

Alaska Alternate Assessment
2007 – 2008
Reading – Writing – Mathematics
Technical Report

Dillard Research Associates
August 1, 2008

GLOSSARY	7
EXECUTIVE SUMMARY	8
CHAPTER 1: BACKGROUND OF THE ALASKA ALTERNATE ASSESSMENT	
<i>Historical Perspective: Designing the Original Alaska Alternate Assessment</i>	10
<i>The Rationale for Redesigning the Alaska Alternate Assessment</i>	11
<i>Organization of Technical Report</i>	12
CHAPTER 2: FIELD TEST ITEM DATA SUMMARY AND BIAS REVIEW	
<i>Overview</i>	13
<i>Results from 2006-2007</i>	13
<i>Standard Setting</i>	13
<i>Statistical Analyses</i>	13
<i>Reading</i>	14
<i>Writing</i>	14
<i>Mathematics</i>	14
<i>Summary</i>	14
CHAPTER 3: TEST DESIGN AND ITEM/TASK DEVELOPMENT	
<i>Overview</i>	15
<i>Special Considerations for 2008</i>	15
<i>Reading</i>	15
<i>Writing</i>	15
<i>Mathematics</i>	15
<i>Extended Grade-Level Expectations and Proficiency Level Descriptors</i>	16
<i>Alignment Study from 2006-2007</i>	16
<i>Test Blueprint</i>	16
<i>Bias Review</i>	16
<i>Test Design, Development, and Score Reporting Categories Blueprint for Item Writing</i>	18
<i>Background</i>	18
<i>Test and Task Description</i>	18
<i>Teacher and Student Materials</i>	19
<i>Test Design for Students with Significant Cognitive Disabilities</i>	19
<i>Reduction in Complexity</i>	19
<i>Development Steps Toward Reducing Complexity</i>	20
<i>Administration Steps Toward Reducing Complexity</i>	20
<i>Reduction in Depth</i>	21
<i>Categorical Concurrence, Range of Knowledge, and Balance of Representation</i>	21
<i>Reduction in Breadth</i>	22
<i>Test Development</i>	22
<i>Expanded Levels of Support (ELOS)</i>	22
<i>Content Prompts</i>	23
<i>Specifications for Item Writing</i>	24

<i>Background and Overview</i>	24
<i>Item Writing Training</i>	24
<i>Alignment of Test Items to Grade Level Expectations</i>	24
<i>Correct Key Placement</i>	24
<i>Item Distribution by Difficulty</i>	24
<i>Item Characteristics</i>	24
<i>Item Writing Criteria</i>	25
<i>ELOS and Standard Test Items</i>	26

CHAPTER 4: TEST ADMINISTRATION PROCEDURES

<i>Overview</i>	28
<i>Student Population Tested</i>	28
<i>Accommodations</i>	28
<i>Test Administrators</i>	29
<i>Test Administrator Training</i>	29
<i>Scorer Training and Qualification – Online Proficiency</i>	29
<i>Scoring Materials and Process</i>	30
<i>Quality Control of Scoring – Reliability of the Alternate Assessment Administration and Scoring Process: Training to Become a Qualified Assessor</i>	30
<i>Training to become a Qualified Mentor Trainer</i>	31
<i>Qualified Mentor Trainers (QT or Mentors) Additional Responsibilities</i>	31
<i>Quality Assurance of Test Development, Administration, and Scoring</i>	33
<i>Test Administrator Training Agendas – Reading, Writing, Mathematics (October 15-16, 2008)</i>	33
<i>Test Security and Administration QA</i>	35
<i>Content of Training – Power point Slides</i>	35
<i>Teacher Participation Guide</i>	35
<i>Training Manual</i>	35

CHAPTER 5: SCORING

<i>Overview</i>	36
<i>Data Entry</i>	36
<i>ELOS Only</i>	36
<i>Standard Administration With or Without Accommodations AND Then Switched to the ELOS</i>	37
<i>Standard Administration With or Without Accommodations</i>	37

CHAPTER 6: STANDARDS VALIDATION

<i>Overview</i>	39
<i>Proficiency Level Descriptors</i>	39
<i>Plan for Standards Validation Process</i>	39
<i>Test Centered Validation</i>	40
<i>Person Centered Validation</i>	40
<i>Training on Test Administration and Scoring</i>	40

<i>Comparison of Items and Tasks</i>	41
<i>Analysis of Student Profiles on 2008 Tasks</i>	41
<i>Training Presentation</i>	41
<i>Evaluations of Standards Validation</i>	41
<i>Test Centered Analysis by Teachers on Linking Items</i>	43
<i>Item difficulty</i>	43
<i>Reading</i>	43
<i>Writing</i>	43
<i>Mathematics</i>	43
<i>Standards Validation Comparing Linking Items in 2007 and 2008</i>	44
<i>Linking items</i>	44
<i>Reading</i>	44
<i>Writing</i>	44
<i>Mathematics</i>	45
<i>Person Centered Analysis by Teachers on Proficiency Confirmation</i>	45
<i>Reading</i>	45
<i>Writing</i>	45
<i>Mathematics</i>	46
<i>Projected Impact Data</i>	46

CHAPTER 7: REPORTING

<i>Overview</i>	47
<i>Report Types</i>	47
<i>Unofficial Student Report</i>	47
<i>Official Report</i>	47
<i>Parent Guide to Interpretation of the Individual Student Reports</i>	47
<i>Educator Guide to Interpretation of the Individual Student Reports</i>	47
<i>DRA Secure Reporting Website</i>	48

CHAPTER 8: TEST VALIDITY

<i>Overview</i>	49
<i>Data Analysis Summary</i>	49
<i>Scaling 2008 onto 2007 Scale and Analyzing Impact</i>	50
<i>Reliability</i>	52
<i>Validity</i>	56
<i>Proficiency Comparison between 2007-2008</i>	56

CHAPTER 9: PROGRAM IMPROVEMENT

<i>Overall Program Evaluation</i>	62
<i>Summary of Consequences Survey</i>	62
<i>Training and Qualifications</i>	62
<i>Test Administration and Decision Making</i>	62
<i>Results</i>	63
<i>Instructional Relevance</i>	63

<i>Professional Developmental Needs</i>	63
<i>Teacher Demographics and Experiences</i>	63
<i>Recommendations for Future Consideration</i>	64
<i>Field Testing of New Items (Standard and ELOS)</i>	64
<i>Package of Test Booklets and Training of Teachers</i>	64

APPENDICES

<i>Appendix 2_1 – Technical Manual</i>	13
<i>Appendix 2_2 – Standard Setting</i>	13
<i>Appendix 2_3 – Descriptive Statistics in Reading</i>	14
<i>Appendix 2_4 – Descriptive Statistics in Writing</i>	14
<i>Appendix 2_5 – Descriptive Statistics in Mathematics</i>	14
<i>Appendix 3_1 – Development of Extended Grade Level Expectations and Proficiency Level</i> <i>Descriptors</i>	16
<i>Appendix 3_2 – Alaska Alternate Assessment Alignment Study Report</i>	16
<i>Appendix 3_3 – Extended Grade Level Expectations Cross Walks</i>	16
<i>Appendix 3_4 – Reading Writing and Mathematics Bias Review Comments</i>	17
<i>Appendix 4a_1 – 2008 Reading Scoring Protocol – ELOS Items</i>	27
<i>Appendix 4a_2 – 2008 Reading Student Materials – ELOS Items</i>	27
<i>Appendix 4a_3 – 2008 Writing Scoring Protocol – ELOS Items</i>	27
<i>Appendix 4a_4 – 2008 Writing Student Materials – ELOS Items</i>	27
<i>Appendix 4a_5 – 2008 Mathematics Scoring Protocol – ELOS Items</i>	27
<i>Appendix 4a_6 – 2008 Mathematics Student Materials – ELOS Items</i>	27
<i>Appendix 4a_7 – 2008 Reading Scoring Protocol – Standard Items</i>	27
<i>Appendix 4a_8 – 2008 Reading Student Materials – Standard Items</i>	27
<i>Appendix 4a_9 – 2008 Writing Scoring Protocol – Standard Items</i>	27
<i>Appendix 4a_10 – 2008 Writing Student Materials – Standard Items</i>	27
<i>Appendix 4a_11 – 2008 Mathematics Scoring Protocol – Standard Items</i>	27
<i>Appendix 4a_12 – 2008 Mathematics Student Materials – Standard Items</i>	27
<i>Appendix 5_1 – Qualification Forms and Process</i>	33
<i>Appendix 5_2 – Alaska Alternate Assessment Training Report</i>	33
<i>Appendix 4b_1 – Training Power point slides</i>	35
<i>Appendix 4b_2 – Teacher Participation Guide</i>	35
<i>Appendix 4b_3 – Reading Training Manual</i>	35
<i>Appendix 4b_4 – Writing Training Manual</i>	35
<i>Appendix 4b_5 – Mathematics Training Manual</i>	35
<i>Appendix 6_1 – Reading, Writing, and Mathematics PLDs Grade band 3/4</i>	39
<i>Appendix 6_2 - Reading, Writing, and Mathematics PLDs Grade band 5/6</i>	39
<i>Appendix 6_3 - Reading, Writing, and Mathematics PLDs Grade band 7/8</i>	39
<i>Appendix 6_4 - Reading, Writing, and Mathematics PLDs Grade band 9/10</i>	39
<i>Appendix 6_5 – Standards Validation Power point slides</i>	41
<i>Appendix 6_6 – Test Centered Judgments</i>	44
<i>Appendix 6_7 – Linking Items – Task Comparisons</i>	45
<i>Appendix 6_8 – Person Centered Judgments</i>	46
<i>Appendix 6_9 – 2007 Impact Data</i>	46

<i>Appendix 7_1 – Example Unofficial Individual Student Report</i>	47
<i>Appendix 7_2 – Example Official Individual Student Report</i>	47
<i>Appendix 7_3 – Parent Guide to Test Interpretation</i>	47
<i>Appendix 7_4 – Educator Guide to Test Interpretation</i>	48
<i>Appendix 7_5 – DRA Secure Reporting Website User Guide</i>	48
<i>Appendix 8_1 – 2008 Scaled to 2007 with Cut Scores</i>	56
<i>Appendix 8_2 – Reading Item Statistics</i>	57
<i>Appendix 8_3 – Writing Item Statistics</i>	58
<i>Appendix 8_4 – Mathematics Item Statistics</i>	61
<i>Appendix 9_1 – Teacher Survey of Consequential Validity Results</i>	62

FIGURES

<i>Figure 1 – Model of Validation</i>	9
<i>Figure 2 – Linear Equating</i>	50

TABLES

<i>Table 1 – Bias and Sensitivity Review Participants</i>	17
<i>Table 2 – Alignment Dimensions</i>	21-22
<i>Table 3 – October 15 Agenda – New Mentors</i>	34
<i>Table 4 – October 15 Agenda – All Mentors</i>	34
<i>Table 5 – October 16 Agenda – All Mentors</i>	34-35
<i>Table 6 – Standards Validation Participants</i>	39
<i>Table 7 – Standards Validation Evaluation</i>	42
<i>Table 8 – Standards Validation Comments</i>	42-43
<i>Table 9 – Cronbach’s Alpha - Reading</i>	52
<i>Table 10 – Reading - Grade Band 3/4</i>	52
<i>Table 11 – Reading - Grade Band 5/6</i>	52
<i>Table 12 – Reading - Grade Band 7/8</i>	52
<i>Table 13 – Reading - Grade Band 9/10</i>	52
<i>Table 14 – Cronbach’s Alpha – Writing</i>	53
<i>Table 15 – Writing Grade Band 3/4</i>	53
<i>Table 16 – Writing Grade Band 5/6</i>	53
<i>Table 17 – Writing Grade Band 7/8</i>	53
<i>Table 18 – Writing Grade Band 9/10</i>	53
<i>Table 19 – Cronbach’s Alpha – Mathematics</i>	54
<i>Table 20 – Mathematics Grade Band 3/4</i>	54
<i>Table 21 – Mathematics Grade Band 5/6</i>	54
<i>Table 22 – Mathematics Grade Band 7/8</i>	54
<i>Table 23 – Mathematics Grade Band 9/10</i>	55

Glossary

ADV	Advanced Proficient
AA	Alternate Assessment
AAS	Alternate Achievement Standards
AIT	Assessor-In-Training
AT-AAC	Assistive Technology-Augmentative Alternative Communication
AYP	Adequate Yearly Progress
BP	Below Proficiency
CLS	Correct Letter Sequences
CNS	Correct Number Sequences
CWS	Correct Word Sequences
DOK	Depth of Knowledge
DRA	Dillard Research Associates
DTC	District Test Coordinator
EED	Early Education Department
ELOS	Expanded Levels of Support
ExGLEs	Extended Grade Level Expectations
FAQ	Frequently Asked Questions
FB	Far Below Proficiency
GLEs	Grade Level Expectations
IEP	Individualized Education Plan
ISR	Individual Student Report
NA-I	Not Administered-Inappropriate
NT	Not Tested
P	Proficient
PLD	Proficiency Level Descriptor
QA	Qualified Assessor
QT	Qualified Mentor Trainer
RWM	Reading, Writing, Mathematics
SBA	Standards Based Assessment
SEM	Standard Error of Measurement
SPGLEs	Science Performance Grade Level Expectations
STD	Standard
TAC	Technical Advisory Committee
TSA	Test Security Agreement

EXECUTIVE SUMMARY

As elaborated by Messick (1989)¹ a validity argument involves a claim with evidence evaluated to make a judgment. Three essential components of assessment systems: constructs (what to measure), the assessment instruments and processes (approaches to measurement), and use of the test results (for specific populations). To put it simply, validation is a judgment call on the degree to which each of these components is clearly defined and adequately implemented.

Validity is a unitary concept with multifaceted processes of reasoning about a desired interpretation of test scores and subsequent uses of these test scores. In this process, we want answers for two important questions. Regardless of whether the students tested have disabilities, the questions are identical:

1. How valid is our interpretation of a student’s test score?
2. How valid is it to use these scores in an accountability system?

Validity evidence may be documented at both the item and total test levels. We use the *Standards*² (AERA et al., 1999) in documenting evidence on content coverage, response processes, internal structure, and relations to other variables.

This document follows the essential data requirements of the federal government as needed in the peer review.³ The critical elements highlighted in that document (with examples of acceptable evidence) include (a) academic content standards, (b) academic achievement standards, (c) a statewide assessment system, (d) validity, (e) reliability, and (f) other dimensions of technical quality. We address the latter four requirements noted above, with other documents providing essential information on the standards and statewide assessment system (see technical specifications and alignment documents for information on academic content standards and the standard setting document for information on the academic achievement standards). In addressing technical documentation, we first present content evidence, then reliability, and finally address the other three areas noted in the peer review guidance: internal structures, criterion relations, and response processes.

