

The word "Alaska" is written in a large, black, cursive font. Above the letters, a series of small, grey stars are arranged in a curved path, resembling the state flag's design.

Alaska

Comprehensive System of Student Assessment

Technical Report

**Spring 2010
High School Graduation
Qualifying Examination (HSGQE)
and HSGQE Retest**



September 2010

TABLE OF CONTENTS

CHAPTER 1: BACKGROUND OF ALASKA ASSESSMENTS.....	1
CHAPTER 2: TEST DESIGN & ITEM DEVELOPMENT	2
2010 Operational Plan.....	2
Test Development Timeline	2
Item and Test Development Process	2
Item Writer Training	3
Reading Passage Selection	4
Passage Readability.....	4
Item Writing.....	4
Item Content Review.....	7
Bias and Sensitivity Review.....	8
Item Field Test	8
Item Field Test Data Review.....	8
Psychometric Guidelines for Selecting Items.....	9
Proportion Correct.....	9
Average Person Logit.....	9
Item-Total Correlation	9
Fit Statistic	10
Differential Item Functioning (DIF) Analyses.....	10
Item Bank.....	10
Overview	10
Functionality	11
Item Cards and Reporting Options.....	11
Security	11
Quality Assurance	11
Item Bank Summary	12
Spring 2010 HSGQE Operational Forms Construction	12
Steps in the Forms Construction Process	13
DRC Internal Review of the Items and Forms	13

CHAPTER 3: TEST ADMINISTRATION PROCEDURES	14
Overview.....	14
Student Population Tested.....	14
Accommodations.....	14
Test Administrator Training.....	15
Test Security.....	15
Materials.....	15
Packaging and Shipping Materials.....	16
Materials Return.....	16
Box Receipt.....	16
CHAPTER 4: SCORING & STUDENT PERFORMANCE REPORTING.....	17
Document Processing	17
Handscoring of Constructed-Response Items.....	17
Readers.....	17
Rangefinding and Developing Training Material	18
Training the Readers	18
Imaging.....	18
Quality Control of Handscoring.....	19
Data Processing.....	19
Reporting Mockups.....	20
Reporting.....	20
District Reports	21
State Reports	21
CHAPTER 5: FORM ANALYSIS & ITEM CALIBRATION.....	22
Rasch Measurement Models.....	22
Item Statistics.....	23
Form Statistics	25
Frequency Distributions	28
Items.....	28
Persons	29
Cautions for Score Use.....	29

CHAPTER 6: SCALING & EQUATING.....	30
Introduction	30
Pre-Equating	30
Operational Item Calibration.....	30
Item Bank Maintenance.....	31
CHAPTER 7: FIELD-TEST ITEM DATA SUMMARY	32
Field-Test Items	32
Field-Test Item Descriptive Statistics	32
Item Bank Maintenance.....	36
CHAPTER 8: SCALE SCORES & PERFORMANCE LEVELS	37
Rationale.....	37
Description of Scores.....	37
Raw Score	37
Scale Score	37
Transformations	38
Scale Score Summary Statistics	38
Proficiency Levels.....	41
CHAPTER 9: TEST VALIDITY & RELIABILITY.....	43
Introduction	43
Validity	43
Content/Curricular.....	43
Construct Validity	44
Validity Evidence for Different Student Populations	46
Reliability	47
Standard Error of Measurement.....	48
Indicators of Consistency	48
REFERENCES	50

APPENDIX 1: SPRING 2010 HSGQE TEST BLUEPRINT..... 1-1

APPENDIX 2: RUBRICS 2-1

6-Point Extended Constructed-Response (ECR) Scoring Rubric for Writing 2-1

6 Points..... 2-1

5 Points..... 2-1

4 Points..... 2-1

3 Points..... 2-2

2 Points..... 2-2

1 Point 2-2

4-Point Extended Constructed-Response (ECR) Scoring Rubric for Grades 10/10+ Writing..... 2-3

4 Points..... 2-3

3 Points..... 2-3

2 Points..... 2-3

1 Point 2-3

APPENDIX 3: DRC ITEM WRITER ORIENTATION MANUAL 3-1

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

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

Mathematics..... 5-1

Level 1 5-1

Level 2..... 5-1

Level 3..... 5-2

Level 4..... 5-2

Reading..... 5-3

Level 1 5-3

Level 2..... 5-3

Level 3..... 5-3

Level 4..... 5-4

Writing..... 5-5

Level 1 5-5

Level 2..... 5-5

Level 3..... 5-5

Level 4..... 5-6

Source of Challenge Criterion 5-6

APPENDIX 6: UNIVERSALLY DESIGNED ASSESSMENTS.....	6-1
Elements of Universally Designed Assessments.....	6-1
Guidelines for Universally Designed Items	6-3
APPENDIX 7: ITEM REVIEW TRACKING FORM	7-1
Content Review Form	7-1
APPENDIX 8: CONFIDENTIALITY AGREEMENT	8-1
APPENDIX 9: BIAS & SENSITIVITY REVIEW FORM.....	9-1
APPENDIX 10: SAMPLES OF MANUALS	10-1
APPENDIX 11: HSGQE INTER-RATER RELIABILITY	11-1
APPENDIX 12: SAMPLES OF GUIDES TO TEST INTERPRETATION	12-1
APPENDIX 13: OPERATIONAL TEST ITEM ANALYSIS.....	13-1
Mathematics.....	13-1
Reading.....	13-2
Writing.....	13-4
APPENDIX 14: OPERATIONAL TEST ITEM AND THRESHOLD DIFFICULTY MAPS	14-1
Mathematics.....	14-1
Reading.....	14-2
Writing.....	14-3
APPENDIX 15: RAW-TO-SCALE SCORE TABLES.....	15-1
Mathematics.....	15-1
Reading.....	15-4
Writing.....	15-8
APPENDIX 16: FIELD-TEST ITEM ANALYSIS	16-1
Mathematics.....	16-1
Reading.....	16-4
Writing.....	16-6

APPENDIX 17: FIELD-TEST DIFFERENTIAL ITEM FUNCTIONING (DIF)
CLASSIFICATION RULES 17-1

Dichotomous (Multiple-Choice) DIF Classification 17-1
Polytomous (Constructed-Response) DIF Classification 17-1

APPENDIX 18: SUBSCALE SCORE SUMMARY STATISTICS 18-1

Mathematics Subscale Reporting Categories 18-1
Reading Subscale Reporting Categories 18-1
Writing Subscale Reporting Categories 18-1

APPENDIX 19: HSGQE PROFICIENCY DESCRIPTORS OF THE MINIMUM
COMPETENCIES IN ESSENTIAL SKILLS..... 19-1

Mathematics..... 19-1
Reading..... 19-3
Writing..... 19-5

APPENDIX 20: TOTAL SCORE AND SUBSCALE SCORE INTERCORRELATIONS ... 20-1

Mathematics Subscale Reporting Categories 20-1
Reading Subscale Reporting Categories 20-1
Writing Subscale Reporting Categories 20-1

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

2010 OPERATIONAL PLAN

The spring 2010 HSGQE in mathematics, reading, and writing was a single recycled operational form. The form had previously been used in a past spring administration. In addition, each content area form included seven forms of newly developed field test items embedded in the recycled form. The test blueprints for the spring 2010 HSGQE and retest are found in Appendix 1.

