66

* These are double-weighted when determining the total points possible.

Since an individual student’s score is based solely on the core items, the total number of operational points is 50 points for mathematics, 65 points for reading, and 66 points for writing. The total score is obtained by combining the points from the core MC and CR (SCR and ECR) portions of the test as follows:

Student’s Score in Mathematics = 44 MC items plus three 2-point SCR items = 50 points

Student’s Score in Reading = 56 MC items plus three 2-point SCR items plus one 3-point SCR item = 65 points

Student’s Score in Writing = 26 MC items plus two 2-point SCR items plus three double-weighted 4-point ECR items plus one double-weighted 6-point ECR item = 66 points

TEST DEVELOPMENT TIMELINE

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

- Development of items, tasks, and writing prompts.
- Review of items by external committees of educators (content review, bias/sensitivity review).
- Field testing of new mathematics, reading, and writing items in an appended field test in April 2005.
- Review of items by external committees of educators (item review with data).
- Final selection of items used to construct the spring 2006 HSGQE.

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 2005 Field Testing and 2006 Operational Assessment.

Time Frame	Activity
April 2005	Administration of 2005 assessment with field test items
August 2005	Content committee review of field tested items; review for statistical quality
August 2005	Content committee review of newly developed items for 2006 field test
August 2005	Bias/sensitivity committee review of newly developed items for 2006 field test
December 2005	Forms construction of spring 2006 operational test
April 4–6, 2006	2006 operational assessment administration

ITEM AND TEST DEVELOPMENT PROCESS

Aligning the items to the performance standards; 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 the *Principles of Universal Design* (Thompson, Johnstone, & Thurlow, 2002) guided the following steps in the item and test development process:

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

Item Writer Training

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

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

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

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

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

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

Reading Passage Selection

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

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

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

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

Passage Readability

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

Item Writing

To ensure that all test items met the requirements of the approved 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 Performance Standards: There must be a high degree of match between a particular question and the standard it is intended to measure. Item writers were asked to clearly indicate what standard each item was measuring.

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

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

Multiple-choice Item Options and Distractor Rationale/Analysis: Writers were instructed to make sure that each item had only one clearly correct answer. Item writers submitted the answer key with the item. All distractors were plausible choices that represented common errors and misconceptions in student reasoning. The rationale/distractor analysis for each distractor for mathematics was also provided.

Constructed-Response: 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 Bloom et.al.'s *Taxonomy of Educational Objectives, Handbook I: The Cognitive Domain* (1956). As each item was written, the writer classified one of four cognition levels: recall, application, analysis, or evaluation for each item. The writers were instructed to write items so that the pool of items would represent a distribution of items across cognitive levels, as required by the test and item specifications.

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

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

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

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

Editorial Review of Items

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

Guidelines for Reviewing Assessment Items

A good item should

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

Item Content Review

Prior to the 2006 field testing, all newly developed test items were submitted to content committees for review. The content committees consisted of Alaska teachers and subject-area supervisors from school districts throughout Alaska. The primary responsibility of the content committee was to evaluate items with regard to quality and content classification, including grade-level appropriateness, estimated difficulty, depth of knowledge, and source of challenge. They also suggested revisions and made recommendations for reclassification of items to different grade levels, if appropriate. The committee also reviewed the items for adherence to the principles of universal design, including language demand and issues of bias, fairness, and sensitivity. At the culmination of the item content review, all items that were presented to the committee for review were either accepted as presented or were revised.

The content review was held August 1–4, 2005. Committee members were selected by EED, and EED-approved invitations were sent to them by DRC. The committee consisted of 20 educators, seven for the reading and mathematics content areas, and six for the writing content area. EED also selected internal staff members for attendance. The meeting commenced with an overview of the test development process. Training was also provided by DRC senior staff members. Training included how to review items for technical quality and content quality, including depth of knowledge and adherence to principles of universal design. In addition, training provided committee members with the procedures for item review, including the use of tracking review forms to be used during the item content review.

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

1. Performance Standard Alignment
2. Difficulty Level (classified as Easy, Medium, or High)
3. Depth of Knowledge (classified as Recall, Application, Strategic Thinking)
4. Correct Answer
5. Quality of Graphics
6. Appropriate Language Demand
7. Freedom from Bias (classified as Yes or No)
8. Overall Judgment (classified as Approved, Accept with Revisions, Move to another grade level, or Rewrite)

Security was addressed by adhering to a strict set of procedures. Items in binders did not leave the meeting rooms and were accounted for at the end of each day before attendees were dismissed. All attendees, with the exception of EED staff, were required to sign a Confidentiality Agreement (Appendix 8). All materials not in use at any time were kept in secure meeting rooms. During lunch and breaks, if meeting rooms were unused, they were locked or closely

monitored by DRC personnel. While not in use by DRC, the meeting rooms were locked and unavailable to anyone other than one DRC person and the Chief of Security of the meeting facility. Rooms were attended to only under strict supervision by DRC personnel. Secure materials that did not need to be retained after the meeting were deposited in secure barrels, and their contents were shredded under supervision of a DRC employee.

Bias and Sensitivity Review

Prior to field testing, all newly developed test items were also submitted to a Bias and Sensitivity Committee for review. This took place on August 1–2, 2005. The committee’s primary responsibility was to evaluate passages and items as to acceptability with regard to bias and sensitivity issues. They also made recommendations for changes or deletion of items in order to remove the area of concern. At the culmination of the bias and sensitivity review, all items that were presented to the committee for review were either accepted as presented or were revised prior to the field test.

The bias/sensitivity committee was composed of 10 men and women who represented the diversity of Alaska students, one of whom had special education expertise. The committee was trained by a DRC test development director to review items for bias and sensitivity issues using a generic Fairness in Testing Manual developed by DRC. This manual was subsequently revised specifically for the Alaska program.

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

Item Field Test

Items being field tested were appended to forms 1–14 for the spring 2006 administrations.

