Oregon Department of Education

2013–2014 Technical Report

Oregon's Alternate Assessment System

Peer Review Documentation: Sections 1-7

Oregon's Alternate Assessment System Technical Report: Peer Review Documentation: Sections 1-7

This technical report is one of a series that describes the development of Oregon's Statewide Assessment System. The complete set of volumes provides comprehensive documentation of the development, procedures, technical adequacy, and results of the system.

TABLE OF CONTENTS

Peer Review Critical Elements Reference Table

1.1 (a) Has the State formally approved/adopted, by May 2003, challenging academic content standards in reading/language arts and mathematics that – • cover each of grades 3-8 and the 10-12 grade range, *or* • if the academic content standards relate to grade ranges, include specific content expectations for each grade level? AND (b) Are these academic content standards applied to *all* public schools and students in the State?
1.2 Has the State formally approved/adopted, academic content standards in science for elementary (grades 3-5), middle (grades 6-9), and high school (grades 10-12)? This must be completed by school year 2005-2006.
1.3 Are these academic content standards challenging? Do they contain coherent and rigorous content and encourage the teaching of advanced skills?
1.4 Did the State involve education stakeholders in the development of its academic content standards?
2.1 Has the State formally approved/adopted challenging academic achievement standards in reading/language arts and mathematics for each of grades 3 through 8 and for the 10-12 grade range? These standards were to be completed by school year 2005-2006. Has the State, through a documented and validated standards-setting process, approved/adopted modified academic achievement standards for eligible students with disabilities? If so, in what subjects and for which grades? Has the State approved/adopted alternate academic achievement standards for students with the most significant cognitive disabilities? If so, in what subjects and for which grades? Note: If alternate or modified academic achievement standards in reading/language arts or mathematics have not been develop/adopted and approved, then the alternate assessments for all students with disabilities must be held to grade-level academic achievement standards.

2.2

Has the State formally approved/adopted academic achievement descriptors in science for each of the grade spans 3-5, 6-9, and 10-12 as required by school year 2005-06?

Has the State formally approved/adopted academic achievement cut scores in science for each of the grade spans 3-5, 6-9, and 10-12 as required by school year 2007-08?

Has the State formally approved/adopted modified academic achievement standards in science? If so, for which grades?

Has the State formally approved/adopted alternate academic achievement standards for students with the most significant cognitive disabilities in science? If so, for which grades?

Note: If alternate or modified academic achievement standards in science have not been adopted and approved, then all students with disabilities must be held to grade-level academic achievement standards.

2.3

1. Do these academic achievement standards (including modified and alternate academic achievement standards, if applicable) include for each content area –

(a) at least three levels of achievement, including two levels of high achievement (proficient and advanced) that determine how well students are mastering a State's academic content standards and a third level of achievement (basic) to provide information about the progress of lower-achieving students toward mastering the proficient and advanced levels of achievement; *and*
descriptions of the competencies associated with each achievement level; *and*

(b) assessment scores ("cut scores") that differentiate among the achievement levels and a rationale and procedure used to determine each achievement level?

2. If the State has adopted either modified or alternate achievement standards, has it developed guidelines for IEP teams to use in deciding when an individual student should be assessed on the basis of modified academic achievement standards in one or more subject areas, or assessed on the basis of alternate achievement standards?

2.4

With the exception of students with disabilities to whom modified or alternate academic achievement standards apply, are the grade-level academic achievement standards applied to *all* public elementary and secondary schools and *all* public school students in the State?**

**OSEP guidance and NCLB requirements indicate that a student placed in a private school by a public agency for the purpose of receiving special education services must be included in the State assessment and their results attributed to the public school or LEA responsible for the placement.

2.5
How has the State ensured alignment between challenging academic content standards and the academic achievement standards?

If the State has adopted modified academic achievement standards, how has the State ensured alignment between its grade-level academic content standards and the modified academic achievement standards?

If the State has adopted alternate academic achievement standards, how has the State ensured alignment between its academic content standards and the alternate academic achievement standards?

2.6
For each assessment, including alternate assessments, provide documentation of the standard setting process. Describe the selection of panelists, methodology employed, and final results.

How did the State document involvement of diverse stakeholders in the development of its academic achievement standards and its modified and/or alternate achievement standards, if any?

If the State has adopted alternate or modified academic achievement standards, did the State's standards-setting process include persons knowledgeable about the State's academic content standards and special educators who are knowledgeable about students with disabilities?

Section 3.1. In the chart below indicate your State's current assessment system in reading /language arts and mathematics in grades 3 through 8 and for the 10-12 grade range using the abbreviations to show what type of assessments the State's assessment system is composed of: (a) criterion-referenced assessments (**CRT**); or (b) augmented norm-referenced assessments **(ANRT**) (augmented as necessary to measure accurately the depth and breadth of the State's academic content standards and yield criterion-referenced scores); or (c) a combination of both across grade levels and/or content areas. Also indicate your current assessment system in science[1] that is aligned with the State's challenging academic content and achievement standards at least once in each of the grade spans 3-5, 6-9, and 10-12. A State may have assessments in reading or language arts depending on the alignment to the State's content standards; both are not required. Please indicate, using the abbreviations shown, the grades and subject areas with availability of native language assessment **(NLA**) or various alternate assessments (**AA-GLAS** for an alternate assessment for students with disabilities based on grade-level standards; **AA-LEP** for an alternate assessment for students with limited English proficiency based on grade-level standards, **AA-MAS** for an alternate assessment for eligible students with disabilities based on modified academic achievement standards; and/or **AA-AAS** for an alternate assessment for students with the most significant cognitive disabilities based on alternate achievement standards).

[1] Science assessments are not due until the 2007-08 school year.

3.2
If the State's assessment system includes assessments developed or adopted at both the local and State level, how has the State ensured that these local assessments meet the same technical requirements as the statewide assessments?

(a) How has the State ensured that all local assessments are aligned with the State's academic content and achievement standards?
(b) How has the State ensured that all local assessments are equivalent to one another in terms of content coverage, difficulty, and quality?
(c) How has the State ensured that all local assessments yield comparable results for all subgroups?
(d) How has the State ensured that all local assessments yield results that can be aggregated with those from other local assessments and with any statewide assessments?

How has the State ensured that all local assessments provide unbiased, rational, and consistent determinations of the annual progress of schools and LEAs within the State?

3.3
If the State's assessment system employs a matrix design—that is, multiple forms within a content area and grade level-- how has the State ensured that:

(a) All forms are aligned with the State's academic content and achievement standards and yield comparable results?
(b) All forms are equivalent to one another in terms of content coverage, difficulty, and quality?

All assessments yield comparable results for all subgroups?

3.4
How has the State ensured that its assessment system will provide coherent information for students across grades and subjects?

(a) Has it indicated the relative contribution of each assessment to ensure alignment to the content standards and determining adequate yearly progress?

(b) Has the State provided a rational and coherent design that identifies all assessments, including those based on alternate achievement standards and modified achievement standards if any, to be used for AYP?

(c) If the State assessment system includes alternate assessments based on alternate or modified achievement standards, has the State provided IEP Teams with a clear description of the differences between assessments based on grade-level achievement standards, assessments based on modified academic achievement standards and assessments based on alternate achievement standards, if applicable, including any effects of State and local policies on the student's education resulting from taking an alternate assessment based on alternate or modified academic achievement standards?

3.5
If its assessment system includes various instruments (e.g., the general assessment in English and either a native-language version or simplified English version of the assessment), how does the State demonstrate comparable results and alignment with the academic content and achievement standards?

3.6
How does the State's assessment system involve multiple measures, that is, measures that assess higher-order thinking skills and understanding of challenging content?

3.7
Has the State included alternate assessment(s) for students whose disabilities do not permit them to participate in the general assessment even with accommodations?