TEST DEVELOPMENT TIMELINE

A series of major test development activities took place in 2009 and 2010, which culminated in the administration of the operational HSGQE assessment. These key activities included the:

- Development of items, tasks, and writing prompts.
- Review of items field tested in 2010 by external committees of educators (content review, bias/sensitivity review).
- Embedded field testing of the new mathematics, reading, and writing items in April 2010 recycled operational form.
- Update of the Alaska Item Bank.
- Preparation of the selected recycled operational form.

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 *Universal Design* (Thompson, Johnstone, & Thurlow, 2002) guided the following steps in the item and test development process:

- Analyze the performance standards and test blueprint.
- Analyze item specifications and style guides.
- Select qualified item writers.
- Develop item-writing workshop training materials.
- Train test development specialists and item writers to write items.
- Write items that match the standards, are free of bias, and address fairness and sensitivity concerns.
- Conduct and monitor internal item reviews and quality processes.

- Prepare passages and items for review by committees of Alaska educators (content and bias/sensitivity).
- Select and assemble items for field testing.
- Field test items, scoring of the items, and analysis of the data.
- Review items and associated statistics after field testing, including bias statistics.
- Update item bank.

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

All item writers were provided with 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 Orientation Manual*, *Fairness in Testing Manual*, *Depth of Knowledge Levels*, and *Universally Designed Assessments* 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 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 and/or edited by both committees of Alaska educators. The final selection of passages to be field tested was based on the specific requirements of the HSGQE 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 the 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 target content test blueprint and item specifications and were adequately distributed across subcategories and levels of difficulty, item writers were asked to document the following specific information as each item was written.

Alignment to the Alaska Performance Standards: There must be a high degree of match between a particular question and the performance standard it is intended to measure. Item writers were asked to clearly indicate what performance 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.

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

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

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

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

Readability: For mathematics item development, writers were instructed to pay careful attention to the readability of each mathematics item to ensure that the focus was on the concepts; not on reading comprehension. As a result, the goal for each mathematics writer was to write items that were, to the greatest degree possible, independent of the assessment of reading. Subject areas such as mathematics contain many content-specific vocabulary terms. These terms make it impossible to use the standard methods available for determining the reading level of test questions. Wherever it was practical and reasonable, every effort was made to keep the vocabulary one grade level below the tested grade level. Resources writers used to verify the vocabulary level were the *EDL Core Vocabularies* (Taylor et.al., 1989) and the *Children's*

Writer's Word Book (Mogilner, 1992). In addition, every mathematics test question was taken before committees comprised of Alaska grade-level experts in the field of mathematics education. They reviewed each question from the perspective of the students they teach, and they determined the validity of the vocabulary used.

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

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

Editorial Review of Items

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

Guidelines for Reviewing Items Selected for Forms

A good item should

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

Item Content Review

Prior to field testing items embedded in the 2010 form, all newly developed test items were submitted to content committees for review. The content committees consisted of Alaska teachers and subject-area supervisors from school districts throughout Alaska. The primary responsibility of the content committee was to evaluate items with regard to quality and content classification, including grade-level appropriateness, estimated difficulty, depth of knowledge, and source of challenge. They also suggested revisions, if appropriate. The committee also reviewed the items for adherence to the principles of universal design, including language demand and issues of bias, fairness, and sensitivity.

The content review was held August 5, 2009. Committee members were selected by EED, and EED-approved invitations were sent to them by DRC. The committee consisted of 18 educators, divided evenly between the content areas. EED also selected internal staff members for attendance. The meeting commenced with an overview of the test development process. Training was provided by DRC senior staff members. Training included how to review items for technical quality and content quality, including depth of knowledge and adherence to principles of universal design. In addition, training included providing committee members with the procedures for item review, including the use of tracking review forms to be used during the item content review.

DRC test development specialists in mathematics, reading, and writing facilitated the review of items. Committee members 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.

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

Security was addressed by adhering to a strict set of procedures. Items in binders did not leave the meeting rooms and were accounted for at the end of each day before attendees were dismissed. All attendees, with the exception of EED staff, were required to sign a Confidentiality Agreement (Appendix 8). All materials not in use 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 items within the 2010 form, all newly developed test items were also submitted to a Bias and Sensitivity Committee for review. This review took place in Alaska on August 5, 2009. The committee's primary responsibility was to evaluate passages and items as to acceptability with regard to bias and sensitivity issues. They also made recommendations for changes or deletion of items in order to remove the area of concern. The bias/sensitivity committee was composed of 10 Alaska educators who represented the diversity of Alaska students. The committee members were trained by a DRC test development lead to review items for bias and sensitivity issues using a Fairness in Testing Manual developed by DRC (Appendix 4). This manual was customized specifically for the Alaska program.

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 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.

All mathematics, reading, and writing items were read by all of the committee members. Each member noted bias and/or sensitivity comments on review forms (Appendix 9). All comments were then compiled and the actions taken on these items were recorded by DRC.

Item Field Test

Items being field tested were embedded in the forms for the spring 2010 administration.

Item Field Test Data Review

Following the spring 2009 administration, which contained 7 forms of field test items, the following field test statistical analyses were completed:

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

Item analysis results were reviewed by DRC psychometricians to identify any items that were not performing as expected. These items were flagged so DRC test development specialists were made aware of potential areas of concern. For example, in the case of multiple-choice items,

DRC test development specialists checked to make sure that the key for each item was correct and that none of the other response options were plausible. In the case of items where large values of DIF occur, DRC test development specialists reviewed each item flagged to consider whether or not a feature of the item may have caused a problem and/or contributed to the DIF. Under the guidance of DRC psychometricians, DRC test development specialists will determine which of the flagged items are to be reviewed by a committee of Alaska educators to determine whether or not the item is appropriate for use. These items will join the pool of field-tested items scheduled to be reviewed at a formal data review meeting in 2010. Additional guidelines concerning the review of item analysis results for the item-selection process are provided below.

PSYCHOMETRIC GUIDELINES FOR SELECTING ITEMS

Proportion Correct

The proportion correct, or p -value, is the proportion of the total group of test takers answering the MC item correctly. The proportion for an item will show how difficult the item was for the students who took that field-test form. In general, MC items with a proportion somewhat higher than half the difference between the chance level and 1.00 should be recommended for selection first, and the range for selection should be between 0.40–0.90. When necessary to meet the test blueprint or other test specifications, items that fall outside this range may be used, albeit sparingly. The overall form was constructed to a mean p -value target range of 0.63 to 0.67, with special care taken to select items that were at or near the cut score.

Average Person Logit

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

Item-Total Correlation

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

Fit Statistic

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

Differential Item Functioning (DIF) Analyses

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

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

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

ITEM BANK

Overview

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

Functionality

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

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

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

Item Cards and Reporting Options

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

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

Security

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 any 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 numbers of eligible items in the item bank available to be used on a new operational form are presented in Table 2–1. Because it was a recycled form, no items were used from the bank to build the 2010 HSGQE form. The item summary table for each content area shows eligible items after the fall 2007 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 2010 SBA Technical Report.