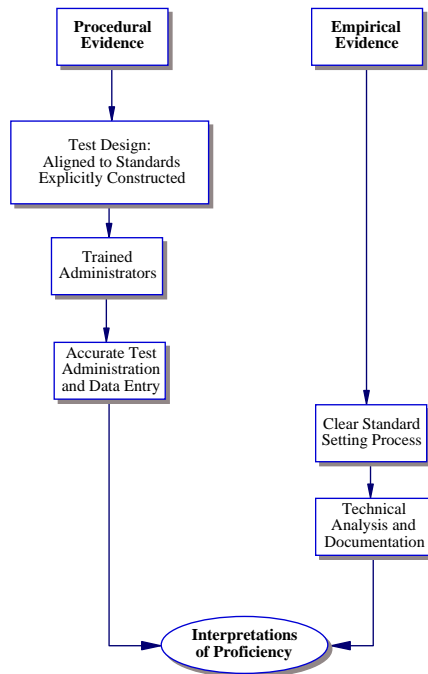
In the end, both procedural and empirical evidence are brought to bear for supporting the claim that students with significant cognitive disabilities are achieving at various levels of proficiency on the alternate assessment.

¹ Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: American Council on Education.

² American Educational Research Association (AERA), American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.

³ U. S. Department of Education (2004). *Standards and Assessments Peer Review Guidance: Information and Examples for Meeting Requirements of the No Child Left Behind Act of 2001*.

Figure 1. Model of Validation



We associate both types of evidence with each of the criteria in the peer review.

- 4.1a. Purpose of assessment and types of uses and decisions (Chapter 1)
- 4.1b. Intended and unintended consequences (Chapter 9)
- 4.1c. Scoring and reporting consistent with content standards (Chapter 8)
- 4.1d. Test and item scores related to internal or external variables (Chapters 7-8)
- 4.1e. Standard setting: select judges, describe methods, and report results (Chapter 6)

- 4.2a. Reliability of the scores (Chapters 4-5 and 8)
- 4.2b. Reporting of conditional SEM (Chapters 7 and 8)
- 4.2c. Evidence of generalizability for all relevant sources (Chapter 8)

- 4.3a. Assessment is fair and accessible with accommodations (Chapters 3-4)
- 4.3b. Linguistic accommodations for ELL (Chapters 3-4)
- 4.3c. Bias review of items (Chapters 2-3)
- 4.3d. Accommodations yield meaningful scores (Chapter 7)

- 4.4a. Consistency of test forms over time (Chapter 3)
- 4.4b. Comparability of on-line and paper-pencil (Not applicable)
- 4.4c. Clear criteria for administration, scoring, analysis, and reporting (Chapters 4-5)
- 4.4d. Monitoring and improving the quality of the assessment (Chapter 9)

- 5.1. Alignment (See Chapter 3)

CHAPTER 1: BACKGROUND OF THE ALASKA ALTERNATE ASSESSMENT

Historical Perspective: Designing the Original Alaska Alternate Assessment

The original design of Alaska's Alternate Assessment, a Student Portfolio, was intended to provide an accountability measure that was consistent with state standards, individualized, performance-based, used independent and reliable scoring, and could be integrated with curriculum and the student's Individualized Education Program (IEP).

The need for developing an alternate assessment was in line with the requirements of the Goals 2000 and Improving America's Schools Act, the Individuals with Disabilities Education Act of 1997, as well as Alaska's Quality Schools Initiative which supported high standards, statewide assessments, and improved results for all students. The goal was to encourage states and districts to move in the direction of inclusive, standards-based IEPs for students with disabilities, including students with the most severe disabilities.

The development of Alaska's Alternate Assessment Student Portfolio was a collaborative effort between the Alaska Department of Education and Early Development, CTB-McGraw Hill, and members of the Statewide Alternate Assessment Stakeholder's Committee. The assessment was developed as part of Alaska's Comprehensive System of Student Assessments. Students were to participate in the Alternate Assessment in grades 3, 6, and 8 at the same time their peers were taking the Benchmark exams. In high school, eligible students participated in the Alternate Assessment in grade 11.

The development process included a Pilot Study with teacher-parent teams, which was completed in February of 2000. The assessment was field tested in the fall of 2000 with students in grades 3 and 11. Full implementation was scheduled for the 2001-2002 school year for all eligible students. The Alaska Alternate Assessment Student Portfolio remained in place through the 2005-2006 school year.

Scoring of the Student Portfolios was accomplished at scoring sessions. The department facilitated scoring sessions, with Alaska teachers trained as table leaders and scorers. Eventually, the department contracted the scoring to Data Recognition Corporation (DRC). Independent raters were trained using exemplars of each score and a scoring rubric. Scorers evaluated the evidence and data presented against the dimensions of the scoring rubric:

- Student Skill - how well the student performs the objective, to what extent the student is independent or requires prompts and assistance, and how much progress over time is evident.
- Generalization - the extent to which the objective is demonstrated in more than one environment or situation with different people (3-4 settings required).
- Appropriateness - the extent to which objectives were age-appropriate, challenging, authentic, and meaningful for the student.

The portfolio evidence (data collection and other evidence) was rated in each dimension and numerically rated (1-4) as: Advanced, Proficient, Below Proficient, Far Below Proficient. Clearly articulated rules that further explained how to score the evidence against the dimensions of the scoring rubric guided scorers in their evaluation.

The Rationale for Redesigning the Alaska Alternate Assessment

There was pressure to change the format and test window of the Alternate Assessment from the teachers as well as from the department. Teachers wanted an assessment test window that more closely matched the general education assessment window; the Portfolio assessment window was 6-9 months. The state also conducted a survey of teachers as to their experiences, both positive and negative, with the portfolio. Teachers administering the portfolio assessment felt that the assessment inaccurately measured student abilities and instead measured the teacher's ability to assemble a convincing portfolio. Intended consequences included increased inclusion of students with significant cognitive disabilities in general education classrooms as well as teacher awareness of state content standards and the need to develop IEP goals and objectives that aligned with these standards. Despite ongoing training in how to write IEP objectives that aligned with the content standards, a review during scoring sessions of the objectives written by IEP teams indicated need for a more defined set of content standards, as many objectives were unaligned with the original content standards.

The state had developed the Grade Level Expectations (GLEs) for general education assessments. Before the development of the GLEs, the state content standards (called Performance Standards) were by age span. The Alternate Performance Standards had to be changed to reflect the change in the general education academic standards, which would resolve the issue of the overly broad Alternate standards. The existing proficiency level descriptors for the Alternate were universal descriptors, and the department wanted to develop grade-level proficiency level descriptors for the Alternate. The department assembled teams of content and special education experts, as well as other stakeholders, for the purpose of developing Extended Grade Level Expectations (ExGLEs) and grade-level Proficiency Level Descriptors (PLDs).

The state contracted with Dr. Gerald Tindal to conduct a Reliability and Validity Study in 2005. The evaluation determined that there was a need for revision of the Student Portfolio in order to meet the requirement for high technical quality required in the No Child Left Behind legislation. The study results recommended that standardized performance tasks be included in the portfolio to stabilize the comparability of assessment results between students. Additionally, the department felt a new standard setting was needed as the portfolio had undergone some revisions since its inception. A more reliable system for training teachers in the field was one of the department goals.

In the fall of 2005, the department issued a Request for Proposals and awarded a contract to Dillard Research Associates to secure a standardized performance-task alternate assessment for students with significant disabilities that included an online test administrator training program to provide greater reliability in the administration and scoring of the assessment. The goals of these changes were to:

- ensure that students are accessing the academic content standards at their grade level by the use of the Extended Grade Level Expectations which are aligned to the test items;
- assess student's achievement based on the academic content reflected in the new grade level proficiency level descriptors;
- provide timely instructional feedback; and
- meet the NCLB condition that Alternate Assessments include the same technical adequacy required of general assessments.

These new Alternate Assessments are standardized, performance tasks administered and scored by assessors who undergo a multiple step qualification process. IEP teams make a determination whether a student is eligible to take the Alternate Assessment by following the guidelines in Alaska's *Participation Guidelines For Alaska Students in State Assessments*, September 2007 edition. After administering the assessments one-on-one to a student, assessors enter student demographic information and scores into an online scoring and reporting system. An unofficial student report is immediately generated for the purpose of providing instructional feedback and guidance to IEP teams. Official student reports that have had the demographic information checked for accuracy and have been assigned proficiency levels are mailed by the department to districts in the summer.

The Reading, Writing, and Mathematics Student Portfolio Alaska Alternate Assessments were approved in 2006 by the United States Department of Education through the peer review process. The science assessment is currently being submitted for approval.

Organization of Technical Report

In the remainder of this technical report, the following topics are addressed: (a) test design and item/task development, (b) field testing and item analysis to direct current test and technical documentation, (c) test administration procedures, (d) scoring, (e) standard setting, (f) reporting, (g) technical documentation, and (h) program improvement.

CHAPTER 2: FIELD TEST ITEM DATA SUMMARY

Overview

This chapter presents the item data for the field test items on the 2006-2007 test. All items for the 2006-2007 test were new, or field test, items. This chapter gives an overview of the item data on the 2006-2007 test. The appendices contain the descriptive statistics for all items and tasks.

Results from 2006-2007

All tasks displayed high to very high internal consistency reliability coefficients, with only a few exceptions. In reading, two tasks reflected high-moderate levels. In writing, one task reflected moderate levels. In mathematics, many of the coefficients below the high-moderate level appeared in tasks involving skills far below or far above grade level.

Appendix 2_1 contains 2006-2007 Alaska Alternate Assessment Technical Documentation. This documentation provides domain-sampling plans, administration summaries, item and task level functioning, subject area strand analysis, and summaries, classification, and validity on standard setting, impact data, and scaling analysis.

< *Appendix 2_1* >

Standard Setting

A standard setting process was held May 1-3, 2007 to set the proficiency levels for each grade band 3/4, 5/6, 7/8 and 9/10 in reading, writing and mathematics. Each subject-area committee consisted of eight professionals, including special education teachers, curriculum specialists and one special education administrator. The standard setting process involved training on test administration, review of the materials and tasks, review of the proficiency level descriptors (PLDs), establishment of proficiency levels for each task and each level, review of an example data set, confirming the reasoning of each task, devising a combined judgment for each proficiency level, completing a standard setting form for each grade level, evaluating the standard setting process, and reviewing impact data.

Appendix 2_2 contains a comprehensive document of all materials used in the standards setting. These materials consist of, agendas, participant lists, sample booklets, and forms.

< *Appendix 2_2* >

Statistical Analyses

Overall, the tasks in all three subject areas showed high levels of internal consistency reliability. All items in all three subject areas exhibited an appropriate range of mean scores, indicating an appropriate range from easy to difficult among the items. The items in all three subject areas also exhibited adequate levels of inter-item correlations, indicating that the items within each task appear to measure related – but not too similar – skills and understandings. Some items showed

higher standard deviations than means, indicating that those items may need clarification in future versions.

Reading

Overall, the tasks showed high levels of internal consistency reliability, an appropriate range of mean scores, and adequate levels of inter-item correlation within tasks. Some low inter-item correlations existed in Tasks 1 and 2 (identify pictures/representation of objects and identify signs and symbols). Task 4 (comprehend oral text) included two items with standard deviations above the mean scores. Task 12E (comprehend printed text) had one item with the standard deviation above the mean.

Writing

Overall, the tasks showed generally high levels of internal consistency reliability, an appropriate range of mean scores, and adequate levels of inter-item correlation within tasks. Tasks 3 and 6 (identify/copy sentences and write sentences from dictation) had high inter-item correlations. Task 7 (sentence mechanics) had low inter-item correlations.

Mathematics

Overall, the tasks showed high levels of internal consistency reliability, an appropriate range of mean scores, and adequate levels of inter-item correlation within tasks. Task 3 (identify shapes) had two items with higher standard deviations than means, and some low inter-item correlations. Tasks 6, 9, and 15 (manipulate mathematical concepts – measurement, number line, and manipulate mathematical concepts - quantity) each had one item with a higher standard deviation than mean. Tasks 7 and 14 (identify money and manipulate mathematical concepts – take away) had some low inter-item correlations. Task 11 (order numbers) had only one item so no statistics were computed. Tasks 13, 16, 17, and 18 (calendar, manipulate mathematical concepts – fractions, count money, and manipulate mathematical concepts – place value) each had three items with higher standard deviations than means, and some low inter-item correlations. Tasks 20, 21, and 22 (computation – addition facts, computation – subtraction facts, and mixed computation – addition and subtraction) did not have individual items recorded.

Summary

As a whole, the tests in each subject area performed very well, with high internal consistency reliability, appropriate ranges of mean scores, and adequate levels of inter-item correlation within tasks.

Appendix 2_3 through 2_5 contain descriptive statistics of all grades for each subject area: reading, writing, and mathematics.

< Appendix 2_3 through 2_5 >

CHAPTER 3: TEST DESIGN AND ITEM/TASK DEVELOPMENT

Overview

In this chapter, general information on test design and item development is given. The Extended Grade-Level Expectations are introduced, as well as the test specifications and blueprint. Development of items and tasks are summarized along with an explanation of the process of test construction.

Special Considerations for 2008

The 2008 test was constructed by closely aligning the items to the Extended Grade-Level Expectations (ExGLEs). Along with this alignment, the assessments were also grade banded to appropriately assess the ExGLEs.

Reading

The reading assessment is designed to measure essential reading skills. The tasks measure the degree to which students with significant cognitive disabilities are learning to read at the symbol, word, and text levels. The tasks increase in complexity with each grade band and include: identification of pictures, symbols, and letters in the alphabet, identification of own name, distinguishing sounds, generating sounds of letters, reading simple words to more complex words, reading sentences, reading text, comprehending text, obtaining information, and identification of root words.

Writing

The writing assessment is designed to measure skill acquisition in written language development for students with significant cognitive disabilities. The tasks measure the degree to which students with significant cognitive disabilities are learning to write using letters, words, and connected sentences. The tasks increase in complexity with each grade and include the following: copy letters, copy words, copy sentences; write their name, write words from dictation, sentence mechanics, write a sentence, write a story, and revise writing.

Mathematics

The mathematics assessment is designed to measure the degree to which students with significant cognitive disabilities have developed numerical understanding. The tasks measure the degree to which students with significant cognitive disabilities are learning to use numbers and mathematical symbols as well as solve problems. The tasks increase in complexity with each grade and include: copying numbers, identifying numbers on a number line, counting, identifying same and different, identifying and matching shapes, reading and writing numbers, counting objects, simple and double digit addition, subtraction, and multiplication, reproducing and extending simple patterns and identifying skip patterns, reading and creating simple graphs, identifying measurement, counting and identifying money, identifying perimeter, identifying fractions, labeling a set as none or zero, understanding symbols, identifying place value, ordering numbers, rounding numbers, and identifying lines of symmetry.

Extended Grade Level Expectations and Proficiency Level Descriptors

Development of the Proficiency Level Descriptors was conducted by the Alaska Education Department. The development process consisted of construction of the proficiency level descriptors by EED staff followed by review and feedback from committees of teachers and stakeholders. The development report provides background information on proficiency level descriptors, outlines the process used, and provides agendas and presentation slides used during the development process.

Appendix 3_1 contains a report explaining the development of the Extended Grade Level Expectations (ExGLEs) and the Proficiency Level Descriptors (PLDs). The report contains historical background, overviews, and the process used in development.

< *Appendix 3_1* >

Alignment Study from 2006 – 2007

An alignment study was conducted in April of 2007. Items in reading, writing, and mathematics were analyzed in accordance to how closely they aligned with the ExGLEs. The Alignment Report summarizes the process used and presents the alignment results of each subject area.