4.1
For each assessment, including all alternate assessments, has the State documented the issue of **validity** (in addition to the alignment of the assessment with the content standards), as described in the *Standards for Educational and Psychological Testing* (AERA/APA/NCME, 1999), with respect to *all* of the following categories:

(a) Has the State specified the purposes of the assessments, delineating the types of uses and decisions most appropriate to each? *and*
(b) Has the State ascertained that the assessments, including alternate assessments, are measuring the knowledge and skills described in its academic content standards and not knowledge, skills, or other characteristics that are not specified in the academic content standards or grade-level expectations? *and*
(c) Has the State ascertained that its assessment items are tapping the intended cognitive processes and that the items and tasks are at the appropriate grade level? a*nd*
(d) Has the State ascertained that the scoring and reporting structures are consistent with the sub-domain structures of its academic content standards (i.e., are item interrelationships consistent with the framework from which the test arises)? a*nd*
(e) Has the State ascertained that test and item scores are related to outside variables as intended (e.g., scores are correlated strongly with relevant measures of academic achievement and are weakly correlated, if at all, with irrelevant characteristics, such as demographics)? a*nd*
(f) Has the State ascertained that the decisions based on the results of its assessments are consistent with the purposes for which the assessments were designed? *And*
(g) Has the State ascertained whether the assessment produces intended and unintended consequences?

4.2
For each assessment, including all alternate assessments, has the State considered the issue of **reliability**, as described in the *Standards for Educational and Psychological Testing* (AERA/APA/NCME, 1999), with respect to *all* of the following categories:

(a) Has the State determined the reliability of the scores it reports, based on data for its own student population and each reported subpopulation? *and*
(b) Has the State quantified and reported within the technical documentation for its assessments the conditional standard error of measurement and student classification that are consistent at each cut score specified in its academic achievement standards? *and*
(c) Has the State reported evidence of generalizability for all relevant sources, such as variability of groups, internal consistency of item responses, variability among schools, consistency from form to form of the test, and inter-rater consistency in scoring?

4.3
Has the State ensured that its assessment system is fair and accessible to all students, including students with disabilities and students with limited English proficiency, with respect to each of the following issues:

(a) Has the State ensured that the assessments provide an appropriate variety of accommodations for students with disabilities? *and*
(b) Has the State ensured that the assessments provide an appropriate variety of linguistic accommodations for students with limited English proficiency? *and*
(c) Has the State taken steps to ensure fairness in the development of the assessments? *And*
(d) Does the use of accommodations and/or alternate assessments yield meaningful scores?

4.4
When different test forms or formats are used, the State must ensure that the meaning and interpretation of results are consistent.

(a) Has the State taken steps to ensure consistency of test forms over time?
(b) If the State administers both an online and paper and pencil test, has the State documented the comparability of the electronic and paper forms of the test?

4.5
Has the State established clear criteria for the administration, scoring, analysis, and reporting components of its assessment system, including all alternate assessments, and does the State have a system for monitoring and improving the on-going quality of its assessment system?

4.6
Has the State evaluated its use of accommodations?

(a) How has the State ensured that appropriate accommodations are available to students with disabilities and students covered by Section 504, and that these accommodations are used in a manner that is consistent with instructional approaches for each student, as determined by a student's IEP or 504 plan?
(b) How has the State determined that scores for students with disabilities that are based on accommodated administration conditions will allow for valid inferences about these students' knowledge and skills and can be combined meaningfully with scores from non-accommodated administration conditions?
(c) How has the State ensured that appropriate accommodations are available to limited English proficient students and that these accommodations are used as necessary to yield accurate and reliable information about what limited English proficient students know and can do?
(d) How has the State determined that scores for limited English proficiency students that are based on accommodated administration circumstances will allow for valid inferences about these students' knowledge and skills and can be combined meaningfully with scores from non-accommodated administration circumstances?

5.1 Has the State outlined a coherent approach to ensuring alignment between each of its assessments, or combination of assessments, based on grade-level achievement standards, and the academic content standards and academic achievement standards the assessment is designed to measure? Has the State outlined a coherent approach to ensuring alignment between each of its assessments, or combination of assessments, based on modified achievement standards and the academic content standards and academic achievement standards the assessment is designed to measure? Has the State outlined a coherent approach to ensuring alignment between each of its assessments, or combination of assessments, based on alternate achievement standards and the academic content standards and academic achievement standards the assessment is designed to measure?
5.2 Are the assessments and the standards aligned **comprehensively**, meaning that the assessments reflect the full **range** of the State's academic content standards? Are the assessments as cognitively challenging as the standards? Are the assessments and standards aligned to measure the depth of the standards? Does the assessment reflect the degree of cognitive complexity and level of difficulty of the concepts and processes described in the standards? If the State has implemented an alternate assessment based on modified academic achievement standards, does the assessment reflect the full range of the State's academic content standards for the grade(s) tested? What changes in cognitive complexity or difficulty, if any, have been made for assessments based on modified academic achievement standards? If the State has implemented an alternate assessment based on alternate academic achievement standards, does the assessment show a clear link to the content standards for the grade in which the students tested are enrolled although the grade-level content may be reduced in depth, breadth or complexity or modified to reflect pre-requisite academic skills?
5.3 Are the assessments and the standards aligned in terms of both **content** (knowledge) and **process** (how to do it), as necessary, meaning that the assessments measure what the standards state students should both know and be able to do? What changes in test structure or format, if any, have been made for assessments based on modified academic achievement standards?
5.4 Do the general assessments and alternate assessments based on modified achievement standards if any, reflect the same **degree and pattern of emphasis** as are reflected in the State's academic content standards?
5.5 Do the assessments yield scores that reflect the full range of achievement implied by the State's academic achievement standards?
5.6 Assessment results must be expressed in terms of the achievement standards, not just scale scores or percentiles.

5.7 What ongoing procedures does the State use to maintain and improve alignment between the assessments and standards over time?
6.1 1. Do the State's participation data indicate that all students in the tested grade levels or grade ranges are included in the assessment system (e.g., students with disabilities, students with limited English proficiency, economically disadvantaged students, race/ethnicity, migrant students, homeless students, etc.)? 2. Does the State report separately the number and percent of students with disabilities assessed on the regular assessment without accommodations, on the regular assessment with accommodations, on an alternate assessment against grade-level standards, and, if applicable, on an alternate assessment against alternate achievement standards and/or on an alternate assessment against modified academic achievement standards?

6.2

1. What guidelines does the State have in place for including all students with disabilities in the assessment system?

(a) Has the State developed, disseminated information on, and promoted use of appropriate accommodations to increase the number of students with disabilities who are tested against academic achievement standards for the grade in which they are enrolled?

(b) Has the State ensured that general and special education teachers and other appropriate staff know how to administer assessments, including making use of accommodations, for students with disabilities and students covered under Section 504?

2. If the State has approved/adopted modified or alternate academic achievement standards for certain students with disabilities, what guidelines does the State have in place for placing those students in the appropriate assessment?

(a) Has the State developed clear guidelines for IEP Teams to apply in determining which students with disabilities are eligible to be assessed based on modified or alternate academic achievement standards?

(b) Has the State informed IEP Teams that students eligible to be assessed based on alternate or modified academic achievement standards may be from any of the disability categories listed in the IDEA?

(c) Has the State provided IEP Teams with a clear explanation of the differences between assessments based on grade-level academic achievement standards and those based on modified or alternate academic achievement standards, including any effects of State and local policies on the student's education resulting from taking an alternate based on alternate or modified standards?

(d) Has the State ensured that parents are informed that their child's achievement will be based on modified or alternate academic achievement standards and of any possible consequences resulting from LEA or State policy (e.g., ineligibility for a regular high school diploma)?