Table 2–1. Eligible HSGQE Items

Mathematics Items

Standard	MC	CR
M1	26	0
M2	21	0
M3	35	5
M4	30	4
M5	34	2
M6	29	2

Reading Items

Standard	MC	CR
R4.1	9	0
R4.2	16	2
R4.3	20	0
R4.4	2	0
R4.7	3	6
R4.8	2	0

Writing Items

Standard	MC	CR
W4.1/2	46	15
W4.3	59	7
W4.4	74	10

SPRING 2010 HSGQE OPERATIONAL FORMS CONSTRUCTION

The spring 2010 HSGQE in mathematics, reading, and writing was comprised of 1 recycled form with 7 forms of embedded field test items. The recycled form had already been 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. It also 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 previous test forms construction process to ensure that the HSGQE is of high quality, legally defensible, and meets 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 the psychometric guidelines for operational forms construction.
3. DRC psychometricians analyzed item statistics for the field tested items and provided DRC test development specialists with characteristics for each item.
4. DRC test development specialists received all item cards and verified that each item image had its correct item characteristics and psychometric data.
5. DRC test development specialists reviewed all items in the operational pool and made an initial selection of items according to test blueprint guidelines and psychometric guidelines.
6. DRC test development specialists created item-mapping charts for the test.
7. Final recommendations for items selected for the operational forms were prepared for review by senior test development staff.
8. Based upon senior review, suggested replacements were made by DRC test development specialists, if necessary.
9. Operational forms were prepared for psychometric review and approval.
10. Based upon psychometric review, suggested replacements were made by DRC test development specialists, if necessary.
11. Operational forms were prepared for EED review and approval.

The Spring 2010 test form was reviewed and approved by EED. Feedback was provided on the selected embedded field test items and a complete review done to ensure the recycled operational form remained consistent to its original administration.

DRC INTERNAL REVIEW OF THE ITEMS AND FORMS

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

CHAPTER 3: TEST ADMINISTRATION PROCEDURES

OVERVIEW

All students in grade 10 are required to take the spring administration of the Grade 10 SBA-HSGQE test, even if they have previously taken the HSGQE. A student who has failed to pass the entire HSGQE, or a student who is a junior or senior and has never taken the HSGQE, as determined by the district, may participate in the spring administration of the HSGQE (4 AAC 06.755). The test administration window was April 6 through April 8, 2010. Schools administered the reading test on April 6, the writing test on April 7, and the mathematics test on April 8, 2010. A District Test Coordinator was assigned at every school district. DRC distributed the testing materials to each District Test Coordinator (DTC).

STUDENT POPULATION TESTED

Districts submitted their enrollment, accommodated materials counts, and updates to district contact information via DRC's Online Enrollment System November 9 through December 8, 2009. Districts also submitted their precode files via DRC's Online Precode System, January 4 through January 19, 2010. Districts with 30 or more schools and 9,000 or more students were given the option to submit their enrollment files directly to DRC by December 8, 2009. Anchorage and Fairbanks took advantage of the opportunity to submit a file and then use DRC's Online Enrollment System for review and verification of their data. In addition, these larger districts were allowed to submit their precode files via DRC's Online Precode System by February 19, 2010 with precode and district/school labels arriving in these districts by March 18, 2010.

The enrollment and documents processed counts were as follows:

Table 3–1. Project Counts

District Count	School Count
54	255
Enrollment Count	Processed Count
HSGQE: 10,154	HSGQE: 9,400
HSGQE Retest: 5,452	HSGQE Retest: 2,709

ACCOMMODATIONS

Appropriate accommodations were available for students with disabilities while taking the assessments. These accommodations were required to be documented in an Individualized Education Program (IEP) or in a 504 plan. Refer to the Participation Guidelines for examples of acceptable accommodations (http://www.eed.state.ak.us/tls/Assessment/participation_guidelines/ParticipationGuidelinesSept2007.pdf).

TEST ADMINISTRATOR TRAINING

DTCs were trained February 23–24, 2010 by EED and DRC. The training focused on test materials receipt, distribution and return procedures, and general testing information. DTCs scheduled training sessions with test administrators during March and April 2010.

TEST SECURITY

The Grade 10 SBA-HSGQE and the HSGQE Retest materials are considered secure materials. According to Alaska test security regulation 4 AAC 06.765, all test materials must be kept secure. No portion of test materials may be photocopied or duplicated at any time. Except for the person testing, no person, including test administrators, is permitted to read test items on the Grade 10 SBA-HSGQE and the HSGQE Retest prior to, during (except for the student testing), or after administration. Teachers, proctors, test administrators, or any testing personnel may not read test items aloud, silently, to themselves, or to another individual, unless specifically required to provide a documented accommodation to an individual or student group. Parents/guardians may not read test items under any circumstances.

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

Prior to the first test administration of the school year, DTCs signed and sent their District Test Coordinator Test Security Agreements to EED.

MATERIALS

The following materials were produced for this administration:

- *District Test Coordinator's Manual*
- *Test Administration Directions*
- Seven Forms (with varied embedded field test items) of Form D Reading Test Books – grade 10
- Seven Forms (with varied embedded field test items) of Form D Writing/Mathematics Test Books – grade 10
- HSGQE Retest Test Books—Form D15
- Large Print Test Books
- Braille Test Books
- HSGQE audiotapes for writing and mathematics
- HSGQE Retest audiotapes for reading, writing, and mathematics
- HSGQE Sign Language DVDs for writing and mathematics
- HSGQE Retest Sign Language DVDs 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 8, 2010, as scheduled.

District ancillary materials were packed in the last box, which was labeled, “District Materials Enclosed.” Boxes were filled 75-percent full to allow districts space to return their materials after they had expanded due to being removed from shrink-wrap and used by students.

DRC overage was shrink-wrapped in groups of three. All secure materials were barcoded, packaged by range sheet, and shrink-wrapped. DRC barcoded all accommodated materials and shrink-wrapped large print and Braille test books.

DRC entered, packed, and shipped requests for additional materials March 8–24, 2010. DRC processed 13 additional materials requests for this administration.

Materials Return

Districts returned all materials via Assessment Distribution Services on April 13, 2010 and most materials arrived at DRC’s warehouse on April 19, 2010.

Districts were instructed to place an orange HSGQE Grade 12 Express label on all boxes containing used grade 12 documents. All districts used orange, district-specific DRC return shipping labels on all return boxes.

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 Assessment Distribution Services. All boxes bearing an orange HSGQE Grade 12 Express label were expedited through box receipt. The Materials Processing team scanned each box using the Operations Materials Management System (OpsMMS) box receipt system and, as soon as box receipt was complete, notified EPM of any schools that did not return a box. 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 did not match, the books were reconciled to ensure accurate numbers. Accommodated testing materials were also checked in securely using the security barcodes on the documents.

The Materials Processing team produced a preliminary Missing Materials Report and performed quality checks based on this report. The report was then forwarded to EPM, who checked for the missing materials on the security checklists and in recorded correspondence from the districts. If sufficient documentation regarding a material was found, the item was removed from the Missing Materials Report.

DRC used its Image Scanning System to scan the HSGQE test books. Scanning of test books was completed on April 27. All predefined editing and validating rules were followed.

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 were available for review. DRC staff used the rubrics and live student responses to build anchor sets and training materials for each item assessed. Writing constructed-response items were scored using "generic" (i.e., 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 were selected and the rubric and scoring guidelines applied. DRC staff moved from item to item until a sufficient number of scored responses were compiled to construct training materials. Responses that were particularly relevant (in terms of the scoring concepts they illustrate) were annotated for use in the scoring guide. The scoring guide for each item served as the readers' constant reference. An anchor set and a training set were created for each field test item. For operational items, these materials would be enhanced with the addition of further training sets and qualifying sets.

