

**Alaska Alternate Assessment
2007 – 2008
Science Technical Report**

Dillard Research Associates
September 15, 2008

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>Pilot Test Development</i>	13
<i>Summary of Alternate Assessment-Extended Science Pilot Test Development Schedule</i>	13
<i>Statistical Analyses</i>	14
CHAPTER 3: TEST DESIGN AND ITEM/TASK DEVELOPMENT	
<i>Overview</i>	15
<i>Organization of Science Extended Grade-Level Expectations</i>	15
<i>Test Blueprint</i>	16
<i>Item/Task Development</i>	20
<i>Item Construction and Review</i>	20
<i>Bias and Sensitivity Review</i>	20
<i>Test Design: Development, and Score Reporting Categories for Item Writing</i>	22
<i>Background</i>	22
<i>Test and Task Description</i>	23
<i>Teacher and student materials</i>	23
<i>Test Design for Students with Significant Cognitive Disabilities</i>	23
<i>Reduction in Complexity</i>	24
<i>Development Steps Toward Reducing Complexity</i>	24
<i>Administration Steps Toward Reducing Complexity</i>	25
<i>Reduction in Depth</i>	25
<i>Categorical Concurrence, Range of Knowledge, and Balance of Representation</i>	25
<i>Test Development</i>	26
<i>Expanded Levels of Support (ELoS)</i>	27
<i>Content Prompts</i>	27
<i>Specifications for Item Writing</i>	28
<i>Background and Overview</i>	28
<i>Item Writer Training</i>	28
<i>Alignment of Test Items to Grade Level Expectations</i>	28
<i>Correct Key Placement</i>	28
<i>Item Distribution by Difficulty</i>	28
<i>Item characteristics</i>	28
<i>Item writing criteria</i>	29
<i>Internal Review of the Items and Forms</i>	30

CHAPTER 4: TEST ADMINISTRATION PROCEDURES

Overview32
Student Population Tested32
Accommodations32
Test Administrators33
 Test Administrator Training33
 Scorer Training and Qualification – Online Proficiency34
 Scoring Materials and Process34
 Quality Control of Scoring – Reliability of the Alternate Assessment Administration and Scoring Process: Training to Become a Qualified Assessor.....34
 Training to become a Qualified Mentor Trainer35
 Qualified Mentor Trainers (QT or Mentors) Additional Responsibilities36
 Quality Assurance of Test Development, Administration, and Scoring.....37

CHAPTER 5: SCORING

Overview39
Data Entry.....39
 ELOS Only40
 Standard Administration With or Without Accommodations AND Then Switched to the ELOS40
 Standard Administration With or Without Accommodations.....40

CHAPTER 6: STANDARD SETTING

Outcomes from Science Standard Setting42
 Proposed Cut Scores.....42
 Impact Data42
 All Grades Proficiency Classification42
 Proficiencies at Four Levels43
Setting Standards in Science44
 Judges Participating in Science Standard Setting44
 Overview of Standard Setting Process.....45
 Agenda – Alternate Assessment Overview.....46
 Evaluation of Standard Setting Workshop - Tabular results47
 Comments about Standard Setting Process47
 Description of Cut Point Scores and Performance Levels48
 Grade level Cut Point Scores and Performance Levels.....48
 Grade 4 – cut point scores summary49
 Grade 4 – teachers discussion49
 Grade 8 – cut point scores summary49
 Grade 8 – teachers discussion50
 Grade 10 – cut point scores summary50
 Grade 10 – teachers discussion51

<i>General Standard Setting Issues (by strand – test)</i>	51
<i>Proficiency Level Descriptors – Initial Development and Grade Levels</i>	52
<i>Initial development</i>	52
<i>Grade 4 Proficiency Level Descriptors</i>	53
<i>Advanced Level</i>	53
<i>Proficient Level</i>	53
<i>Below Proficient Level</i>	53
<i>Far Below Proficient Level</i>	53
<i>Grade 8 Proficiency Level Descriptors</i>	54
<i>Advanced Level</i>	54
<i>Proficient Level</i>	54
<i>Below Proficient Level</i>	54
<i>Far Below Proficient Level</i>	54
<i>Grade 10 Proficiency Level Descriptors</i>	55
<i>Advanced Level</i>	55
<i>Proficient Level</i>	55
<i>Below Proficient Level</i>	55
<i>Far Below Proficient Level</i>	55

CHAPTER 7: REPORTING

<i>Overview</i>	56
<i>Report Types</i>	56
<i>Unofficial Student Report</i>	56
<i>Official Report</i>	56
<i>Parent Guide to Interpretation of Individual Student Reports</i>	56
<i>Educator Guide to Interpretation of Individual Student Reports</i>	56
<i>DRA Secure Reporting Website Overview</i>	57

CHAPTER 8: TECHNICAL DOCUMENTATION

<i>Overview</i>	58
<i>Reliability</i>	59
<i>Validity</i>	60
<i>Grade 4 Results</i>	61
<i>Grade 8 Results</i>	62
<i>Grade 10 Results</i>	63
<i>Response Processes: Analysis of ELOS</i>	65
<i>Response Processes: Relations Among Science and Reading, Writing, and Mathematics</i>	66
<i>Summary</i>	68

CHAPTER 9: PROGRAM IMPROVEMENT

<i>Overall Program Evaluation</i>	69
<i>Summary of Consequences Survey</i>	69
<i>Training and Qualifications</i>	69
<i>Test Administration and Decision Making</i>	69

<i>Accessibility and Results</i>	70
<i>Instructional Relevance</i>	70
<i>Professional Developmental Needs</i>	70
<i>Teacher Demographics and Experiences</i>	70
<i>Recommendations for Future Consideration</i>	71
<i>Science Grade 4 Comments on Item Revisions</i>	71
<i>Science Grade 8 Comments on Item Revision</i>	72
<i>Field Testing of New Items (Standard and ELOS)</i>	72
<i>Package of Test Booklets and Training of Teachers</i>	72

APPENDICES

<i>Appendix 2_1 – Pilot Test Scoring Protocols and Student Materials</i>	14
<i>Appendix 2_2 – Science Pilot Assessment Results</i>	14
<i>Appendix 3_1 – Alaska Science ExGLE book</i>	15
<i>Appendix 3_2 – Forms Used in Bias Review</i>	21
<i>Appendix 3_3 – Alignment of Science Tasks and Items with EXGLEs</i>	31
<i>Appendix 4_1 – Teacher Participation Guide</i>	32
<i>Appendix 4_2 – Qualification Process: Forms and Procedures</i>	37
<i>Appendix 4_3 – Quality Assurance in Test Development and Administration</i>	37
<i>Appendix 4_4 – Alaska Alternate Assessment Training Report</i>	38
<i>Appendix 6_1 – Standard Setting Booklets for Grade 4</i>	46
<i>Appendix 6_2 – Standard Setting Booklets for Grade 8</i>	46
<i>Appendix 6_3 – Standard Setting Booklets for Grade 10</i>	46
<i>Appendix 6_4 – Three Rounds of Teacher Judgments</i>	48
<i>Appendix 6_5 – Methods for Establishing Proficiency Level Descriptors and Extended Grade Level Expectations</i>	52
<i>Appendix 7_1 – Unofficial Student Report</i>	56
<i>Appendix 7_2 – Official Report</i>	56
<i>Appendix 7_3 – Parent Guide to Test Interpretation</i>	56
<i>Appendix 7_4 – Educator Guide to Test Interpretation</i>	57
<i>Appendix 7_5 – DRA Secure Reporting Website</i>	57
<i>Appendix 8_1 – Statistical Data from Science Assessments</i>	64
<i>Appendix 8_2 – Response Processes: Analysis of ELOS</i>	65
<i>Appendix 9_1 – Consequences Survey</i>	71

FIGURES

<i>Figure 1 – Model of Validation</i>	9
<i>Figure 2 – Grade 4 Results by Percentage</i>	62
<i>Figure 3 – Grade 8 Results by Percentage</i>	63
<i>Figure 4 – Grade 10 Results by Percentage</i>	64

TABLES

<i>Table 1 – Alternate Assessment-Extended Science Pilot Test Development Schedule</i>	13
<i>Table 2 – Science ExGLEs</i>	15
<i>Table 3 – Grade 4 Test Blueprint</i>	17
<i>Table 4 – Grade 8 Test Blueprint</i>	18

<i>Table 5 – Grade 10 Test Blueprint</i>	<i>19</i>
<i>Table 6 – Bias and Sensitivity Review Participants.....</i>	<i>20</i>
<i>Table 7 – Bias and Sensitivity Review Comments – Elementary School Level</i>	<i>21</i>
<i>Table 8 – Bias and Sensitivity Review Comments – Middle School Level.....</i>	<i>21-22</i>
<i>Table 9 – Bias and Sensitivity Review Comments – High School Level.....</i>	<i>22</i>
<i>Table 10 – Alignment Dimensions</i>	<i>26</i>
<i>Table 11 – Science Proposed Cut Scores</i>	<i>42</i>
<i>Table 12 – Grade 4, 8, 10 Total Prof Class (below/above) (raw score)</i>	<i>42</i>
<i>Table 13 – Grade 4 Total Proficiency Classification (raw score)</i>	<i>43</i>
<i>Table 14 – Grade 8 Total Proficiency Classification (raw score)</i>	<i>43</i>
<i>Table 15 – Grade 10 Total Proficiency Classification (raw score).....</i>	<i>43</i>
<i>Table 16 – Science Standard Setting Participant Information</i>	<i>44-45</i>
<i>Table 17 – April 24 Agenda.....</i>	<i>46</i>
<i>Table 18 – April 25 Agenda.....</i>	<i>47</i>
<i>Table 19 – Results of Science Standard Setting.....</i>	<i>47</i>
<i>Table 20 – Grade 4 Cut Scores Summary.....</i>	<i>49</i>
<i>Table 21 – Grade 8 Cut Scores Summary.....</i>	<i>50</i>
<i>Table 22 – Grade 10 Cut Scores Summary.....</i>	<i>51</i>
<i>Table 23 – Standard Administration Results</i>	<i>58</i>
<i>Table 24 – Cronbach’s Alpha for Grades 4, 8, 10.....</i>	<i>60</i>
<i>Table 25 – Grade 4 Results.....</i>	<i>61</i>
<i>Table 26 – Grade 4 Results by Strand</i>	<i>61</i>
<i>Table 27 – Grade 8 Results.....</i>	<i>62</i>
<i>Table 28 – Grade 8 Results by Strand</i>	<i>63</i>
<i>Table 29 – Grade 10 Results.....</i>	<i>64</i>
<i>Table 30 – Grade 10 Results by Strand</i>	<i>64</i>
<i>Table 31 – Grade 4 Correlations Among Science and Reading, Writing, and Mathematics Scores Overview.....</i>	<i>66</i>
<i>Table 32 – Grade 8 Correlations Among Science and Reading, Writing, and Mathematics Scores.....</i>	<i>67</i>
<i>Table 33 – Grade 10 Correlations Among Science and Reading, Writing, and Mathematics Score.....</i>	<i>68</i>

Glossary

A	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
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.