3. If the State has adopted modified academic achievement standards, do the guidelines include all required components?

(a) Criteria for IEP Teams to use to determine which students with disabilities are eligible to be assessed based on modified academic achievement standards that include, at a minimum, each of the following?

- The student's disability has precluded the student from achieving grade-level proficiency as demonstrated by objective evidence of the student's academic performance; and
- The student's progress to date in response to appropriate instruction, including special education and related services designed to address the student's individual needs, is such that, even if significant growth occurs, the IEP Team is reasonably certain that the student will not achieve grade-level proficiency within the year covered by the student's IEP; and
- The student's IEP goals for subjects assessed by the statewide system are based on the academic content standards for the grade in which the student is enrolled.

(b) Has the State informed IEP Teams that a student may be assessed based on modified academic achievement standards in one or more subjects?

(c) Has the State established and monitored implementation of clear and appropriate guidelines for developing IEPs that include goals based on content standards for the grade in which a student is enrolled?

(d) Has the State ensured that students who are assessed based on modified academic achievement standards have access to the curriculum, including instruction, for the grade in which the students are enrolled?

(e) Has the State ensured that students who take an alternate assessment based on modified academic achievement standards are not precluded from attempting State diploma requirements?

(f) Has the State ensured annual IEP Team review of assessment decisions?

4. Has the State documented that students with the most significant cognitive disabilities are, to the extent possible, included in the general curriculum?

6.3
What guidelines does the State have in place for including all students with limited English proficiency in the tested grades in the assessment system?
(a) Has the State made available assessments, to the extent practicable, in the language and form most likely to yield accurate and reliable information on what these students know and can do?
(b) Does the State require the participation of every limited English proficient student in the assessment system, unless a student has attended schools in the US for less than 12 months, in which case the student may be exempt from one administration of the State's reading/language arts assessment?
(c) Has the State adopted policies requiring limited English proficient students to be assessed in reading/language arts in English if they have been enrolled in US schools for three consecutive years or more?

6.4
What policies and practices does the State have in place to ensure the identification and inclusion of migrant and other mobile students in the tested grades in the assessment system?

7.1
Does the State's reporting system facilitate appropriate, credible, and defensible interpretation and use of its assessment data?

7.2
Does the State report participation and assessment results for all students and for each of the required subgroups in its reports at the school, LEA, and State levels? In these assessment reports, how has the State ensured that assessment results are not reported for any group or subgroup when these results would reveal personally identifiable information about an individual student?

7.3
How has the State provided for the production of individual interpretive, descriptive, and diagnostic reports following each administration of its assessments?
(a) Do these individual student reports provide valid and reliable information regarding achievement on the assessments in relation to the State's academic content and achievement standards?
(b) Do these individual student reports provide information for parents, teachers, and principals to help them understand and address a student's specific academic needs? Is this information displayed in a format and language that is understandable to parents, teachers, and principals and are the reports accompanied by interpretive guidance for these audiences?
(c) How has the State ensured that these individual student reports will be delivered to parents, teachers, and principals as soon as possible after the assessment is administered?

7.4
How has the State ensured that student-level assessment data are maintained securely to protect student confidentiality?

7.5
How has the State provided for the production of itemized score analyses so that parents, teachers, and principals can interpret and address the specific academic needs of students?

Overview

This volume provides updated documentation of the alternate assessment in Oregon, its design and development, the technical characteristics of the instruments, and its use and impact in providing proficiency data on grade level state standards as part of the mandates from No Child Left Behind (NCLB).

Section 1: Content Standards

1.1 - 1.4 Content Standards

The Oregon Extended assessment, Oregon's alternate assessment based on alternate achievement standards (AA-AAS), is linked directly to the state's challenging, coherent, and rigorous general education content standards in reading, writing, mathematics, and science. The assessments were administered in the 2013-14 school year in grades 3-8 and once in the high school (10-12) grade band according to the following schedule:

Content Area	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 11
Reading	X			X			X
Writing		*			*		X
Mathematics	X	X	X	X	X	X	X
Science			X			X	X

*The Oregon State Legislature discontinued the writing assessments for the 2010-12 biennium in grades 4 and 7 due to budget shortfalls. The discontinuation extended through the 2013-14 school year, as well.

The Oregon Department of Education (ODE) has not yet developed or adopted extended/expanded grade band expectations for students with significant cognitive disabilities. The instructional targets for this group are thus established by the grade level standards for all Oregon students that are reduced in terms of depth, breadth, and complexity by practitioners, including teachers and item writers. Oregon general and special education teachers have reviewed all test items for: 1) linkage to the Oregon general education content standards, 2) accessibility for students with significant cognitive disabilities, 3) sensitivity, and 4) bias. All operational items met the established criteria.

Section 2: A Single Statewide Assessment of Challenging Academic Achievement Standards Applied to all Public Schools and LEAs

2.1 & 2.2 Content Standards and Alternate Achievement Standards (AAS)

The Oregon Extended assessment (OrExt), Oregon's Alternate Assessment based on Alternate Achievement Standards (AA-AAS), is part of the Oregon Statewide Assessment System. The Oregon Extended is administered to Oregon students with the most significant cognitive disabilities in grades 3-8 and 11. The State Board of Education in Oregon adopted the Common Core State Standards (CCSS) in English language arts and mathematics on October 28, 2010. The OrExt links to those standards in English language arts and mathematics. Oregon adopted the Next Generation Science Standards (NGSS) on March 6, 2014. An OrExt that is linked to the NGSS is planned for the 2018-19 school year. Until that time, the OrExt will remain linked to existing Oregon science standards. Results from the English language arts and math administrations are included in calculations of participation and performance for Annual Measureable Objectives (AMO) – a provision of NCLB. Science participation is also included as part of the Title 1 Assessment System requirements, and is administered in grades 5, 8, & 11.

All academic achievement standards for the OrExt assessment have been submitted for peer review in prior years. Oregon continues to use the same AAS for implementation of our AA-AAS. The AAS was developed by Oregon general and special education teachers and were in place in reading, writing, and mathematics in 2005-06. The AAS for science were in place in 2007-08, but revised in 2010-11 to link to newly adopted science standards.

Oregon is planning to develop and field test an entirely new AA-AAS in the 2014-15 school year, which will be founded in the CCSS and NGSS, but also linked to current OR science standards.

2.3 Levels of Achievement & Cut Scores

The alternate achievement standards in Oregon are composed of four levels (though only three are required). In descending order, they are 1) Exceeds, 2) Meets, 3) Nearly Meets, and 4) Does Not Yet Meet. The Exceeds and Meets levels denote an achievement level that represents high achievement, while the bottom two levels represent achievement that is not yet proficient. The procedures followed to develop Oregon's alternate achievement standards were consistent with Title 1 assessment system requirements, including the establishment of cut scores, where relevant. In order to define four levels of proficiency, Oregon set three cut scores across all subject areas: 1) to separate Exceeds from Meets, 2) to separate Meets from Nearly Meets, and, 3) to separate Nearly Meets from Does Not Yet Meet.

2.4 Same Standards Applied to All

This expectation applies only to general education assessments, by definition.

2.5 Alignment Between AAS and Content Standards

Peer reviewers have received historical documentation that the Oregon Extended assessments are linked to Oregon's academic content standards, promote access to the general education curriculum, and reflect professional judgment of the highest learning standards possible. Alignment documentation is ongoing as field test items become operational and can be found within section 4.3c.

2.6 Standard Setting

The *2010-11 Science Technical Report* submitted with Oregon's Peer Review submission included evidence of the standard setting process. Oregon has experienced no changes with regard to our academic achievement standards since that time and thus no additional evidence is being submitted as part of this technical report.

Section 3: A Single Statewide System of Annual High-Quality Assessments

3.1 Grades and Content Assessed

The evidence for this section, the statewide assessment chart, is not included as part of this technical report.

3.2 Local Assessments

Oregon administers statewide assessments and does not therefore need to establish comparability with local assessments.