Item Field Test Data Review

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

- Proportion selecting 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 well have caused a problem and/or contributed to the DIF. DRC test development specialists then determined which of the flagged items were reviewed by a group of Alaska educators to determine whether or not the item was appropriate for use. In most 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, regarded as statistically acceptable. Likewise, items of extremely poor statistical quality were easily regarded as unacceptable and needed no further review. However, there were some items that DRC test development specialists 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 August 1–4, 2005 and included content committees composed from 20 Alaska educators. EED also selected internal staff members for attendance. 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 reasons that 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, instructional issues, etc.) and a decision regarding acceptance. DRC test development specialists facilitated the statistical review of the items.

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

For the items that were taken to Data Review in August of 2005, Table 2–5 shows the results.

Table 2–5. Items at Data Review

August 2005 Data Review

Subject	Grade	Accept	Accept with Revisions	Accept Total	% Accept	Reject	Total
Reading	10	154	0	154	92%	14	168
Mathematics	10	161	0	161	88%	23	184
Writing	10	56	0	56	79%	15	71

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 or rubric, 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

Many of the viewing options in the item bank are based on read-only privileges. 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 is presented in Table 2–6. The item summary table for each content area shows eligible items after the fall 2005 HSGQE Retest was built. Items doubled coded to both a HSGQE performance standard and a SBA Grade Level Expectation will appear in both Item Bank Summary tables in this document and the 2006 SBA Technical Report.

Table 2–6. Eligible HSGQE Items

Mathematics Items

Standard	MC	CR
M1	27	0
M2	24	2
M3	56	2
M4	20	4
M5	19	4
M6	22	3

Reading Items

Standard	MC	CR
R4.1	30	0
R4.2	75	7
R4.3	68	3
R4.4	7	3
R4.7	21	4
R4.8	12	1

Writing Items

Standard	MC	CR
W4.1/2	63	26
W4.3	76	0
W4.4	52	3

PSYCHOMETRIC GUIDELINES FOR SELECTING ITEMS

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

The proportion correct or *p*-value is the proportion of the total group of test takers answering the question correctly. The proportion for an item will show how difficult the item was for the students who took that field test form. In general, multiple-choice 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 sparingly. The overall forms are constructed to a target mean of 0.65.

Average Person Logit

The average person logit is the average measure for the persons selecting that response. The total average person logit is the average ability of the persons attempting that item, which can vary from field test form to field test form. The average person logit for the correct response should be greater than every other response. The difference between the average person logit for the correct responses and the incorrect responses is an indication of the discrimination of the item. The larger the difference, the more discriminating the item. The 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 (correct (1) or incorrect (0)) 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 answered the item incorrectly. Item discrimination is an important statistic in the forms construction process, and the higher the average value the more reliable the test. Items with item-total correlations of 0.35 or greater are 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. However, items with point-biserial values between 0.20 and 0.35 for a standards-based assessment are considered only if the inclusion of such items was necessary to satisfy specific content cells of the detailed test blueprint.

Fit Statistic

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

Differential Item Functioning (DIF Analyses)

Fairness in testing is an important consideration. Several methods were employed during the development of items to ensure that items function similarly for all groups of students. Some of the methods are qualitative in nature and are part of the development process itself. Several statistical methods can quantify the functioning of an item. DIF analysis is conducted on all field test items to determine whether an item favors one group of students over another.

DIF procedures examine the possibility that an item's characteristics may negatively affect the performance of select groups of students. Although the terms item bias and DIF are often used interchangeably, DIF does not necessarily imply unfairness. Rather, 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. The judgment of fairness is based on whether or not the difference in difficulty is believed to be related to the construct being measured, which is directly dependent on the purpose for which an assessment is being used.

DRC utilized 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 multiple-choice 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 constructed response items with more than two score categories.

FINAL SELECTION OF ITEMS AND SPRING 2006 HSGQE OPERATIONAL FORMS CONSTRUCTION

The spring 2006 HSGQE in mathematics, reading, and writing was comprised of 15 forms each containing the same set of core items. The test forms for the spring 2006 HSGQE 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 the utmost care and precision, and all forms reflected 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 blueprint, including the number of items per domain or reporting category for each content-area test.
2. DRC psychometricians provided DRC test development specialists with an overview of 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.
9. Operational forms were prepared for psychometric review and approval.
10. Based upon psychometric review, suggested replacements were made by DRC test development specialists if necessary.
11. Operational forms were prepared for EED review and approval.

Construction of the Operational Forms

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

Guidelines for Placing Items into Forms

- Forms will include an adequate objective coverage, as required by the detailed test blueprint.
- No item in a form will “clue” another item on that same form.
- “Clang” will be avoided (i.e., distractors should be unique from one another).
- Forms will be ethnically diverse, both in terms of artwork and in terms of names.
- Forms will target an equal representation of genders, both in terms of artwork and names.
- Forms will include a wide range of topics and a variety of questions.
- Correct answer distributions will follow guidelines (app. 25% A, 25% B, 25% C, and 25% D).
- Overall form will be within the target proportion range of .63–.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

STUDENT POPULATION TESTED

Districts submitted their enrollment and accommodated materials counts, and updates to district contact information via DRC's Online Enrollment System January 9–24, 2006. Districts also submitted their precode files January 9–31, 2006. Districts with 30 or more schools and 9000 or more students were given the option to submit their enrollment files directly to DRC by January 24, 2006. Mat-Su, Anchorage, Fairbanks, and Kenai 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 by February 24, 2006, with precode and district/school labels arriving in these districts by March 14, 2006. Several districts did not submit enrollment counts and/or precode files by the deadline. DRC receive estimated enrollment counts and precode information for these districts from EED.

The enrollment and documents processed counts were as follows:

Table 3–1. Project Counts

District Count	School Count
53	276
Enrollment Count	Processed Count
HSGQE: 11,064	HSGQE: 10,114
HSGQE Retest: 5,989	HSGQE Retest: 4,486

SPIRALING PLAN

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

TESTING SCHEDULE

The spring 2006 HSGQE & Retest was administered April 4–6, 2006. The reading test was administered on April 4, the writing test on April 5, and the mathematics test on April 6.