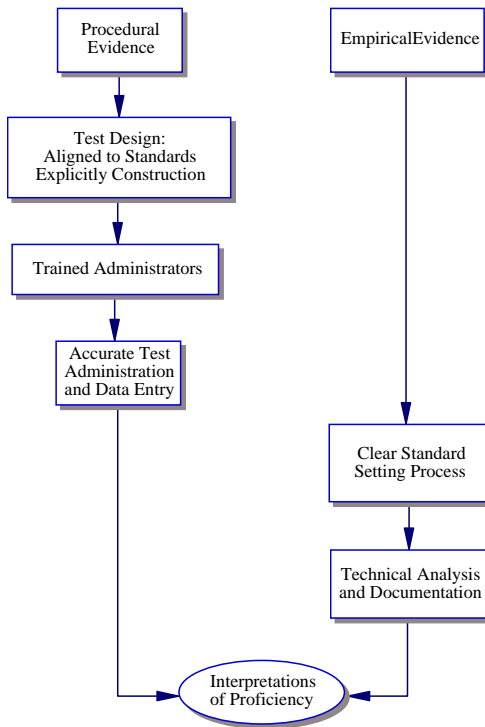
In the end, both procedural and empirical evidence are brought to bear for supporting the claim that students with significant cognitive disabilities are achieving at various levels of proficiency on the alternate assessment.

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 AND BIAS REVIEW

Pilot Test Development

In the winter of 2007, an alternate assessment developed in Oregon was brought into Alaska for field-testing. Teachers were trained in the winter to administer and score the sub-tests from the total battery. A science pilot test was conducted in the spring of 2007. Originally, the test was called the Extended Science Alternate Assessment and later changed to Alaska Alternate Assessment in Science. The Extended Science Pilot test was comprised of eleven tasks, each containing 5 items. Teachers were provided with both an administration and scoring booklet, and a booklet of student materials. All tasks included graphics depicting science facts, concepts, or principles.

Summary of Alternate Assessment-Extended Science Pilot Test Development Schedule

Table 1. Alternate Assessment-Extended Science Pilot Test Development Schedule

Task	Deliverable Date
Science bias review conducted	Oct. 30, 2006
Final science scoring protocol and bias review summary	Jan. 10, 2007
Final version of individual score report	Jan. 10, 2007
Test blueprints - read, write, math, science (incl. ELOS)	Feb. 1, 2007
Alignment study conducted (read, write, math, science)	Feb. 5-7, 2007
Design and develop extended science training modules	Feb. 8, 2007
Science pilot test training	Feb. 8, 2007
Alignment study report due (including Accessibility Study)	Mar. 15, 2007
Alternate Assessment Science Pilot Test Window	Feb. 12-April 13, 2007
Technical studies presented to TAC	May 3, 2007
Inter-rater reliability study from proficiency test	
Internal consistency from operational test	
Content-related evidence (blueprint and alignment)	
Criterion-related evidence	
Standard setting	
National TAC meeting	May 9-10, 2007
Report for Peer review	May 9-10, 2007
Draft Contract with Statement of Work, Schedule, and Deliverables	May 28, 2007
Data tapes to EED with results and proposed cut scores for reading, writing, mathematics to EED for AYP	Jun. 14, 2007
Alt. assessment web reports posted (as pdf): Individual student reports and district reports (perhaps school and classroom reports)	Jun. 14, 2007
2007-2008 Contract and Statement of Work to AK	Jun. 14, 2007
Contracts Office	
Recommendations for Proposed Cut Scores to State Board	Jun. 30, 2007
Second batch of data tapes and results with adopted cut scores for reading, writing, and mathematics to EED for AYP	Aug. 14, 2007
Secure materials available for download	Feb. 15, 2008

Twenty-three teachers participated in the science pilot, testing a total of twenty-six students. The pilot consisted of field test administration of the Extended Science test items to students of all grade levels. The participating teachers provided student scores as well as feedback about the test

items. Not all students took all tasks, so the statistics were based on only a few students and items.

Appendix 2_1 contains the *pilot test scoring protocols and student materials*. These materials provide teachers explicit guidance in administration of the test, both general and specific language to use when testing the students, possible accommodations, a place to mark student responses, and example scores for various responses. Student materials are bundled separately and formatted for teachers to break up into individual items with cards so that students are not distracted or overwhelmed by too much information.

< Appendix 2_1 >

Statistical Analyses

Because teachers administered only subtests, no formal statistical analyses could be conducted on the results. Rather, descriptive statistics and qualitative comments were used to analyze the functioning of the science alternate assessment. In addition, Cronbach's alpha was computed for each of the 11 tasks. In general, the tasks functioned adequately with the standard deviation less than the mean and reliability coefficients well above .70s. The results of the Extended Science pilot were very encouraging.

Comments from the participating teachers indicated possible changes to make the test more accessible to students such as accommodations involving real objects and color flashcards, and making the items more specific to the Alaskan culture and environment.

Appendix 2_2 contains the 2007 *Science pilot assessment results* document that presents the pilot items, statistical results by item, and comments made by pilot participants. The results indicated that (a) no obvious administration problems were apparent from the manner in which the materials were organized and (b) students could respond to the various formats with every task having at least some students receiving full credit.

< Appendix 2_2 >

CHAPTER 3: TEST DESIGN AND ITEM/TASK DEVELOPMENT

Overview

The general education Science Performance Grade Level Expectations (SPGLEs) articulate the skills students need to learn and be able to do at end of a given grade level.

Organization of Science Extended Grade-Level Expectations

Table 2. Science ExGLEs

Strand – Grade 4-8-10
SB-Understand the concepts, models, theories, universal principles, and facts that explain the physical world.
SB-Understand the concepts, models, theories, universal principles, and facts that explain the physical world.
SB-Understand the concepts, models, theories, universal principles, and facts that explain the physical world.
SB-Understand the concepts, models, theories, universal principles, and facts that explain the physical world.
SC-Understand the concepts, models, theories, facts, evidence, systems, and processes of life science.
SC-Understand the concepts, models, theories, facts, evidence, systems, and processes of life science.
SC-Understand the concepts, models, theories, facts, evidence, systems, and processes of life science.
SD- Understand the concepts, processes, theories, models, evidence, and systems of earth and space sciences.
SD- Understand the concepts, processes, theories, models, evidence, and systems of earth and space sciences.
SD- Understand the concepts, processes, theories, models, evidence, and systems of earth and space sciences.
SD- Understand the concepts, processes, theories, models, evidence, and systems of earth and space sciences.
SE- Understand the relationships among technology and science.
SE- Understand the relationships among technology and science.

Science **Extended** Grade Level Expectations (ExGLEs) were developed for students in grades 4, 8, and 10 to accompany Alaska’s Science Performance Standards/Grade Level Expectations (SPGLEs). The science alternate assessment was developed in alignment with the Extended Grade Level Expectations (ExGLEs) that had been developed by EED.

The ExGLEs were developed to avoid including statements of curricular activities, instructional strategies, or value-laden concepts and understandings. The ExGLEs are reduced in complexity and breadth in order to provide access to the general education curriculum for students with significant cognitive disabilities.

Appendix 3_1 contains the *Alaska Science ExGLE* book. This appendix document provides organization of the ExGLEs, the ExGLEs by grade cluster, and draft Proficiency Level Descriptors (PLDs).

< Appendix 3_1 >

Test Blueprint

Dillard Research Associates designed tasks that were directly aligned with each ExGLE. In each grade to be tested (4, 8, and 10), both scoring protocols and student materials were developed. The scoring protocols are documents that specify all administration procedures and provide scoring boxes, while the student materials provide documents that are used directly with the students.

For 2007-2008, reduction in breadth for the alternate assessment was achieved systematically by the process of selecting eligible content (standards) for development based on the vertical alignment of content standards. A two-step process was used in which item writers first examined the eligible content (standards) within a particular grade. The second step was to examine the eligible content (standards) across grades.

This process involved selecting eligible content (standards) that are not only less complex, but are concepts, terms, and knowledge that are simpler or place less demand on student behaviors. Specifically, test items are written to address eligible content (standards) that demand these less complex behaviors such as recognition, recall, and comprehension skills. These skills are isolated into distinct behaviors such as identify, state, define, label, conclude, group, restate, review, or translate.

Content items are scored on a scale of 0 – 2 points, with 0 reflecting a completely incorrect response, and 2 reflecting a completely correct response. Although partial credit is used with the reading, writing, and mathematics tests (and therefore teachers are familiar with this type of scoring), no partial credit is given in the science assessment. The scoring guide is available on the scoring protocol to help teachers score student responses.

Alternate Assessment items on the assessment are not arranged according to difficulty. The test is arranged in tasks that align to content standards and strands and, within tasks, six content items are arranged progressively through a series of standards or through a sequence of related items; the order of placement within tasks is not based on item difficulty. In fact, teachers and test administrators may start the student at any location on the test that they deem appropriate/easiest for the student.

The following tables contain a crosswalk of Science Strands, Science Content Standards, Science ExGLEs, and science tasks for each of the grade levels tested (4, 8, and 10).

Table 3. Grade 4 Test Blueprint

Strand – Grade 4	Content Standard	ExGLE	Concepts of Physical Science	Concepts of Life Science	Concepts of Earth Science	History and Nature of Science, Science and Technology
SB-Understand the concepts, models, theories, universal principles, and facts that explain the physical world.	B1-Concepts of Physical Science	[3/4] SB1.1 I	1.4: Items 1-2 = 4 pts.			
SB-Understand the concepts, models, theories, universal principles, and facts that explain the physical world.	B1-Concepts of Physical Science	[3/4] SB2.1	1.4: Item 3 = 2 pts.			
SB-Understand the concepts, models, theories, universal principles, and facts that explain the physical world.	B1-Concepts of Physical Science	[3/4] SB3.1	1.4: Items 4-5 = 4 pts.			
SB-Understand the concepts, models, theories, universal principles, and facts that explain the physical world.	B1-Concepts of Physical Science	[3/4] SB4.1	1.4: Item 6 = 2 pts.			
SC-Understand the concepts, models, theories, facts, evidence, systems, and processes of life science.	C1-Concepts of Life Science	[3/4] SC1.1		2.4: Items 1-2 = 4 pts.		
SC-Understand the concepts, models, theories, facts, evidence, systems, and processes of life science.	C1-Concepts of Life Science	[3/4] SC2.1		2.4: Items 3-4 = 4 pts.		
SC-Understand the concepts, models, theories, facts, evidence, systems, and processes of life science.	C1-Concepts of Life Science	[3/4] SC3.1		2.4: Items 5-6 = 4 pts.		
SD- Understand the concepts, processes, theories, models, evidence, and systems of earth and space sciences.	D1-Concepts of Earth Science	[3/4] SD1.1			3.4: Item 1 = 2 pts.	
SD- Understand the concepts, processes, theories, models, evidence, and systems of earth and space sciences.	D1-Concepts of Earth Science	[3/4] SD1.2			3.4: Items 2-3 = 4 pts.	
SD- Understand the concepts, processes, theories, models, evidence, and systems of earth and space sciences.	D1-Concepts of Earth Science	[3/4] SD2.1			3.4: Items 4-5 = 4 pts.	
SD- Understand the concepts, processes, theories, models, evidence, and systems of earth and space sciences.	D1-Concepts of Earth Science	[3/4] SD3.1			3.4: Item 6 = 2 pts.	
SE-Understand the relationships among science, technology, and science.	E1-Science and Technology	[3/4] SE2.1				4.4: Items 1-3 = 6 pts.
SE-Understand the relationships among science, technology, and science.	E1-Science and Technology	[3/4] SE3.1				4.4: Item 4 = 2 pts.
SG-Understand the history and nature of science	G1-History and Nature of Science	[3/4] SG1.1				4.4: Item 5 = 2 pts.
SG-Understand the history and nature of science	G1-History and Nature of Science	[3/4] SG2.1				4.4: Item 6 = 2 pts.