3.3 Matrix Design

Oregon does not employ a matrix design. The Oregon Extended uses two versions of the same test, the Standard version and the Scaffold version.

3.4 Coherent Information

Oregon has provided documentation of 3.4(a), (b), and (c) in prior submissions. ORExt assessment results continue to be used for AMO calculations in English language arts and math.

3.5 Comparable Results

Though possible to translate into any language of instruction as an accommodation, the ORExt assessment is published exclusively in English. Form comparability based upon language is therefore not required.

3.6 Multiple Measures

The ORExt assessment is built upon a variety of items that address a wide range of performance expectations rooted in the CCSS, NGSS, and Oregon science content standards. The challenge built into the test design is based first upon the content within each standard in English language arts, mathematics, and science. That content is reduced in terms of depth, breadth, and complexity in a manner that is verified by Oregon general and special education teachers to develop assessment targets that are appropriate for students with the most significant cognitive disabilities. The scaffold and standard versions of the assessment are designed to provide access to all students, including these two disparate groups. Our assessments utilize universal design principles in order to include all students in the assessment process, while effectively challenging the higher performing students.

3.7 Alternate Assessments

Oregon has implemented an AA-AAS. Documentation of the procedures by which the current assessments and achievement standards were developed has already been submitted. Oregon does not have, nor does it plan to develop, an alternate assessment based on modified achievement standards (AA-MAS).

Section 4: Technical Quality

4.1 Validity

As elaborated by Messick (1989)[2], the validity argument involves a claim with evidence evaluated to make a judgment. Three essential components of assessment systems are necessary: (a) constructs (what to measure), (b) the assessment instruments and processes (approaches to measurement), and (c) 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? and (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*[3] (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 process.[4] The critical elements highlighted in Section 4 in that document (with examples of acceptable evidence) include (a) academic content standards, (b) academic achievement standards, (c) a statewide assessment system, (d) reliability, (e) validity, and (f) other dimensions of technical quality.

Given the content-related evidence that we present related to test development, administration, and scoring, the response processes related to the levels of independence, the reliability information reflected by adequate coefficients for tasks and tests, and finally, the relation of tasks within and across subject areas (providing criterion-related evidence), we conclude that the alternate assessment judged against alternate achievement standards allows valid inferences to be made on state accountability proficiency standards.

4.1(a) Content

In this technical report, data is presented to support the claim that Oregon's AA-AAS provides the state technically adequate student performance data to ascertain proficiency on grade level state content standards for students with significant cognitive disabilities – which is its defined purpose. The AA-AAS are linked to grade level academic content; generate reliable outcomes at the test level; include all students; have a cogent internal structure; and fit within a network of relations within and across various dimensions of content related to and relevant for making proficiency decisions.

---

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

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

[4] 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*

4.1(b) Knowledge and Skills

The ORExt assessments have been determined to link to grade level academic content, as specified for all tested subject areas in May 2008. This was documented and submitted within the *2007-08 Technical Report*. Subsequent alignment studies were implemented in mathematics and science when Oregon adopted new general education content standards in those two content areas after the 2007-08 school year. Alignment documentation in mathematics was submitted in the *Oregon Alternate Assessment 2011 Alignment Study in Mathematics*, completed on February 12, 2011. In the area of science, linkage to grade level content has been documented in the *Oregon Alternate Assessment 2011 Alignment Study in Science*, completed on May 4, 2011. These studies were both required because Oregon adopted new general education content standards in mathematics and science, respectively.

Because the assessments demonstrate appropriate linkage to Oregon's general education content standards and descriptive statistics demonstrate that each content area assessment is functioning as intended, it is appropriate to assume that these standards define the expectations that are being measured by the Oregon Extended assessments. See *Appendix D*, providing correlational statistics.

4.1(c) Cognitive Processes

Evidence of content coverage is concerned with judgments about "the adequacy with which the test content represents the content domain" (AERA et al., 1999, p. 11)[7]. As a whole, the ORExt is comprised of sets of items that sample student performance on the intended domains. The expectation is that the items cover the full range of intended domains, with a sufficient number of items so that scores credibly represent student knowledge and skills in those areas. Without a sufficient number of items, the potential exists for a validity threat due to construct under-representation (Messick, 1989)[5].

Our foundation of validity evidence from content coverage comes in the form of test blueprints or test specifications. Among other things, the *Standards* (AERA et al., 1999)[7] suggest specifications should "define the content of the test, the number of items on the test, and the formats of those items" (Standard 3.3, p. 43).[6]

All items and tasks are linked to grade level standards and a prototype was developed using principles of universal design[7] with traditional, content-referenced multiple-choice item writing techniques[8]. The most important component in these initial steps addressed language complexity and access to students using both receptive, as well as expressive,

---

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

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

[7] Johnstone, C., Thompson, S., Bottsford-Miller, N., & Thurlow, M. (2008). Universal design and multimethod approaches to item review. *Educational Measurement: Issues and Practice, 27*(1), 25-36. doi: 10.1111/j.1745-3992.2008.00112.x

[8] Halydyna, T. , & Rodriquez, M. C. (2013). *Developing and validating test items*. New York, NY: Routledge.

communication. Additionally, both content breadth and depth were addressed. We developed two test forms of each grade level test, a standard and a scaffold version. The scaffold administration utilizes a more accommodated approach that allows for students with very limited attentional resources to access the same test content as their peers who participate in the standard version. The test is designed to be comparable across multiple disabilities, with prerequisite skills and test type accounting for most of the variance. Any differences between the assessments are thus deemed to be construct-irrelevant (see *Appendices A-H*). In each task, we generally increased the depth of knowledge from the first to the last item.

We developed the test iteratively by developing items and tasks, piloting them, reviewing them, and editing successive drafts. We used a combination of existing panels of veteran teachers who have worked with the Oregon Department of Education (ODE) in various advising roles on testing content in general education, using the same processes and criteria, as well as the introduction of newer teachers who are qualified as we proceed to remain relevant. Behavioral Research and Teaching (BRT) personnel conducted the internal reviews of content. After the internal development of prototype items, all reviews then involved Oregon content experts with significant training and K-12 classroom experience. The first level review was to ensure universal design and incorporated two experts to represent the blind and deaf communities. Finally, subsequent reviews were conducted to ensure appropriate administration and scoring, all of which was completed as part of training.

Due to the substantive evidence that has been documented, evidence suggests that the ORExt items are tapping the intended cognitive processes and that the items and tasks are at the appropriate grade level through the alignment studies documented above, including reviews of linkage, content coverage, and depth of knowledge.

4.1(d) Scoring and Reporting

The primary purpose of the ORExt assessment is to yield technically adequate performance data on grade level state content standards for students with significant cognitive disabilities in English language arts, mathematics, and science at the test level. All scoring and reporting structures mirror this design and have been shown to be reliable measures at the test level (see *Appendix B*).

4.1(e) Criterion

Pre-requisite skills assessments are designed to allow teachers to use various levels of support, as appropriate. Because the ORExt assessment first documents the student's access skill (pre-requisite skill) to assist teachers in presenting the content items, pre-requisite skills are assessed to provide the necessary supports for appropriate test administration (with four levels: full physical support, partial physical support, prompted support, and no support). Content prompts were designed to document students' skill and knowledge on grade level academic content standards. There are also two test administration types that Individualized Educational Program (IEP) teams select: (a) standard or (b) scaffold. Both types address the same content and only differ in the amount of scaffolding they provided to access the target skill (content prompt).