MATERIALS

The following materials were produced for this administration:

- *District Test Coordinator's Manual*
- *Test Administration Directions*
- Form D Reading Test Books – 14 versions
- Form D Writing/Mathematics Test Books – 14 versions
- HSGQE Retest Test Books – 1 version
- Large Print Test Books
- Braille Test Books
- HSGQE audiotapes for writing and mathematics
- HSGQE Retest audiotapes for reading, writing, and mathematics
- Ancillary materials – rulers, protractors, large print and Braille rulers, large print and Braille protractors, precode labels, district/school labels, “Do Not Score” labels, return shipping labels, return materials instruction packets, 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 6, 2006 as scheduled.

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

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

DRC provided EED with a Point of Delivery Report on March 20, 2006. 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 6 through April 3, 2006. DRC processed 33 additional materials requests for this administration.

Materials Feedback

DRC did not send rulers and protractors for the HSGQE Retest. Anchorage alerted DRC to this on April 5, the night before the mathematics exam. On April 6, DRC discovered that protractors were not required for any of the Retest items and rulers were only required for one sample item. DRC identified those districts that may have been short on rulers and protractors. DRC and EED made calls to those DTCs to let them know that rulers and protractors were not needed for the Retest.

Materials Return

Districts returned all materials via Manna Distribution Services on April 11 and materials began arriving at DRC's warehouse on April 13. All materials arrived at DRC by April 17 with the exception of Iditarod. Iditarod's material did not arrive at DRC until April 28, 2006. The district did not drop it off with the 3rd party carrier until April 26—fifteen days after the deadline. DRC was not able to include results for the grade 12/12+ students with the grade 12/12+ reports for the rest of the state.

Districts were instructed to place an orange Grade 12/12+ HSGQE Express label on all boxes containing grade 12/12+ documents. All districts used orange DRC return shipping labels. DRC return shipping labels were district specific and included a line for district test coordinators to indicate how many boxes they were returning to DRC.

Box Receipt

As materials arrived, DRC's Materials Processing team (MAT) checked the bill of lading to ensure that the number of boxes received matched the number signed for by the DTC and Manna Distribution Services. The Materials Processing team scanned each box using the Operations Materials Management System (OpsMMS) box receipt system and notified EPM of any schools 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 and schools that did not return materials.

CHAPTER 4: SCORING & STUDENT PERFORMANCE REPORTING

DOCUMENT PROCESSING

All secure materials were scanned 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. There were sixteen missing materials from ten districts for this administration.

DRC used its Image Scanning System to scan the HSGQE test books. Scanning of test books was completed on April 26. All editing and validating rules were followed.

Several challenges were encountered during the scanning phase.

- About 50% of the writing skills checklists and math reference sheets were left in the test books. DRC's Document Processing team (DP) had to remove them before scanning.
- Several students put the tear-out pages back in the book in a different spot.

Despite these challenges, DRC was able to complete scanning on time with no impact on the schedule.

HANDSCORING OF CONSTRUCTED-RESPONSE ITEMS

For the Alaska assessments, DRC employed a variety of score-point scales for scoring short constructed-response (SCR) and extended constructed-response (ECR) items.

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

Readers

The scorers for the Alaska HSGQE were selected from DRC's larger pool of available professional test scorers. All of our readers for the Alaska HSGQE 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 are scoring. For example, mathematics scorer candidates must successfully solve a DRC mathematics problem and show all steps necessary to reach the correct answer. Reader candidates are asked to respond to a DRC writing topic.

Rangefinding and Developing Training Material

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

Training the Readers

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

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

After the scoring guide and all training sets were discussed, readers of operational (common) 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 all HSGQE responses.

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' 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 was 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.

All SCR operational (common) items were independently scored by two readers. If the scores were in exact agreement, that score was the "score of record." If the two scores were not in exact agreement, the response received another independent reading and all three scores were compared for an exact match which would stand as the score of record. This process continued with multiple independent reads until there were two scores in exact agreement. All ECR operational (common) items were also scored by two independent readers. If the scores were in exact agreement, that score stood as the score of record. If the scores were adjacent (e.g., a 3 and a 2), the higher score stood as the score of record. If the scores were non-adjacent (e.g., a 1 and a 3), the response was forwarded to an expert scorer for a third independent reading. If the third score was in exact or adjacent agreement with either of the first two scores, that score stood as the score of record. If all three scores were non-adjacent (e.g., 0, 2, and 4), the response was forwarded to a scoring supervisor for resolution scoring, which served as the score of record. The same scoring rules were applied to field test items.

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. The inter-rater reliability statistics are included in Appendix 11.

DATA PROCESSING

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

DRC provided EED with the student file so corrections and updates could be applied. After the demographic information was updated, the student file was scored against the appropriate answer key, indicating correct and incorrect responses. Correct responses were designated by converting the numeric value into an alpha value (e.g., 1 becomes A, 2 becomes B). Incorrect responses remained numeric. In addition, the original response string was stored for data verification and auditing purposes.

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

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

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 Report Delivery System and state-level data files and layouts to the FTP site.

Report Mockups

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 fonts 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 the Software Quality Assurance staff 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 any 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 sample reports was reviewed by EED prior to producing live reports for districts and schools.

REPORTING

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 the *Guides to Test Interpretation* are provided in Appendix 12 and are also available on EED's Web site.

Grade 12/12+ student reports were provided electronically as scheduled on April 26, 2006. All HSGQE reports were provided electronically as scheduled on May 12, 2006. All paper copies of HSGQE reports were delivered to districts as scheduled by May 19, 2006.

An erasure analysis will be delivered to EED by August 31, 2006.

District Reports

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

State Reports

- Student Data File
- Abbreviated Student Data File

CHAPTER 5: FORM ANALYSIS & ITEM CALIBRATION

INTRODUCTION

An assessment program, no matter how well planned, implemented, and executed, is of no value unless the results can be communicated quickly, accurately, and clearly. DRC strongly believes that the Rasch (1960) methodology provides the best solution for realizing these goals. DRC has conducted several studies to determine which model is most suitable for large-scale assessment. The results of the studies were concise and decisive: The Rasch methodology. . .