Table 4. Grade 8 Test Blueprint

Strand – Grade 8	Content Standard	ExGLE	Concepts of Physical Science	Concepts of Life Science	Concepts of Earth Science	Science and Technology
SB-Understand the concepts, models, theories, universal principles, and facts that explain the physical world.	B1-Concepts of Physical Science	[7/8] SB1.1	1.8: Item 1 = 2 pts.			
SB-Understand the concepts, models, theories, universal principles, and facts that explain the physical world.	B1-Concepts of Physical Science	[7/8] SB3.1	1.8: Item 2 = 2 pts.			
SB-Understand the concepts, models, theories, universal principles, and facts that explain the physical world.	B1-Concepts of Physical Science	[7/8] SB4.1	1.8: Items 3-4 = 4 pts.			
SB-Understand the concepts, models, theories, universal principles, and facts that explain the physical world.	B1-Concepts of Physical Science	[7/80] SB2.1				
SC-Understand the concepts, models, theories, facts, evidence, systems, and processes of life science.	C1-Concepts of Life Science	[7/8] SC1.1	1.8: Items 5-6 = 4 pts.			
SC-Understand the concepts, models, theories, facts, evidence, systems, and processes of life science.	C1-Concepts of Life Science	[7/8] SC2.1		2.8: Items 1-2 = 4 pts.		
SC-Understand the concepts, models, theories, facts, evidence, systems, and processes of life science.	C1-Concepts of Life Science	[7/8] SC3.1		2.8: items 3-4 = 4 pts.		
SD- Understand the concepts, processes, theories, models, evidence, and systems of earth and space sciences.	D1-Concepts of Earth Science	[7/8] SD1.1		1.8: Items 5-6 = 4 pts.		
SD- Understand the concepts, processes, theories, models, evidence, and systems of earth and space sciences.	D1-Concepts of Earth Science	[7/8] SD1.2			3.8: Item 1 = 2 pts.	
SD- Understand the concepts, processes, theories, models, evidence, and systems of earth and space sciences.	D1-Concepts of Earth Science	[7/8] SD2.1			3.8: Items 2-3 = 4 pts.	
SD- Understand the concepts, processes, theories, models, evidence, and systems of earth and space sciences.	D1-Concepts of Earth Science	[7/8] SD3.1			3.8: Items 4-5 = 4 pts.	
SD- Understand the concepts, processes, theories, models, evidence, and systems of earth and space sciences.	D1-Concepts of Earth Science	[7/8] SD4.1			3.8: Item 6 = 2 pts.	
SE- Understand the relationships among technology, and science.	E1-Science and Technology	[7/8] SE2.1				4.8: Items 1-5 = 10 pts.
SE- Understand the relationships among technology and science.	E1-Science and Technology	[7/8] SE3.1				4.8: Item 6 = 2 pts.

Table 5. Grade 10 Test Blueprint

Strand – Grade 10	Content Standard	ExGLE	Concepts of Physical Science	Concepts of Life Science	Concepts of Earth Science	Science and Technology
SB-Understand the concepts, models, theories, universal principles, and facts that explain the physical world.	B1-Concepts of Physical Science	[9/10] SB1.1				
SB-Understand the concepts, models, theories, universal principles, and facts that explain the physical world.	B1-Concepts of Physical Science	[9/10] SB2.1	1.10: Items 1-2 = 4 pts.			
SB-Understand the concepts, models, theories, universal principles, and facts that explain the physical world.	B1-Concepts of Physical Science	[9/10] SB3.1	1.10: Items 4-5 = 4 pts.			
SB-Understand the concepts, models, theories, universal principles, and facts that explain the physical world.	B1-Concepts of Physical Science	[9/10] SB4.1	1.10: Items 5-6 = 4 pts.			
SC-Understand the concepts, models, theories, facts, evidence, systems, and processes of life science.	C1-Concepts of Life Science	[9/10] SC1.1		2.10: Items 1-2 = 4 pts.		
SC-Understand the concepts, models, theories, facts, evidence, systems, and processes of life science.	C1-Concepts of Life Science	[9/10] SC2.1		2.10: Items 3-4 = 4 pts.		
SC-Understand the concepts, models, theories, facts, evidence, systems, and processes of life science.	C1-Concepts of Life Science	[9/10] SC3.1		2.10: Items 5-6 = 4 pts.		
SD- Understand the concepts, processes, theories, models, theories, evidence, and systems of earth and space sciences.	D1-Concepts of Earth Science	[9/10] SD1.1				
SD- Understand the concepts, processes, theories, models, evidence, and systems of earth and space sciences.	D1-Concepts of Earth Science	[9/10] SD1.2			3.10: Items 1-2: 4 pts.	
SD- Understand the concepts, processes, theories, models, evidence, and systems of earth and space sciences.	D1-Concepts of Earth Science	[9/10] SD2.1			3.10: Item 3 = 2 pts.	
SD- Understand the concepts, processes, theories, models, evidence, and systems of earth and space sciences.	D1-Concepts of Earth Science	[9/10] SD3.1			3.10: Items 4-6 = 6 pts.	
SE- Understand the relationships among technology and science.	E1-Science and Technology	[9/10] SE2.1				4.10: Items 1-3 = 6 pts.
SE- Understand the relationships among technology and science.	E1-Science and Technology	[9/10] SE3.1				4.10: Items 4-6 = 6 pts.

Item/Task Development

Item Construction and Review

The new test was constructed in 2007-2008 and entitled the Alternate Assessment in Science. The new assessment in science consisted of three individual grade level assessments that were closely aligned to the ExGLEs. Items were created and went through an extensive review process by DRA and EED, including feedback from the field in the form of a bias and sensitivity review. Throughout the assessment content review, items were cross-walked to the ExGLEs and Content Standards, and the assessment was minimized from 50 items to 24 items per grade level.

Bias and Sensitivity Review

Initially, a bias review was held in person in Anchorage on October 30, 2006 as part of the pilot study. For the actual assessment used in the 2007-2008 administration, a new 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 second 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 6. 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 science test (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_2 includes an example of the *forms used in the bias review*: test security agreement for participants, explanation of the process, crib sheet, cover sheet, Instrument 1 – Linguistic Complexity Rubric for Universal Design Item-Task Development, and Instrument 2 – Bias and Sensitivity Review Checklist.

< Appendix 3_2 >

The following comments were collected during and after the panel. The comments were then carefully reviewed and items were modified to reflect suggested changes from the bias and sensitivity review.

Table 7. Bias and Sensitivity Review Comments – Elementary School Level

Science Task – Elementary School	Comments
2.4 - Structure and Properties of Matter	
3.4 - Chemical and Physical Changes	#1 - Answer in wrong order. #5 - burning leaves in directions to match pictures
4.4 - Fundamental Forces and Motions	Wagon - unfamiliar, Bigger than car
5.4 - Interaction of Energy and Matter	#3 - sunscreen lotion used for skiing
6.4 - Organism Characteristics and Needs	
7.4 - Classification, Life Cycle	#4 - directions: chicks/pictures: frogs + answer sheet
8.4 - Interdependence of Organisms in Environment	
9.4 - Survival, Structure, Function	#2 - pictures: it's s/b its
10.4 - Structure of Earth and Material Use	#1 - what s/b where #3 - seeds in dirt (maybe not clear that water is most correct). #5 - box of tissues directions and picture.
11.4 - Weather and Seasons	#3 - Directions don't match pictures. #4 - Directions don't match pictures. #5 - Directions don't match pictures. For all - answers match pictures.

Table 8. Bias and Sensitivity Review Comments – Middle School Level

Science Task – Middle School	Comments
2.7 - Changes of State	#3 - picture of water, not ocean
3.7 - Force, Mass, and Motion	#2 - picture "pushing" directions "with a" #3 - add: all made of same material to directions. #4 - directions say wagons are "different" they look same. #5 - Pictures don't show 'person'
4.7 -Fundamental Forces and Motions	#3 - wrong correct answer (pg. 8). #4 - Does student materials show correct answer?
5.7 - Types of Energy / Transformations	Sunscreen/dark skin/alternate substance
6.7 - Organisms / Structures	
7.7 - Energy, Flow, Photosynthesis / Organisms	What is Item 1-3 for ? (plant picture). #3 - more correct

	'utility pole' #4 - Worms in picture, not in directions
8.7 - Heredity	#4 - differs from high school - male is curly?
9.7 - Evolution, Selection, and Adaptation	#2 – worm sin picture, not directions
10.7 - Scientific Inquiry and Understanding	#1 - Answer could be "like" - all boys jumping. #2 - How can there be 1 correct answer? #5 - "a small dog"
General Comments	Label ALL pages "high school" "middle school" or "elementary school"

Table 9. Bias and Sensitivity Review Comments – High School Level

Science Task High School	Comments
2.10 - Changes of State	
3.10 - Force, Mass, and Motion	Directions (pg. 3) are not in order of task pictures. The wagon tongue and reins wrong. "Table" for "it" (#5 - pg. 4 scoring = additional words)
4.10 - Force / Gravity	Student pg. missing Item 3. Wording for answer Item 5 (pg. 6)
5.10 - Types of Energy / Transformations	Windmill a difficult choice? 'Waves' not really in pictures.
6.10 - Organisms / Structures	Question: glass of H ₂ O/picture: water. Pictures don't match directions or scoring sheet. Direction: carton/ picture: gallons
7.10 - Energy, Flow, Photosynthesis / Organisms	
8.10 - Heredity	Directions - genes - dominant "override"? #1 - dark spots - are clear on student pictures. #2 - directions don't mention atoms - can you "see" genes? #5 - wording diff. between directions and student sheets
9.10 - Evolution, Selection, and Adaptation	#2 - directions "5 frogs"? #5 - wording different between directions and student sheets
10.10 - Scientific Inquiry and Understanding	#1 - Graph in student materials is fish, directions say plants.
11.10 - The Earth, Space, and Resources	

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

Background

This document explains the specifications used when the Science 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 Science Alternate Assessments. Following this brief description of terms, a more thorough presentation of the test design is presented. 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 primarily a selected response format, with each item having a single correct answer that is selected from among presented choices. For all items, teachers assign a score of 0 (no credit) or 2 (full credit). 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 are organized into consumable workbooks by subject area comprised of two pages per task (one page for administration and 1 page for scoring). Student materials are 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, 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 are 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 reduced during development in a variety of ways, reduction in depth of knowledge assessed, breadth of content standards covered, and complexity of content required.