Training the Readers

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

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

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

Short constructed response 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.

Extended constructed response 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.

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.

Reporting Mockups

DRC worked with EED to determine appropriate file layouts. The layouts included field names, field descriptions, and field values. DRC posted district-level data files and layouts to the DRC Report Delivery System and state-level data files and layouts to an FTP site that EED can access.

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 Analysts for accuracy and consistency. EED reviewed the mock-ups as part of the Functional Specifications Document for Reporting review. 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, and so DRC prioritized the scoring and reporting of the district's student documents.

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 as outlined under the "District Reports" and "State Reports" headings that follow. DRC also produced Parent/Student and Teacher/Staff versions of the *Guide to Test Interpretation*. Samples of these guides are provided in Appendix 12 and are also available on EED's Web site.

Grade 12 student reports were provided electronically as scheduled on April 28, 2010. All HSGQE reports were provided electronically as scheduled on May 14, 2010. All paper copies of HSGQE reports were delivered to districts as scheduled on or before May 21, 2010.

The erasure analysis, which included all straggler (late) returns, was delivered to EED on August 13, 2010.

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
- DVDs

CHAPTER 5: FORM ANALYSIS & ITEM CALIBRATION

RASCH MEASUREMENT MODELS

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

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

The standard Rasch calibration procedure sets the mean difficulty of the items on any unanchored calibration at zero. Any item with a p -value lower than the mean receives a positive logit and any item with a p -value higher than the mean receives a negative logit. Consequently, the logits for any calibration, whether it is a third-grade reading test or a high-school mathematics test, relate to an arbitrary origin defined by the average of item difficulties for that form. The average third-grade reading item will have a logit of zero; the average high-school mathematics item will have a logit of zero in unanchored calibrations. This common logit scale describes 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. Multiple-choice (MC) items, scored either right or wrong, were calibrated using the familiar form of the dichotomous Rasch model. Constructed-response (CR) items were calibrated using another model in the Rasch family, Master's partial-credit model (Wright & Masters, 1982). The latter model parameterizes each threshold needed to obtain the maximum score on the task. Consequently, there is one item difficulty parameter for each of the $n - 1$ score transitions (0/1, 1/2, etc.), or thresholds. While the partial-credit model is a non-trivial extension of the simple logistic Rasch model, an MC item may be thought of as a partial-credit task with only one threshold.

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

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

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

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

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

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

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

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

The unconditional, joint maximum likelihood (UCON) estimation procedure estimates the person parameters (i.e., ability) simultaneously with the item parameters (i.e., difficulty). The UCON procedure was accomplished using WINSTEPS Version 3.69.1.14 (Linacre, 2010). This calibration software is commercially available and widely used in the testing industry and is considered the industry standard for Rasch calibration.

ITEM STATISTICS

Appendix 13 provides item-level statistics by content area for the spring 2010 HSGQE operational assessments. These statistics (i.e., logit, standard error, fit, p -value, item-total correlation, and proportion of omits) represent the item characteristics most commonly used to determine whether an item functioned in an appropriate manner. Table 5–1 presents the mean or median of these statistics within each content area.

The logit column in the table and appendix provides the estimated item difficulty for the item. The standard error (SE) column gives the asymptotic standard error associated with the item difficulties.

The Rasch fit statistics are used to determine how well items conform to the requirements of the Rasch measurement model. The items were analyzed for scale comparability by examining the residuals between observed and expected scores for the persons and items (Smith, 2000; Mead, 1978). This process investigated the underlying construct measured by a test by analyzing the patterns of item covariation within the scale. For example, when local dependence is exhibited, it may indicate violations of unidimensionality, thus introducing sources of variability that are unrelated to the construct being measured. Even if some minor item dependence existed in the CR item formats, they were likely to have minor influence on scores (Stout, 1987). A standardized weighted total fit (OUTFIT z -std) statistic was computed for each item. This fit statistic quantifies the sum of the squared difference of the observed item performance from the expected performance for all persons. Items may not fit the Rasch model for several reasons, all of which relate to students responding to items in an unexpected way. In many cases, the reason behind why students respond in unexpected ways to a particular item is unclear. However, it is possible to determine possible causes of an item's misfit by re-examining the item and its distractors. Content specialists examined items with large fit statistics and confirmed that each item had only one correct answer and was properly written.

The p -value for an MC item is the percent (or proportion) of all students that responded to an item correctly. The p -value for a CR item represents the average score earned divided by the maximum number of points for that item. For the spring 2010 HSGQE forms, the range of CR item scores is from 0–2 or 0–4 points in mathematics, 0–2, 0–3, or 0–4 in reading, and 0–2, 1–4, or 1–6 in writing.

The item-total correlation (PtBis or Corr.) provides a measure of internal consistency of the responses. It assesses how well each item measures the trait defined by the set of items as a whole. Typically, students with high proficiency (i.e., those that perform well on the HSGQE content-area test overall) would be expected to answer items correctly, and students with low proficiency (i.e., those that perform poorly on the HSGQE content-area test overall) would be expected to answer items incorrectly. If these expectations are met, the item-total correlation between the item and the total test score will be high and positive, indicating that the item is a good discriminator between high-ability and low-ability students. An item-total correlation value above 0.30 is usually considered acceptable. An item-total correlation value below 0.30 indicates that an item may not be measuring what it was intended to measure, and should be reviewed. DRC content specialists reviewed all items with item-total correlations below 0.30 and verified that each item was acceptable as written and scored. As seen in Table 5–1, the median item-total correlations for MC and CR items exceeded the 0.30 criterion.

The omits column represents the proportion of persons leaving the item blank for MC items and the proportion of persons with blanks or other condition codes for CR items. The non-scorable codes are recoded as 0 points during item calibration.

Table 5–1. Summary of Operational Item Analysis

Content Area	Item Type	Mean Logit	Mean SE	Mean Fit	Mean <i>p</i> -value	Median PtBis or Corr	Mean Omits
Mathematics	MC	0.01	0.02	-0.87	0.70	0.39	0.00
	CR	1.86	0.02	-5.80	0.36	0.54	0.08
Reading	MC	0.10	0.03	-1.75	0.74	0.43	0.00
	CR	0.38	0.02	8.83	0.62	0.42	0.02
Writing	MC	0.41	0.02	3.07	0.68	0.31	0.00
	CR	0.82	0.01	-6.57	0.54	0.62	0.03

FORM STATISTICS

Tables 5–2 through 5–7 contain summary descriptive statistics for student performance and item difficulty, including mean score, standard deviation, and minimum and maximum scores by content area. These statistics were generated using WINSTEPS v3.69.1.14 (Linacre, 2010) and illustrate student and item performance. The top halves of the student summary tables provide descriptive statistics for persons (i.e., students) measured. The column labeled “Measure” provides the mean and standard deviation of the estimated student proficiency measures. The “Model Error” column presents similar information for the asymptotic standard errors.