1. is a legally defensible model. In a widely publicized and carefully scrutinized legal case in Texas, the Rasch model was successfully defended. This precedent has not been set using more complex models.
2. is extremely efficient. Since the statistical analysis is less time consuming and less complex with the Rasch model, data turnaround time can be shortened.
3. has interpretation ease. There exists a one-to-one correspondence between Rasch-generated scores and number correct scores. This correspondence results in data that are easier to explain to parents, student, educators, and policy makers.

RASCH MEASUREMENT MODELS

Scale scores for the HSGQE were computed using the family of Rasch measurement models. The advantage of using Rasch models in scaling is that all of the items measuring performance in a particular content area can be placed on a common difficulty scale. With the Rasch family of measurement models, in contrast to more complex models, the number correct score is a sufficient statistic for estimating person ability. This allows the Rasch difficulty values for the individual items to be used in computing a Rasch ability level for any raw score point on any test constructed from these items, eliminating the need for pattern scoring.

Rather than percent correct, the Rasch model expresses item difficulty (and student ability) in units commonly referred to as logits. In the simplest case, a logit is a transformed p -value with the average p -value becoming 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 level 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 and item measures to determine the probability that a person will respond correctly to any test item. This comparison is not possible in the percent correct metric. It is impossible to predict how well a person who answered 80% of the items correctly will perform on an item answered correctly by 80% of the persons.

The standard Rasch calibration procedure arbitrarily sets the mean difficulty of the items on any form at zero. Any item with a p -value lower than the average item on the form receives a positive logit difficulty and any item with a p -value higher than the average receives a negative logit. Consequently, the logits for any calibration, whether it is a third grade reading test or a high school science test, relate to an arbitrary origin defined by the average of item difficulties

for that form. The average third grade reading item will have a logit of zero; the average high school science item will have a logit of zero. This logit scale applies to both item difficulties and student abilities.

Because both dichotomous and polytomous items were part of the HSGQE assessments, DRC utilized a mixed-model item calibration approach that placed both item types onto a common scale. MC items scored either right or wrong were calibrated using the familiar form of the dichotomous Rasch model. 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, a multiple-choice item may be thought of as a partial-credit task with only one threshold.

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

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

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

In this model $\pi_{ni0} + \pi_{ni1} < 1$ since more than two response categories are available, and δ_{i1} , while still the difficulty of the first threshold in item i , is one of several threshold difficulties for the item. Finally, as person n must make one of the four possible scores on item i ,

$$\pi_{ni0} + \pi_{ni1} + \pi_{ni2} + \pi_{ni3} = 1.$$

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,

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

It gives the probability of person n scoring x on the m_i threshold of item i as a function of the person's measure (β_n) and the threshold difficulties of the m_i thresholds, in item i . The observation x is a count of the successfully completed item thresholds.

The unconditional, joint maximum likelihood (UCON) estimation of item difficulties is accomplished with WINSTEPS (Linacre, 2006). This calibration software is commercially available and widely used in the testing industry. WINSTEPS is considered the industry standard for Rasch calibration.

In addition, to assess overall model fit to the data, the items were analyzed for scale comparability by examining the residuals between observed and expected scores for the persons and items (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).

ITEM STATISTICS

Appendix 13 provides item level statistics by subject area for the spring 2006 HSGQE assessments. These statistics represent the item characteristics most commonly used to determine whether an item functioned in an appropriate manner. The item mean for a multiple-choice item is synonymous with the item p -value. It is the percent of all students that responded to an item correctly. The p -value for constructed-response item represents the average score earned divided by the maximum number of points for that item. For the HSGQE spring forms, this score can range from 0 to 2 or 0 to 4 points in mathematics, 0 to 2, 0 to 3, or 0 to 4 in reading, and 0 to 2, 1 to 4, or 1 to 6 in writing. 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 nonscorable codes are recoded as 0 points during item calibration.

The Rasch fit statistics are used to determine how well items conform to the requirements of the Rasch measurement model. The major significance of a Rasch model is that the item difficulty parameter estimates are independent of the ability distribution (and all other characteristics of the examinees). 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 the cause of an item's misfit by re-examining the item and its distracters. For instance, if several high ability students miss an easy item, re-examination of the item may show that it actually has more than one correct response or that the students recorded the answers to two items in the same item grid.

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 items as a set. Typically, students with high ability (i.e., those that perform well on the HSGQE overall) would be expected to get items correct, and students with low ability (i.e., those that perform poorly on the HSGQE overall) to get items incorrect. If these expectations are met, the item-total correlation between the item and the total test score will be high and positive, indicating that the item is a good discriminator between high ability and low ability students. An item-total correlation value above 0.30 is 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.

FORM STATISTICS

Appendix 14 contains summary descriptive statistics for measured student ability and item difficulty, including minimum and maximum scores, mean score, and standard deviation by subject. These statistics were generated using WINSTEPS and illustrate student and item performance. The top half of the person summary tables provides descriptive statistics for persons measured. The column labeled “Measure” provides the mean and standard deviation of the estimated student ability measures. The column labeled “Model Error” presents the standard error associated with these values when the assumptions of the Rasch model hold.

The top half of the item summary tables provides the same descriptive statistics outlined above, with the exception that items are the unit of analysis rather than examinees. 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 standard error associated with the estimated mean and standard deviation values when the requirements of the Rasch model hold.

The bottom half of the tables contains Real Root Mean Square Error (RMSE), which corresponds to a worst case estimate, and Model RSME, that corresponds to a best case estimate. The reliability estimates included are a variation of Coefficient Alpha called the index of “person separation reliability,” a measure of score reliability. This is a refined measure of internal consistency reliability. This index, along with the adjusted standard error and separation estimates needed to compute it, is also provided in the tables. The adjusted standard deviation is an estimate of the “true” standard deviation, which adjusts for potential measurement error by removing it from the standard deviation estimate (Wright and Masters, 1982):

$$SA_i^2 = SD_i^2 - V(MSE_i).$$

The item/person separation (SEP) value then provides this adjusted standard deviation in RMSE units. It is calculated by finding the ratio of the adjusted standard deviation to the RMSE (i.e., SA/MSE). The person/item reliability is computed using:

$$R_i = \frac{SA_i^2}{SD_i^2}.$$

At the bottom of the tables, the “traditional” 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.