Appendix 3_2 contains the Alignment Report commissioned by Dillard Research Associates and conducted by Meagan Karvonen of Western Carolina University and Patricia Almond of University of Oregon. The report consists of introduction and summaries to alignment, as well as the results of the alignment study for all subject areas: reading, writing, mathematics, and science.

< *Appendix 3_2* >

Test Blueprint

Crosswalks were created mapping the 2006-2007 test items to the 2007-2008 test items for all subject areas. The crosswalks list the ExGLE (though the heading on the crosswalk document incorrectly refers to the GLEs), the task and item number, and the point value associated with each item for each testing year.

Appendix 3_3 contains crosswalks by grade band for reading, writing, and mathematics.

< *Appendix 3_3* >

Bias Review

A bias and sensitivity review of the 2008 secure tests was conducted November 6th and November 8th, 2007. The reviews began with reading and writing on Nov. 6th, and concluded with math and science on Nov. 8th. The purpose of the review was to examine the bias of each item of the assessment and to assess if the format of the items affected the performance of the student in a negative manner. Reviewers were given examples to focus on such as: translations to Braille and sign language, simplified language, response demands, access versus target skills, accommodations versus modifications, race-ethnicity, gender bias, cultural bias, language bias, and value in the community.

Twelve participants from Alaska and two specialists with the deaf and blind community from Oregon were recruited based on their previous experience with the alternate assessment and this population of students. All reviewers were qualified assessors and held certification in special education.

Table 1. Bias and Sensitivity Review Participants

Participant	District	Position
Williams, Joel	Lower Kuskokwim School District	Special Ed – Emotionally Disturbed
Harvey, Sandra	Nome Public Schools	Elementary Education, Special Education
Kaasa, Dan	Kenai Peninsula, Borough School District	Elementary Education, Special Ed – Cognitively Impaired
Lytle, Kelly	MatSu School District	Special Ed – Cognitively Impaired
Soles, Jeanne	Aleknagik – Southwest Region School District	Special Education, Elementary Education
Feliciano, Regina	Chignik Lagoon School District	Special Education, Elementary Education
Macklin, Karen	Sitka School District	Elementary Education, Special Education, Special Ed – Learning Disability, Principal, Director of Special Education
Manning, Terry	Fairbanks North Star Borough Schools	Special Education, Special Ed – Learning Disability, Special Ed – Emotionally Disturbed, Elementary Education
McCall, Bonnie	North Slope Borough Schools	Elementary Education, Special Education
Robbins, Terri	Ketchikan	Special Ed – Learning Disability, Special Ed – Mentally Handicapped, Special Ed – Emotionally Disturbed
LaFever, Lyne	Yukon Flats School District	School Counselor, Special Education
Gentz, Janet	Oregon School for the Blind	Consultant
Boston, Eleni	Willamette Educational Service District – Oregon	Teacher of Deaf and Hard of Hearing

Each reviewer signed a test security agreement (TSA) and was then sent materials, including the proposed tests (both scoring protocols and student materials, as well as directions for conducting the review). They provided feedback in two ways: (a) each teacher recorded their feedback on a spreadsheet, and (b) they participated in an audio conference discussing the issues at hand. All feedback was reviewed by Dillard Research Associates (DRA) and incorporated into the secure tests.

Appendix 3_4 contains comments from the bias review. The comments are organized into comments made to accommodate the deaf and blind community, and if any, what changes were made to the assessment test.

< Appendix 3_4 >

Test Design, Development, and Score Reporting Categories Blueprint for Item Writing

Background

This document explains the specifications used when the Alternate Assessments were designed. Test specifications such as these are used to establish the guidelines by which test content may be selected and test items written. They lead to a "test blueprint" that lays out for the test item writers, typically Alaska teachers, contracted researchers, and specialists, item format and the expectations of coverage for each category.

The content of these specifications reflects the skill expectations outlined in the Extended Grade Level Expectations. Item development for the alternate assessment allows for reductions in depth, breadth, and complexity to the standards to allow access to a proportion of the population significantly impacted by cognitive disabilities, these reductions (described below) constitute a refinement in the alignment process that is referred to as "linking."

"For alternate assessments in grades 3 through 8 based on alternate achievement standards, the assessment materials should show a clear link to the content standards for the grade in which the student is enrolled although the grade-level content may be reduced in complexity or modified to reflect pre-requisite skills. For each grade, the State may define one or more alternate achievement standards for proficiency" (p. 15).⁴

Test and Task Description

The following terminology (in italics below) is necessary to understand the Alternate Assessment. Following this brief description of terms, a more thorough presentation of the test design is presented. The test design and specifications were applied in four areas of (a) reading, (b) writing, (c) mathematics, and (d) science. Each test includes both teacher *administration and scoring protocols* and *student materials*.

Each test is comprised of *tasks*, in turn comprised of several *items*. Two forms of test items are included in two separate booklets for each subject area: (i) standard content items to ascertain students knowledge on extended grade level expectations, and (ii) extended levels of support items which are reduced in complexity to ascertain the optimal manner for test administration for low functioning students who are unable to participate meaningfully in the standard content items.

Alaska's Alternate Assessment prompts use both a selected response format and a generated response format, with each item having a single correct answer that can either be selected from presented choices or that is defined according to a scoring rubric. For most items, teachers score each response according to a rubric and assign a score of 1 (partial credit) or 2 (full credit).

⁴ U. S. Department of Education (April, 2004). Standards and Assessments Peer Review Guidance: Information and Examples for Meeting Requirements of the No Child Left Behind Act of 2001. Author.

Students receive a score based on the number of prompts answered correctly compared to the total number of questions on the form. Students are not penalized for guessing.

Teacher and Student Materials

In 2007-2008 administration and scoring protocols for teachers were organized into consumable workbooks by subject area comprised of two pages per task (one page for administration and one page for scoring). Student materials were designed to promote optimal access via use of white space, font size, and graphics.

Test Design for Students with Significant Cognitive Disabilities

Alaska's alternate assessments were developed in accordance with federal regulations provided in December 2003 that allow for assessments developed for students with significant cognitive disabilities to be measured against an expectation of performance that differs in complexity from the grade-level achievement standard. In addition, federal guidance allows for items on these alternate assessments to demonstrate adequate links to grade-level content standards where direct alignment to the standards cannot be achieved without impacting student access to the information assessed. These allowances were intended to support increased access of students who would otherwise be unable to meaningfully participate in a statewide assessment even with accommodations. The alternate assessments were simplified during development in a variety of ways, reduction in depth of knowledge assessed, breadth of content standards covered, and complexity of content required.

Reduction in Complexity

Reduction in complexity in development of the alternate assessment refers to the series of steps that were taken to increase the cognitive accessibility of an item by analyzing and removing potential barriers for the population of students with significant cognitive disabilities. This process was used during development of test items (for both scoring and administration and concurrently for student materials).⁵

The use of simple language and direct sentences for all prompts was a critical component of removing the complexity from Alternate Assessments.

Simplified language was used in all test texts⁶; for example, proper nouns were-replaced to the extent possible, direct sentences were used with few dependent clauses, and the number of words reduced. Teacher scripted language and the student materials were concurrently developed to ensure alignment.

The general layout of the item was considered from the view of readability and legibility. To the extent possible for 2007-2008, all specific administration directions for items within a task were

⁵ Tindal, G. (2006). Alignment of Alternate Assessments using the Webb System: Report 2. In Council of Chief State School Officers (Ed.), *Aligning Assessment to Guide the Learning of All students: Six Reports*. Washington, D.C.: Author.

⁶ Abedi, J. (2002). Standardized achievement tests and English language learners: psychometrics issues. *Educational Assessment*, 8(3),231-257.

formatted on a single page of the Scoring Protocol for the teacher (for ease of administration). Student materials included items organized into ‘cards’ so teachers could cut them out to manipulate the distractors and mask interfering stimuli. Most items were displayed with 22-point font. All pictures were constructed for minimal complexity using black and white primarily (with few shades of gray).

Using the same general profile of the students that participate in this assessment, the test developers created items (beyond "plain language" items) to address the following:

Development Steps Toward Reducing Complexity

Select the most appropriate word with the least number of syllables.

Reduce number of words used in items, directions, and passages.

Use independent clause structure instead of dependent clause structure in passages.

Develop prompts with minimal wording.

Ensure more opportunities for modeling.

Provide more examples when possible.

Create clear (not tricky) distractors.

Provide explicit textual information with reduced requirements for extended inference so that all information is direct and literal and does not need to be pulled from various sources.

Provide rules rather than exceptions.

Use careful sequencing so that potentially similar/confusing information is not presented adjacent to similar information.

Provide multiple-choice options for items when possible or appropriate for item construction.

Administration Steps Toward Reducing Complexity

Employ appropriate pacing in the administration directions.

Supply performance-neutral praise statements for teacher to use regularly throughout assessment (e.g. You are working so hard!).

Provide additional wait time following the presentation of an item signals and cuing.

Provide alternative means of demonstrating accuracy "raise your hand/nod/blink."

Reduction in Depth

In the analysis of the Alternate Assessment, *depth-of-knowledge* (DOK) was judged at four levels: “(a) recall of fact, information, or procedure; (b) skill in using information, conceptual knowledge, or procedures of two or more steps; (c) strategic thinking, reasoning, developing a plan or sequence of steps, complexity, more than one possible answer, requiring less than 10 minutes to do; and, (d) extended thinking, requiring an investigation, time to think and process multiple conditions of the problem or task, and requiring more than 10 minutes to do non-routine manipulations” (Tindal, 2006, p. 38)⁷. In an alignment study conducted by Karvenon and Almond (2007)⁸ and submitted as part of the previous peer-review, the test was formally analyzed for DOK; this information was used to guide all item adaptations for the 2007-2008 version. The following table lists the original Webb definition, Tindal’s adaptation for use with students with the most significant cognitive disabilities, and the operational definition used in the Alaska Alignment Study (last column).

Categorical Concurrence, Range of Knowledge, and Balance of Representation

Alaska’s extended grade level expectations are organized according to a hierarchical structure. Strands and attributes are at the broadest category and are comprised of individual grade level expectations. Reduction in breadth of standards coverage is a component of item development for this population.

Table 2. Alignment Dimensions

	Original Web Definition	<i>Tindal’s Adaptation for students with significant disabilities</i>	<i>Alaska Alignment Study operational definition</i>
<i>Dimension to Evaluate</i>	Observation Rating or Checklist	<i>Behavioral Event</i>	<i>Performance Assessment</i>
Assessment sampling plan (test blueprint for the assessment).	The sampling plan is the setting in which observations are to take place <u>and</u> the types of students who are being observed.	The sampling plan is the type of documents present in the portfolio <u>and</u> the types of students who are being selected for portfolio review.	With performance assessments, the sampling plan may be implicit in the tasks or explicit in the manner in which items are constructed.
Depth of Knowledge: First, determine the depth of knowledge for the standard or objective. Second, analyze the depth of knowledge for the alternate assessments.	Rate the standard and the notes or pictures on a 4-point scale of DOK <i>given the environment</i> in which behavior is observed.	Rate the standard and the work sample products on a 4-point scale of DOK using only the evidence in the portfolio.	Rate the standard and the behavioral samples on a 4-point scale of DOK.

⁷Tindal, G. (2006). Alignment of alternate assessments using the Webb system, in *Aligning Assessment to Guide the Learning of All Students*. Washington, D. C. Council of Chief State School Officers.

⁸ Karvonen, M., & Almond, P. (2007). Alternate Assessment Alignment Study Report to the Alaska Department of Education and Early Development.

<p>Categorical Concurrence: (a) Are the behaviors in the assessment access or target skills? (b) Do the ‘target skills’ match standards and objectives?</p>	<p>For each standard, note the prevalence of objectives with observations of <i>target skills in the environment</i>.</p>	<p>For each standard, note the prevalence of objectives with work samples as <i>target skills</i>.</p>	<p>For each standard, note the prevalence of objectives with tasks as <i>target skills</i> as displayed for each task.</p>
<p>Range of Knowledge: Ascertain matched standard objectives and one target skill.</p>	<p>Proportion of standard objectives with <i>targeted skills</i> in the notes and pictures.</p>	<p>Proportion of standard objectives with <i>targeted</i> work samples.</p>	<p>Proportion of standard objectives with <i>targeted</i> performance tasks.</p>
<p>Balance of Representation: Ascertain matched standard objectives with at least one target skill.</p>	<p>Proportion of standard objectives with multiple (and varied) behavioral events providing stable inferences from the observations.</p>	<p>Proportion of standard objectives with multiple (and varied) behavioral events providing stable inferences from the judgments.</p>	<p>Proportion of standard objectives with multiple task scores providing stable inferences from the (sub)totals.</p>

Reduction in Breadth

Reduction in breadth of standards coverage is a component of item development for this population. A one-to-one correspondence is present with the tasks and items in relation to the grade level expectations with the caveat that all strands/attributes were equally addressed in the proportion of points accumulated for the total test, which was fixed at 100 points. This system allows the proficiency standards from the first year (which used a weighting algorithm to equalize the differential points across strands/attributes) to be comparable to the second year with no weighting needed.

Test Development

The Alternate Assessment was developed using the following general process. Items were developed within a grade band so that similar items could be grouped into tasks. A total of two to eight items were developed for each task. Scoring was developed for partial credit (1 point) or full credit (2 points). All grade band expectations were addressed. The architecture of each task followed the same format with two types of tests: (a) extended levels of support for students who scored zero on three items across at least three tasks, and (b) a second section focusing directly on content prompts.

Expanded Levels of Support (ELOS)

For students who performed extremely low, the ELOS items ensured participation and allowed assessors to ascertain their level of independence. The ELOS items were oriented toward subject matter constructs necessary for interacting with problems used in assessing performance. The items were developed to allow maximum participation of students with the most significant cognitive disabilities, and provide minimal access to grade level content material. These items also were used to understand what level of support was necessary for the student to interact with the assessment materials.

ELOS items were scored at four levels: ‘*Independent*’ (coded 4) signifies that the student can successfully complete the item with no assistance. ‘*Verbal-visual-gestural*’ (coded 3) indicated

that the student required some assistance to orient or focus, but once engaged can complete the item successfully. *'Partial physical prompting'* (coded 2) indicated that the student needed a physical prop or prompt to successfully complete the item. *'Full physical'* assistance (coded 1) was given when the teacher provided a 'hand-over-hand' to ensure the student's success on the item. Teachers were directed to begin with independent level and only provide further levels of support when the student hesitated for an extended period of time or did not respond.

Teachers also could mark that an item was 'inappropriate' (I) or the student refused to respond (R). In the ELOS, it was NOT possible to code the task as too difficult (only Inappropriate [I] or Refuses [R]). ELOS items were specifically designed to test 'pre-requisite' skills and therefore, it was important to document those that were present and those that were not present. Furthermore,

the scoring system (rubric) used with ELOS was universal and could accommodate all responses.

Content Prompts

Teacher materials

were comprised of general administration directions and specific item wording while specific student materials were developed to match each item. Item level test development followed common test construction procedures with specific guiding principles as outlined previously supporting each subject level. All content prompts were presented with scripted directions and scoring keys as described in each of the subject areas in this document. The content prompts were either scored as 0, 1, or 2 points Teachers also had the discretion to score content prompts as 'inappropriate'. These terms are explained in detail in the administration manual. The total points possible for the content prompts across all tasks was 100 points.