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

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

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

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

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

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

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

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

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

$$G_I = SA_I / RMSE_I .$$

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

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

It can also be calculated using only the separation value:

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

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

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

Table 5-2. Mathematics—Summary of 10,579 Measured Students

	Raw Score	Count	Measure	Model Error
Mean	32.8	47.0	1.16	0.38
SD	9.5	0.0	1.21	0.13
Max.	50	47	5.82	1.83
Min.	1	47	-4.05	0.31
	RMSE	True SD	Person Separation	Person Separation Reliability
Real	0.41	1.14	2.77	0.88
Model	0.40	1.14	2.85	0.89

SE of Student Measure Mean = 0.01

Student Raw Score-to-Measure Correlation = 0.97

Coefficient Alpha Student Raw Score Reliability = 0.91

Table 5–3. Mathematics—Summary of 47 Measured Items

	Raw Score	Count	Measure	Model Error
Mean	7381.6	10579.0	0.13	0.02
SD	1521.0	0.0	0.89	0.00
Max.	9657	10579	2.67	0.03
Min.	2949	10579	-1.51	0.02
	RMSE	True SD	Item Separation	Item Separation Reliability
Real	0.03	0.89	35.42	1.00
Model	0.02	0.89	36.16	1.00

SE of Item Measure Mean = 0.13

Table 5–4. Reading—Summary of 9,929 Measured Students

	Raw Score	Count	Measure	Model Error
Mean	47.4	60.0	1.43	0.35
SD	11.8	0.0	1.13	0.13
Max.	65	60	5.71	1.83
Min.	0	60	-5.41	0.26
	RMSE	True SD	Person Separation	Person Separation Reliability
Real	0.39	1.06	2.74	0.88
Model	0.37	1.07	2.86	0.89

SE of Student Measure Mean = 0.01

Student Raw Score-to-Measure Correlation = 0.96

Coefficient Alpha Student Raw Score Reliability = 0.92

Table 5–5. Reading—Summary of 60 Measured Items

	Raw Score	Count	Measure	Model Error
Mean	7837.6	9929.0	0.12	0.03
SD	2412.7	0.0	0.74	0.00
Max.	19554	9929	1.90	0.04
Min.	4356	9929	-1.74	0.01
	RMSE	True SD	Item Separation	Item Separation Reliability
Real	0.03	0.74	27.10	1.00
Model	0.03	0.74	28.06	1.00

SE of Item Measure Mean = 0.10

Table 5–6. Writing—Summary of 10,540 Measured Students

	Raw Score	Count	Measure	Model Error
Mean	42.2	36.0	1.38	0.32
SD	10.6	0.0	1.03	0.07
Max.	66	36	7.15	1.84
Min.	2	36	-2.83	0.26
	RMSE	True SD	Person Separation	Person Separation Reliability
Real	0.35	0.97	2.79	0.89
Model	0.33	0.98	2.97	0.90

SE of Student Measure Mean = 0.01

Student Raw Score-to-Measure Correlation = 0.98

Coefficient Alpha Student Raw Score Reliability = 0.90

Table 5–7. Writing—Summary of 32 Measured Items

	Raw Score	Count	Measure	Model Error
Mean	12341.7	10540.0	0.53	0.02
SD	9390.8	0.0	0.60	0.01
Max.	37286	10540	1.60	0.03
Min.	4771	10540	-0.98	0.01
	RMSE	True SD	Item Separation	Item Separation Reliability
Real	0.02	0.60	26.91	1.00
Model	0.02	0.60	27.89	1.00

SE of Item Measure Mean = 0.10

FREQUENCY DISTRIBUTIONS

Items

Appendix 14 provides frequency distributions of all HSGQE item difficulties, including the thresholds for CR items. Each item sequence number is shown to the right of its corresponding logit, which represents the lowest possible value for that row. When more than one item falls in the logit range, the items are arranged from lowest to highest logit value. For instance, as seen in Figure 14–1 of the appendix, the logit value for Mathematics Item 23 is between 1.1 and 1.3, and it is also lower than the logit value for Item 32, which is located to the right on the same line. In addition, each CR item sequence number is displayed to the right of its corresponding logit for each possible threshold.

Persons

Appendix 15 provides frequency distributions of raw scores and scale scores by content area for the spring 2010 HSGQE administration. The columns in these tables present each raw score, scale score, scale score asymptotic standard error, frequency count, frequency percent, cumulative frequency, and cumulative percent. The range of reported scale scores for the HSGQE is 100 through 600.

CAUTIONS FOR SCORE USE

As with any assessment, student scores at the minimum or maximum ends of the score range will have large standard errors of measurement and should be viewed cautiously. For instance, if the maximum score for the 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 level of performance, 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.

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 achieve the same score on subsequent testing than it is for the students in the middle of the distribution. If a student who scored 8 out of 40 on a test were to take the same test again, there would be 32 opportunities to correctly answer an item that had been incorrect. There would only be eight opportunities to incorrectly answer items that were answered correctly the first time. If an item is answered differently, it is more likely to increase the student's score than to decrease it. The converse of this is also true for students with very high scores; the next time they test they are more likely to achieve a lower score, and this lower score may be a result of regression toward the mean rather than an actual loss in achievement. Regression toward the mean is a phenomenon apparent with all tests, and caution should be taken when interpreting any scores at extreme of the distribution.

CHAPTER 6: SCALING & EQUATING

INTRODUCTION

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

The spring 2010 HSGQE operational test in mathematics, reading, and writing was constructed by DRC to meet approved HSGQE test blueprints.

An identical form of the spring 2010 HSGQE was initially administered operationally in spring 2006, thus there was no need for equating or calibrating the items.

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. The spring 2010 HSGQE had been originally administered operationally and pre-equated in spring 2006. In the case of the spring 2006 HSGQE, all operational items had been previously field tested in the spring of 2005 (The spring 2005 field test items were appended to the operational form.). 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

Operational item calibration for the spring 2010 HSGQE was conducted in spring 2006, when the current form was initially administered, and reconfirmed following the spring 2010 administration. 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). The fact that these HSGQE field tests were appended means that there is always the potential for changes in administration item position to impact the item difficulties.

Because the spring 2010 test form was pre-equated in spring 2006, 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 2010 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 MC items as well as polytomously scored CR tasks required the use of a partial-credit model. The Newton-Raphson iterative procedure was used to obtain precise ability estimates:

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

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

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

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

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

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

ITEM BANK MAINTENANCE

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

CHAPTER 7: FIELD-TEST ITEM DATA SUMMARY

FIELD-TEST ITEMS

After a newly written item has passed committee review, it is field tested. For the spring 2010 HSGQE, seven field-test forms were administered. Each form contained the same 47 operational mathematics test items, 58 operational reading test items, and 32 operational writing test items. In addition, each form embedded 10 to 15 multiple-choice (MC) items and 1 or 2 constructed-response (CR) items, depending on the content area. Field-test items do not count towards an individual student’s score. Only the operational test items count towards the individual’s score. For the spring 2010 HSGQE, an additional operational form was used for retesters that did not include any field-test items.

Each student taking the HSGQE for the first time took the field-test items, providing a diverse sample of student performance on each field-test item. In addition, because students did not know that field-test items were embedded, no differential motivation effects were expected to have occurred.

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 as the operational test.

FIELD-TEST ITEM DESCRIPTIVE STATISTICS

Appendix 16 provides field-test item statistics by content area for the spring 2010 HSGQE. These statistics represent the item characteristics most commonly used to determine whether an item functioned in an appropriate manner and are the same as those defined in Chapter 5 for operational items.