⁴U. 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.

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 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.

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.

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. A one-to-one correspondence is present with the tasks and items in relation to the grade level expectations: All strands/attributes are equally addressed in the proportion of points accumulated for the total test, which was fixed at 48 points.

Table 10. Alignment Dimensions

	<i>Original Web Definition</i>	<i>Tindal’s Adaptation for students with</i>	<i>Alaska Alignment Study operational</i>
--	--------------------------------	----------------------------------------------	-------------------------------------------

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.

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

Test Development

The Alternate Assessment was developed using the following general process. Similar items were grouped into tasks, with 6 items for each task. Scoring was developed for partial credit (1 point) or full credit (2 points) in all other subject areas, though no partial credit was allowed in science. All grade level expectations were addressed. The architecture of each task followed the same format with two types of items: (a) standard items focusing directly on content prompts, and (b) extended levels of support (ELOS) items for students who scored zero on three successive items across three successive tasks.

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.

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.

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.				

‘*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]), along with the levels of materials (ascertain attention). 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.

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]), along with the levels of materials (ascertain attention). 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 are 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 (though only 0 or 2 points were used for the science assessment). Content prompts were generally oriented toward functional issues that students need 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, the content addressed weather and seasons, as well food and movement.

Specifications for Item Writing

Background and Overview

The Science 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 science test 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 linked to the extended grade level expectations using a rigorous process at three points during the test item development: Initially in the design of the items, during an initial alignment study, and as part of the bias and content review.

Correct Key Placement

During the development process, item writers rotated the correct key to avoid fixed responding.

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 was not based on item difficulty. Teacher administrators had the option to start the student at any location on a test that they deemed appropriate and easiest for the student.

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.

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

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 and 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.

Internal Review of the Items and Forms

Prior to release in the field, but after all training materials and practice tests had been developed, the drafts of student materials and scoring protocols were revised to better align the tasks and

items with the Alaska Extended Grade Level Expectations and to reduce the test length to 24 items and 48 points.

Appendix 3_3 contains a spreadsheet *aligning the science task and item numbers with the Extended Grade Level Expectations*, Alaska Strands, and comments from EED indicating suggested edits during the item review process.

< Appendix 3_3 >

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 can move to ELOS items for instructional feedback the first year and may move directly to ELOS items in the following year.

The Expanded Levels of Support (ELOS) test items were developed for very low functioning students and administered to students who met the 3 consecutive error rule (e.g. failure on 3 items and 3 tasks). This rule is explained in its entirety in Appendix 4_1: 2007-08 Teacher Participation Guide (p. 13).

CHAPTER 4: TEST ADMINISTRATION PROCEDURES

Overview

The Alaska Science Alternate Assessment is individually administered in grades 4, 8, and 10, in a one-on-one setting with the test administrator and student. The setting and the amount of time devoted to testing is entirely flexible and decided upon by the administrator, who best understands what is most appropriate and beneficial to the student.

For each grade level, there are four tasks, with six items each. The scoring protocol is a concise eight pages per grade level, with each task devoting one page to prompting directions, followed by one page for scoring. Each test has a total point value of 48. Each item contains a scripted statement in which teachers present the item and the student is directed to select or generate a response. All tasks include graphics depicting a science fact, concept, or principle. These materials allow for student responses to be either verbal or nonverbal.

The necessary scoring protocol and materials are available for download on the training website. The website was built with 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, plus test administration rules, Teacher Participation Guide, and the Science Training Manual for Administration and Scoring.

<Appendix 4_1>

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 Science Training Manual 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 the QA in how to present 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 goes through, and 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. Although there is less grey area in scoring the subject area of Science, ample practice is still provided through 11 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, he or she must contact the Dillard Research Associates (DRA) helpdesk to have their account reset, thus requiring the Assessor-in-Training to retake all proficiency tests. A total of 256 teachers reached proficiency on the science tests the first time.

Scoring Materials and Process

Scoring materials 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 two for a correct response, or a zero for an incorrect response. There is no partial credit awarded on the Science Assessment. 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 breaking down the scores by task and item. The unofficial report 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. Qualified Assessors receive a multiple step training in order to qualify as a test administrator. Each district is 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 has completed the 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 going on to become 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. The preferred qualifications for mentors are: 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 approximately 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
- 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. Training is conducted by EED and the test vendor. 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. Training of mentors is supervised by both EED and the vendor.

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 4_2 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 4_2 >

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. See appendix 4_2, which contains the Test Security Agreement and the updated Test Security and Online Test Security Guidelines.

Procedures were established for *ensuring quality for both test development and test administration*.

<Appendix 4_3>

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 4_4 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 4_4 >

CHAPTER 5: SCORING

Overview

For the Alaska Alternate Science Assessment, scoring items is objective and straightforward as this assessment is designed around the concept of scoring a student’s independent response. If the student produces a correct response, two points are awarded. If the student is incorrect, no points are given. Once the scores have been recorded, the QA 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 consists of 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 will 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 has been entered and the “Submit Scores” option is 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. Also, all items marked "I" need to have a reason given. 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 will appear and indicate 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. 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. Also, all items marked "I" need to have a reason given. 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 is entered incorrectly 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 will appear and indicate 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 SETTING

Outcomes from Science Standard Setting

Participants were trained on the standard setting process, and then within their groups analyzed the science assessments and possible scores, proposing a set of cut scores. Groups were then given actual student data to cross reference with their proposed cut scores. Final decisions were made at the end of day two and presented to all participants for review. Outcomes from the standard setting in science included proposed cut scores, review of impact data, and finalized cut scores.

Proposed cut scores

The following table presents finalized cut scores. To obtain a proficiency level of advanced, proficient, below proficient, or far below proficient in the Alaska Alternate Assessment in Science, a student must obtain a score as follows:

Table 11. Science Proposed Cut Scores

Proficiency Level	Grade 4	Grade 8	Grade 10
Science: Advanced	44 or above	44 or above	44 or above
Science: Proficient	24-43	29-43	26-43
Science: Below proficient	12-23	16-28	18-25
Science: Far Below Proficient	11 or below	15 or below	17 or below

Impact Data

After initial cut scores were proposed, each group received impact data of actual student scores from the Alternate Assessment in Science. They then continued to assign each student’s proficiency level based on their proposed cut scores, and decide if the cut scores they had chosen were compatible with the given data. The following tables present the student impact data for each grade level in accordance to group decisions on final cut scores.

All Grades Proficiency Classification

Table 12. Grade 4, 8, 10 Total Proficiency Classification (below/above) (raw score)

	Gr. 4 Freq.	Gr. 4 Percent	Gr. 8 Freq.	Gr. 8 Percent	Gr. 10 Freq.	Gr. 10 Percent
far below/ below	19	30.6	22	41.5	19	22.6
proficient/ advanced	43	69.4	31	58.5	65	77.4
Total	62	100.0	53	100.0	84	100.0

Proficiencies at Four Levels

Table 13. Grade 4 Total Proficiency Classification (raw score)

Level	Frequency	Percent
far below	13	21.0
below	6	9.7
proficient	26	41.9
advanced	17	27.4
Total	62	100.0

Table 14. Grade 8 Total Proficiency Classification (raw score)

Level	Frequency	Percent
far below	13	24.5
below	9	17.0
proficient	11	20.8
advanced	20	37.7
Total	53	100.0

Table 15. Grade 10 Total Proficiency Classification (raw score)

Level	Frequency	Percent
Far below	14	16.7
below	5	6.0
proficient	50	59.5
advanced	15	17.9
Total	84	100.0

Setting Standards in Science

This section provides information and discussion on the process and outcomes of the standard setting in science. The standard setting was conducted April 24-25, 2008 at the Talking Books library in Anchorage, AK. There were a total of twenty-one participants, divided into groups of 7 for each grade level. Teachers were recruited based on their experience in science content at each of the grade levels or their experience in special education, particularly with students having significant cognitive disabilities.

The groups first reviewed test administration using the scoring protocol and student materials for their assigned grade level. Next, the three groups reviewed the Extended Grade Level Expectations (ExGLEs) for the grade level they were assigned. Once they were familiar with the ExGLEs, they reviewed a crosswalk document that mapped relevant tasks to each ExGLE. The Proficiency Level Descriptors were reviewed, and participants worked independently to make decisions on where the cut scores should lie for each task. They then discussed their rationale as a group and came to a group consensus. After cut scores were determined, each group was given an example impact data set to confirm their reasoning of each task. As a group they devised a combined across task judgment for each proficiency level and completed the standard setting form for their grade level.

Judges Participating in Science Standard Setting

Table 16. Science Standard Setting Participant Information

Last	First	District	Degrees and Credentials
Baker	Craig	Kodiak	B.S. Biology, Biology, General Science
David	Vickie		B.S., M.A., K-12 SpEd - Resource, L.D., and Adaptive
Decker	Laurie	Anchorage SD	B.S. Biology and general science, M.Ed Special Education (U of O).
Diemer	Susan	Anchorage SD	M.S. SpEd - U of O 1989, B.E.D. - Elem Ed UAA 1985, AK cert. reg. Ed K-8, AK cert. SpEd K-12
Hagberg	Anita	Mat-Su	1) B.S. Degree - Health Education, double minor - psychology and sociology. SpEd teaching endorsement, 2) Associates Degree - Psychology, counseling rehab. 3) Basic law enforcement training, Deputy Sheriffs (Forensics)
Hubbard	Joanna	Anchorage SD	B.A. Dartmouth College - Environmental and evolutionary biology. M.S. Montana State University - science education, K-8 Alaska regular teaching cert.
Hughes	Shayne	Galena	Sec. Education: highly qualified in Biology and general science. Endorsements are in: chemistry, physics, geology, biology
Hutchins	John	Mat-Su	B.A.E. Social Studies, SpEd/Regular Ed, Highly qualified in secondary and elementary education
Jones	David	KGBSD	MAT Elementary, MED Educational Leadership
Kaasa	Dan	Kenai Peninsula Borough	Assistive Technology and Augmentative/Alternative Communication Specialist. Certification: EIEd /SpEd K-12 - B.S. Education
Koenfal	Anne	FNSBSD	B.A. in SpEd, Intensive Resource Teacher grades 2nd and 6th.
Kurzbard	Harvey	FNSBSD	M.Ed - Liberal Studies

Owens	Theresa	NW Arctic B or SD	Director of SpEd and Assessment, AK - Teaching credential K-12 P.E., Adapted P. E. K-12, SpEd Resource K-8, Type B - Principal and SpEd Director. B.S. (P.E. Willamette Univ.), Ed.M (Education/APE/Oregon State), M.S. (Ed Leadership, National University)
Paden	Jon	Unalaska City School District	B.A. Chemistry, M.A.T Secondary, M.Ed Technology
Russin	Alex	LYSD	B.A. English, M.Ed - Ed leadership, Math/Science background
Sheppard - Gillam	Lori		B.S. - biology/Lewis & Clark College. M.S. - Geosciences/ Mississippi State University. Type A teaching certificate - U of AK Anchorage
Street	Stacey	KIBSD	B.S. SpEd Severe/Profound K-12
Whittstock	Bridgett	Petersburg City Schools	B.A. Japanese Lang. & Lit., Single Subject Teaching Credential for Japanese Lang. & Lit. M.Ed, major in SpEd, Learning/Severely handicapped teaching credentials K-12, Alaska Type A teacher certification
Williams	Joel	Lower Kuskokwim	B.A. Music, M.S. SpEd
Wilson	Susan	Anchorage SD	B.A. - Elementary Education, Math minor, Sped Endorsement: Graduate Studies, several certifications including: crisis prevention intervention (CPI), Linda Mood Bell (Handwriting without tears, slingerland, etc.)