The 'content prompts' were generally oriented toward functional issues that students needed to be successful in their general home, school, and community environments. Nevertheless, only grade level academic content expectations were used in developing the content. For example, in mathematics, as much as possible, the tasks addressed situations requiring the use of numbers such as temperature, dressing, using a bus schedule, traveling to a movie, using money, or solving a problem at home, school, or in the community.

Levels of Independence Scoring Rubric				
A- Already has this skill	1 - Full Physical Contact for response <i>(e.g., hand over hand)</i>	2 - Partial Physical Contact for response <i>(e.g., nudge or adjust body)</i>	3 - Visual: Materials Movement <i>(e.g., move into line of vision)</i> - Verbal: Auditory Statement <i>(e.g., more than repeat prompt)</i> - Gesture: Hand Signal <i>(e.g., tap table, pick up card)</i>	4 - Independent: No contact and no prompting
I – Inappropriate/ Inaccessible based on the nature of the student's disability (*)				
R – Student refuses to complete				
(*) In a text box located in the online scoring and reporting system, the Qualified Assessor must provide an explanation about why this item was inappropriate or inaccessible based on the student's disability.				

Specifications for Item Writing

Background and Overview

The Alternate Assessments were individually administered using paper and pencil or any necessary assistive device identified in the student's Individualized Educational Program (IEP). The administrator followed a script and compiled student responses, scored them, and entered them manually into a database for analysis.

Item Writer Training

Dr. Gerald Tindal, Ph.D. developed the Alternate Assessment tests using traditional item writing specifications (See Downing, 2006)⁹. Dr. Tindal has published in the area of assessment for students with disabilities over the past 25 years. Dr. Tindal worked with Steve Jonas, a special educator in Oregon who has over 25 years experience working with students with disabilities.

Alignment of Test Items to Grade Level Expectations

Test items in the Alternate Assessments were carefully linked to the extended grade level expectations using a rigorous process at three points during the test item development.

Correct Key Placement

During development, item writers were instructed to rotate the correct key for their items during item authoring to ensure randomness.

Item Distribution by Difficulty

Items on the assessment were not arranged according to difficulty. The test was arranged in tasks that aligned to the expectations (strand/attributes and objectives); however, the order of item placement (within tasks) was not based on item difficulty. Teacher administrators were directed to begin with the first task and proceed through successive tasks; they could administer more advanced tasks if they thought the student was capable of responding successfully.

Item characteristics. Item writing and passage selection were guided by the following principles:

Had one correct response option, contained plausible distractors that represented feasible misunderstandings of the content, and provided options that were grammatically parallel in structure and length.

Represented the range of cognitive complexities and included challenging items for students performing at all levels.

Was appropriate for students in the assigned grade and population in terms of reading level, vocabulary, interest, and experience.

⁹ Downing, S. (2006a). Twelve steps for effective test development. In S. Downing and T. Haladyna (Eds.). Handbook of test development, pp 3-26. Mahwah, NJ: Lawrence Erlbaum.

Was embedded in a real-world context when possible (i.e. when contextual information can be provided in a non-distracting manner without introducing the need for complex cognitive processes).

Did not provide answers or hints to other items in the set or test.

Was in the form of questions or sentences that required completion.

Used clear language and not be worded in the negative unless doing so provided substantial advantages in item construction.

Was free of absolute wording, such as "always" and "never," and have qualifying words (e.g., least, most, except) printed in small caps for emphasis.

Reflected the diversity of students in Alaska.

Did not involve death, violence, drug and alcohol abuse, criminal activities, or the occult.

Was free of ethnic, gender, political, and religious bias.

Used appropriate type size for the grade level, ranging from 18 - 24 font (Tahoma).

Used selections similar in format to excerpts from content textbooks, literature, or practical reading tasks for this population.

Had content organized with a definite beginning, middle, and end and a sense of completeness, were of high interest and appropriate readability for the grade level and population, and were of appropriate length for the grade level: Elementary Grade Bands (90 words on average) Middle/High Grade Bands (140 on average)

Was free of ethnic, gender, political, and religious bias.

Did not represent material that is so widely anthologized or taught that students may have already been exposed to the content.

Did not provide answers or hints to other items on the test.

Where possible, included material about Alaska or the Pacific Northwest.

Item writing criteria. The criteria adopted for writing items were as follows:

To the extent possible, each task included items with a range of difficulty that was approximately the same across strands/attributes.

Test items were in the form of questions or sentences that required completion.

When items were multiple-choice (not "Yes" or "No"), each item had no less than three and no more than four answer choices.

Answer choices were arranged with sufficient white space on the page to ensure that there was no opportunity for distraction or confusion of responses.

Except in translation items (name to numeral or numeral to name), numbers were expressed as numerals.

When possible and not overly distracting, the text of the question was repeated on the student materials, in appropriately sized font ranging from 18 - 24 (Tahoma).

Commas were used in numbers with four or more digits.

Answer choices included units, as appropriate.

Decimal numbers less than 1 were written without leading zeros.

Computations required in test items were not so complicated that they took an inordinate amount of time to complete. Instead, reasoning within the context of the items was emphasized.

Test items were not worded in the negative ("Which of these is NOT ..."), except in rare instances when it offered substantial advantages for the item construction or representation of the construct.

When creating answer choices: "None of the above", "All of the above", and "There is not enough information to tell" were not used as an answer choice.

Test items were appropriate for students in the assigned grade and population in terms of reading level, interests, and experience.

Test items generally did not contain extraneous information.

Many items had a corresponding graphic display or student visual.

Graphic displays and response options appeared in the student materials and were identified for the administrator in the scoring protocol and/or administration manual.

Fractions were represented in a manner consistent with current research for the special education population (i.e. graphically or numerically with a horizontal line or both).

To the extent possible, the representation did not interfere with the construct under assessment.

Students were told in the test directions to choose the best answer from among the choices.

Test items were free of age, gender, ethnic, religious or disability stereotypes or bias.

Shading was minimized and used only to make a figure's size, shape or dimensions clear, and not solely for artistic effect.

ELOS and Standard Test Items

The final copy of the test was bundled in two parts: (a) Teacher Scoring Protocol and (b) Student Materials. The Standard form of the test was administered to students with traditional forms of communication. Administrators are directed to give the standard test items first to ALL students.

If they meet the 3 consecutive zeros in 3 items in 3 tasks, they were directed to administer ELOS items for instructional feedback.

Appendix 4a_1 – 4a_6 contain the ELOS items. The Expanded Levels of Support (ELOS) test items are for very low functioning students and administered to students who met the three consecutive error rule (e.g. failure on 3 items and 3 tasks).

< Appendix 4a_1 through Appendix 4a_6 >

Appendix 4a_7 – 4a_12 contain the standard administration items. There is one Standard Assessment per subject area. The Standard Assessments include both the Scoring Protocol, which contains administration directions and scoring for each item, and Student Materials, which contain materials for the student to use during administration.

< Appendix 4a_7 through Appendix 4a_12 >

CHAPTER 4: TEST ADMINISTRATION PROCEDURES

Overview

This chapter presents specific test administration information regarding the student population tested, administrators and training, and test security and online administration. The process and requirements of becoming a qualified test administrator are explained as well as the materials needed for training and test administration.

The necessary scoring protocol and materials are available for download on the training website. The website has different levels of access, so that these secure testing materials are only available to individuals reaching the level of Qualified Assessor (QA) or Qualified Mentor Trainer (QT).

The *Teacher Participation Guide* includes all information for teachers to participate in the Alternate Assessment, including administration information and rules, and requirements for becoming and maintaining Qualified Assessor and Qualified Mentor-Trainer status.

Student Population Tested

This test is reserved for those students with the most significant cognitive disabilities and up to 1% of the student population may be considered proficient on this assessment if they achieve proficiency. The decision of which students will participate in the Alternate Assessment is the result of a discussion between the student's IEP team and school district. This discussion is guided by the eligibility criteria for students with significant cognitive disabilities published in the Participation Guidelines for Alaska Students in State Assessments, September 2007 edition, pages 7-9, available on the Alaska Department of Education and Early Development website at: http://www.eed.state.ak.us/tls/assessment/alternate_optional.html.

Accommodations

The Alternate Assessment allows for many accommodations to be granted during administration. The Training Manuals provides helpful examples of accommodations that administrators may be most likely to utilize during actual administration. However, these are but a sampling of possible accommodations, and should not be considered an exhaustive list. Ultimately, it is up to the administrator to decide which accommodations are appropriate for their student, based on accommodations listed in the student's IEP. This discussion is guided by the eligibility criteria for students with significant cognitive disabilities published in the Participation Guidelines for Alaska Students in State Assessments, September 2007 edition, pages 7-9, available on the Alaska Department of Education and Early Development website at: http://www.eed.state.ak.us/tls/assessment/alternate_optional.html.

There is a certain amount of flexibility for Qualified Assessors (QA) in presenting the student materials. In addition to altering the materials for an allowable accommodation (e.g., increasing the text size of student materials), all QAs may substitute real life objects for those represented in the materials. For example, an actual glass of water may be used in lieu of the drawing of a glass of water provided in the materials, if the QA feels it would be beneficial.

Test Administrators

Only school personnel may administer the Alternate Assessment. This includes both teachers and paraprofessionals. In order to become a QA, individuals must go through online training, pass proficiency tests, and administer a practice assessment, which is then reviewed by their Qualified Mentor-Trainer (QT). Each QT must go through this training, as well as additional in person training provided annually by the Department of Education, in order to serve as a valuable resource to QAs.

Test Administrator Training

The bulk of training occurs on the website <http://ak.k12test.com>. Assessors-in-Training (AIT) go through a series of vignettes designed to familiarize them with both appropriate testing and scoring techniques. These training vignettes familiarize Assessors-in-Training with the wide variety of tasks they will encounter on the Alternate Assessment, and show videos demonstrating all the nuances needed in a proper administration. Following the training exercises, Assessors-in-Training must pass a series of brief proficiency tests related to the different tasks in each content area, as well as tests on general administration. The next section contains information on scoring, specifically training to become a Qualified Assessor and Qualified Mentor Trainer with in-depth details on training procedures.

After Assessors-in-Training complete all training and proficiency tests successfully, they must administer a practice test and have it reviewed by their QT. These individuals have been appointed by the Special Education Director or Superintendent to be the primary point of contact for the Alternate Assessment Program Manager. Once the Assessor-in-Training has completed these tasks, the QT upgrades their account to the status of QA. In subsequent years, QAs must complete only refresher proficiency tests to keep their certificate and maintain QA status. At the beginning of the 2008 test window there were 201 QAs.

The additional responsibilities of a QT necessitated additional training, which was held on October 15 -16, 2007 in Anchorage. This training provided more in-depth information on the creation of and changes to the 2007-2008 Alternate Assessments. Considerable time also was spent exploring the updated training website. New QTs had to complete all the training that a QA completed; returning QTs completed brief refresher training tutorials and proficiency tests. New QTs also had to train a protégé and be approved by DRA. At the beginning of the 2008 testing window, there were 41 QTs.

Scorer Training and Qualification – Online Proficiency

In order to ensure that valid and reliable test scores are being recorded, thorough training is required for all QAs. As described in the previous chapter, QAs must complete online training and proficiency tests, which focus on proper scoring in all of the different task types. Ample practice is provided through training vignettes and corresponding proficiency tests. Only after passing these tests does an individual become a QA and begin administering the test to students.

The tests contain administration videos with up to five questions regarding testing techniques and scoring. There are 57 tests total. Eighty percent correct is required to pass a proficiency test and trainees have 10 opportunities per test to pass. If an Assessor-in-Training fails 10 tests, they

must contact the Dillard Research Associates (DRA) helpdesk to have their account reset, thus requiring the Assessor-in-Training to retake all proficiency tests.

Scoring Materials and Process

Scoring directions are located in the scoring protocol. Each task has its own scoring page, which comes after the administration protocol for the task. Scores are marked on the page, next to each item, by writing in either a '2' for a correct response, a '1' for a partially correct response, or a '0' for an incorrect response. After the assessment has been fully administered, the QA logs onto the training website to record the student's scores. These scores are then reported in an unofficial report that shows percentage correct on tasks completed and does not reflect raw scores or proficiency levels.

Quality Control of Scoring – Reliability of the Alternate Assessment Administration and Scoring Process: Training to become a Qualified Assessor

A cadre of Qualified Assessors (QA) completed administration and scoring of the Alaska Alternate Assessment. QAs received multiple-step training in order to qualify as a test administrator. Each district was encouraged to also have a Qualified Mentor Trainer (QT) who has completed additional training and can train and mentor other school personnel in developing the skills to reliably administer the Alternate Assessment.

In order to ensure score reliability, a multiple step process is in place to develop competent, knowledgeable test administrators and scorers. A standard approach to administration and scoring leads to fair assessments, comparable scores between assessors and across settings, and provides an accurate picture of what the student knows and can do.

Step 1-Orientation and Online Training

Training is provided under the guidance of a qualified mentor trainer. Assessors-in-training (AIT) are given an orientation to the Alternate Assessment by a mentor or the Department of Education. Next, AITs register themselves on the online system and receive a password. They then complete a self-paced series of training modules offered online. These modules include: an overview of the task, instruction on how to administer the task with both text and video provided, instruction on how to score the task with both text and video provided, and finally the AIT takes a proficiency module for that task. After scoring 80% or above, the AIT is done with training for that particular task. The training modules are accompanied by a duplicate hard copy training manual that is available on the training website for download and printing.

The training modules consist of 11 reading modules, 10 writing modules, 21 math modules, 11 science modules, and 8 administration modules.

Step 2-Administering Practice Tests

The assessor-in-training now downloads the practice tests in reading, writing, mathematics, and science and prepares the materials. A test consists of scoring protocols and student materials. The AIT administers and scores the tests to a student. It is recommended that the assessor locate a student who may or may not have a mild learning disability at approximately a fourth grade level. The overarching goal is to administer all the tasks in the test in order to become

comfortable and fluent handling the student materials and scoring protocols while administering and scoring the test. Additionally, the AIT is required to read and sign a Test Security Agreement and keep it on file with the District Test Coordinator (DTC).

Step 3-Evaluation of Scoring Protocols

The scoring protocols are given to a Qualified Mentor Trainer (QT) to evaluate and score. Scoring Protocols for AIT who are becoming mentors are scored by the test vendor. The AIT receives additional training if necessary, and may be required to resubmit their scoring protocols until they receive a passing score.

Step 4-Certificate of Qualification

Qualified Mentor Trainers issue Certificates of Achievement for Qualified Assessors (QA), and change the status of the AIT in the online system to Qualified Assessor. The QA now has access to the secure test materials and to the Scoring and Reporting data entry section of the assessment.

Step 5-Maintaining Qualifications

The requirements for maintaining the Qualified Assessor status are to attend any trainings the district's Qualified Mentor Trainer may require, complete the designated refreshing skills to maintain familiarity with the tasks, and sign an updated Test Security Agreement (TSA).

Training to Become a Qualified Mentor Trainer

The purpose of the Alternate Assessment Mentor Program is to prepare district level trainers who train district personnel in correct test administration procedures for the Alternate Assessment. Mentors are available through the year to answer questions and assist district personnel. They are the first point of contact in the district for the state's Alternate Assessment Program Manager. Additionally, mentors act as an advisory group for the Alternate Assessment. Mentors should be a certified teacher in the State of Alaska with a special education endorsement and have experience with low-incidence disabilities. The state encourages every district to have at least one Qualified Mentor Trainer and one Qualified Assessor. The state currently has 41 trained mentors representing 39 of 54 total districts with 10-15 new mentors to be trained in Fall 2008.