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

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

Table 7–1. 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 the $\hat{\Delta}_{MH}$ value, one of three severity classification categories was assigned (i.e., A, B, C). Rules for the classification are found in Appendix 17. The A category represents negligible DIF. The B category indicates moderate potential DIF; that is, that one group outperformed the other group once the effects of differences in skill levels between the two groups have been removed. The C category indicates that there is large potential DIF. The plus (+) and minus (-) signs that follow the DIF category indicate which group is favored by the item. The minus sign indicates that the reference group outperformed the focal group once the skill level differences between the groups have been accounted for. The plus sign indicates that the focal group outperformed the reference group once the skill level differences between the groups have been removed.

The analysis on CR items was based on the SMD procedure (Zwick & Thayer, 1996). SMD takes into account the natural ordering of the response levels of the item. In contrast to the MH procedure, this summary statistic compares the means of the reference and focal groups, adjusting for differences in the distribution of each group’s members across the four ability stratifications. Data were organized into a two-by- T -by-four contingency table shown in Table 7–2, 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 SMD statistic was calculated using the equation:

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

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

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

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

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

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

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

One of three severity classification categories was then assigned (i.e., A, B, C). Appendix 17 provides rules for classification.

A summary of DIF results for field-test items is presented in Table 7–3.

Table 7–3. Field-Test Differential Item Functioning (DIF) Summary

Subject	Form	Reference/Focal	Number of Items per DIF Classification					
			A-	A+	B-	B+	C-	C+
Mathematics	1	Male/Female	3	10	1	0	1	0
		White/Alaska Native	12	4	0	0	0	0
	2	Male/Female	8	7	0	0	1	0
		White/Alaska Native	8	5	2	1	0	0
	3	Male/Female	9	6	0	1	0	0
		White/Alaska Native	11	2	2	0	1	0
	4	Male/Female	10	5	1	0	0	0
		White/Alaska Native	9	7	0	0	0	0
	5	Male/Female	7	7	1	1	0	0
		White/Alaska Native	9	5	2	0	0	0
	6	Male/Female	11	3	1	1	0	0
		White/Alaska Native	5	8	2	1	0	0
	7	Male/Female	6	7	1	1	0	1
		White/Alaska Native	11	3	2	0	0	0
Reading	1	Male/Female	5	5	0	0	0	1
		White/Alaska Native	6	2	2	0	1	0
	2	Male/Female	2	5	0	2	0	2
		White/Alaska Native	3	2	3	0	2	0
	3	Male/Female	3	6	1	1	0	0
		White/Alaska Native	5	1	4	1	0	0
	4	Male/Female	3	6	0	2	0	0
		White/Alaska Native	8	2	1	0	0	0
	5	Male/Female	2	9	0	0	0	0
		White/Alaska Native	6	0	0	0	5	0
	6	Male/Female	4	5	0	1	0	1
		White/Alaska Native	6	2	1	0	2	0
	7	Male/Female	2	7	0	1	0	1
		White/Alaska Native	5	2	3	0	1	0
Writing	1	Male/Female	2	8	1	1	0	1
		White/Alaska Native	4	4	4	0	1	0
	2	Male/Female	6	5	0	1	0	1
		White/Alaska Native	5	3	3	0	2	0
	3	Male/Female	3	7	0	1	0	2
		White/Alaska Native	6	3	2	0	2	0
	4	Male/Female	2	9	0	1	0	1
		White/Alaska Native	8	2	2	0	0	1
	5	Male/Female	4	8	0	0	0	1
		White/Alaska Native	7	1	2	2	1	0
	6	Male/Female	4	7	0	0	0	2
		White/Alaska Native	8	0	0	0	5	0
	7	Male/Female	5	7	0	0	0	1
		White/Alaska Native	5	2	2	1	3	0

ITEM BANK MAINTENANCE

Following field-test item calibration and analysis, the item bank was then updated with the new item information. Selected field-test items were then made available for Data Review Committee final appraisal. Once approved, the operational portion of subsequent forms can be constructed from the calibrated item bank.

Item data review for the field-test items administered in spring 2010 was conducted in summer 2010.

CHAPTER 8: SCALE SCORES & PERFORMANCE 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 content areas are identical with a minimum of 100 and a maximum of 600. However, the proficient cut score varies across the three content areas and scores cannot be compared directly across the content areas.

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 (MC) items answered correctly plus the number of points earned on constructed-response (CR) items on a content-area assessment. By itself, the raw score has limited utility; it can only be interpreted in reference to the total number of items on a content-area assessment, and raw scores should not be compared across reporting categories or administrations.

Scale Score

Since a given raw score may not represent the same skill level on every test form, all statewide assessment score reports include scale scores. Scale scores are statistical conversions of raw scores that adjust for slight shifts in item difficulties and permit valid comparison across all test administrations within a particular content area. 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 content-area skill to answer correctly. This depends on the difficulty of the specific questions used on each form. To be fair to students and to permit valid comparison of test scores across administrations, the skills represented by each score point must remain consistent from year to year.

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

TRANSFORMATIONS

As previously discussed, raw scores were transformed into logits in the initial calibration. Logits in turn were mathematically transformed into scale scores to provide a more convenient metric for reporting. To maintain consistency from administration to administration, the minimum scale scores necessary for proficiency were set at 328 for mathematics, 287 for reading, and 304 for writing. Table 8–1 provides the equations and minimum logits used for each transformation. These equations were applied to the overall test as well as to each reporting subscale. Refer to Appendix 14 to locate the logit cut scores compared to item difficulties for each content area.

Table 8–1. Transformation Equations

Content Area	Equation	Logit Cut
Mathematics	Scale Score = $59.8444 (\text{logit} + 0.0046) + 301.0335$	0.4460
Reading	Scale Score = $69.3854 (\text{logit} + 0.3630) + 228.1892$	0.4788
Writing	Scale Score = $55.3838 (\text{logit} + 0.5011) + 229.4855$	0.8378

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

SCALE SCORE SUMMARY STATISTICS

This section includes scale score descriptive information for each overall content-area assessment. Subscale descriptive statistics can be found in Appendix 18. Histograms of the overall-test scale scores are also provided in Figures 8–1 through 8–3. Recall that, as explained in Chapter 5, if the student achieves a scale score of 600, it cannot be determined whether the student would have achieved a higher scale score if that score were possible. All that is known is that the student’s scale score, as revealed by this test, is at least 600. In this manner, extreme scale scores may vary from one administration to the next even if the number of items tested does not, making comparisons of students that score at the extreme ends of the score distribution difficult. The spikes in the histograms occur when two raw score points fall in the same scale score range represented by the bar.