Perhaps the best model for understanding criterion-related evidence comes from Campbell and Fiske (1959)[9] in their description of the multi-trait, multi-method analysis. [we translate the term 'trait' to mean 'skill']. In this process (several) different traits are measured using (several) different methods to provide a correlation matrix that should reflect specific patterns supportive of the claim being made (that is, provide positive validation evidence). Sometimes, these various measures are of the same or similar skills, abilities, or traits, and other times they are of different skills, abilities, or traits. We present data that quite consistently reflect higher relations among tasks **within** an academic subject than **between** academic subjects. We also present data in which performance on content prompts is totaled within categories of disability, expecting relations that would reflect appropriate differences (see Tindal, McDonald, Tedesco, Glasgow, Almond, Crawford, & Hollenbeck, 2003).[10]

As mentioned in section 4.1b, our assessments appear to be measuring separate constructs, as intended (see *Appendix D)*, provided the Pearson correlation statistics. *Appendices E-H* provide regression model analyses demonstrating that student performance is primarily associated with the task, not the level of support provided (as determined by the Prerequisite Skills task), the test type (Standard versus Scaffold), or the student's disability. In addition, demographic calculations demonstrate that the assessments are not biased toward any ethnic groups (see *Appendix A)*.

4.1(f) Decisions

As mentioned above in section 4.1a, data are presented to support the claim that Oregon's AA-AAS provides the state technically adequate student performance data to ascertain proficiency on grade level state content standards for students with significant cognitive disabilities – which is its defined purpose. The AA-AAS are linked to grade level academic content; generate reliable outcomes at the test level; include all students; have a cogent internal structure; and fit within a network of relations within and across various dimensions of content related to and relevant for making proficiency decisions.

4.1(g) Consequential

ODE implemented a research survey program to address the need to document the consequences, both intended and unintended, of the ORExt Assessments. The research questions have been framed based upon current consequential validity approaches for alternate assessments in the literature, as well as issues that are of specific value in Oregon. The survey generated 600 responses (17% of requested respondents). The sample was 77% female and represented all regions of the state, as well as age ranges. The survey included a range of quantitative and qualitative components.

---

[9] Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multi-trait, multi-method matrix. In W. A. Mehrens & R. L. Ebel (Eds.), *Principles of educational and psychological measurement: A book of selected readings* (pp 273-302). Chicago, IL: Rand McNally & Company.

[10] Tindal, G., McDonald, Tedesco, M., Glasgow, A., Almond, P., Crawford, L., & Hollenbeck, K. (2003). Alternate assessments in reading and math: Development and validation for students with significant disabilities. *Exceptional Children*, 69(4), 481-494.

The quantitative results demonstrate that Oregon educators have an equivocal impression regarding the social impact that the ORExt is having on students with significant cognitive disabilities, as measured by our survey. Respondents generally responded that they mildly disagreed that positive impacts could be associated with the ORExt; at the same time, they also conveyed that no negative consequences are associated with the ORExt.

The qualitative results revealed three areas in which educators felt that positive consequences were associated with the ORExt: 1) Participation: SWSCDs were included in the statewide assessment system (not excluded because of their disability and could feel good about taking an assessment just like their general education peers, 2) Data: educators who used the data found that it can be helpful for a variety of applications, including placement, IEP development, adaptation of curriculum and instruction, and 3) One-to-one time with the instructor: some felt that the time that students got to spend with their teachers was very beneficial, providing them with a lot of positive adult feedback; others remarked that the test administration process itself also provided useful information to the careful observer. There were also three primary domains that educators felt were negative aspects of the ORExt: 1) Time: educators felt that the time spent testing could better be spent working on instructional goals related to the students' Individualized Education Programs (IEPs); 2) Challenge: educators felt both that the test was too difficult and too easy, depending upon the students whom they served; several remarked that the content on the assessment does not match what is being taught in the classroom, and 3) Data: educators shared frustration that they do not receive score reports for the ORExt. Complete results from the survey can be found in *Appendix 2.5*.

4.2(a) Score Reliability

Three types of analyses are presented in *Appendix B*: (a) inter-item correlations, (b) internal consistency for each task in each subject area for every grade level, and (c) total test reliabilities. The tests are quite reliable at the total test levels.

4.2(b) Standard Error of Measure

The average SEM associated with each cut score for 2013-14 student data is presented in the table below:

Key:

SEM = *Standard Error of Measure associated with the cut score to the left; averaged to the hundredths' place.*

DNYM = Does Not Yet Meet (not included as the lowest level of proficiency)

NM = Nearly Meets

M = Meets

E = Exceeds

READING

Grade	NM	*SEM*	M	*SEM*	E	*SEM*
3	**97**	*2.0*	**103**	*2.0*	**113**	*2.6*
4	**101**	*1.9*	**107**	*2.1*	**116**	*3.0*
5	**105**	*2.0*	**110**	*2.3*	**119**	*3.5*
6	**97**	*2.1*	**103**	*2.1*	**116**	*3.2*
7	**98**	*2.1*	**106**	*2.2*	**117**	*3.4*
8	**102**	*2.1*	**112**	*2.7*	**120**	*3.8*
11	**101**	*2.2*	**109**	*2.5*	**121**	*3.6*

WRITING

Grade	NM	*SEM*	M	*SEM*	E	*SEM*
11	**98**	*2.1*	**103**	*2.3*	**122**	*4.6*

MATHEMATICS

Grade	NM	*SEM*	M	*SEM*	E	*SEM*
3	**97**	*1.8*	**104**	*1.8*	**112**	*2.5*
4	**98**	*1.9*	**106**	*1.9*	**114**	*2.7*
5	**101**	*1.9*	**110**	*1.9*	**118**	*2.9*
6	**99**	*1.7*	**103**	*1.9*	**110**	*2.9*
7	**101**	*1.7*	**102**	*1.8*	**107**	*2.2*
8	**101**	*1.7*	**105**	*1.8*	**110**	*2.1*
11	**99**	*1.8*	**106**	*1.9*	**115**	*3.2*

SCIENCE

Grade	NM	*SEM*	M	*SEM*	E	*SEM*
5	**100**	*1.8*	**107**	*2.0*	**114**	*2.7*
8	**95**	*1.9*	**101**	*1.7*	**113**	*2.7*
11	**98**	*1.7*	**103**	*1.9*	**109**	*2.5*

4.2(c) Generalizability

Oregon has reported evidence of generalizability for all relevant sources, including analyses by demographic groups, to ensure that items are functioning consistently across demographic groups (see *Appendix G & H)*. The internal consistency of item responses is analyzed within *Appendix B*. Tindal and Yovanoff conducted a study in 2007-08 wherein a sample of students was administered both the standard versions and the scaffold versions of the assessment demonstrated that students performed similarly irrespective of test version at the task and test level. This documentation was submitted in the *2007-08 Technical Report*.

4.3(a) Accommodations

The Oregon Extended assessments are designed according to universal design principles and utilize a simplified language approach (see *Appendix 1.5*). They are also provided in two versions, the standard and the scaffold, to support access to the tests. The Oregon Extended assessments can be ordered in both Large Print and Braille (contracted and non-contracted) versions, as well. Oregon has ensured that the Oregon Extended assessments provide an appropriate variety of accommodations for students with disabilities. The state has provided guidance regarding accommodations in presentation, response, setting, and timing in the *Accommodations Manual 2013-14: How to Select, Administer, and Evaluate Accommodations for Oregon's Statewide Assessments* (see *Appendix 2.4*). Accommodations that are used in Oregon are also analyzed at the test level to ensure that they are indeed leveling the playing field and not providing any particular advantage or disadvantage to any defined group.

The state also developed a training and proficiency program for sign language interpretation of its assessments in the 2013-14 school year. The training process (http://orschools-ode-accomm.ziptrain.com) included videos of interpreters administering items to students, materials that support appropriate administration (i.e., transcripts and PowerPoint slides that supplement the video administrations and the current ODE accommodations manual), and proficiency testing to support standardized interpretation for Oregon's assessments, including the ORExt. Participants reviewed examples of appropriate interpreted administration, along with commentary, as well as non-examples. A 15-item proficiency test was administered, with an 80% required for passing (12/15 items correct). The site was used to train 78 participants. Three participants took two attempts to pass the profiency test. One participant took three attempts to pass the proficiency test. The overall average score on the proficiency test was 92%.