FREQUENCY DISTRIBUTIONS

Items

Appendix 15 provides frequency distributions of all HSGQE item difficulties, including the thresholds for constructed-response items.

Persons

Appendix 16 provides frequency distributions of raw scores and scale scores by subject area for the 2006 spring administration. The columns in these tables present each raw score, scale score, frequency count, frequency percent, cumulative frequency, and cumulative percent. The range of scale scores for the HSGQE is set to 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 HSGQE in reading is 600 and a student achieves this score, it cannot be determined whether the student would have achieved a higher scale score if that score were possible. All that is known is that the student's ability estimate, 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 HSGQE have been fixed so they do not change between administrations.

Analyses of scores of students at extreme ends of the distribution should also be undertaken cautiously because of a phenomenon known as regression toward the mean. It is more difficult for the students with very high or very low scores to maintain their score on subsequent testing than it is for the students in the middle of the distribution. If a student who scored 38 out of 40 on a test were to take the same test again, there would be 38 opportunities to incorrectly answer an item that had been correct. There would only be two opportunities to correctly answer items that were missed the first time. If an item is answered differently, it is more likely to decrease the student's score than to increase it. The converse of this is also true for students with very low scores; the next time they test they are more likely to achieve a higher score, and this higher score may be a result of regression toward the mean rather than an actual gain in achievement. Regression toward the mean is a phenomenon apparent with all tests, and caution should be taken when interpreting any scores at extremes of the distribution.

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 2006 operational HSGQE test in mathematics, reading, and writing were developed by DRC to meet approved HSGQE test blueprints.

PRE-EQUATING

In the pre-equating process, a newly developed test is linked to a set of items that were used previously on one or more test forms. In the case of the spring 2006 HSGQE, all operational items had been previously field tested in the spring of 2005. This allows for the new test's scale score to be equated to previous administrations. This procedure is known as common item equating. The quality of HSGQE equating from administration to administration is very high because of this common item equating design.

OPERATIONAL ITEM CALIBRATION

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

Because the spring 2006 test form was pre-equated, the raw score to scale score conversion was determined solely by the item and threshold difficulties estimated from the spring 2005 field test administration. Data from the spring 2006 operational administration were used only to confirm the original field test item and threshold difficulties.

The calibrated item and threshold difficulties from the spring 2005 field test were used to obtain Rasch person ability estimates and asymptotic standard errors of measurement for each possible raw score value for the overall test, as well as each subscale/reporting standard. The generation of this raw score-to-Rasch ability was accomplished through application of the fundamental formulas in the Rasch measurement model (Wright and Masters, 1982).

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

$$b_r^{(t+1)} = b_r^t - \frac{r - \sum_i \sum_{k=0}^m k P_{rik}^{(t)}}{- \sum_i \left[\sum_{k=0}^m k^2 P_{rik}^{(t)} - \left(\sum_{k=0}^m k P_{rik}^{(t)} \right)^2 \right]}, \quad r=1, M-1,$$

where b_r^t is the estimated ability of the student with score r after t iterations, $M=mL$, 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:

$$SE(b_r) = \left[\sum_i \left[\sum_{k=0}^m k^2 P_{rik}^{(t)} - \left(\sum_{k=0}^m k P_{rik}^{(t)} \right)^2 \right] \right]^{-1/2}.$$

The iteration was terminated using the default WINSTEPS convergence criteria.

CHAPTER 7: FIELD TEST ITEM DATA SUMMARY

FIELD TEST ITEMS

Once a newly constructed item had passed committee review it was ready for field testing. For instance, for the 2006 HSGQE reading test, there were 14 different field test forms containing the same 60 operational test items. In addition, each form also appended 10 multiple-choice questions and 1 constructed-response item. The field test items do not count towards an individual student’s score. Only the operational test items that were common across all test forms counted towards the individual score. A fifteenth form was designed especially for retesters that did not include any field test items.

The 14 forms were spiraled at the student level for the seven largest school districts and at the district level for the remaining districts in the state so that a large representative sample of test takers responded to the field test items. This spiraling design provided a diverse sample of student performance on each field test item. In addition, because students did not know that field test items were appended no differential motivation effects were expected.

After the assessment was administered, the operational items were then used as anchors for transforming the field test item parameters to the same logit scale that was previously established.

FIELD TEST ITEM DESCRIPTIVE STATISTICS

Appendix 17 provides item level field test statistics by subject area for the spring2006 HSGQE. These statistics represent the item characteristics most commonly used to determine whether a field test item functioned in an appropriate manner.

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

The MH procedure, as implemented by DRC, compared the observed and expected totals of a two-by-two-by-four contingency table (Holland & Thayer, 1986) shown in Table 7–1. The contingency table contrasts a focal group with a reference group by item response (correct/incorrect) by four ability levels (quartiles of the total test score). Males and Caucasians were considered the reference groups for the gender and ethnicity comparisons.

Table 7–1. Mantel-Haenszel Contingency Table

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

An odds-ratio,

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

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

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

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

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

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

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

From this value one of three severity classification categories was assigned (A, B, C). Rules for the classification are found in Appendix 18. The A category represents negligible DIF. In this case the performance of the two groups on the item was not statistically different. The B category indicates moderate DIF, that is to say, that one group outperformed the other group once differences in skill levels between the two groups have been removed. The C category indicates that there is large DIF. C DIF is statistically significant. The plus (+) and minus (-) signs that follow the DIF category indicate which group is favored by the item. The minus sign indicates that the reference group outperformed the focal group once the skill level differences between the groups have been removed. The plus sign indicates that the focal group outperformed the reference group once the skill level differences between the groups have been removed.