Table 8–2. Content Area Scale Score Information

	Grade 10		
	Mathematics (n=8984)	Reading (n=8978)	Writing (n=8990)
Mean	381.72	359.50	340.24
Standard Error of Mean	0.74	0.79	0.59
Median	376	361	341
Mode	398	411	341
Standard Deviation	69.94	74.98	55.98
	Grade 11		
	Mathematics (n=1058)	Reading (n=672)	Writing (n=1088)
Mean	315.92	295.48	300.12
Standard Error of Mean	1.73	3.21	1.51
Median	308	280	294
Mode	291	251	281
Standard Deviation	56.28	83.32	49.85
	Grade 12		
	Mathematics (n=479)	Reading (n=261)	Writing (n=440)
Mean	311.24	287.99	297.93
Standard Error of Mean	2.67	4.66	2.43
Median	303	280	289
Mode	291	232	294
Standard Deviation	58.38	75.27	51.02

Figure 8–1. Mathematics Scale Score Frequencies

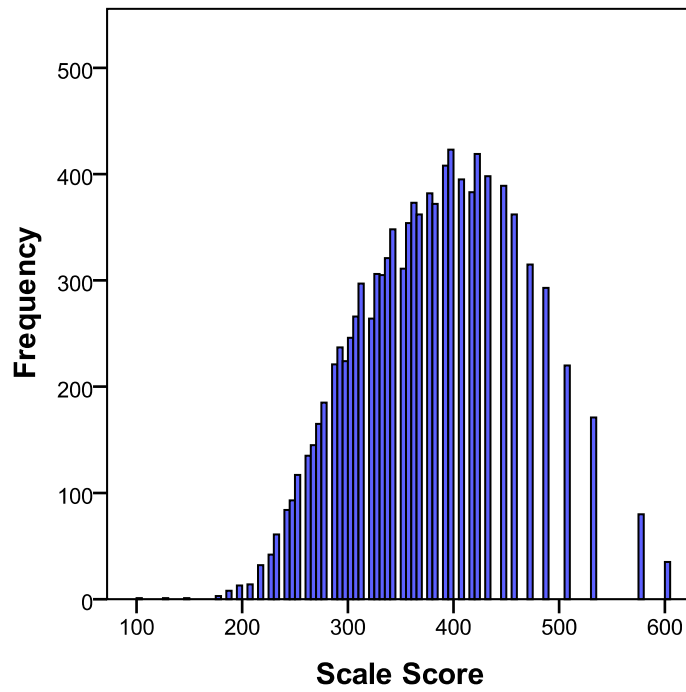


Figure 8–2. Reading Scale Score Frequencies

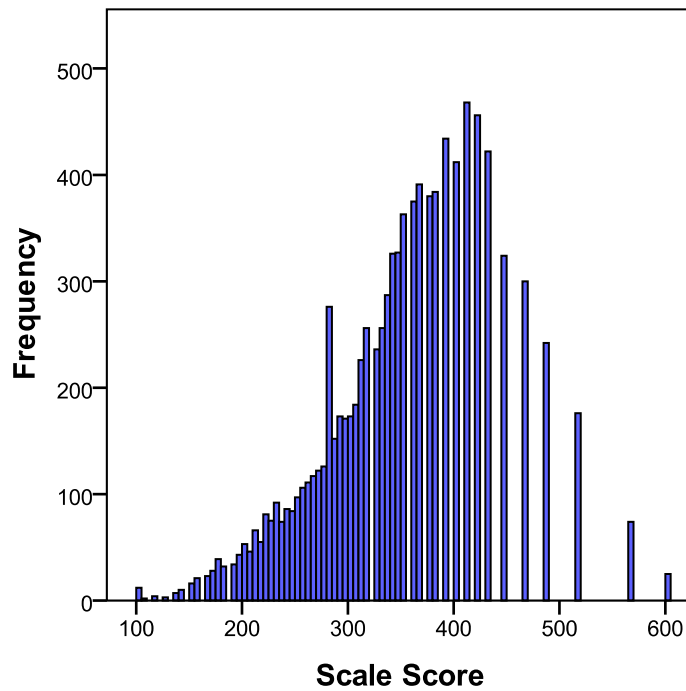
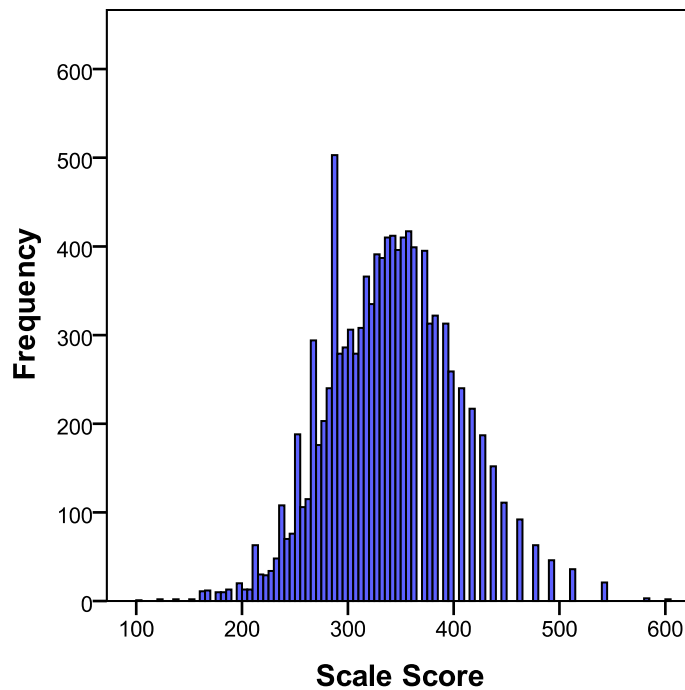


Figure 8–3. Writing Scale Score Frequencies



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 19 provides detailed information about the proficiency level as well as the Proficiency Level Definitions and Descriptors in each content area tested.

Table 8–3 provides the distribution of students in each of the proficiency levels for all content areas. Note that the last column, “All Content Areas,” only pertains to grade 10 students taking the HSGQE for the first time and “Not Proficient” under the “All Content Areas” column indicates the number of students who were classified as “not proficient” in at least one of the three content areas.

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

		Mathematics		Reading		Writing		All Content Areas	
	Grade	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
Not Proficient	10	2054	22.86	1433	15.96	2322	25.83	3053	34.96
	11	718	67.86	347	51.64	640	58.82	-	-
	12	340	70.98	141	54.02	280	63.64	-	-
	All	3161	29.88	1941	19.54	3263	30.94	-	-
Proficient	10	6930	77.14	7545	84.04	6668	74.17	5681	65.04
	11	340	32.14	325	48.36	448	41.18	-	-
	12	139	29.02	120	45.98	160	36.36	-	-
	All	7419	70.12	7993	80.46	7282	69.06	-	-

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 2010 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 content area. Reliability is defined as the consistency of measures. The ability to measure consistently is necessary, but not sufficient, condition for making valid interpretations of the results.

VALIDITY

Content/Curricular

The HSGQE is a criterion-referenced assessment. 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 unacceptable for any reason, the committee can either suggest revisions (e.g., reclassification, rewording) or elect to eliminate the item from the field-test item pool. Items that are approved by the review committee are later embedded in operational 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

For the spring 2010 HSGQE, Alaska educators provided valuable input on the alignment of the items and the statewide content standards during item development. Items were written specifically to measure the objectives and specifications of the content standards for the HSGQE. Because many different people with different backgrounds wrote the items, the process included a built-in system of checks and balances for item development and review that reduced single-source bias. This direct input from educators offers evidence regarding the content validity of the HSGQE. See Chapter 2 for details regarding the content review process.

Developer Input

EED and DRC staff have a history of test-building experience, including content-related expertise that they contributed to the spring 2010 forms. 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, & NCME, 1999) provides the following list of possible sources:

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

Evidence of Construct Validity

The collection of construct-related evidence is a continuous and ongoing process. Two current metrics of construct validity for the spring 2010 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, excluding that item score. Conceptually, if an item has a high item-total correlation (i.e., 0.40 or above), it indicates that students who performed well on the test overall usually answered the item correctly and students who performed poorly on the test overall usually answered the item incorrectly. That is, the item did a good job discriminating between high-scoring and low-scoring 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 2010 HSGQE can be found in Appendix 13. The majority of items have item-total correlations of at least 0.30 (80.58% of items). These high item-total correlations provide evidence for construct validity.