4.3(b) Linguistic Accommodations

The ORExt assessments provide an appropriate variety of linguistic accommodations for students with limited English proficiency. They also use a simplified language approach in test development in order to reduce language load of all items systematically (see *Appendix 1.5*). Any given student's communication system may include home signs, school signs, English words, and Spanish words, for example. The ORExt assessment can be translated or interpreted by a qualified administrator in the student's native language. Administrators

are allowed to translate/interpret the test directions. Assessors can adapt the assessment to meet the needs of the student, while still maintaining standardization due to systematic prompts and well-defined answers . As mentioned above, the state has also developed a training and proficiency program for sign language interpretation of its assessments, which was implemented in the 2013-14 school year.

4.3(c) Fairness

The state has taken steps to ensure fairness in the development of the assessments, including an analysis of each field test item by Oregon teachers not only for linkage to standards, but also for access, sensitivity, and bias. This process increases the likelihood that students are receiving instruction in areas reflected in the assessment, and also that the items are not biased toward a particular demographic or sub-group (see *Appendix 1.5*).

Oregon Extended Assessment Field Testing 2012-13

Field testing was conducted in reading, writing, mathematics, and science, in the 2013-14 school year. Field testing in writing was conducted only in grade 11, as that was the only grade in which the writing assessment was administered due to budgetary constraints.

The field test development plan followed a three-year implementation strategy in English language arts and mathematics that steadily transitioned the OrExt toward items and assessments that are now linked with the CCSS. The implementation strategy for science follows an independent course that is defined by the State of Oregon's 2009 Science Standards.

2013-14 Field Test Development Plan

The table below provides an overview of the field test implementation plan for 2013-2014:

Extended Assessment Subject Area	**Field Test Location on 2013-14 Test**
Reading grades Elementary (3, 4, 5) Middle School (6, 7, 8) High School grade 11	Tasks 2 & 3 (total of 10 items per grade level/band)
Writing (Language) grade 11	Tasks 8 & 9 (total of 10 items per grade level/band)
Math grades 3, 4, 5, 6, 7, 8, 11	Item 6 in Tasks 2 – 9 (total of 8 items per grade level)
Science grades 5, 8, 11	Item 6 in Tasks 2 – 9 (total of 8 items per grade level)

A total of 10 items were developed for each grade band/level for the English language arts assessments, while eight Mathematics and Science items were developed for each grade band/level.

2011-14 Field Test Development Plan

We have now reached the end of a three-year implementation plan. Thus, all current items link to the CCSS in English language arts and mathematics, while all science items are linked to Oregon's current science standards. The table below provides an overview of the three-year CCSS implementation plan, including years 2012, 2013, and 2014:

Content Area/Grade	Projected number of CCSS-aligned field test items developed by 2014	Total Number of Items on Operational Test (excluding Prerequisite Skills)
Reading Grades 3-5	30	50
Reading Grades 6-8	30	50
Reading Grade 11	30	50
READING TOTAL	**90**	**150**
Writing Grade 11	30	50
Writing Grade 7	20	50
Writing Grade 4	20	50
WRITING TOTAL	**70**	**150**
Math Grade 3	24	48
Math Grade 4	24	48
Math Grade 5	24	48
Math Grade 6	24	48
Math Grade 7	24	48
Math Grade 8	24	48
Math Grade 11	24	48
MATH TOTAL	**168**	**336**
Science Grade 5	24	48
Science Grade 8	24	48
Science Grade 11	24	48
SCIENCE TOTAL	**72**	**144**

Distributed Item Review & Data Analysis 2013-14

The Oregon Department of Education contracted with Behavioral Research and Teaching (BRT) to develop field test items in reading, writing, math, and science for the 2013-14 spring test administration. BRT employed a multi-stage development process to ensure that test items were linked to relevant content standards, were accessible for students with significant cognitive disabilities, and that any perceived item biases were eliminated. This review process included 33 reviewers with an average of 22.3 years of experience in education (see *Appendix 1.4*).

4.3(d) Meaningful Scores

While accommodations are truly built into the test design to a large degree, as described above, the use of accommodations on the ORExt assessments does not appear to interfere with the constructs being measured and therefore the scores yielded by such administrations are deemed to be comparably useful to an administration without accommodation. ODE plans to collect specific accommodations information in future years (e.g., by coding which accommodations were used during testing for each assessment during data entry), which will support further research in this area.

4.4(a) Test Form Consistency

The ORExt assessments are provided in two versions, the standard and the scaffold. Both versions use the same prompts, while the scaffold version provides additional supports to redirect students with limited attention to tasks at hand. A study conducted in the 2007-08 school year demonstrated that comparable results are achieved irrespective of the version administered. The results are thus deemed to be comparable.

4.4(b) Test Form/Format Comparability

The ORExt assessments are administered only in a paper and pencil format.

4.5 Clear Criteria

The ORExt assessments are administered according to the administration, scoring, analysis, and reporting criteria established in the General Administration and Scoring Manual (see *Appendix 1.2*). Test security policies and consequences for violation are addressed in the Test Administration Manual on an annual basis (see *Appendix 2.3*). The state's accommodations manual clearly delineates which accommodations can be administered for which assessments (see *Appendix 2.4*). Oregon requests and receives feedback regarding its assessment system in the form of training evaluations. An established ODE contact person is available to assist with policy-related questions, while BRT provides a HelpDesk related to the training and proficiency website. All technical assistance is documented and reviewed for patterns that can be used to make systematic improvements from year to year (see *Appendix 1.1h*). The state's training program for test administration is complex; it is described below.

Oregon Extended Assessment Training 2013-14

The Oregon Department of Education (ODE) provided three direct statewide trainings for new Qualified Trainers (QTs) and Qualified Assessors (QAs) via a face-to-face trainings in Hillsboro and Salem and a regionally-hosted webinar trainings in the Redmond/Medford and Pendleton areas. The schedule for the regional trainings, as well as relevant training information, is provided below:

Date	Who/Team	Location
11-07-2013	Team: Brad Lenhardt & Gerald Tindal Contact: Joan Steiner joans@nwresd.k12.or.us	NWRESD Hillsboro, OR
11-12-2013 (combined webinar)	Team: Brad Lenhardt, Gerald Tindal, & Dan Farley Contact: Catherine Kelly catherine.kelly@hdesd.org Marian Gerstmar marian_gerstmar@soesd.k12.or.us	HDESD- Redmond, OR SOESD- Medford, OR
11-14-2013	Team; Brad Lenhardt, Gerald Tindal, & Dan Farley Contact: Eleni Boston eleni.boston@wesd.org	Willamette ESD Salem, OR
11-21-2013	Team: Brad Lenhardt, Gerald Tindal, Dan Farley Contact: Mary Apple mary.apple@imesd.k12.or.us	Inter-Mountain ESD Pendleton, OR

The Oregon Extended assessment can be administered only by trained Qualified Assessors (QAs). Qualified Assessors who also receive direct instruction from ODE and BRT may become Qualified Trainers (QTs) who are certified to train local staff using the train-the-trainers model. Training for new assessors must be completed on an annual basis. Assessors who do not maintain their respective certifications for any given year must re-train if they choose to enter the system again.

The tables below contain data from the Oregon Extended Assessment Training and Proficiency Website (http://or.k12test.com/). This website is a component of required training for assessors of the Extended Assessment. All assessors need to complete some

form of training each year to retain their status for administering the Extended Assessments.