Overview of Standard Setting Process

The standard setting process used for the Alternate Assessment in Science followed the procedures described by Jaeger and Mills (2004)¹⁰, an integrated judgment procedure for setting standards on complex large-scale assessments. This system is essentially a test-centered approach that requires judges to make decisions about proficiency levels in relation to a set of items relative to performance standards.

- Step 1. Train on test administration: Review both scoring protocols and student materials.
- Step 2. Review Extended Grade Level Expectations (ExGLEs).
- Step 3. Review cross walk document and relevant tasks; consider all tasks in the grade level.
- Step 4: Review draft proficiency level descriptors (PLDs).
- Step 5. Establish proficiency levels for each task and for each proficiency level (advanced, proficient, below, and far below).
- Step 6. Confer with partners and articulate a rationale.
- Step 7. Review an example data set to confirm the reasoning of each task.
- Step 8. As a group, devise a combined (across task) judgment for each proficiency level.
- Step 9. Complete a Standard Setting Form for the grade level.
- Step 10. Review Impact Data.

¹⁰ Jaeger, R. M., & Mills, C. N. (2001). An integrated judgment procedure for setting standards on complex large-scale assessments (Chapter 11, pp. 313-338). In G. J. Cizek (Ed.) *Setting performance standards: concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum.

Unlike the Angoff (1971)¹¹ procedure that relies on making judgments about the probability of a proficient person on a performance standard being able to pass an item, the integrated judgment uses a number of items for making this judgment. One of the problems with the Angoff method is the variation among the item types and the likelihood of obtaining inconsistent results. Another problem is that it requires a complex cognitive judgment: being able to imagine a person at the performance standard and then estimate the probability of answering an item correctly.

The system proposed by Jager and Mills (2001) first presents judges an opportunity to complete responses to a *set of test items*, thus maximizing the test-based information to be used in setting standards. Second, judges are asked to classify work with respect to the performance standard directly, instead of imagining an examinee’s likelihood of passing an item. Third, judges directly use the scale of the performance standard rather than an intermediate item difficulty scale.

One of the advantages of this approach is its directness in linking performance to proficiency, reducing the chain of inferences in the process. And very importantly “the procedure permits judges to make compensatory judgments about the examinee’s responses to items by letting strong performances on some test items compensate for relatively weak performances on others” (p. 316).

The following three appendices (6_1 through 6_3) contain the *Standard Setting Booklets* for grades 4, 8, and 10. These booklets document the standard setting process.

< Appendix 6_1 through 6_3 >

Agenda – Alternate Assessment. This is an agenda for the Science Standard Setting Session, which took place on April 24-25, 2008 at Talking Books Library in Anchorage, Alaska.

Table 17. April 24 Agenda

8:00-9:00	Table Leader Meeting and Continental breakfast
9:00-9:30	Welcome, Introductions, Overview & purpose of standard setting committee: Importance of taking the test, test administration rules, training in integrated judgment method, how to use impact data to make changes, rewriting PLDs.
9:30-10:15	Taking the AA Science test together, (grades 4, 8, 10) Administration rules and administering tasks in all grades
10:15-10:30	Break
10:30-11:00	Training on Process: Integrated Judgment
11:00-11:45	Round 1 of proficiency judgments
11:45-1:00	Lunch
1:00-2:00	Complete Round 1 of proficiency judgments
2:00 – 3:30	Round 2 of proficiency judgments
3:30 – 4:00	Debrief and review

¹¹ Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.). *Educational Measurement* (2nd edition, pp. 508-600). Washington, DC: American Council on Measurement.

Table 18. April 25 Agenda

8:00 – 8:30	Continental breakfast
8:30 – 10:00	Impact analysis
10:00 – 11:00	Review of Achievement Level Descriptors (PLDs)
11:00 – 12:00	Summary and evaluation of standard setting
12:00	Adjourn

Evaluation of Standard Setting Workshop – Tabular results

Table 19. Results of Standard Setting in Science

Statistic	TrainAdmin	IntJudgOrient	StdSetBkltOrg	10StepsUsfl	JudgCert
Average	3.9	3.85	3.825	3.8	3.7
Freq 1s	0	0	0	0	0
Freq 2s	0	0	0	0	1
Freq 3s	2	3	2	4	4
Freq 4s	18	17	17	16	15
Prct 1s	0%	0%	0%	0%	0%
Prct 2s	0%	0%	0%	0%	5%
Prct 3s	10%	15%	10%	20%	20%
Prct 4s	90%	85%	85%	80%	75%

Comments about standard setting process. The following comments reflect general issues participants noted about the standard setting process as a whole, or suggestions for future improvements on the text booklet or standard setting process.

The student book (language/cues) did not match administration manual.

Nice cross section to work with. Uncomfortable with the vague nature of the process. I was told by my table leader that there was no right or wrong method of doing this. I don't agree, there should be a distinct methodology. Time was insufficient it seemed, I felt very rushed.

Group consensus was achieved. Table leaders should not also serve as group members.

The process was useful but cumbersome - however I understand the need for it.

Table leader did well keeping us on track of validating other opinions. Handouts extremely helpful and sharing of data was very understandable. Ample opportunities for questions, whole process made me feel very good about my individual judgment and group judgments.

Good explanation of process. Objectives were clearly stated. More charts available to groups, i.e. students just below "cut" scores.

Excellent group at 4th grade table. Change proficiency level descriptors to match test.

It was good to have such raw data available to groups for review, i.e. 4 students got score of 10, 2 students got score of 12, etc. Flexibility on proficiency descriptors and welcome to comments on these and GLE's very useful! Need printed copies of everything for all to look at, do it ahead of time, groups will want it.

Good info and training provided. Well organized.

It would be good to have an overall consensus of % proficient. Otherwise there could be discrepancy from grade to grade (easier - harder).

Learning curve was appropriate for the time allotted. More snacks :)

Went great. Thanks for having me be part of it.

I have a much deeper appreciation and awareness for how the cut scores are established. I am also appreciative of the fact that science content folks were included with this process.

More item data would have been helpful - but as scientists, we're always looking for more data. Sometimes less is enough.

Very collaborative, good discussion and logical process. The group was a good mix of SpEd, Reg. Ed, and science.

Very helpful - great understanding of the AA - very helpful for a SpEd teacher/assessor

A learning process - struggles but fun. Learned a new term "uneasy alliance." Item analysis would have been nice. A crab lunch served!!

Process brought out team to a very reasoned comfort zone.

I am absolutely certain of both my judgments and those of my team. Good word all the way around. Possibly access to more data.

Ample time to discuss. Good data.

Description of Cut Point Scores and Performance Levels

The following paragraphs summarize the decisions made at the standard setting meeting. In the three rounds of decision-making, teachers made their initial value determination independently, then they conferred as a group, and finally, they made a decision after reviewing initial impact data.

The *three rounds of teacher judgments* are contained in Appendix 6_4. The final judgments and discussion is included in the text below.

< Appendix 6_4 >

Grade Level Cut Point Scores and Performance Levels

Grade 4 – cut point scores summary. After analysis of the student data, the cut score for the grade 4-Advanced category was modified. The original proposed cut score was at 40 points, distributed as follows: Concepts of Physical Science – 10 points, Concepts of Life Science – 10 points, Concepts of Earth Science – 12 points, Science and Technology – 8 points. The finalized cut scores were modified as follows: Concepts of Life Science – 12 points, Science and Technology – 10 points. These two changes raised the Combined Advanced Proficiency category to 44 points. All other categories: Proficient, Below, and Far Below, remained the same as the proposed cut scores. The following table presents the finalized cut scores for Grade 4 Science.

Table 20. Grade 4 Cut Scores Summary

Gr. 4 Final	Descriptor and weight	Concepts of Physical Science	Concepts of Life Science	Concepts of Earth Science	Science and Technology	Combined Proficiency
		12 points	12 points	12 points	12 points	48 points
	Advanced	10	12	12	10	44
	Proficient	6	6	8	4	24
	Below	2	4	4	2	12
	Far Below	0	0	0	0	0

Grade 4 teachers – discussion. The group is feeling that there should be more students in the proficient category and less in advanced. This would better reflect a typical distribution of standard test scores - we were wanting to look at the distribution of students in the (4.4) History/Tech area - this was the most difficult area and we initially set the advanced score at 8. We discussed the issue that perhaps the "advanced" students would be those that were able to answer these questions and therefore we should raise the advanced to 10 which would raise the total min score for advanced to 42. We were wondering how this would change the % of students in advanced and proficient. The group was comfortable with the cut scores for below and proficient. 40 = 33.9, If changed to 42% - 33, 44% - 30, 46% - 16. Discussions - the test does not test higher level thinking that would match the advanced proficiency level descriptors. The general feeling was that the proficient group should be the largest and the extreme ends should be more equal. Some concern about making it too difficult to be advanced - given the small number of items and the small number of students who took the test. Decision was made to change Advanced to 44. This is still a high % in advanced. We had a concern of making advanced 46 and only 1 item incorrect. Since these scores will be used in future years, overall the group was ok with 44 as the cut for advanced. Looked at data of numbers of students in each group. Continue to be ok with other cut scores - *Group feels strongly that proficiency level descriptors need to be re-written to reflect the A.A. test. This is what the parent's/community will see.

Grade 8 – cut point scores summary. Grade 8 finalized cut scores reflected one change made to the Advanced category after analysis of student data. The proposed cut scores for the Advanced category were as follows: Concepts of Physical Science – 10 points, Concepts of Life Science – 11 points, Concepts of Earth Science – 10 points, Science and Technology – 11 points, creating a total of 42 points. Due to the scoring for this assessment, the student may only receive 2 points total for each item, which made a score of 11 points on any task impossible. The finalized cut scores were modified as follows: Concepts of Life Science – 12 points, Science and Technology – 12 points. These two changes raised the Combined Advanced Proficiency category to 44 points. All other categories: Proficient, Below, and Far Below, remained the same as the proposed cut scores. The following table presents the finalized cut scores for Grade 8 Science.