Qualified Mentor Trainers (QT or Mentors) Additional Responsibilities

- Attend Mentor Training annually
- Become certified as a Qualified Assessor and a Qualified Mentor Trainer
- Annually refresh skills to maintain qualifications
- Conduct training for district personnel using materials provided by EED and the test vendor, Dillard Research Associates (DRA)
- Become familiar with eligibility criteria and test security
- Become familiar with the Extended Grade Level Expectations (EXGLEs)
- Answer staff questions about the alternate assessment
- Assist the District Test Coordinator in identifying students eligible for the Alternate Assessment
- Act as primary district contact for Alternate Assessment Program Manager
- Provide feedback on the Alternate Assessment as requested by EED and the test vendor

Step 6-Attend Annual Mentor Training

After completing steps 1-5 above and receiving a Qualified Assessor certificate, Mentors-in-training attend an Annual Mentor Training. EED and the test vendor conduct training. Mentors complete an Implementation Plan (sample form included below) annually which must be approved and signed by the district Special Education Director. The purpose of the Implementation Plan is to help Mentors develop a plan to coordinate the training of school personnel to the Qualified Assessor level, and to assist District Test Coordinators in identifying students eligible for the Alternate Assessment. EED and the vendor supervise training of mentors.

Step 7-Training a Protégé

Mentors train a protégé by providing an orientation to the Alaska Alternate Assessment, supervising the protégé's progress in completing the online training and proficiencies, and providing ongoing support. After completing and passing all the required online training and proficiency modules, the Mentor ensures that the protégé selects a student and administers the practice test and signs a Test Security Agreement.

Step 8-Evaluation of Protégé's Scoring Protocols after administering practice test

The Qualified Mentor Trainer evaluates their protégé's scoring protocols, has the protégé correct any errors, supervises any necessary retraining, then submits the scoring protocols containing the mentor-in-training scoring and feedback to EED who ensures all necessary components are included, then submits the scoring and feedback to the test vendor for evaluation of the mentor's ability to score another's work.

Step 9-Certificate of Qualification

EED notifies the QT when the vendor has approved their evaluation of the protégé's scoring protocols. While the QT issues a Certificate of Achievement to their protégé as a new Qualified Assessor and changes their status in the online system, EED also issues a Certificate of Achievement to the mentor-in-training, designating them as a new Qualified Mentor Trainer, and changes the new mentor's status in the online system. This change in status provides the QT access to view their district mentor information and grants the ability to make status changes.

Step 10-Training District Personnel

The QT may now implement training of district personnel selected to become Qualified Assessors. The QT will use the same procedures as with their protégé. The QT will evaluate the scoring protocols of the assessors-in-training, but will not submit these to EED or the test vendor. The QTs function as evaluators, make all status changes in the secure online system for their district, and issue Certificates of Achievement.

Step 11-Maintaining Qualifications

Qualified Mentor Trainer must attend Annual Mentor training, complete the designated refreshing skills to maintain familiarity with the tasks, and sign an updated Test Security Agreement kept on file with EED and the District Test Coordinators.

Resources Available to Qualified Assessors and Qualified Mentor-Trainers:

Annual training, training manuals, access to a HelpDesk maintained by the test vendor, coaching by Qualified Mentor Trainers, peer support, retraining available on the online test site, Program Manager for the Alternate Assessment.

Appendix 5_1 contains the *forms and procedures used in the qualification process*:

1. Qualified Assessor, Qualified Mentor-Trainer Qualification Sequence
2. Scoring Protocol Review Sheet (Used by the test vendor to evaluate scoring protocols for the mentors-in-training, and by Qualified Mentor Trainers to evaluate protégés)
3. Alternate Assessment Test Security and Online Test Security and Agreement for Testing Personnel, Qualified Assessors, and Qualified Mentor-Trainers
4. Alternate Assessment District Implementation Plan

< Appendix 5_1 >

Quality Assurance of Test Development, Administration, and Scoring

During training, all participants are required to sign and return a test security agreement. This document reiterates the message from training: test security is of the utmost importance in obtaining valid and reliable scores. As such, Qualified Assessors must keep all materials in a secure location. Following the administration, all testing materials should be placed in the student's file for at least one year. These procedures *ensure quality in both test development and test administration*.

Upon certification, and change of status in the online system, Qualified Assessors are able to access the secure test and data entry section of the assessment system.

The effect of this training was documented in the Alaska Alternate Assessment Training Report. Appendix 5_2 contains the *Alaska Alternate Assessment Training Report* comprised of summaries of the in-person trainings, the online training, and all requirements of becoming a Qualified Mentor Trainer (QT) or a Qualified Assessor (QA), and a Frequently Asked Questions that summarize thorny issues posing problems in the field.

< Appendix 5_2 >

Test Administrator Training Agendas – Reading, Writing, Mathematics (October 15-16, 2008)

Facilitators: Aran Felix, Alternate Assessment Program Manager; Jerry Tindal, Project Director;
 Director: Steve Jonas, Trainer for Dillard Research Associates (DRA)
 Participants: Alaska Alternate Assessment Mentors (new and returning)
 Goals: Mentors will participate in a two-day training session and will meet the following objectives:

1. Assessment Refresher: Refresher training for the Alaska Alternate Assessments: Reading, Writing, Mathematics, and Science
 - a. Scoring Protocols and Student Materials
 - b. Standardized Administration
 - c. Data entry and student reports
 - d. ELOS Skills
 - e. Test administration schedule for 2008
2. Qualified Assessor and Qualified Mentor Trainer Roles: Learn about the requirements to become a *Qualified Assessor* and the role that mentors will play as *Qualified Mentor Trainers*.
3. Science: Learn what the science scoring protocol and student materials look like, practice how to administer the assessment, and materials preparation.

Table 3. October 15 Agenda – New Mentors

Day One Monday October 15, 2007 8:30 – 12:00 <u>New Mentors</u>	
<i>Time</i>	<i>Topic</i>
8:30	<i>Coffee, Tea, Water, Snacks</i>
9:00	Sign in, Introductions, Overview of Morning Training
9:20	The Alternate Assessment System: Why are we here, new assessment news, eligibility criteria for students
9:30	Becoming a Qualified Assessor and Qualified Mentor
9:45	Role of Qualified Mentor and working with Protégés
10:15	Components of the system

Table 4. October 15 - All Mentors

Day One Monday October 15, 2007 1:00 – 4:30 <u>All Mentors</u>	
<i>Time</i>	<i>Topic</i>
1:00-1:30	Introductions, Housekeeping, Overview of Training, UAS credit, Expectations, sign-in sheets, recruit for Bias Review
1:30-2:00	Distribute, assemble manuals, and begin review of contents, Policy & guidance, calendar, Test Security, review/sign/turn in Test Security Agreements
2:00-2:45	Review Expectations and responsibilities of mentors (QA/QT path) and Implementation Plans (discuss/turn in)
2:45-3:00	<i>Break</i>
3:00-3:30	Standard Setting results, AYP Impacts, Cut Scores
3:30-3:45	Review Educator Guide to Interpreting Student Reports for AA
3:45-4:00	Use of assessment reports by Assessors: Incorporating results into PLEP/PLAAFP and IEP goals

Day One Monday October 15, 2007 1:00 – 4:30 <u>All Mentors</u>	
<i>Time</i>	<i>Topic</i>
4:15-4:30	Wrap-Up, Discussion & Questions
4:30	<i>Adjourn</i>

Table 5. October 16 Agenda – All Mentors

Day Two Tuesday October 16, 2007 8:00 – 4:30 <u>All Mentors</u>	
<i>Time</i>	<i>Topic</i>
8:00	<i>Coffee, Tea, Water, Snacks</i>
8:30-9:30	What’s New, Test Administration Refresher: Reliability, Validity, Standardized Administration Procedures & Rules, Administration Codes, Data Entry
9:30-10:30	New look of Online System, Review of Reading, Math, Writing
10:30-10:45	<i>Break</i>
10:45-11:45	Standard Administration and New ELOS Test Items
11:45	<i>Lunch</i>
1:00	Introduction of Science Scoring Protocols and Student Material, overview of tasks, student materials preparation
2:00	Demonstration of science administration and Mentors practice of two assigned tasks.
2:30	<i>Break</i>
2:45	Continue with practice of science tasks
3:15	Participants share results of task administration and use of student materials with whole group
4:00	Summary of training, review of mentor responsibilities, Discussion & Questions
4:30	<i>Adjourn</i>

Test Security and Administration QA

During training, all participants are required to sign and return a test security agreement. This document reiterates the message from training: test security is of the utmost importance in obtaining valid and reliable scores. As such, QAs must keep all materials in a confidential location, and refrain from discussing specifics of the test with others. Following the administration, all testing materials should be placed in the student’s file for at least one year. Teachers could not access the secure test booklet unless they had passed the training requirements (passing all proficiency tests and for assessors in training, administration and submission of a practice test). After completion of all requirements, they could then access the secure test.

Content of Training – Power point Slides

Power point slides were used for the two-day training in Anchorage, Alaska. Participants were also given hand outs of the slides so they could take notes as the slides were viewed.

< *Appendix 4b_1* >

Teacher Participation Guide

A comprehensive Teacher Participation Guide is available for download on the training website. This Guide contains all information and requirements about the specific training components.

< *Appendix 4b_2* >

Training Manual

A training manual for each subject area is available for download on the training website. These manuals are supplementary to the web training, containing the same training on each task as the website. The hard copy manuals were created for assessors to download and print.

< *Appendix 4b_3 – 4b_5* >

CHAPTER 5: SCORING

Overview

For the Alaska Alternate Assessments in Reading, Writing, and Mathematics, scoring varied by item and task. Partial credit is available for most items; the scoring protocol contains specific item scoring guidelines. Once the scores have been recorded, the QA or QT logs onto the training website, goes to the Data Entry tab, and enters the student's scores. This produces an unofficial report that compiles the student's scores on all items in all content areas.

Data Entry

The data entry tab of the Alaska Alternate Assessment website has two functions: Student Setup and Enter Scores. When entering data, the assessor must first select the Student Setup tab to enter all student demographic information. The required fields are: State ID, District ID, Student first and last name, Grade, District, School, and birthday. If the State ID is not entered a pop-up menu will appear indicating, "The State ID must be between 1 and 10 digits." If first or last name are not entered, a pop-up menu will appear indicating the first and last name are also required; "The first name must be between 1 and 20 characters."

The Grade, District, School, and Birthday fields all contain drop down boxes containing the most current state approved list of Alaska school districts and schools throughout the state. If a district or school is not selected, an error box will appear indicating a district and school must be selected; "You must select a District from the dropdown menu." The Student Setup page does not save or let the user continue until all errors have been fixed and all required information has been entered.

After student demographic information has been entered, the assessor must enter student scores by selecting the "Enter Scores" option. The Enter Scores page contains a list of all students, and selections for reading, writing, math, and depending on their grade level, science. Each subject area selection contains a drop down box with the following administration conditions: Reading (or other subject area) Tested, Absent, Long Term Illness, Suspension, Late Entry. All administration conditions other than Late Entry default and fill each subject area with the same administration condition as the state requires that the student fall under these categories during the entire testing window, therefore unable to take any subject area test. Late Entry may be selected for only one subject area and scores entered for the other areas.

To enter student scores, the assessor must select "(Subject) Tested" and click on the link underneath. The assessor must then select in which items the student participated; ELOS only, Standard Administration with or without accommodations AND then switched to the ELOS, or Standard Administration with or without accommodations.

ELOS Only

First the assessor name, date of assessment, and teacher name are required at the top of the screen. Then the student scores for each item are entered. Each ELOS task contains the options: A, I, R, or point values from 1 – 4. These values indicate:

A – Already has this skill,

- I – Inappropriate/Inaccessible based on the nature of the student’s disability,
- R – Student refuses to complete,
- 1 – Full Physical Contact for response (*e.g. hand over hand*),
- 2 – Partial Physical Contact for response (*e.g. nudge or adjust body*),
- 3 – Visual: Materials Movement (*e.g., move into line of vision*), Verbal: Auditory Statement (*e.g., more than repeat prompt*), Gesture: Hand Signal (*e.g., tap table*)
- 4 – Independent: No contact and no prompting

If I – Inappropriate/Inaccessible is selected, a response box appears where the assessor is required to indicate the reason for this selection. After all data have been entered and the “Submit Scores” option has been selected, a pop-up box appears asking the assessor to indicate they have adhered to the Three Task-Fifteen Item Rule before final submission; “Condition of Data Entry: Before submitting these data, please verify that the data you entered adheres to the Three Task-Fifteen Item Rule.” All items marked "I" must have a rationale. Press "OK" to record or "Cancel" to review.

Standard Administration With or Without Accommodations AND Then Switched to the ELOS

First the assessor name, date of assessment, and teacher name are required at the top of the screen. Then the student scores for each item are entered. The Standard tasks are listed first, which contain point values for each item. If the assessor administered standard items and then switched to ELOS, each task must adhere to the Three Task-Three Item Minimum Rule. If this rule is not followed, a pop-up box will appear indicating where the error occurred; “Warning: This data entry does not adhere to the Three Task-Three Item Minimum Rule. Please review the rule and then complete data entry. Item 2 in Task 1.34A was incomplete.” This pop-up box appears, indicating each item that contains an error until all errors are fixed. The assessor must go back and correct all errors before scores may be submitted. ELOS tasks are listed after the standard tasks. After all ELOS scores have been entered and standard scores have been entered correctly, a pop-up box appears asking the assessor to indicate they have adhered to the Three Task-Fifteen Item Rule before final submission; “Condition of Data Entry: Before submitting these data, please verify that the data you entered adheres to the Three Task-Fifteen Item Rule.” All items marked "I" must have a rationale. Press "OK" to record or "Cancel" to review.

Standard Administration With or Without Accommodations

First the assessor name, date of assessment, and teacher name are required at the top of the screen. Then the student scores for each item are entered. Each task contains point values for each item. The assessor must adhere to the Three Task-Three Item Minimum Rule. If data are entered incorrectly a pop-up box appears, indicating where the error occurred with a “Warning: This data entry does not adhere to the Three Task-Three Item Minimum Rule.” Please review the rule and then complete data entry. Item 2 in Task 1.34A was incomplete.” This pop-up box appears, indicating each item that contained an error until all errors are fixed. The assessor must go back and correct all errors before scores may be submitted.

After all scores have been correctly entered and submitted, the last step the assessor must do is check the “Mark as Complete” box on the Enter Scores screen. The student’s data will not be officially submitted until the assessor has indicated that each record has been completed.

CHAPTER 6: STANDARDS VALIDATION

Overview

A summary of the Standards Validation in Reading, Writing, and Mathematics is discussed in this chapter. The process and outcomes are discussed. Background information on participants is given as well as all materials used in the standards validation.

Proficiency Level Descriptors

The initial proficiency level descriptors were developed by committees of content experts, special educators, and other stakeholders. The standard setting participants reviewed the PLDs and any suggestions for change were submitted to EED upon conclusion of the standard validation.