New assessors, or returning assessors who needed further training again in 2013-14, were required to pass five proficiencies with a score of 80% or higher. These five proficiencies were in Administration, Reading, Math, Writing, and Science. Returning QAs or QTs for the 2012-13 school year only needed to pass a Refresher Proficiency, again with a score of 80% or higher. The tables below contain data on the number of assessors (participants) in each of the five proficiencies, as well as the Refresher Proficiency. Included in the data is the number of attempts needed to attain a passing score as well as the average passing score of the participants.

The number of assessors on the Oregon Extended Assessment Training and Proficiency Website:

Assessor in-Training - 1,038
Qualified Assessors - 1,275
Qualified Trainers - 159

301 Test Participants – Administration Proficiency

Number of Participants	Percentage of Participants	Attempts to Pass	Average Passing Score
227	75%	1	92%
58	19%	2	90%
14	5%	3	91%
1	>1%	4	95%
1	>1%	5	80%

291 Test Participants – Reading Proficiency

Number of Participants	Percentage of Participants	Attempts to Pass	Average Passing Score
281	97%	1	95%
8	3%	2	88%
2	>1%	3	95%

287 Test Participants – Math Proficiency

Number of Participants	Percentage of Participants	Attempts to Pass	Average Passing Score
285	99%	1	98%
2	1%	2	95%

284 Test Participants – Writing Proficiency

Number of Participants	Percentage of Participants	Attempts to Pass	Average Passing Score
234	82%	1	91%
32	11%	2	87%
15	5%	3	91%
1	>1%	4	80%
1	>1%	5	85%
1	>1%	6	80%

283 Test Participants – Science Proficiency

Number of Participants	Percentage of Participants	Attempts to Pass	Average Passing Score
282	99%	1	99%
1	1%	2	100%

1,149 Test Participants – Refresher Proficiency

Number of Participants	Percentage of Participants	Attempts to Pass	Average Passing Score
1,084	94%	1	93%
46	4%	2	88%
19	2%	3	93%

An additional analysis was compiled which compared the average score of participants on their first attempt of a proficiency to the final passing score of the proficiency subject area. All final attempt average scores were higher than attempt one average scores.

Comparison of Attempt One Scores to Passing Score on Final Attempt

Subject	Attempt 1 Average Score	Number of Participants	Final Attempt Passing Score	Number of Participants
Administration	85%	306	91%	301
Reading	94%	292	95%	291
Math	98%	288	98%	287
Writing	87%	288	89%	284
Science	98%	284	99%	283
Refresher	91%	1,159	92%	1,149

Assessors had the most difficulty passing the Administration assessments. However, it was a small number of participants who struggled to pass these assessments by the second attempt, with a range of 0% to 5% of Assessors who did not pass after two attempts. One Assessor required six attempts before s/he was able to pass the Writing proficiency.

Training and Test Appendices

Topic	File Name
Slides for training new qualified assessors, new qualified trainers, and returning assessors	App1.1a_QATraining2013_14
Slides for orienting assessors to the use of the Training and Proficiency website	App1.1b_ORExtendQTTrng2013_14
A handout which reviews general questions about the Oregon Extended assessment program	App1.1c_ORExtFAQs2013_14
The final test administration calendar for all Oregon assessments	App1.1d_TestSchd2013_14
Sample agenda for training new qualified assessors using the train-the-trainers model	App1.1e_QT_Training_Agenda2013_14
Provides assessors and trainers instructions regarding how to access the online training and proficiency website	App1.1f_ExtAssessAccessInstr2013_14
Provides qualified trainers with a list of duties associated with their training responsibilities	App1.1g_TrainerResponsibilities2013_14
Help Desk log and evaluation report	App1.1h_HelpDeskLog2013_14
General Administration and Scoring Manual	App1.2_ExAssessAdminMan2013_14
Sample test items in RWMS	App1.3_RWMSSampleItems2013_14
Report summarizing the results of the field test item reviews conducted with Oregon teachers	App1.4DIRReport2013_14
Describes how items are reduced in depth, breadth, and complexity, as well as how item bias is reduced/eliminated.	App1.5_ OORExtReduceDepthBreadthComplex2013_14
Mock-up individual student report, demonstrating that the reports contain cut scores and ALDs	App1.6_OORExtend_StudentReport_Mock_Up
Guidance developed by ODE to assist IEP teams in making appropriate assessment decisions	App2.1_AssessDecisionMakingGuidelines
Guidance for assessors for data entry process	App2.2_DataEntryGuide2013_14
ODE Test Administration Manual (TAM), providing test security and administration requirements for all Oregon assessments,(see *Appendix I*, page I-1)	App2.3_TAM2013_14
Provides ODE's guidance and expectations related to accommodations.	App2.4_ODEAccomMan2013_14
Provides information regarding the perceived social consequences of OORExt implementation	App2.5_OORExtConsValidRept2013_14

Appendix 1.1a-b
Appendix 1.1a & 1.1b are the PowerPoint trainings that were used by ODE and BRT trainers to train new qualified assessors (QAs) and qualified trainers (QTs) in four regionally hosted webinar trainings in November 2013. QTs also used the package to train New Qualified Assessors for the 2013-14 school year. The training provides participants with the information needed to pass proficiency tests as part of the requirements to become a QA for the Oregon Extended Assessments and was delivered by QTs throughout the state. The training package addresses the following topics:

- What's new in 2013-14
- 2014 Test Window
- Eligibility – which students take AA-AAS?
- Standard Administration/Scaffold Administration?
- Student Confidentiality & Test Security
- Test Administration (Physical & Logistic)
- Scoring & Data Entry
- Reports & Sharing Results with Parents
- 2014 Field Testing Plan
- Navigating the Training and Proficiency website
- Resources

Appendix 1.1c
Appendix 1.1c is a document that provides general test administration (which students should take the ORExt assessment, why, etc.) questions and answers.

Appendix 1.1d
Appendix 1.1d is the test calendar for the entire Oregon statewide assessment program, including the OAKS, the ORExt, the ELPA, and the NAEP.

Appendix 1.1e
Appendix 1.1e is a sample agenda that ODE makes available to QTs around the state to train their respective new QAs as they implement the train-the-trainers model used by the Oregon Extended assessment.

Appendix 1.1f
Appendix 1.1f is the list of instructions provided to new QAs and QTs regarding how to access the online training and proficiency website.

Appendix 1.1g
Appendix 1.1g is the list of responsibilities associated with being a QT for the ORExt assessment

Appendix 1.1h
Appendix 1.1h is the report that summarizes all of the technical assistance questions garnered from the field this year. Efforts are made to find any patterns that our team may use to improve training for the following year.

Appendix 1.2
Appendix 1.2 is ODE's General Administration and Scoring Manual for 2013-14. The manual establishes ODE's expectations regarding the test window, utilizing the ORExt training and proficiency website, using the sign language interpreter training and proficiency website, and informing parents. It also provides the following information for stakeholders, including educators and parents:

- Overview of the Extended Assessments
- Assessing a Student
- Scoring
- Decision Making
- Information for Teachers.

The manual provides three appendices that provide guidance regarding the provision of supports, parent questions and answers, and a glossary.

Appendix 1.3
Appendix 1.3 provides stakeholders with visual representation of the structure of the ORExt. Sample tasks/items are conveyed, including both Prerequisite and Content Prompts. There are standard and scaffold administration tasks represented in reading, writing, math, and science. The appendix shows what a QA would view during test administration (Scoring/teacher's Protocol) as well as what students would view as the QA asks the test questions (Student Materials). Stakeholders can see the structure of each task/item, as well as how the items are scored. They can also gather an idea about the types of formats that are used for answer choices that are included within the Student Materials documents.

Appendix 1.4
Appendix 1.4 is a document that summarizes the process and participants used to review ongoing field test items for the ORExt using the Distributed Item Review (DIR) website, supported by a webinar training and ongoing technical assistance.