Table 21. Grade 8 Cut Scores Summary

Gr. 8 Final	Descriptor and weight	Concepts of Physical Science	Concepts of Life Science	Concepts of Earth Science	Science and Technology	Combined Proficiency
		12 points	12 points	12 points	12 points	48 points
	Advanced	10	12	10	12	44
	Proficient	6	8	7	8	29
	Below	3	4	4	5	16
	Far Below	0	0	0	0	0

Grade 8 teachers – discussion. Looking for a material break in data distribution would be a justification (rationale) for changing the cut scores. Group is in consensus on this. Discussion about whether there is another path for these students since they aren't on a diploma track. Down the road will there be a basics skills diploma? Group wants to also look at how kids scored on each item. 2 Questions: 1) What if we lowered proficiencies what would that do to the numbers going down to 17 would pull in 4% into proficient. What was the one question that every are missed that would justify changing (was in the "plate" question). No natural breaks in the science groupings. 2) Which needed happen if we adjusted the A/P score break from 42 to 44? Would drop 15% into the level and drop A level to approx. 28%. So if we changed the group levels, the A and P groups would be more in line with where the state should be. Also would be changing the passing score. Going back to question 1 - Asked Jerry about item analysis and he said we don't have that info now. Paul came over and gave us numbers of at a certain score level. The group seems to want to leave the proficiency score to 27 but would feel more comfortable if there was a concrete reason to do so. Group decision - Review the A but score to 44 and leave the P cut score to 29. With further discussion we decided to say with the P score at 29. That these students don't fit a full shaped curve of a mean of 29.8 that would justify moving the P score downward.

Grade 10 – cut point scores summary. The Grade 10 group averaged their individual proposed cut scores to come to a consensus for the original proposed cut scores. After analyzing the student data, they discussed and made several modifications to the Advanced category to finalize their cut scores. The proposed cut scores for the Advanced category were as follows: Concepts of Physical Science – 10 points, Concepts of Life Science – 10 points, Concepts of Earth Science – 8 points, Science and Technology – 10 points, creating a total of 38 points. After reaching a final consensus, the finalized cut scores are as follows: Concepts of Physical Science – 10 points, Concepts of Life Science – 12 points, Concepts of Earth Science – 10 points, and Science and Technology – 12 points. These changes raised the Combined Advanced Proficiency category to 44 points. All other categories: Proficient, Below, and Far Below, remained the same as the proposed cut scores. The following table presents the finalized cut scores for Grade 10 Science.

Table 22. Grade 10 Cut Scores Summary

Gr. 10 Final	Descriptor and weight	Concepts of Physical Science	Concepts of Life Science	Concepts of Earth Science	Science and Technology	Combined Proficiency
		12 points	12 points	12 points	12 points	48 points
	Advanced	10	12	10	12	44
	Proficient	6	6	6	8	26
	Below	4	4	4	6	18
	Far Below	0	0	0	0	0

Grade 10 teachers – discussion. The grade 10 teachers averaged their original scores to reach round 1 consensus. They then made their judgments for round 2 and discussed until reaching a consensus.

General Standard Setting Issues (by strand – test)

The following are general notes taken on the discussion and issues each group addressed. Similar concerns were expressed throughout the group, which are noted below, as well as a few notes on the process in which each group participated.

Concepts involved (i.e., weight, density, and matter with the floating rock

Number of items per concept (i.e., force)

Distractors and item characteristics

Random error and the probability of guessing (also giving Ss the benefit of the doubt); option placements

Range values available as they made categorical decisions of proficiency

What students must know (by item) and concept (e.g., ameba and paramecium); short changing students when expectations are low

Cultural and geographic issues (student background and living conditions interacting with items, both content and format); uniquely Alaskan focus

Item difficulty (i.e., plate tectonics) and misunderstandings (i.e. dinner plates); type of items by category

Uneasy alliance between general and special education views, particularly with respect to disabilities

Swapped items and points by stand and category proficiencies (it looked and sounded like Texas Hold ‘em)

Vocabulary (both construct relevant and access)

The (in)dependence of item and strands, which resulted in differential point values by category

Kid focused outcomes (consequences of percentages in various categories); the political value of percentages by category;

Administration conditions (i.e., read aloud nature of presentation with pointing response options) and its influence on access

Final Review and Adjustment During Round 3

Number and percentage of students at each proficiency category

Number of students just above and just below the proposed cut scores

Item difficulty

Strand and total values and the iterative swapping up and down of cut scores

Proficiency Level Descriptors – Initial Development and Grade Levels

Initial development. EED provided groups of teachers and stakeholders with lists of indicators that the groups formed into grade level PLDS. The standard setting participants reviewed the PLDs and any suggestions for change were submitted to EED upon conclusion of the standard setting.

Page 2 of Appendix 6_5 provides a summary of *methods and procedures used to develop the Proficiency Level Descriptors and Extended Grade Level Expectations for Science*. This appendix also includes the agenda, science committee participants, and the Power Point® of training for the Alternate Assessment Work Group *Developing Science Alternate Achievement Standards*.

< Appendix 6_5 >

Grade 4 Science – Proficiency Level Descriptors

Advanced Level

The student demonstrates a highly developed conceptual understanding of the processes and content of science by identifying or demonstrating an understanding of: cause-and-effect (e.g., when more water is added to a full glass, the water will spill out); the concept that living things reproduce; basic characteristics of common objects (rock is hard, etc.); states of matter of water; living and non-living things; a variety of Earth's features and features in the natural world; types of weather; the relationship of plants and animals to their habitats; tools/materials and their uses; what materials found on earth are used for; transfer of energy (e.g., electricity can be turned on and off by a switch); and ways objects can move.

Proficient Level

The student demonstrates a basic conceptual understanding of the processes and content of science by identifying or demonstrating an understanding of: cause-and-effect (e.g., when more water is added to a full glass, the water will spill out); the concept that living things reproduce; basic characteristics of common objects (rock is hard, etc.); states of matter of water; living and non-living things; a variety of Earth's features and features in the natural world; types of weather; the relationship of plants and animals to their habitats; tools/materials and their uses; what materials found on earth are used for; transfer of energy (e.g., electricity can be turned on and off by a switch); and ways objects can move.

Below Proficient Level

The student shows a partial understanding of the processes and content of science by identifying or demonstrating an understanding of: cause-and-effect (e.g., when more water is added to a full glass, the water will spill out); the concept that living things reproduce; basic characteristics of common objects (rock is hard, etc.); states of matter of water; living and non-living things; a variety of Earth's features and features in the natural world; types of weather; the relationship of plants and animals to their habitats; tools/materials and their uses; what materials found on earth are used for; transfer of energy (e.g., electricity can be turned on and off by a switch); and ways objects can move.

Far Below Proficient Level

The student did not display a minimal understanding of science processes or content as described in the extended grade level expectations.

Grade 8 Science – Proficiency Level Descriptors

Advanced Level

The student demonstrates a highly developed conceptual understanding of the processes and content of science by identifying or demonstrating an understanding of: the physical changes commonly found in the environment; the concept that organisms differ from one species to another; features of geophysical events; the earth, sun, and moon; seasonal characteristics; the uses of technology; simple descriptors to relate information about an object; familiar electronic devices; directional movement of objects; the stages of life cycles; the connection between living organisms and their environment; and tools to their function.

Proficient Level

The student demonstrates a basic conceptual understanding of the processes and content of science by identifying or demonstrating an understanding of: the physical changes commonly found in the environment; the concept that organisms differ from one species to another; features of geophysical events; the earth, sun, and moon; seasonal characteristics; the uses of technology; simple descriptors to relate information about an object; familiar electronic devices; directional movement of objects; the stages of life cycles; the connection between living organisms and their environment; and tools to their function.

Below Proficient Level

The student shows a partial understanding of the processes and content of science by identifying or demonstrating an understanding of: the physical changes commonly found in the environment; the concept that organisms differ from one species to another; features of geophysical events; the earth, sun, and moon; seasonal characteristics; the uses of technology; simple descriptors to relate information about an object; familiar electronic devices; directional movement of objects; the stages of life cycles; the connection between living organisms and their environment; and tools to their function.

Far Below Proficient Level

The student did not display a minimal understanding of science processes or content as described in the extended grade level expectations.

Grade 10 Science – Proficiency Level Descriptors

Advanced Level

The student demonstrates a highly developed conceptual understanding of the processes and content of science by identifying or demonstrating an understanding of: the basic characteristics of matter, including identifying objects as liquid, solid, or gas; the way in which objects get energy; how the states of water affect weather; purpose of different animal adaptations; the classification of animals as herbivores, carnivores, and omnivores; the characteristics of the solar system; the movement of objects; inherited traits; how the Earth's surface can change as a result of geological activity; and tools and their purposes.

Proficient Level

The student demonstrates a basic conceptual understanding of the processes and content of science by identifying or demonstrating an understanding of: the basic characteristics of matter, including identifying objects as liquid, solid, or gas; the way in which objects get energy; how the states of water affect weather; purpose of different animal adaptations; the classification of animals as herbivores, carnivores, and omnivores; the characteristics of the solar system; the movement of objects; inherited traits; how the Earth's surface can change as a result of geological activity; tools and their purposes; and the characteristics of the solar system.

Below Proficient Level

The student shows a partial understanding of the processes and content of science by identifying or demonstrating an understanding of: the basic characteristics of matter, including identifying objects as liquid, solid, or gas; the way in which objects get energy; how the states of water affect weather; purpose of different animal adaptations; the classification of animals as herbivores, carnivores, and omnivores; the characteristics of the solar system; the movement of objects; inherited traits; how the Earth's surface can change as a result of geological activity; tools and their purposes; and the characteristics of the solar system.

Far Below Proficient Level

The student did not display a minimal understanding of science processes or content as described in the extended grade level expectations.

CHAPTER 7: REPORTING

Overview

Five reports are available that depict the outcomes from the Science 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 Assessment in Science measures, 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 is 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: TECHNICAL DOCUMENTATION

Overview

The study of validity is greatly aided by following the *Standards for Educational and Psychological Testing* (American Educational Research Association--AERA, American Psychological Association—APA, and the National Council on Measurement in Education — NCME, 1999)¹². The *Standards* book was developed by experts in testing to help test sponsors provide the highest quality testing programs possible.

In the process of validation, we engage in four activities: (a) defining what students learn, (b) stating the validity argument, (c) making the claim for validity, and (d) gathering the validity evidence to support the argument and claim. At the final stage of the validation process, a qualified evaluator takes into consideration the logic of the argument and its plausibility, the claim, and the evidence in support of the claim. A summative judgment is made and, usually, recommendations are made for how the testing program could be improved to strengthen supporting evidence and eliminate threats to validity uncovered in the evidence gathering.