< Appendix 6_1 through 6_4 >

Plan for Standards Validation Process

On April 22, 2008, a standards validation workshop was held in Anchorage. During the standards validation process, 12 teachers were presented a systematic process for comparing items and tasks from 2007 to 2008 using both a task-centered and person-centered approach. In the following table, the individuals are listed with their assignment, location, qualification (as a trainer or assessor), other positions held, and their endorsements.

Table 6. Standards Validation Participants

Name	Assign	Location	QA-QT	Position	State of AK Endorsements
Reading Group					
Vickie David	¾	NW Arctic	QT	Teacher	Sped-ED-Resource-Adapt PE
Mary Ann Christensen	5/6	Ketchikan			Sped-LD, Elem
Terry Manning	7/8	Fairbanks	QT	SPED Coordinator	Sped-LD-ED, Elem, Admin
John Hutchins	9/10	Mat Su		teacher	Sped-LD, History, social sciences
Writing Group					
Ann Konefal	¾	Fairbanks	QA		Sped-Cog.Impaired
Jeanne Soles	5/6	SW Region	QT	parent, teacher	Sped, Elem, Japanese
Dan Kaasa	7/8, Sci 10	Kenai	QT	Assist.Tech	Sped-Cog.Impaired
Nan Koentopp	5/6	Craig	QT		Sped-Cog.Impaired, Educ Tech
Math Group					
Bridgett Wittstock	¾	Petersburg	QT	Sped Director	Sped, Elem, Japanese
Theresa Owens	5/6	NW Arctic	QT	LEP,SPED Director, DTC	Sped-Adapt PE, Sped Dir, principal, PE
Joel Williams	7/8, Sci 8	LKSD	QT	Teacher	Sped-ED-Resource-Adapt PE
Stacey Street	9/10, Sci 10	Kodiak	QT	Teacher	Sped

The 2007 Alaska Alternate Assessment used the same tasks for all grade bands and then, during standard setting, applied only those tasks aligned with the Extended Grade Level Expectation (ExGLE) *for that grade band*. N.B. An assumption was made that all strands should be equally reflected in the final score. Therefore, because the tasks were unevenly distributed across the strands, task points were weighted to create a 100-point test that evenly reflected the strands within a grade band.

In 2008, the number of points from tasks was equally distributed across all the strands within a 100-point test. As done the year before, an assumption was made that all strands would be equally reflected in the final score. To accomplish this outcome, a considerable number of new items were created, particularly for strands that had been under-represented in the previous test (for tasks that were over-represented in 2007, the task was reduced and items were not used). Addition of the new items precluded the step used in 2007 for weighting scores.

To ensure that the same standards applied from 2007 to 2008 and to prevent significant changes in the impact of the standards, two strategies were used for judging comparability of results from 2007 to 2008.

*Test Centered Validation*¹⁰

1. Teachers compared the items common to both 2007 and 2008 tests, and judged the items' relative difficulty by noting the features and formats of the items within the tasks.
2. The values of linking items were compared, noting changes from 2007 to 2008 in the mean and standard deviation.

*Person Centered Validation*¹¹

1. A sample of students was identified from 2007 in each achievement level. N.B. Students who scored just above and just below each cut score were selected in each grade band.
2. The scores and last year's test responses were presented to teachers and they confirmed the students' proficiency categories.

Three major components of the validation process included (a) training on the 2007 and 2008 assessment, (b) comparison of items and tasks and review of item statistics for each grade band, and (c) analysis of students' performance in each grade band to determine/confirm proficiency using the standards adopted last year.

Training on Test Administration and Scoring

1. The scoring protocols and student materials used in 2007 and in 2008 were reviewed.

¹⁰ Adapted from Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996). *Standard setting: A bookmark approach*. Paper presented at the Council of Chief State School Officers National Conference on Large-Scale Assessment, Boulder, CO.

¹¹ Adapted from Cohen, A. S., Kane, M. AT., & Crooks, T. J. (1999). A generalizable examinee-centered method for setting standards on achievement tests. *Applied Measurement in Education*, 12(4), 367-381.

2. Administration of a few tasks was practiced, with notes taken about the range of items and manner of interaction with the student.

Comparison of Items and Tasks

1. Teachers reviewed specific items that had been used in both the 2007 and 2008 tests.
2. Teachers compared the items on their difficulties and judged their similarity, recording any specific notes in the response booklet.
3. Teachers made their judgments independently and then conferred with table participants after completing a grade level analysis of items and tasks.

Analysis of Student Profiles on 2008 Tasks

1. Teachers reviewed the levels of performance on each task and noted the pattern across tasks.
2. Teachers made judgments independently and then conferred with table participants after they had completed all items and tasks.

After all components were complete, an open discussion occurred about the standards and the degree to which they applied for 2008 using information from all three sources (training on the tests, comparison of the items and tasks, and analysis of student performances). Finally, the validation study was evaluated.

Training Presentation

A slide presentation included opening remarks and an overview of the training session. The presentation highlighted the major differences and the linking items between the 2007 and 2008 tests. It provided an overview of the standards validation process. The presentation explained the basis and process for both the test-centered and person-centered components of the standards validation. Finally, the presentation introduced the evaluation of the standards validation process.

< Appendix 6_5 >

Evaluation of Standards Validation

At the end of the day, teachers evaluated the process for validating the standards by rating five statements on a four-point scale of: completely disagree [1] to completely agree [4]; they also wrote comments about issues that supported the process as well as comments about improving the process.

1. Training on the test administration and scoring was a useful component for validating the standards.
2. Comparing items and tasks was a useful component for validating the standards.
3. Comparing student performance was a useful component for validating the standards.
4. I was adequately prepared to make judgments.
5. I am quite certain of my judgments.

Table 7. Standards Validation Evaluation

Tchr #	1. Test AdminTrn	2. Comp Item/Tsk	3. Stdnt Profiles	4. Prep Jdgmnt	5. Certn Jdgmnt
1	4	4	4	3	3
2	4	4	4	3	3
3	4	4	4	3	3
4	4	4	4	3	2.5
5	4	4	4	4	4
6	4	4	4	4	4
7	4	4	4	3	3
8	4	4	4	3	4
9	4	4	4	4	4
10	4	4	3	3	4
11	4	4	4	3	3
12	2	4	3	3	4
Average	3.8	4.0	3.8	3.3	3.5

Freq 1s	0	0	0	0	0
Freq 2s	1	0	0	0	0
Freq 3s	0	0	2	9	5
Freq 4s	11	12	10	3	6

Percent 1s	0%	0%	0%	0%	0%
Percent 2s	8%	0%	0%	0%	0%
Percent 3s	0%	0%	17%	75%	42%
Percent 4s	92%	100%	83%	25%	50%

Table 8. Standards Validation Comments

Comments
Very well organized
The small group discussion was essential to decision making
Once again - I feel strongly about Sevrina and Gerry Tindal's addition to this group. I feel their organization, professionalism, humor, and understanding of this process is a wonderful asset to the state's department.
Good discussions, I would like more certain of my judgments w/ more data (additional year). This is a good process just needs another year of data and comparing tests w/o significant changes I think.
Lots of opportunities for questions, good discussions, good to let us make decisions before additional information shared that could of affected our decision making. Good to review the tests thoroughly beforehand.
The team approach is most helpful. I enjoyed coming together as a group, being able to discuss. I don't know how you could improve it - Oh! Maybe have a form on a jump drive that we can use to put our comments in. This may help you all later so you can cut and paste our input. I think it would have been helpful to have seen all the students scores.
Seems on track. Certainly no major issues. Based on data presented no changes required.

Thorough, direct and to the point, my kind of work group. The student scoring information agreed with our critique of actual test comparison items.
Continued improvement in questions being relevant. More time to discuss and to determine change.
Very useful.
Good learning experience
Since I had given the test, I didn't feel I needed this but if someone hadn't given it, this would be beneficial.
Visuals would help improve.

Test Centered Analysis by Teachers on Linking Items

Item difficulty. Teachers provided quantitative ratings on the differences in item difficulty between 2007 and 2008 items as well as qualitative judgments on what they perceived to be the essential differences (for example, the item shape, frequency, range, etc.). All ratings and comments are noted in the response booklets that had been used as the stimulus materials. The ratings are noted by hatch marks appearing below the scale. In general, the 2007 test was viewed as more difficult.

Reading. Four teachers rated reading tasks. Out of 11 tasks, the teachers rated 2 tasks as being more difficult on the 2007 test, 8 tasks as more difficult in 2008, and 1 task as equal in 2007 and 2008. Viewing the tests as a whole, teachers rated the 2007 test more difficult in 3/4 and the 2008 tests more difficult in 5/6, 7/8, and 9/10. At least three teachers agreed on each rating.

Writing. Four teachers rated writing tasks. Out of 9 tasks, the teachers rated 1 task as more difficult in 2007, 4 tasks as more difficult in 2008, 1 task with split votes for more difficult in 2008 and equal in 2007 and 2008, and 3 tasks as equal in 2007 and 2008. Viewing the tests as a whole, teachers rated the 3/4 test equal in 2007 and 2008, the 5/6 test with split votes for equal in 2007 and 2008 and more difficult in 2008, and the 7/8 and 9/10 tests as more difficult in 2008. Other than the two split votes mentioned, at least three teachers agreed on each rating.

Mathematics. Four teachers rated mathematics tasks. Out of 24 tasks, teachers rated 11 tasks as more difficult in 2007, 6 tasks with split votes for more difficult in 2007 and equal in 2007 and 2008, 3 tasks with split votes between all three categories, 3 tasks with split votes for more difficult in 2007 and more difficult in 2008, and 1 task as more difficult in 2008. At least 3 teachers agreed on 12 out of 24 tasks. Viewing the tests as a whole, teachers rated the 3/4, 5/6, and 7/8 tests as more difficult in 2007, and the 9/10 test with split votes between all three categories. Except for the split vote on 9/10, all four teachers agreed on the summary ratings of the 3/4, 5/6, and 7/8 tests.

In general, there was consistent agreement on the ratings of items from one test as more difficult than the linking items from the other test, and on the ratings of grade-band tests. Teachers rated a most of the 2008 reading and writing tests as more difficult than the 2007 tests, but most of the 2007 mathematics tests as more difficult than the 2008 tests. Teachers tended to agree more often on the overall difficulty of each grade-band test. The highest incidence of disagreement came in mathematics. Appendix 6_6 contains the test centered judgments.

Standards Validation Comparing Linking Items in 2007 and 2008

Linking items. For linking items, teachers were provided comparisons of 2007 and 2008 count, averages, and standard deviations. When the items were directly comparable, the item statistics generally reflected very similar difficulty between 2007 and 2008 items, with the more recent version (2008) perhaps somewhat more difficult at each grade level. Items in cells with highlight (red font in gray cell) are similar in administration but different in scoring, making their comparability difficult to judge.

Reading. On the 3/4 test, 12 out of 23 items on the 2008 test with direct links to 2007 test items were more difficult on the 2008 test, 5 out of 17 items on the 2007 test with direct links to 2008 test items were more difficult on the 2007 test, and 1 pair of linked items had the same average. In general, the difficulties were similar, with a maximum difference in averages between 2007 and 2008 of 0.43.

On the 5/6 test, 7 out of 16 items on the 2008 test with direct links to 2007 test items were more difficult on the 2008 test, 2 out of 10 items on the 2007 test with direct links to 2008 test items were more difficult on the 2007 test, and 1 pair of linked items had the same average. In general, the difficulties were similar, with a maximum difference in averages between 2007 and 2008 of 0.5.

On the 7/8 test, 6 out of 14 items on the 2008 test with direct links to 2007 test items were more difficult on the 2008 test, 3 out of 8 items on the 2007 test with direct links to 2008 test items were more difficult on the 2007 test, and 1 pair of linked items had different scoring. In general, the difficulties were similar, with a maximum difference in averages between 2007 and 2008 of 0.36.

On the 9/10 test, 4 out of 8 items on the 2008 test with direct links to 2007 test items were more difficult on the 2008 test, and 2 out of 6 items on the 2007 test with direct links to 2008 test items were more difficult on the 2007 test. In general, the difficulties were similar, with a maximum difference in averages between 2007 and 2008 of 0.55.

Writing. On the 3/4 test, 12 out of 19 items on the 2008 test with direct links to 2007 test items were more difficult on the 2008 test, 1 out of 20 items on the 2007 test with direct links to 2008 test items were more difficult on the 2007 test, and 6 pairs of linked items had different scoring. In general, the difficulties were similar, with a maximum difference in averages between 2007 and 2008 of 0.43.

On the 5/6 test, 3 out of 5 items on the 2008 test with direct links to 2007 test items were more difficult on the 2008 test, 1 out of 6 items on the 2007 test with direct links to 2008 test items were more difficult on the 2007 test, and group of linked items had different scoring. In general, the difficulties were similar, with a maximum difference in averages between 2007 and 2008 of 0.46.

On the 7/8 test, 3 out of 3 directly linked pairs of items had different scoring.

On the 9/10 test, 2 out of 2 directly linked pairs of items had different scoring.

Mathematics. On the 3/4 test, 5 out of 15 items on the 2008 test with direct links to 2007 test items were more difficult on the 2008 test, 3 out of 8 items on the 2007 test with direct links to 2008 test items were more difficult on the 2007 test, 1 item had the same average, and several items had different scoring. In general, the difficulties were similar, with a maximum difference in averages between 2007 and 2008 of 0.39.

On the 5/6 test, 3 out of 18 items on the 2008 test with direct links to 2007 test items were more difficult on the 2008 test, 8 out of 15 items on the 2007 test with direct links to 2008 test items were more difficult on the 2007 test, and several items had different scoring. In general, the difficulties were similar, with a maximum difference in averages between 2007 and 2008 of 1.44 on one item, and 0.24 on the other items.

On the 7/8 test, 9 out of 16 items on the 2008 test with direct links to 2007 test items were more difficult on the 2008 test, 4 out of 15 items on the 2007 test with direct links to 2008 test items were more difficult on the 2007 test, 1 pair of items had the same average, and one group had different scoring. In general, the difficulties were similar, with a maximum difference in averages between 2007 and 2008 of 0.22.

On the 9/10 test, 2 out of 11 items on the 2008 test with direct links to 2007 test items were more difficult on the 2008 test, 8 out of 10 items on the 2007 test with direct links to 2008 test items were more difficult on the 2007 test, and 1 item had different scoring. In general, the difficulties were similar, with a maximum difference in averages between 2007 and 2008 of 0.35.

Among all grade bands in all content areas, the difficulties of linked items were very similar between the 2007 and 2008 tests. Most differences in means on comparable items between the two tests were very slight and even negligible. In reading and writing, the 2008 test tended to be slightly more difficult, and in mathematics the 2007 test was slightly more difficult. Appendix 6_7 contains the linking items through task comparisons.

< Appendix 6_7 >

Person Centered Analysis by Teachers on Proficiency Confirmation

Each subject area team placed each of 10 students – who had been sampled from the entire data files in a stratified manner – as proficient, just below proficient, and just above proficient. The results reflect uniform agreement with the cut scores of 2007.

Reading. Four judges categorized each selected student as below proficient, proficient, or advanced. In the 3/4 grade band, all four judges agreed on all 10 students. In the 5/6 grade band, all four judges agreed on 6 out of the 10 students, and three out of four judges agreed on the other 4 students. In the 7/8 and 9/10 grade bands, all four judges agreed on all 10 students in each grade band.