In addition to item-total correlations, Rasch fit statistics also provide good evidence of construct validity. The Rasch model requires unidimensional data. Therefore, statistics showing that the items fit the measurement model also provide evidence of construct validity. Fit statistics for the spring 2010 HSGQE can be found in Appendix 13. In this administration, 71.22% of item fit statistics are below +5.00, indicating good construct validity.

Intercorrelations

A third indicator of construct validity is the intercorrelations between the content area total scale scores and the subscale reporting category scale scores. This information is contained in Appendix 20. In addition, intercorrelations between the scale scores for the three content area total scale scores are presented.

Validity Evidence for Different Student Populations

The primary evidence for the validity of the 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 subpopulation of students over another subpopulation. 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. Table 9–1 presents the student demographic information for the HSGQE.

Table 9–1. Summary of Student Demographics

Demographics	Mathematics		Reading		Writing	
	N	%	N	%	N	%
ALL STUDENTS	10580	100.00	9934	100.00	10545	100.00
ETHNICITY						
White (Caucasian)	5527	52.24	5302	53.37	5504	52.20
African American	420	3.97	370	3.72	404	3.83
Hispanic	584	5.52	527	5.31	569	5.40
Asian/Pacific Islander/Native Hawaiian	941	8.89	882	8.88	942	8.93
Alaska Native and American Indian	2556	24.16	2347	23.63	2585	24.51
Two or more races	551	5.21	505	5.08	540	5.12
Unknown	1	0.01	1	0.01	1	0.01
SOCIOECONOMIC STATUS						
Not Low Income	6448	60.95	6143	61.84	6426	60.94
Low Income	4132	39.05	3791	38.16	4119	39.06
ENGLISH PROFICIENCY STATUS						
English Proficient	9252	87.45	8762	88.20	9213	87.37
Limited English Proficient	1328	12.55	1172	11.80	1332	12.63
MIGRANT STATUS						
Non-Migrant	9957	94.11	9336	93.98	9912	94.00
Migrant	623	5.89	598	6.02	633	6.00
SPECIAL EDUCATION STATUS						
Regular Education	9243	87.36	8786	88.44	9203	87.27
Individualized Education Plan	1337	12.64	1148	11.56	1342	12.73
GENDER						
Female	5202	49.17	4783	48.15	4955	46.99
Male	5378	50.83	5151	51.85	5590	53.01
ACCOMMODATIONS						
Total	2	100.00	5	100.00	5	100.00
Braille	1	50.00	3	60.00	2	40.00
Large Print	1	50.00	2	40.00	3	60.00

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

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

$$X = T + E, \quad (1)$$

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

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

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

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

$$\Sigma p(1 - p), \quad (3)$$

where p is the proportion correct for each item.

The reliability index used for the 2010 administration of the HSGQE was Coefficient Alpha (Cronbach, 1951):

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

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

Acceptable α values generally range in the high 0.80s to low 0.90s. When there is no error, the reliability index is the true score variance divided by the true score variance, which is one. Tables 5–2 through 5–7 provide Coefficient Alpha for each content area. As can be seen in the tables, mathematics, reading, and writing have Coefficient Alphas of 0.91, 0.92, and 0.90, respectively. These high α values provide evidence for good reliability.

Standard Error of Measurement

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

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

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

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

Ability estimates from scores near the center of the test are known with greater precision than are abilities associated with extremely high or low scores. The expression for computing the asymptotic standard error via WINSTEPS was provided in Chapter 6. This value is then transformed to the HSGQE scale to obtain the final SEM for each raw score. These values for the spring 2010 HSGQE are provided in the raw-to-scale score tables in Appendix 15. In addition, person separation reliability and item separation reliability values, which use these asymptotic standard errors are provided in Tables 5–2 through 5–7. Person separation reliability is the Rasch equivalence of reliability described in Equation 2.

Indicators of Consistency

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

Decision Consistency Index

Method I

In a personal communication to DRC from Dr. Huynh Huynh on the DRC/South Carolina project, an extension of the two-parameter beta-binomial model (Huynh, 1976) to polytomous constructed-response items was detailed. The extension was used in these computations. Table 9–2 depicts the general framework of binary decisions.

Table 9–2. 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} is the proportion of examinees consistently classified as proficient on the basis of test scores obtained from both Form X and Form Y and p_1 is the proportion of examinees classified on the basis of test scores obtained from either form.

Method II

To solve the problem of a complex assessment (i.e., including partial credit items), Livingston and Lewis (1995) proposed a consistency index that first requires the calculation of an effective test length, n . This calculation transforms the original raw score random variable from $X = 0, \dots, K$ into a new random variable $X' = 0, \dots, n$, where n is the number of dichotomous, locally independent, equally difficult items required to produce a raw score of the same reliability. Then, using the transformed observed distribution X' , parameters are estimated for a four-parameter beta-binomial model where the conditional error distribution is assumed to be binomial. The X' distribution is then converted back onto the original X scale using interpolation. This method is designed only to estimate a contingency table, not a full bivariate distribution which means the probability of a consistent decision by chance, and subsequently kappa, cannot be estimated.

The results of both consistency analyses are presented in Table 9–3.

Table 9–3. Decision Consistency Indices

Content Area	Huynh (1976)		Livingston and Lewis (1995)
	Consistency Index	κ	Consistency Index
Mathematics	0.88	0.71	0.89
Reading	0.92	0.74	0.93
Writing	0.86	0.68	0.88

REFERENCES

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (2nd ed.). Washington, DC: American Educational Research Association.
- Bloom, B. S., Englehart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. 1956. *Taxonomy of Educational Objectives: The classification of educational goals: Handbook 1: Cognitive Domain*. New York: Longman, Green, and Co.
- Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.
- Holland, P., & Thayer, D. (1986, April). *Differential item performance and the Mantel-Haenszel procedure*. Paper presented at the American Educational Research Association annual meeting, San Francisco, CA.
- Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement*, *13*, 253–64.
- Linacre, J. M. (2010). WINSTEPS Rasch measurement (Version 3.69.1.14). [Computer program]. Chicago: WINSTEPS.com.
- Linacre, J. M. (2009). WINSTEPS Rasch measurement (Version 3.68). [Computer program]. Chicago: WINSTEPS.com.
- Linn, R., & Gronlund, N. (1995). *Measurement in assessment and teaching* (7th ed.). Upper Saddle River, NJ: Prentice-Hall.
- Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, *32*, 179–197.
- Mead, R. J. (1978). Examining residuals from the Rasch model. *Proceedings of the 1978 conference on adaptive testing*. Minneapolis, MN: University of Minnesota.
- Mogilner, A. (1992). *Children's Writer's Word Book*. Cincinnati, OH: Writer's Digest Books.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research (Expanded edition, 1980). Chicago: University of Chicago Press.
- Smith, R. M. (2000). Fit analysis in latent trait measurement models. *Journal of Applied Measurement*, *1*, 199–218.
- Stout, W. (1987). A non-parametric approach to assessing latent trait unidimensionality. *Psychometrika*, *52*, 589–617.
- Taylor, S. E., Frackenpohl, H., White, C. E., Nieroroda, B. W., Browning, C. L., & Brisner, E. P. (1989). *EDL Core Vocabularies in Reading, Mathematics, Science, and Social Studies*. Orlando, FL: Steck-Vaughn Company.

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

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

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

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