In previous sections, the standards and aligned tasks are presented as evidence of what students are expected to learn. In addition, the manner in which we developed the test, including expanded levels of support (ELOS) is part of our validity argument about access so students can show their proficiency: Student’s disability does interfere with the assessment of that student’s learning. Our primary claim, therefore, is that the test scores are valid in making judgments about proficiency on a large-scale test that has been reduced in breadth, depth, and complexity. In this section, we present the evidence that supports this claim. The evidence we present is both procedural and statistical/empirical.

Nearly 200 students took the Alternate Assessment in Science with valid scores on the standard administration.

Table 23. Standard Administration Results

	Grade	Frequency	Percent
Valid	4	62	31.2
	8	53	26.6
	10	84	42.2
	Total	199	100.0

We first present reliability data by addressing the internal consistency of the test items. Cronbach’s alpha was used to document this internal consistency; it was computed for each task and is displayed in the table below for each grade level.

¹² American Educational Research Association (AERA), American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: AERA

Reliability

In the previous section, we presented information about the training and administration of all qualified assessors (QAs). Certainly this is an important dimension of reliability. In this section, however, we confine ourselves to the items.

Reliability requires quantifying the measurement error associated with (a) observed behaviors, and (b) numeric scores assigned to our observations. The situation becomes complex when observed behaviors depend on the sampling of items and how items ‘stimulate’ observed behavior.

One of the most traditional conceptualizations is in terms of the *true score*. Unfortunately, we never actually know a student’s true score; it must be estimated from the *observed score*, which provides imperfect information. Therefore, in addition to the observed score, we must theorize an *error score*. A very simple concept of observed score, true score, and error score is captured in equation: *observed score = true score + error score*.

The concept of error scores is at the heart of reliability. The goal of good measurement design is to minimize the error component. Note, in our simple model, error is thought to occur randomly. The importance of random error may be recognized if we use an assessment repeatedly to measure the same individual. The obtained measures would not be identical. In fact, they will be more or less variable, depending on the reliability of our instrument. Our best estimate of the examinee true score will be the average of the repeated measures. The variability around the mean is our theoretical concept of error, also called error variance.

Another conceptualization of measurement accuracy is developed in terms of the standard error of measurement (SEM). As described above, the concept of random error around the true score results from administering repeated parallel forms, the SEM of a measure is essentially the average deviation of error scores around the true score. As with reliability, SEM (σ_e) can be estimated in terms of correlated observations x_1 and x_2 .

$$\sigma_e = \sigma_x \sqrt{r_{x_1x_2}} \sqrt{1 - r_{x_1x_2}} \quad (5)$$

According to this equation, as the correlation of parallel forms increases, the standard error of measurement diminishes to zero. It is very important to keep in mind that our measures are estimations, and that theoretically, each time the assessment is administered we will obtain a different measure. The reliability or error in measurement determines how different.

Samples of performance items and tasks are prepared in the spirit of parallel forms. That is, the items and tasks are ideally comparable to the extent that a student would not perform differently among them as they are all within a range of difficulty. As with any sampling of items, the sample of tasks is apt to be more or less variable with respect to difficulty and representation of the performance domain. Using multiple tasks are alternate forms for either comparing an individual over time or one examinee to another, the extent to which the tasks differ is of obvious consequence.

To ascertain the item and task consistency (lack of random error), we used Cronbach’s alpha, which is a measure of internal consistency, which we also report in the section on validity as a

reflection of the internal structures of the tasks. Basically, this statistic calculates the item difficulties and inter-item relations. Later, when we describe the reports generated after data entry, the student’s score is reported with one SEM extended below and above this level to provide a known probability that the score is within this band. We used 1 SEM, which provides a 67% probability.

Table 24. Cronbach’s Alpha for Grades 4, 8, 10

Cronbach’s Alpha	Task 1	Task 2	Task 3	Task 4
Grade 4	.770	.812	.847	.703
Grade 8	.646	.867	.764	.876
Grade 10	.671	.696	.520	.777

These estimates are for each task (comprised of only 6 items); therefore, they represent the lower bound (particularly as no decision is made at the task level). **For the total test, the reliability of the Science Alternate Assessment is .93 for grade 4, .92 for grade 8, and .89 for grade 10.**

Validity

In this section, the first analysis presents, for each grade level, descriptive statistics for each task and the degree to which items within the task inter-correlate as reflecting the internal structure of the test. The *Standards* (AERA et al., 1999, pp. 13-15)¹³ call for a study on internal structure as part of test validation. For valid test score interpretations and validity generalization, it is expected that (a) the items show some level of internal consistency (Standard 1.11); (b) the internal structure of the test remains stable across major reporting groups (p.15); and (c) the internal structure of the test remains stable across alternate (and hopefully equivalent) forms of the same test (pp. 51-52). Inter-task correlations are expected to be positive and moderate. High inter-task correlations are not desirable because the strands may essentially reflect very similar types of skills or abilities.

The validity claim for achievement requires that items be both related to the standards but also within reach of the students. For each grade level, we present the mean (average) and then identify any items that appeared to be difficult.

In the narrative, tasks were associated with the following specific strands:

Task 1 – Concepts of Physical Science

Task 2 – Concepts of Life Science

Task 3 – Concepts of Earth Science

Task 4 – History and Nature of Science

¹³ American Educational Research Association (AERA), American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.

For each task, difficult items refer to those with a mean less than 1 (on a 2-point scale). The task and item statistics include only those students who took ALL items within a task (and therefore has a lower student count than the total test at each grade level). The means are slightly higher than the total test because this group of students is likely to be higher performing than students who only took a subset of items within a task (but appeared in the count for the total test).

Grade 4 Results

Task 1 had a mean of 8 points with two difficult items (4 and 6) and an average inter-item correlation .36 (ranging from .26 to .53).

Task 2 had a mean of 8 points with one difficult item (6) and an average inter-item correlation of .43 (ranging from .41 to .64).

Task 3 had a mean of 9 points with no difficult items and an average inter-item correlation of .51 (ranging from .18 to .81).

Task 4 had a mean of 8 points with 1 difficult item (5) and an average inter-item correlation of .30 (ranging from .08 to .69).

The average inter-item correlation was .38, reflecting an adequate degree of relation without being too low and therefore signifying no consistency or being too high and therefore signifying redundant information. The average performance on the entire scale was about 33 points (out of 48 total) with a standard deviation of 13 points.

Table 25. Grade 4 Results

Scale Statistics			
Mean	Variance	Std. Deviation	N of Items
32.79	177.283	13.315	24

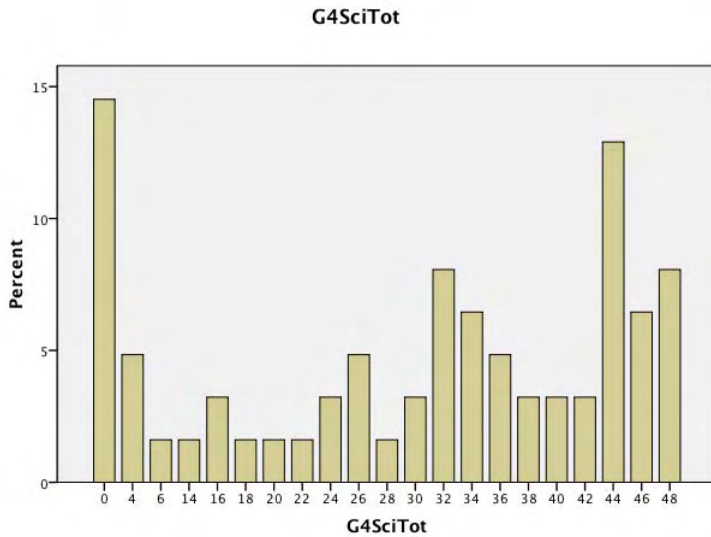
For the group of students with scores across all strands, the strands were consistently difficult.

Table 26. Grade 4 Results by Strand

Strand Statistics			
	Mean	Std. Deviation	N
G4PhysTot	7.34	4.076	58
G4LifeTot	7.10	4.372	58
G4EarthTot	8.52	4.281	58
G4TechTot	7.28	3.852	58

The strands were highly inter-correlated with an average of .81 and a range of .76 to .85, reflecting a high degree of internal consistency. The following distribution was obtained.

Figure 2. Grade 4 Results by Percentage



Grade 8 Results

Task 1 had a mean of 8 points with one difficult item (5) and an average inter-item correlation .26 (ranging from .11 to .46).

Task 2 had a mean over 9 points with no difficult items and an average inter-item correlation of .53 (ranging from .29 to .73).

Task 3 had a mean of 9 points with no difficult items and an average inter-item correlation of .37 (ranging from .20 to .51).

Task 4 had a mean of 9 points with no difficult items and an average inter-item correlation of .55 (ranging from .35 to .85).

The average inter-item correlation was .43, reflecting an adequate degree of relation without being too low and therefore signifying no consistency or being too high and therefore signifying redundant information. The average performance on the entire scale was about 36 points (out of 48 total) with a standard deviation of 13 points.

Table 27. Grade 8 Results

Scale Statistics			
Mean	Variance	Std. Deviation	N of Items
36.05	176.474	13.284	24

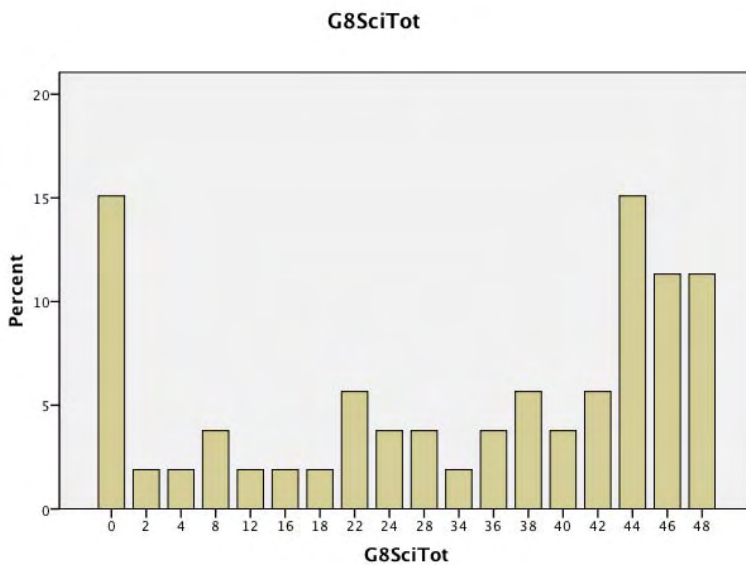
For the group of students with scores across all strands, the strands were consistently difficult.

Table 28. Grade 8 Results by Strand

Strand Statistics			
	Mean	Std. Deviation	N
G8PhysTot	8.09	3.265	46
G8LifeTot	8.87	4.293	46
G8EarthTot	8.57	3.857	46
G8TechTot	8.74	4.245	46

The strands were highly inter-correlated with an average of .82 and a range of .77 to .92, reflecting a high degree of internal consistency. The following distribution was obtained.

Figure 3. Grade 8 Results by Percentage



Grade 10 Results

Task 1 had a mean over 9 points with one difficult item (6) and an average inter-item correlation .30 (ranging from .01 to .60).