Writing. Four judges categorized each selected student as below proficient, proficient, or advanced. In the 3/4 grade band, all four judges agreed on 5 out of the 10 students, three out of four judges agreed on 2 students, and three out of three voting judges agreed on the other 3 students. In the 5/6 grade band, all four judges agreed on all 10 students. In the 7/8 grade band, all four judges agreed on 2 students, two out of three voting judges agreed on 2 students, two out

of two voting judges agreed on 5 students, and the two voting judges were split on one student. No data were collected for the 9/10 grade band.

Mathematics. Four judges categorized each selected student as below proficient, proficient, or advanced. In the 3/4 grade band, all four judges agreed on 5 out of the 10 students, and three out of four judges agreed on the other 5 students. In the 5/6 grade band, all four judges agreed on 7 out of 10 students, and three out of four judges agreed on the other 3 students. In the 7/8 grade band, all four judges agreed on 4 students, three out of four judges agreed on 3 students, two out of four judges agreed on 2 students, and two out of three voting judges agreed on 1 student. In the 9/10 grade band, all four judges agreed on 4 students, and three out of four judges agreed on 6 students.

In general, judges tended to agree unanimously more often than they had any disagreement. Even in the absence of a unanimous agreement, three out of four judges agreed more often than they had two-to-two or other split votes. The judges seemed to strongly agree on the cut scores from 2007. Appendix 6_8 contains the person centered judgments.

< Appendix 6_8 >

Teachers were provided the potential results of using the 2007 cut-scores on the 2008 results as a final step in the process. The percentages of students meeting and exceeding the standards increased substantially using the cut scores from 2007 to the scores in 2008.

Projected Impact Data

Impact Data from the spring of 2007 was given to participants for grades 3 through 10 in reading, writing, and mathematics. The student data were projected in two categories: Advanced/Proficient, and Below/Far Below Proficient. Appendix 6_9 contains the 2007 Impact Data.

< Appendix 6_9 >

CHAPTER 7: REPORTING

Overview

Five reports are available that depict the outcomes from the Alternate Assessment. There is a 6th set of reports located on the EED website, containing the following: Statewide Results, District wide results, District by Ethnicity and Gender, District by Special Populations, and School wide (<http://www.eed.state.ak.us/tls/assessment/results/results2008.html>).

Report Types

Report Types are - Student Reports: Official and Unofficial, Guides to Test Interpretation: Parent and Educator, Secure Reporting Website for districts, and EED Assessment Result Statewide Results, District wide results: District by Ethnicity and Gender, District by Special Populations, and School wide.

Unofficial Student Report

This report is available immediately after student scores are entered into the online data entry system. The report contains student demographic data, percentages correct, and bar charts depicting percentages correct.

< *Appendix 7_1* >

Official Report

Official individual student reports are mailed by EED to individual districts the summer following the testing period. These reports contain student demographic data, percentage correct, a chart depicting the score possible and score earned as well as a bar representing where the student's score lies in accordance to proficiency levels. A description of chart interpretation is given, as well as a description of the PLDs for the student's grade level.

< *Appendix 7_2* >

Parent Guide to Interpretation of the Individual Student Report

This report consists of four pages explaining how to read the individual student reports. The first two pages explain the purpose of testing, what the Alternate Assessments measure, components of the Alternate Assessment, and a guide to reading the individual student report. The last two pages contain an example of the individual student report. These reports are example reports; no actual student data are given.

< *Appendix 7_3* >

Educator Guide to Interpretation of the Individual Student Report

The educator guide is similar to the parent guide but goes on to provide further information on conditions of administration, unofficial student reports, task descriptions, and cut score ranges. This guide contains explanations and examples of all individual student reports, official, unofficial, and ELOS.

< *Appendix 7_4* >

DRA Secure Reporting Website

This website is secure for districts. Each district has its own logon and password and is able to view and print only student reports for that district.

< *Appendix 7_5* >

CHAPTER 8: TEST VALIDITY

Overview

After the close of the 2008 testing window, student data were downloaded from the website and analyzed. The following chapter explains the summary of data analysis, as well as reliability and validity of the 2008 test items.

Data Analysis Summary

On *April 6, 2008*, the test window closed for data entry. A total of 59 teachers showed a total of 44 incomplete records in their data entry; of 59 teachers, 10 had 2 incomplete records, two had 3 incomplete records, two had 4 incomplete records, and one had 7 incomplete records. The remaining 44 teachers had 1 incomplete record. Of the 59 teachers, most of them (48) had not already submitted any other records as complete while the other 11 had previously submitted complete records for at least 1 and up to 9 students.

Nevertheless, the data were extracted from the database to assemble the syntax code in the statistical software (Statistical Package for the Social Sciences[®] – Version 16). In this step, the extracted data file also was recoded with efficient variable names and computed variables were calculated. This **first** analysis of the data conducted by Dr. Tindal and the results were then submitted to the DRA programmer (Aaron Glasgow) to check and correct the values for all recoded and computed variables.

On *April 8, 2008*, the following email was sent to all mentors:

Dear Mentors - **URGENT**. Please read the email below and check the attached excel spreadsheet to determine if you, or any of your Qualified Assessors, have not marked the "Record Complete" checkbox. The AA test website will be opened briefly to correct student data. If the assessor is not available to correct their student's data entry, please contact me at once with this information, and we research this and make the corrections for the assessor. This could have tremendous impact on district data and on standard setting! Thank you so much for your help with this. Aran Felix

Dear Qualified Assessor:

One or more of your student's records lack the required data entry that indicates the assessment process is complete.

For a student's record to be scored and counted for AYP you **MUST**:

- 1) Login into the <http://ak.k12test.com> website.
- 2) Go to your student records and confirm that you have marked the checkbox located under **Record Complete** if you did complete the assessment for the student(s). See picture below. (*)
- 3) The Alternate Assessment web system will be reopened for your convenience Wednesday, April 9, and Thursday, April 10, **ONLY**.

Problems? Contact Jerry Tindal (gerald.tindal@mac.com) **AND** Aran Felix (aran.felix@alaska.gov) immediately.

NOTE: No new students may be entered into the system.

(*) What if you assessed a student, but the student was unable to complete all the tasks in that section? That counts as assessing the student. Please check the box. If you did not assess a student, do not check the box. Fill out the Reasons Not Tested under student demographic section.

On *April 19*, a second data extraction was completed to run the analyses for the standards validation in reading, writing, and mathematics (to be held on April 23) and standard setting in science (to be held on April 24-25). Dr. Tindal completed this **second** analysis of the data to provide item and task analyses as well as projections of proficiency status (.sav file V3). Dr.

Yovanoff, a statistical consultant with DRA, reviewed the syntax code for this analysis, made several changes to provide a more efficient code, and analyzed the data for a **third** time (.sav file V5). These results were used in the presentation to the Alaska TAC on April 30, 2008.

On April 30, a third data extraction was completed because two students had to be hand entered by Aaron Glasgow. Dr. Tindal analyzed the data files for the **fourth** time to match the results to the previous analyses; minor differences were noted in the grade levels for which the two students had been added (grades 6 and 10). These final results have been submitted on May 18, 2008.

The first step in the data analysis was calculation of variable totals using the strand weights. Although the final files submitted to EED on the contract deadlines had been finalized by DRA but not reviewed or approved by EED, several significant changes were made in the scoring protocols and student materials. The final files from DRA had been completely grade-banded with item totals equal to 100 points for each test (to remove the need for differential use of items and weighting of points). These changes were made in late December and therefore required some use of weighting so that the 2007 scores could be comparable using the 100-point scale. Generally, the weights were very insignificant in adjusting the meaning of student performance but importantly critical in maintaining score comparability.

The data were analyzed by calculating item and task statistics for in reading, writing, mathematics, and science. The information from this analysis includes the difficulty and variance of both items and tasks. All of these analyses were done using the 2008 scales (which, as noted above, were weighted slightly to equal 100 points). This allowed the data to be used for standards validation and eventually establish comparable point values and cut-score impact data for 2007 and 2008.

Scaling 2008 onto 2007 Scale and Analyzing Impact

The last analysis presents the 2008 test scaled to the 2007 test so that the cut scores from 2007 could be applied to the 2008 scores. In scaling the test, a mean linear equating was used following the procedures described by Kolan and Brennan¹² and Petersen, Kolan, and Hoover¹³ in which the mean and standard deviation are used to apply the new scale (2008) to the old scale (2007).

Figure 2. Linear Equating

Linear equating

$$ly(c) = y = \frac{\sigma(Y)}{\sigma(X)}x + \left[\bar{Y} - \frac{\sigma(Y)}{\sigma(X)}\bar{X} \right]$$

Dr. Yovanoff from DRA analyzed the data by combining the 2007 AYP file with the 2008 data file, changing the 2008 scores so they were comparable, and then applying the 2007 cut-scores (from Appendix A). Three analyses were completed to present results displaying: (a) descriptive

¹² Kolen, M.J. & Brennan, R.L. (1995). *Test equating methods and practices*. New York: Springer.

¹³ Petersen, N.S., Kolen, M.J., Hoover, H.D. (1989). Scaling, norming, and equating. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221-262). New York: Macmillan.

statistics (means and standard deviations), (b) percentages at each of four categories and each grade level, and (c) percentages at each of two categories and each grade level.

The results reflect considerable consistency in the percentages attained in most of the various proficiency categories (both the four levels of ‘far below, below, proficient, and advanced as well as the two categories of below and above proficiency). In general, there was a slight to somewhat large positive change in percentage of students performing at a proficient level.

Reliability

Table 9. Cronbach's Alpha - Reading

Cronbach's Alpha	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Task 7
Grade 3/4	.767	.931	.968	.747	.853	.875	.920
Grade 5/6	.896	.906	.853	.860	.882		
Grade 7/8	.873	.793	.944	.884	.867	.856	
Grade 9/10	.933	.753	.847	.830			

Table 10. Reading Grade Band 3/4

Cronbach's Alpha	Task 1 ISS	Task 2 ILS	Task 3 BS	Task 4 Name	Task 5 Jill	Task 6 Annie	Task 7 Jimmy
Grade 3/4	.767	.931	.968	.747	.853	.875	.920

Table 11. Reading Grade Band 5/6

Cronbach's Alpha	Task 1 Words	Task 2 Sent	Task 3 Jill	Task 4 Annie	Task 5 Jimmy
Grade 5/6	.896	.906	.853	.860	.882

Table 12. Reading Grade Band 7/8

Cronbach's Alpha	Task 1 Words	Task 2 ObtInf	Task 3 Sent	Task 4 Hannah	Task 5 Jan	Task 6 City
Grade 7/8	.873	.793	.944	.884	.867	.856

Table 13. Reading Grade Band 9/10

Cronbach's Alpha	Task 1 Decode	Task 2 IDRoot	Task 3 Hannah	Task 4 Jan
Grade 9/10	.933	.753	.847	.830

Table 14. Chronbach's Alpha - Writing

Cronbach's Alpha	Task 1	Task 2	Task 3	Task 4
Grade 3/4	.971	.961		.932
Grade 5/6	.947		.950	.943
Grade 7/8	.949	.894	.950	
Grade 9/10	.901		.844	

Table 15. Writing Grade Band 3/4

Cronbach's Alpha	Task 1 CopyLtr	Task 2 CopyWrđ	Task 3 WrtName	Task 4 WrtWrđDict
Grade 3/4	.971	.961		.932

Table 16. Writing Grade Band 5/6

Cronbach's Alpha	Task 1 CopySent	Task 2 WrtName	Task 3 WrtWrđDict	Task 4 Sent
Grade 5/6	.947		.950	.943

Table 17. Writing Grade Band 7/8

Cronbach's Alpha	Task 1 WriteSentDict	Task 2 RevSent	Task 3 WrtSent
Grade 7/8	.949	.894	.950

Table 18. Writing Grade Band 9/10

Cronbach's Alpha	Task 1 RevSent	Task 2 WrtStory	Task 3 RevWrtg
Grade 9/10	.901		.844

Table 19. Cronbach's Alpha - Mathematics

Cronbach's Alpha	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Task 7	Task 8	Task 9	Task 10	Task 11	Task 12	Task 13	Task 14	Task 15
Grade 3/4	.973	.742		.932	.935										
Grade 5/6	.814	.719			.922	.852	.850	.897	.826	.814	.758				
Grade 7/8	.448	.462	.867			.918	.562	.954	.809	.830	.828	.823	.893	.850	.879
Grade 9/10	.902	.728		.676	.889	.735		.803	.813	.648	.721	.660			

Table 20. Mathematics Grade Band 3/4

Cronbach's Alpha	Task 1 Cpy Num	Task 2 Num Line	Task 3 Count	Task 4 Same Diff	Task 5 ID Shape
Grade 3/4	.973	.742		.932	.935

Table 21. Mathematics Grade Band 5/6

Cronbach's Alpha	Task 1 RdWr Num	Task 2 Num Line	Task 3 Cnt Obj	Task 4 Cnt	Task 5 Simp Add	Task 6 Simp Pat	Task 7 Simp Grph	Task 8 Shrt/ Lng	Task 9 Id Money	Task 10 IDShape	Task 11 Sm/ DfShp	Task 12 ID Perim
Grade 5/6	.814	.719			.922	.852	.850	.897	.826	.814	.758	

Table 22. Mathematics Grade Band 7/8

Cronbach's Alpha	Task 1 RdWr Num	Task 2 Id Frac	Task 3 Num Line	Task 4 IdSkp Pat	Task 5 Count	Task 6 DdAd d Sub	Task 7 Rep ExtSm Pat	Task 8 LblSt No Zero	Task 9 Und Symb	Task 10 RdSm p Grph	Task 11 IdUnt Msmnt	Task 12 Cnt Mny	Task 13 Id Shps	Task 14 Mtch Shps	Task 15 Id Perim
Grade 7/8	.448	.462	.867			.918	.562	.954	.809	.830	.828	.823	.893	.850	.879

Table 23. Mathematics Grade Band 9/10

Cronbach's Alpha	Task 1 ID PlcVlu	Task 2 ID Frctns	Task 3 Ord Num	Task 4 Round Num	Task 5 2DigAdd SubMult	Task 6 RepSimp Pptrns	Task 7 Undstd Syms	Task 8 RdSimp Grphs	Task 9 IDUnts Msmnt	Task 10 Count Money	Task 11 Lines Symty	Task 12 ID Perimtr
Grade 9/10	.902	.728		.676	.889	.735		.803	.813	.648	.721	.660

Validity

Proficiency Comparison between 2007-2008

After close of the 2007-2008 testing window, data were analyzed and compared with 2006-2007 data. The Proficiency Comparison provides information on the count, mean values, variance, and proficiency categories for the two testing years. Appendix 8_1 contains a report outlining the 2008 items scaled to the 2007 items with cut scores.

< Appendix 8_1 >

Results from the 2008 testing window were analyzed by tasks. Comparisons were made based on item difficulty and average inter-item correlation.

Reading Grade 3/4

Task 1 had a mean of 8 points with six difficult items (1, 2, 5, 6, 7, and 8) and an average inter-item correlation .30 (ranging from .03 to .48).

Task 2 had a mean of 7 points with three difficult items (3, 8, and 9) and an average inter-item correlation .58 (ranging from .25 to .86).