The analysis on constructed-response items was based on the SMD procedure (Zwick & Thayer, 1996). SMD takes into account the natural ordering of the response levels of the item. In contrast to the MH procedure this summary statistic compares the means of the reference and focal groups, adjusting for differences in the distribution of each group’s members across the four ability stratifications. Data were organized into a two-by- T -by-four contingency table shown in Table 7–2, where T is the number of score categories and the plus (+) signs denote summation over a particular index.

Table 7–2. SMD Contingency Table

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

The statistic was calculated using the equation:

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

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

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

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

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

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

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

A negative SMD value implies that the focal group has a lower mean item score than the reference group conditional on ability. Z_{SMD} was calculated using the SMD value and its corresponding standard error:

$$\text{Var}_{(SMD)} = \sum_k w_{Fk}^2 \text{Var} \left(\frac{1}{n_{F+k}} F_k - \frac{1}{n_{R+k}} R_k \right).$$

One of three severity classification categories was then assigned (A, B, C).

Summary of DIF results for field test items are provided in Appendix 19.

ITEM BANK MAINTENANCE

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

Item data review for the field test items administered in spring 2006 will be conducted in summer 2006 and will be reported in the fall 2006 and spring 2007 technical reports.

CHAPTER 8: SCALE SCORES & PROFICIENCY LEVELS

RATIONALE

To ensure that student proficiency results are reported using a common scale, EED provides a common scale score system for each HSGQE 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 scale ranges across all subjects are identical with a minimum of 100 and a maximum of 600. However the proficient cut score varies across the three subjects and scores cannot be compared directly across the subjects.

DESCRIPTION OF SCORES

Raw Score

The basic summary statistic on all HSGQE assessments is the raw score. A raw score is reported for each examinee in mathematics, reading, and writing. The raw score is the number of multiple-choice items answered correctly plus the number of points earned on constructed-response items on a subject-area assessment. By itself, the raw score has limited utility; it can only be interpreted in reference to the total number of items on a subject-area assessment, and raw scores should not be compared across tests 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 subject. The scale score range for the HSGQE is from a minimum of 100 to a maximum of 600.

When new test forms are developed, the new set of items will require slightly different levels of subject-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 subject. With scale scores, schools can compare the demonstrated knowledge and ability of groups of students across years. Comparing scale scores on the assessments can help schools determine the impact of instruction and curriculum.

Scale Score Interpretations and Limitations

The scale scores associated with the HSGQE are not vertically equated with the Standards Based Assessments (SBAs) at grades 3–10. Therefore interpretation of individual score differences between the assessments is inappropriate.

Because the scale score are established independently by subject, a comparison of scale scores between subjects is also inappropriate. Each scale score is based on a set of standards that define that content area and on items that operationalize that definition. The appropriate comparison is to compare the student to summary statistics for other students.

TRANSFORMATIONS

The equated student ability measurements in logits were transformed mathematically to a more convenient metric. To maintain consistency from administration to administration, the minimum scale scores necessary for proficiency were 328 for mathematics, 322 for reading, and 275 for writing. Table 8–1 provides the equations used for each transformation. These equations were applied to the overall test as well as to each reporting subscale.

Table 8–1. Transformation Equations

Subject	Equation
Mathematics	$SS = 59.8444(\text{logit} + 0.0046) + 301.0335$
Reading	$SS = 69.3854(\text{logit} + 0.3630) + 228.1892$
Writing	$SS = 55.3838(\text{logit} + 0.5011) + 229.4855$

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

SCALE SCORE SUMMARY STATISTICS

This section includes scale score descriptive information for each overall content area test. Subscale descriptive statistics can be found in Appendix 20. Histograms of the overall test scale scores are also provided.

Table 8–2. Content Area Scale Score Information

	Grade 10		
	Mathematics	Reading	Writing
Mean	384.54	364.17	348.24
Standard Error of Mean	0.75	0.76	0.62
Median	383	368	346
Mode	457	401	352
Standard Deviation	73.42	74.48	60.42
	Grade 11		
	Mathematics	Reading	Writing
Mean	322.94	311.73	297.67
Standard Error of Mean	1.32	1.36	2.33
Median	320	314	289
Mode	308	319	281
Standard Deviation	53.50	60.84	60.12
	Grade 12		
	Mathematics	Reading	Writing
Mean	322.76	311.08	290.62
Standard Error of Mean	2.00	1.93	4.15
Median	320	314	277
Mode	314	319	273
Standard Deviation	56.41	58.97	67.36