Task 2 had a mean over 9 points with no difficult items and an average inter-item correlation of .32 (ranging from .10 to .64).

Task 3 had a mean of 7.5 points with one difficult item (6) and an average inter-item correlation of .16 (ranging from -.02 to .35).

Task 4 had a mean over 9 points with no difficult items and an average inter-item correlation of .39 (ranging from .22 to .62).

The average inter-item correlation was .29, reflecting an adequate degree of relation without being too low and therefore signifying no consistency or being too high and therefore signifying

redundant information. The average performance on the entire scale was about 36 points (out of 48 total) with a standard deviation of 13 points.

Table 29. Grade 10 Results

Scale Statistics			
Mean	Variance	Std. Deviation	N of Items
35.51	114.636	10.707	24

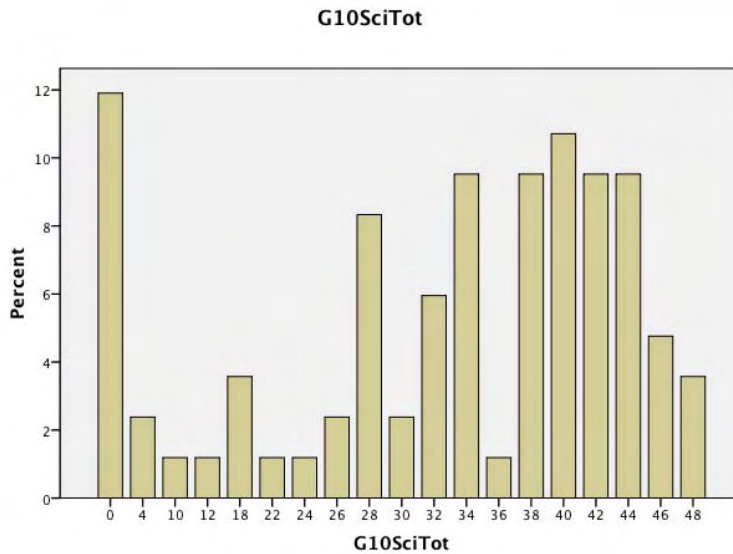
For the group of students with scores across all strands, the strands were consistently difficult.

Table 30. Grade 10 Results by Strand

Strand Statistics			
	Mean	Std. Deviation	N
G10PhysTot	9.23	2.788	75
G10LifeTot	8.83	3.371	75
G10EarthTot	7.20	3.171	75
G10TechTot	9.04	3.652	75

The strands were highly inter-correlated with an average of .65 and a range of .58 to .70, reflecting a high degree of internal consistency. The following distribution was obtained.

Figure 4. Grade 10 Results by Percentage



Appendix 8_1 contains all statistical data from the science assessments in grades 4, 8, and 10.

Response Processes: Analysis of ELOS

Student response processes to most general assessment items (selected response, short constructed response) may be considered through typical item review procedures. However, cognitive processes used in responding to items and tasks are more difficult to assess directly for alternate assessments aligned to alternate achievement standards. We approached the test demands by creating two versions of items: Standard and Expanded Levels of Support (ELOS).

The argument is whether or not students have been able to show as much as they know and can do. For a small group of students, who took both the standard and the ELOS, we can analyze the effect of our stopping rule by analyzing their performance on the standard test administration.

In grade 4, for all science tasks with the exception of three, the majority of students participating in ELOS received scores of 0. In grade 8, for all science tasks, the majority of students participating in ELOS received scores of 0. For many tasks, 100% of students who participated received scores of 0. In grade 10, for all science tasks, the majority of students participating in ELOS received scores of 0. For many tasks 100% of students who participated received scores of 0. At all grade levels, the majority of students received scores of 0 on most science tasks.

Additionally, the mean scores of students participating in ELOS were low, relative to the total possible points. In grade 4, the mean scores for the four strands ranged from 1.60 to 4.80 (out of a total of 12 points per strand). For all strands, the standard deviations were greater than the means, likely due to one unusually high score. In grade 8, the mean scores for the three strands that students completed ranged from 0 to 1.00. One strand had a standard deviation that was higher than the mean, likely due to one unusually high score. In grade 10, the mean scores for the four strands ranged from 0 to .67. Three strands had standard deviations that were greater than the means, likely due to one unusually high score. In general, all strands in all grade levels reflected the low score values of students participating in ELOS.

Most students participating in ELOS received scores that indicate that they are not meeting AYP. In grade 4, two students (representing 13.3% of the total population) scored at the proficient level. The students who received proficient scores are students who took both standard and ELOS test items. The scoring rules state that the standard item scores override the ELOS scores. The remaining three students (representing 20.0% of the total population) scored far below proficiency. In grade 8, all four students who participated in ELOS (representing 26.7% of the total population) scored far below proficiency. In grade 10, all six students who participated in ELOS (representing 40.0% of the total population) scored far below proficiency. With the exception of the two students in grade 4, all students participating in ELOS scored far below or below proficient.

In general, these data indicate that at all grade levels, students who participated in ELOS demonstrated low performance on the science assessment.

Appendix 8_2 contains the analysis of ELOS items and response processes.

<Appendix 8_2>

Response Processes: Relations Among Science and Reading, Writing, and Mathematics

To understand how much science performance was potentially related to skills in other areas, the data files for subject areas were concatenated using the Alaska Student Identification number and a correlation was computed for the raw scores among all areas. As noted in the tables below, science performance was highly inter-correlated with performance in these other subjects.

Table 31. Grade 4 Correlations Among Science and Reading, Writing, and Mathematics Scores
Correlations

		Science	Reading	Writing	Math
Science	Pearson Correlation	1.000	.886**	.787**	.892**
	Sig. (2-tailed)		.000	.000	.000
	N	61.000	61	61	60
Reading	Pearson Correlation	.886**	1.000	.788**	.866**
	Sig. (2-tailed)	.000		.000	.000
	N	61	61.000	61	60
Writing	Pearson Correlation	.787**	.788**	1.000	.896**
	Sig. (2-tailed)	.000	.000		.000
	N	61	61	61.000	60
Math	Pearson Correlation	.892**	.866**	.896**	1.000
	Sig. (2-tailed)	.000	.000	.000	
	N	60	60	60	60.000

** . Correlation is significant at the 0.01 level (2-tailed).

Table 32. Grade 8 Correlations Among Science and Reading, Writing, and Mathematics Scores
Correlations

		Science	Reading	Writing	Math
Science	Pearson Correlation	1.000	.870**	.765**	.871**
	Sig. (2-tailed)		.000	.000	.000
	N	53.000	51	49	48
Reading	Pearson Correlation	.870**	1.000	.861**	.926**
	Sig. (2-tailed)	.000		.000	.000
	N	51	51.000	49	48
Writing	Pearson Correlation	.765**	.861**	1.000	.866**
	Sig. (2-tailed)	.000	.000		.000
	N	49	49	49.000	48
Math	Pearson Correlation	.871**	.926**	.866**	1.000
	Sig. (2-tailed)	.000	.000	.000	
	N	48	48	48	48.000

** . Correlation is significant at the 0.01 level (2-tailed).

Table 33. Grade 10 Correlations Among Science and Reading, Writing, and Mathematics Scores
Correlations

		Science	Reading	Writing	Math
Science	Pearson Correlation	1.000	.810**	.799**	.881**
	Sig. (2-tailed)		.000	.000	.000
	N	84.000	83	82	82
Reading	Pearson Correlation	.810**	1.000	.861**	.862**
	Sig. (2-tailed)	.000		.000	.000
	N	83	83.000	82	82
Writing	Pearson Correlation	.799**	.861**	1.000	.892**
	Sig. (2-tailed)	.000	.000		.000
	N	82	82	82.000	81
Math	Pearson Correlation	.881**	.862**	.892**	1.000
	Sig. (2-tailed)	.000	.000	.000	
	N	82	82	81	82.000

** . Correlation is significant at the 0.01 level (2-tailed).

Summary

It is likely that students’ performance on the Science Alternate Assessment sufficiently reflects what they know and can do. Students are consistent in their performance on the various tasks and the items are correlated with each other; the distribution is adequate; and the number of students reaching proficiency is reasonable. It is likely, however, that when the proficiency cut scores are placed over the score range, a bi-modal distribution results. Given the high relation among science and the other subject areas, it is likely that skill and knowledge in science is a function of students’ general academic proficiency (which likely serve as access skills for science content).

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.

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.

Accessibility and 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.

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

< Appendix 9_1 >

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.

Science Grade 4 Comments on Item Revisions

The team didn't like certain words in the proficiency level descriptors.

Applying: Students are not assessed by applying.

Demonstrating: How are they demonstrating? If it is worded so that they have to demonstrate like it is now, then how will they be held accountable? They could miss all of the demonstrating cause and effect questions and pass all the others and be proficient. It would be nice to have qualifiers to help.

Using: Not actually using anything on the task.

Symbols to represent/data: not on assessment.

Developing: We like the word and add it to some of the descriptors. *Far Below Proficient Definition:* Doesn't reflect the student. Reflects the teacher. It should be in the same terms about the student. Student refused to answer? Student wasn't able to answer? Student wasn't cognitively able to answer question? Etc. Not all students in this category reflect the need for more instruction.

Directions should be consistent: For example task 4.4 #1, #3, and 3.4 #4. Some teachers will read labels for students and some won't. Inconsistent. Would also love it to explicitly say coloring of pictures is OK.

Partial Credit: If it isn't in writing, teachers will stick to what is printed for scoring options. Teachers don't know they can give partial credit instead of 0/2 score for science.

Science Grade 8 Comments on Item Revision

Plates question – presentation of materials, plates = eating dishes to the students. Not a proficient kid concept. Food on a plate. Now introduce plates make mountains – no way! But it is in the ExGLE specifically, plate tectonics, so somehow need to address.

Characteristic of fall is that plants change color.

Science and Technology – microscope – can clarify wording: if you want to see something really small, point to what you would use? Weigh a dog – kids might not have experience with this in their world. Weigh fish that you just caught.

Question 1 of Task 4.8 not a great question, in regards to choices offered. Hat is worn when duck hunting where there are loud noises, glasses at car races and loud noises. Instead ask: if you want to protect your ears, would you wear a hat, earplugs, or sunglasses. If a student got 1 point on every question, then 6 points = basic understanding.

ExGLEs are what all kids need to know no matter where they live. State Science test is a chance for teachers to think about what they are presenting to students and growth opportunity to kids.

My kids need to know what a microscope is, and go to district office and ask “how are you going to support my kids?” Limitations of these children is SO great however, have to keep them in mind when we are looking at these items. Microscope, scale, and thermometer correct – we would have a party.

Moving to 8 is short-changing kids according to one science teacher. More science teachers are now encouraging the AA kids to come to their classroom, kids are getting more exposure.

Accessibility factor – test is not a perfect instrument to begin with, then the kids severe disabilities make it an even less than perfect instrument. Need flexibility in administering items.

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. This change is 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).