Task 3 had a mean of 15 points with five difficult items (2, 4, 5, 6, and 7) and an average inter-item correlation .79 (ranging from .70 to .87).

Task 4 had a mean of 2 points with one difficult item (2) and an average inter-item correlation .61 (ranging from .61 to .61).

Task 5 had a mean of 8 points with three difficult items (6, 7, and 8) and an average inter-item correlation .43 (ranging from .07 to .64).

Task 6 had a mean of 8 points with three difficult items (3, 4, and 8) and an average inter-item correlation .46 (ranging from .33 to .62).

Task 7 had a mean of 9 points with one difficult item (4) and an average inter-item correlation .60 (ranging from .42 to .72).

Reading Grade 5/6

Task 1 had a mean of 10 points with two difficult items (6 and 7) and an average inter-item correlation .52 (ranging from .35 to .80).

Task 2 had a mean of 11 points with four difficult items (2, 3, 4 and 5) and an average inter-item correlation .71 (ranging from .64 to .77).

Task 3 had a mean of 13 points with three difficult items (7, 11 and 12) and an average inter-item correlation .34 (ranging from .11 to .49).

Task 4 had a mean of 13 points with one difficult item (10) and an average inter-item correlation .35 (ranging from -.002 to .63).

Task 5 had a mean of 13 points with no difficult items and an average inter-item correlation .41 (ranging from .09 to .68).

Reading Grade 7/8

Task 1 had a mean of 6 points with no difficult items and an average inter-item correlation .65 (ranging from .53 to .77).

Task 2 had a mean of 4 points with all difficult items (1 through 7) and an average inter-item correlation .35 (ranging from .17 to .60).

Task 3 had a mean of 13 points with two difficult items (2 and 4) and an average inter-item correlation .84 (ranging from .81 to .86).

Task 4 had a mean of 15 points with one difficult item (5) and an average inter-item correlation .40 (ranging from .17 to .55).

Task 5 had a mean of 14 points with one difficult item (5) and an average inter-item correlation .35 (ranging from .06 to .57).

Task 6 had a mean of 14 points with two difficult items (5 and 10) and an average inter-item correlation .33 (ranging from .06 to .67).

Reading Grade 9/10

Task 1 had a mean of 18 points with no difficult items and an average inter-item correlation .65 (ranging from .49 to .79).

Task 2 had a mean of 4 points with three difficult items (1, 4 and 5) and an average inter-item correlation .35 (ranging from .11 to .68).

Task 3 had a mean of 20 points with four difficult items (9, 10, 11 and 12) and an average inter-item correlation .37 (ranging from .10 to .61).

Task 4 had a mean of 18 points with five difficult items (1, 9, 10, 11, and 12) and an average inter-item correlation .35 (ranging from .06 to .66).

< *Appendix 8_2* >

Writing Grade 3/4

Task 1 had a mean of 17 points with no difficult items and an average inter-item correlation .76 (ranging from .62 to .88).

Task 2 had a mean of 21 points with no difficult items and an average inter-item correlation .82 (ranging from .73 to .90).

Task 3 – Write Own Name (not included in writing stats)

Task 4 had a mean of 10 points with all difficult items (1 through 6) and an average inter-item correlation .70 (ranging from .51 to .84).

Writing Grade 5/6

Task 1 had a mean of 7 points with one difficult item (5) and an average inter-item correlation .90 (ranging from .90 to .90).

Task 2 – Write Own Name (not included in writing stats)

Task 3 had a mean of 11 points with all difficult items (1 through 5) and an average inter-item correlation .79 (ranging from .73 to .85).

Task 4 had a mean of 11 points with no difficult items and an average inter-item correlation .85 (ranging from .84 to .87).

Writing Grade 7/8

Task 1 had a mean of 14 points with no difficult items (5 and 10) and an average inter-item correlation .86 (ranging from .85 to .87).

Task 2 had a mean of 21 points with two difficult items (3 and 6) and an average inter-item correlation .57 (ranging from .41 to .82).

Task 3 had a mean of 13 points with no difficult items and an average inter-item correlation .91 (ranging from .91 to .91).

Writing Grade 9/10

Task 1 had a mean of 24 points with no difficult items and an average inter-item correlation .63 (ranging from .50 to .81).

Task 2 – Write a Story (not included in write stats)

Task 3 had a mean of 26 points with no difficult items and an average inter-item correlation .43 (ranging from .16 to .70).

< Appendix 8_3 >

Mathematics Grade 3/4

Task 1 had a mean of 12 points with no difficult items and an average inter-item correlation .82 (ranging from .70 to .91).

Task 2 had a mean of 2 points with one difficult item (2) and an average inter-item correlation .50 (ranging from .33 to .73).

Task 3 – Count (not included in math stats)

Task 4 had a mean of 6 points with three difficult items (2, 4 and 6) and an average inter-item correlation .63 (ranging from .43 to .84).

Task 5 had a mean of 9 points with no difficult items and an average inter-item correlation .83 (ranging from .80 to .86).

Mathematics Grade 5/6

Task 1 had a mean of 6 points with no difficult items and an average inter-item correlation .54 (ranging from .36 to .90).

Task 2 had a mean of 2 points with one difficult item (3) and an average inter-item correlation .47 (ranging from .43 to .50).

Task 3 – Count Objects (not included in math stats)

Task 4 – Count (not included in math stats)

Task 5 had a mean of 10 points with no difficult items and an average inter-item correlation .63 (ranging from .43 to .75).

Task 6 had a mean of 11 points with no difficult items and an average inter-item correlation .52 (ranging from .23 to .82).

Task 7 had a mean of 10 points with no difficult items and an average inter-item correlation .44 (ranging from .22 to .88).

Task 8 had a mean of 4 points with no difficult items and an average inter-item correlation .81 (ranging from .81 to .81).

Task 9 had a mean of 6 points with no difficult items and an average inter-item correlation .54 (ranging from .43 to .70).

Task 10 had a mean of 6 points with no difficult items and an average inter-item correlation .40 (ranging from .10 to .65).

Task 11 had a mean of 2 points with no difficult items and an average inter-item correlation .52 (ranging from .37 to .63).

Mathematics Grade 7/8

Task 1 had a mean of 3 points with no difficult items and an average inter-item correlation .30 (ranging from .30 to .30).

Task 2 had a mean of 1 point with one difficult item (2) and an average inter-item correlation .33 (ranging from .33 to .33).

Task 3 had a mean of 4 points with two difficult items (4 and 5) and an average inter-item correlation .57 (ranging from .40 to .84).

Task 4 – Identify Skip Patterns (not included in math stats)

Task 5 – Count (not included in math stats)

Task 6 had a mean of 10 points with no difficult items and an average inter-item correlation .62 (ranging from .48 to .78).

Task 7 had a mean of 5 points with one difficult item (3) and an average inter-item correlation .35 (ranging from .22 to .60).

Task 8 had a mean of 3 points with no difficult items and an average inter-item correlation .85 (ranging from .72 to .95).

Task 9 had a mean of 1 point with two difficult items (1 and 2) and an average inter-item correlation .68 (ranging from .68 to .68).

Task 10 had a mean of 9 points with two difficult items (1 and 10) and an average inter-item correlation .39 (ranging from .07 to .93).

Task 11 had a mean of 4 points with two difficult items (1 and 3) and an average inter-item correlation .45 (ranging from .30 to .71).

Task 12 had a mean of 2 points with one difficult item (2) and an average inter-item correlation .70 (ranging from .70 to .70).

Task 13 had a mean of 3 points with all difficult items (1 through 4) and an average inter-item correlation .69 (ranging from .59 to .86).

Task 14 had a mean of 6 points with no difficult item and an average inter-item correlation .45 (ranging from .26 to .69).

Task 15 had a mean of 3 points with no difficult items and an average inter-item correlation .70 (ranging from .49 to .81).

Mathematics Grade 9/10

Task 1 had a mean of 4 points with no difficult items and an average inter-item correlation .65 (ranging from .51 to .78).

Task 2 had a mean of 3 points with one difficult item (2) and an average inter-item correlation .39 (ranging from .15 to .57).

Task 3 – Order Numbers (not included in math stats)

Task 4 had a mean of 1 point with all difficult items (1 through 3) and an average inter-item correlation .41 (ranging from .22 to .73).

Task 5 had a mean of 7 points with two difficult items (5 and 6) and an average inter-item correlation .60 (ranging from .41 to .85).

Task 6 – Reproduce and Fill in Simple Patterns (not included in math stats)

Task 7 had a mean of 3 points with two difficult items (3 and 4) and an average inter-item correlation .41 (ranging from .25 to .67).

Task 8 had a mean of 10 points with two difficult items (1 and 10) and an average inter-item correlation .36 (ranging from .05 to .79).

Task 9 had a mean of 5 points with no difficult items and an average inter-item correlation .44 (ranging from .19 to .65).

Task 10 had a mean of 5 points with all difficult items (1, 2 and 3) and an average inter-item correlation .64 (ranging from .36 to .64).

Task 11 had a mean of 6 points with one difficult items (5) and an average inter-item correlation .27 (ranging from .06 to .68).

Task 12 had a mean of 1 point with all difficult items (1 and 2) and an average inter-item correlation .50 (ranging from .50 to .50).

< Appendix 8_4 >

CHAPTER 9: PROGRAM IMPROVEMENT

Overall Program Evaluation

The general training of administrators and mentors is well designed with high quality control of the process. The initial training of mentors along with follow-up training using the web-based proficiency testing results in assurance that teachers are prepared for administration of the test. The measures are functioning as they should and the data entry and reports work well to establish proficiency levels. Most teachers evaluate the alternate assessment system positively.

Summary of Consequences Survey

During the testing window, DRA had created a teacher survey for consequential validity on the website wufoo.com. Teachers were encouraged to take this survey upon completion of their individual test administration to help provide DRA with important information on the testing process. A link was placed on the testing website to directly access the survey. Teachers were offered a \$25 gift certificate to amazon.com upon completion of the survey. After the closing of the testing window, a bulk e-mail was sent to all Qualified Assessors (QAs) and Qualified Trainers (QTs) registered online with the web link for the survey. Out of 274 QAs and QTs, 179 responded to the survey. The results were summarized including responder information, the survey itself, and any noticeable trends in preparation to present to the Technical Advisory Committee.

Appendix 9_1 contains the results of the *Teacher Survey for Consequential Validity*.

< Appendix 9_1 >

Training and Qualifications

The majority of participants agreed, or strongly agreed, with positive statements about the training and qualifications for becoming Qualified Assessors and Qualified Mentor-Trainers. The majority of participants reported that web-based training required between four and eight hours of their time, while the remainder reported that training required eight hours or more. A majority of participants agreed that time spent on the training was well spent, and that the training materials were informative. A majority agreed that their districts provided sufficient time to become proficient, and that the requirements for qualification were clear and reasonable, including administration of the practice test. A majority agreed that the requirements for retaining qualifications were reasonable. Finally, a majority agreed that they felt fully capable of administering the assessment after training.

Test Administration and Decision Making

Large majorities of participants agreed, or strongly agreed, with positive statements about test administration and decision-making. The majority agreed that decision-making was clear for administering standard (STD) or extended levels of support (ELoS) items. Over 70% agreed that the test materials were well organized, while almost 30% disagreed. A majority agreed that they had sufficient time to prepare materials, administer the test, and enter data. A majority agreed

that scoring criteria were clear (though about 25% disagreed) and that the test was easy to administer, including with the use of accommodations used during instruction.

Results

A majority of participants agreed, or strongly agreed, with positive statements about accessibility of test items and relevance of test results. A majority agreed that both standard and extended levels of support (ELOS) items were accessible. Over half of participants agreed or strongly agreed that results accurately represented students' progress on Extended Grade Level Expectations (ExGLEs), but over 40% disagreed or strongly disagreed. A significant majority of participants agreed that the results reports were easy to interpret.

Instructional Relevance

Participants' agreement differed on positive statements about the links between instruction and the alternate assessment. Where there was a majority in agreement with the positive statements in this area, it was usually a smaller majority than in other areas. About 60% agreed that the content on the alternate assessment is closely related to their instruction, that they use the ExGLEs to guide instruction, that they use the ExGLEs to guide writing IEP goals, and that they use the results from the alternate assessment to guide instruction. Over 60% of participants said that, after giving the alternate assessment, they have not: spent more time teaching academic content, provided more accommodations or other supports, or increased academic expectations of students. Roughly half of participants agreed that they learned new information about their students or new skills from administering the alternate assessment. A narrow majority agreed that students with significant cognitive disabilities should be included in the statewide assessment system.

Professional Development Needs

Participants expressed the need for professional development in several areas. A narrow majority of participants expressed a need for professional development in linking language arts and math instruction with content standards and alternate assessments, but over 60% did not feel the need for such professional development in science. A significant majority expressed the need for professional development in balancing academic and functional skills in instruction, in using accommodations, and in using alternative or augmented communication systems. A smaller majority expressed the need for professional development in explaining assessment results to parents.

Teacher Demographics and Experiences

Participating teachers completed several questions capturing demographic information. Their teaching experience ranged from 0 to 37 years, and special education experience ranged from 0 to 33 years, with roughly consistent distributions on those ranges. Most teachers had a higher education degree, 68% with at least a Bachelors and 46% with a Masters. On teaching licensure, 70% had general education licensure, and 93% had special education licensure. About 13 participants held positions other than teacher, including administrator or early childhood educator. About 10% of participants had an English Language Arts endorsement, 3% had Mathematics, 5% had Science, 5% had Health PE, 3% had Fine or Performing Arts, 6% had Social Studies, and 38% had some other endorsement, including Deaf Education and various

other grade specific special education certifications. Over 40% administered the Standard test in Reading to one or more students, while about 9% administered the ELOS test in Reading. About 30% administered the Standard test in Writing to one or more students, while about 8% administered the ELOS test in Writing. Almost 25% administered the Standard test in Mathematics to one or more students, while about 8% administered the ELOS test in Mathematics. Almost 30% administered the Standard test in Science to one or more students, while about 8% had administered the ELOS test in Science. Over 90% reported having a Qualified Trainer in their district. Participants reported typical numbers of students with various categories of disability on their caseloads.

Recommendations for Future Consideration

In the next version of the alternate assessment, we plan to develop a larger bank of items in advance so that EED can field-test them to determine their adequacy. First, we plan to conduct a content review of the cousin items (that are to be field-tested). The items and tasks can be improved by providing more clear measurement of the constructs, as reflected in the comments made in the standard setting meeting.

Field Testing of New Items (Standard and ELOS)

Second, we plan to deploy them within the existing task structures (rather than add a separate field test), so that some items are operational (from the previous year) and other items are field tested. This process can ensure that the items not only function appropriately but also provide comparable items from year to year so that any changes in performance levels reflect improvements in achievement not differences in item difficulties. We also plan to develop additional ELOS items that more closely align with the ExGLES and provide a set of sub-skills that make the content more accessible to the lowest functioning 10% of the 1% group.

Package of Test Booklets and Training of Teachers

Finally, we need to re-package the test and student materials by grade level so that teachers can print only the materials that they need for each grade level. In addition, we will provide individual labeling of student materials to match the task to the materials apply. These changes are likely to require further training, particularly on appropriate accommodations and assistive technology that would make the test items more accessible to students yet be consistent with definition of the constructs being tested (at the item and task levels).