Figure 8–1. Mathematics Scale Score Frequencies – All Students

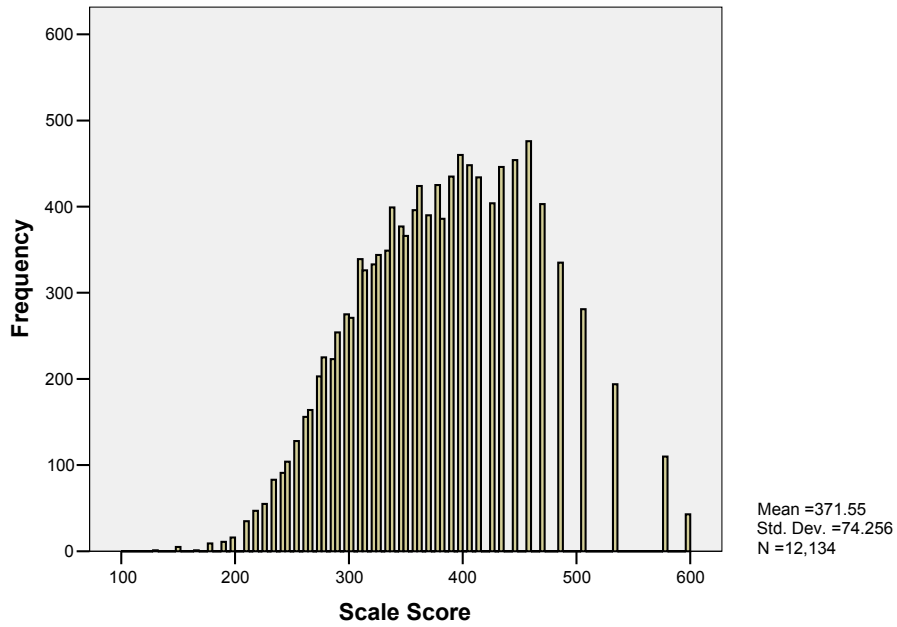


Figure 8–2. Reading Scale Score Frequencies – All Students

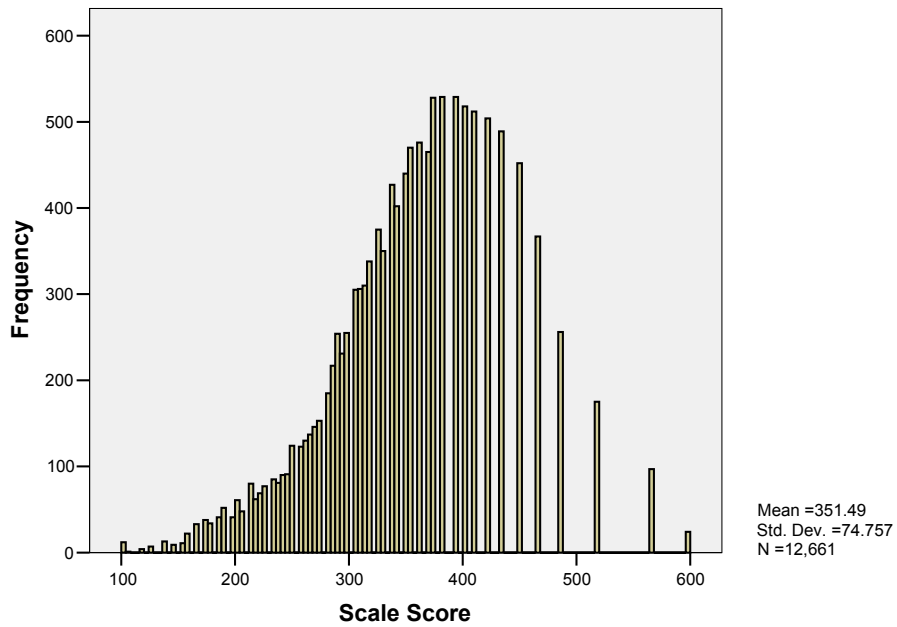
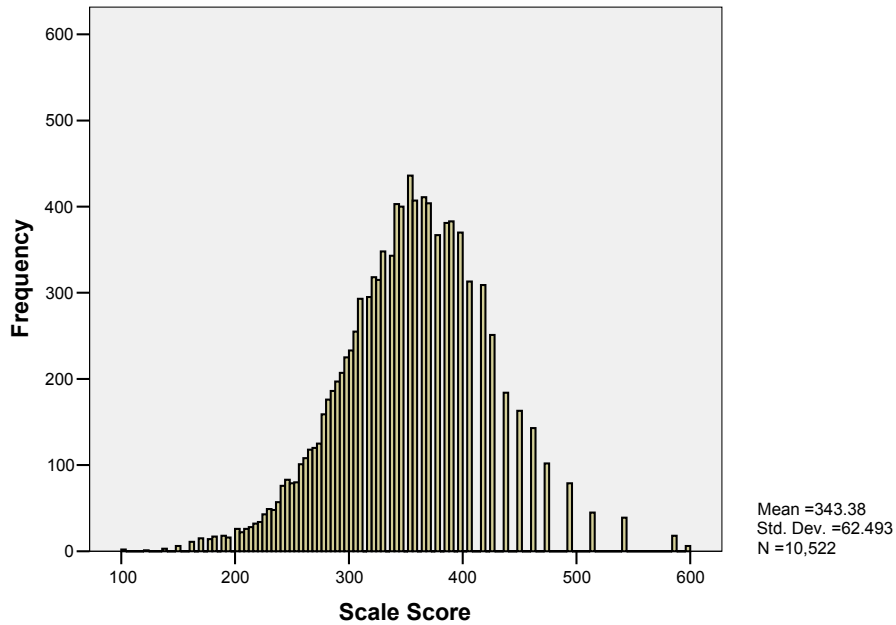


Figure 8–3. Writing Scale Score Frequencies – All Students



PROFICIENCY LEVELS

Information from the HSGQE is used to determine whether graduation requirements have been met in each school and district. Alaska has two levels of achievement; Not Proficient and Proficient.

The Proficient level corresponds to meeting the graduation requirements. Scale score values at each level of proficiency are the same each year. Appendix 21 provides detailed information about the proficiency level as well as the Proficiency Level Definitions and Descriptors in each subject tested.

Table 8–3 provides the distribution of students in each of the proficiency levels for all subjects.

Table 8–3. Student Distribution of the Two Proficiency Levels

	Grade	Mathematics		Reading		Writing	
		Count	Percent	Count	Percent	Count	Percent
Not Proficient	10	2203	22.95	2531	26.28	968	10.12
	11	974	59.25	1128	56.68	248	37.24
	12	457	57.41	543	58.26	126	47.73
	All	3699	30.48	4276	33.77	1358	12.91
Proficient	10	7396	77.05	7100	73.72	8593	89.88
	11	670	40.75	862	43.32	418	62.76
	12	339	42.59	389	41.74	138	52.27
	All	8435	69.52	8385	66.23	9164	87.09

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

Huynh (1976) suggested a beta-binomial model was preferable for determining the consistency of classification. Using this method on a complex assessment requires the dichotomization of that assessment while still using the reliability from the original assessment. Table 8–4 depicts the general framework of binary decisions.

Table 8–4. Binary Decisions—General Framework

Form X \ Form Y	Not Proficient	Proficient	Total
Not Proficient	p_{00}		p_{x0}
Proficient		p_{11}	p_{x1}
Total	p_{y0}	p_{y1}	

From this general framework the reliability index can be computed.

$$\kappa = \frac{p_{11} - p_1^2}{p_1 - p_1^2},$$

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

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

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.

Within the framework of the Rasch measurement model, there is a direct method of calculating decision consistency that does not require the distributional assumptions of the True Score methods. This method, described in Stearns and Smith (2006), is based on the asymptotic standard error estimates for each possible raw score. Given a student's score and the logit value of the three cut scores it is possible to directly test the probability that the student will receive the same proficiency level if tested again. These likelihoods can be summed over all students to determine an overall decision consistency rate.

The results of all three consistency analyses are presented in Table 8–5.

Table 8–5. Decision Consistency Indices

Subject	Huynh (1976)		Livingston and Lewis (1995)	Stearns and Smith (2006)
	Consistency Index	κ	Consistency Index	Consistency Index
Math	0.8893	0.7356	0.8841	0.9497
Reading	0.9159	0.7409	0.9222	0.9039
Writing	0.8882	0.7322	0.8878	0.8991

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 2006 HSGQE 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 subject.

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

VALIDITY

Content/Curricular

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

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 HSGQE results. When items are judged to be inappropriate for any reason, the committee can either suggest revisions (e.g., reclassification, rewording) or elect to eliminate the item from the field test item pool. Items that are approved by the review committee are later embedded in operational HSGQE 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 HSGQE.

Educator Input

Alaska educators provide valuable input on the alignment of the items and the statewide content standards. Items are written specifically to measure the objectives and specifications of the content standards for the HSGQE. Using a varied source of item writers provides a system of checks-and-balances for item development and review that reduces single source bias. Because

many different people with different backgrounds write the items, it is less likely that items will suffer from a bias that might occur if items were written by a single author. This direct input from educators offers evidence regarding the content validity of the HSGQE.

Developer Input

For the items included in the spring 2006 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 HSGQE.

Item to Content Area Match

Expert judgments from educators, test developers, and assessment specialists provide support for the alignment of the HSGQE with the statewide content standards. In addition, because expert teachers in the content areas were involved in establishing the content standards, the judgments of these same expert teachers in the review process provide a measure of content validity. A match between the content standards and the components of the HSGQE 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 HSGQE test blueprint provides this documentation. The blueprints for mathematics, reading, and writing are presented in Appendix 1.

Construct Validity

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

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

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

- Substantial relationships between different methods of measurement regarding the same defined construct.
- Relationships to non-assessment measures of the same defined construct.

Evidence of Construct Validity

The collection of construct-related evidence is a continuous and ongoing process. Two current metrics of construct validity for the HSGQE are item-total correlations and Rasch item fit statistics. An item-total correlation is the correlation between an item and the total test 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 ability and low ability 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 2006 HSGQE can be found in Appendix 13.

Criterion-Related Validity

Although the primary evidence for the validity of the HSGQE lies in the content and construct validity of the test, it is also informative to collect criterion-related validity evidence. The term criterion validity refers to the degree to which a test correlates with one or more outcome criteria. These analyses are beyond the scope of this technical report.

Validity Evidence for Different Student Populations

The primary evidence for the validity of the HSGQE 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 HSGQE 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.

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

RELIABILITY

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

$$X = T + E,$$

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:

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

The variance of the observed measures can be estimated from the variance of the raw scores using the usual formula and the error variance can be estimated by the $\Sigma p(1-p)$, where p is the proportion correct for each item.

The reliability index used for the 2006 administration of the HSGQE was the Coefficient Alpha (Cronbach, 1951):

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

where

k is the number of items,

σ_i^2 is the variance of the set of scores associated with item i , and

σ_X^2 is the variance of the set of observed total scores.

Acceptable α values generally range in the high 0.80s to low 0.90s. When there is no error, the reliability index is the true score variance divided by the true score variance, which is one. Appendix 14 provides Coefficient Alpha for each subject.

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:

$$SEM = \sigma_x \sqrt{1 - \rho_{XX}} ,$$

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

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

Ability estimates from scores near the center of the test are known with greater precision than are abilities associated with extremely high or low scores. The expression for computing the asymptotic standard error via WINSTEPS was provided in Chapter 6. This value is then transformed to the HSGQE scale to obtain the final SEM for each raw score. These values for the spring 2006 HSGQE are provided in the raw-to-scale score tables in Appendix 16.

REFERENCES

- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. 1999. *Standards for educational and psychological testing*. 2nd ed. Washington, D.C.: American Educational Research Association.
- Bloom, B. S., Englehart, M. D., Furst, E. J., Hill, W. H., and Krathwohl, D. R. 1956. *Taxonomy of Educational Objectives: The classification of educational goals: Handbook 1: Cognitive Domain*. New York: Longman, Green, and Co.
- Cronbach, L. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika* 16: 297–334.
- Holland, P., and Thayer, D. 1986. *Differential item performance and the Mantel-Haenszel procedure*. Paper presented at the American Educational Research Association annual meeting, San Francisco, CA.
- Huynh, H. 1976. On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement* 13: 253–64.
- Linacre, J. M. 2006. *WINSTEPS Rasch measurement (Version 3.60.1)*. Chicago: WINSTEPS.com. Computer program.
- Linn, R., and N. Gronlund. 1995. *Measurement in assessment and teaching*. 7th ed. New Jersey: Prentice-Hill.
- Livingston, S. A., and Lewis, C. 1995. Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement* 32:179–197.
- Mead, R. 1978. Examining residuals from the Rasch model. *Proceedings of the 1978 conference on adaptive testing*. Minneapolis, MN: University of Minnesota.
- Mogilner, A. 1992. *Children's Writer's Word Book*. Cincinnati, OH: Writer's Digest Books.
- Rasch, G. 1960. *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research (Expanded Edition, 1980. Chicago: University of Chicago Press).
- Stearns, M., and Smith R. M. 2006. *On the estimation of classification consistency indices for complex assessments*. Paper presented at the Thirteenth International Objective Measurement Workshop. Berkeley, CA.
- Stout, W. 1987. A non-parametric approach to assessing latent trait unidimensionality. *Psychometrika* 52: 589–617.
- Taylor, S. E., Frackenpohl, H., White, C. E., Nieroroda, B. W., Browning, C. L., and Brisner, E. P. 1989. *EDL Core Vocabularies in Reading, Mathematics, Science, and Social Studies*. Orlando, FL: Steck-Vaughn Company.

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

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

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

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