Appendix 1.5
Appendix 1.5 is a document that summarizes the procedures used during item development to reduce item depth, breadth, and complexity. The document also provides more detail regarding how language complexity is addressed and reviewed in an effort to decrease the language load of items and make the test more accessible to all students. The document also discusses ways in which bias is addressed during test development.

Appendix 1.6
Appendix 1.6 is a document that displays the individual student report (ISR) that ODE publishes for students who participate in the ORExt. The mock-up includes cut scores and

achievement level descriptors (ALDs), as well as links to the ODE website for additional information.

Appendix 2.1
Appendix 2.1 is the guidance that ODE has provided to IEP teams to assist them in making appropriate assessment determinations for students with disabilities.

Appendix 2.2
Appendix 2.2 is the guidance that ODE has provided to assessors to walk them through the online data entry process for the ORExt.

Appendix 2.3
Appendix 2.3 is the test administration manual for all assessments in the Oregon statewide assessment system, including the OAKS, the ORExt, and the ELPA.

Appendix 2.4
Appendix 2.4 is the accommodation manual for all assessments in the Oregon statewide assessment system, including the OAKS, the ORExt, and the ELPA. The manual provides guidance regarding use of accommodations in instruction and assessment, as well as implementation strategies and accommodations use evaluation. Each accommodation is coded for use in data analysis related to assessment scores for the OAKS.

Appendix 2.5
Appendix 2.5 is the consequential validity report for the spring 2014 consequential validity study conduced by BRT. The report provides document of the perceptions in the field related to both intended and unitended social consequences of the ORExt.

4.6(a) Appropriate Accommodations Available for SWDs

The state has ensured that appropriate accommodations are available to students with disabilities and students covered by Section 504 by providing guidance and technical support on accommodations (see *Appendix 2.4*). Guidelines regarding use of the accommodations for instructional purposes are included in the document, as all students are expected to receive test accommodations that are consistent with instructional accommodations.

4.6(b) Accommodated SWD Administration Validity

While accommodations are built into the flexibility provided by the ORExt test design and assessment results demonstrate that student performance varies according to their abilities and not other irrelevant factors, Oregon is researching our ability to analyze specific accommodations that have been administered by assessors for the ORExt. This work is in addition to the annual training and proficiency testing efforts related to becoming a qualified assessor and/or qualified trainer for the ORExt.

4.6(c) Appropriate Accommodations Available for LEP Students

The state has ensured that appropriate accommodations are available to students with limited English proficiency by providing guidance and technical support on accommodations (see *Appendix 2.4*) Communication systems for this student population are limited; exposure to multiple languages can make a student's communication system more complex. The ORExt uses universal design principles and simplified language approaches in order to increase language access to test content for all students. In addition, directions and prompts may be translated/interpreted for students in their native language. An analysis of accommodated versus non-accommodated administrations is needed in order to demonstrate that the provision of language accommodations is not providing any advantage to students with limited English proficiency, nor any disadvantage to other participants.

4.6(d) Accommodated LEP Administration Validity

An analysis of accommodated versus non-accommodated administrations is needed in order to demonstrate that the provision of language accommodations is not providing any advantage to students with limited English proficiency, nor any disadvantage to other participants. This type of analysis should be feasible once accommodations information is collected during data entry, currenly planned for the 2015 administration.

Data Analyses

Eight analyses were conducted on Oregon Extended assessment data this year, including analyses of demographics, reliability, descriptive statistics, correlations, and four regression models:

Data Analyses Appendices Table

Topic	File Name
Demographics for participants in 2012-2013 alternate assessment	AppA_Dems
Reliability of items, tasks, and tests in reading, writing, mathematics and science for all grade levels	AppB_Reliab
Descriptive statistics for all tasks in reading, writing, mathematics and science for all grade levels	AppC_Descrpt
Correlations across subject areas	AppD_Corr
Simultaneous regression model using pre-requisite skills as a predictor of scale scores	AppE_Model_1
Simultaneous regression model using pre-requisite skills *and* type of administration as a predictor of scale scores	AppF_Model_2
Sequential regression model using disability, test administration type, and demographics to predict pre-requisite skills total	AppG_Model_3
Sequential regression model using disability, test administration type, and demographics to predict scale scores	AppH_Model_4

Demographics

The full demographics for students taking the ORExt are reported in *Appendix A*. Students race/ethnicity was reported in seven categories: (a) Asian/Pacific Islander, (b) American Indian/Alaskan Native, (c) Black, (d) Hispanic, (e) Multiethnic, (f) White, and (g) Decline/Missing. In each grade, the majority of students' ethnic categories were reported as Hispanic or White.

Reading

Elementary Grade Band. For grades 3-5, approximately 66.8% were male, 54.4% were White, and 30.4% were Hispanic. Approximately 70.0% of all students were administered the Standard version of the test, while the remaining 30.0% were administered the Scaffold version of the test.

Middle Grade Band. For grades 6-8, approximately 65.9% were male, 61.1% were White, and 24.8% were Hispanic. Approximately 63.6% of all students were administered the Standard version of the test, while the remaining 36.4% were administered the Scaffold version of the test.

High School. Approximately 64.7% were male, 64.1% were White, and 22.0% were Hispanic. Approximately 55.0% of all students were administered the Standard version of the test, while the remaining 45.0% were administered the Scaffold version of the test.

Writing

Grade 11. Approximately 66.2% were male, 64.2% were White, and 22.1% were Hispanic. Approximately 55.4% of all students were administered the Standard version of the test, while the remaining 44.6% were administered the Scaffold version of the test.

Math

Grade 3. Approximately 66.1% of students taking the mathematics portion of the Oregon Extended Assessment were male, 52.8% were White, and 30.1% were Hispanic. Approximately 61.5% of all students were administered the Standard version of the test, while the remaining 38.5% were administered the Scaffold version of the test.

Grade 4. Approximately 65.4% were male, 55.1% were White, and 29.1% were Hispanic. Approximately 66.1% of all students were administered the Standard version of the test, while the remaining 33.9% were administered the Scaffold version of the test.

Grade 5. Approximately 64.0% were male, 55.5% were White, and 27.7% were Hispanic. Approximately 69.1% of all students were administered the Standard version of the test, while the remaining 30.9% were administered the Scaffold version of the test.

Grade 6. Approximately 65.3% were male, 58.2% were White, and 27.4% were Hispanic. Approximately 64.3% of all students were administered the Standard version of the test, while the remaining 35.7% were administered the Scaffold version of the test.

Grade 7. Approximately 64.7% were male, 61.3% were White, and 24.1% were Hispanic. Approximately 64.4% of all students were administered the Standard version of the test, while the remaining 35.6% were administered the Scaffold version of the test.

Grade 8. Approximately 63.0% were male, 62.0% were White, and 23.0% were Hispanic. Approximately 59.9% of all students were administered the Standard version of the test, while the remaining 40.1% were administered the Scaffold version of the test.

Grade 11. Approximately 64.3% were male, 63.5% were White, and 22.7% were Hispanic. Approximately 55.4% of all students were administered the Standard version of the test, while the remaining 44.6% were administered the Scaffold version of the test.

Science

Grade 5. Approximately 66.9% of students taking the science portion of the Oregon Extended Assessment were male, 57.1% were White, and 26.6% were Hispanic. Approximately 62.2% of all students were administered the Standard version of the test, while the remaining 37.8% were administered the Scaffold version of the test.

Grade 8. Approximately 62.3% of students taking the science portion of the Oregon Extended Assessment were male, 61.9% were White, and 22.2% were Hispanic. Approximately 56.8% of all students were administered the Standard version of the test, while the remaining 43.2% were administered the Scaffold version of the test.

Grade 11. Approximately 64.6% of students taking the science portion of the Oregon Extended Assessment were male, 63.3% were White, and 23.0% were Hispanic. Approximately 54.8% of all students were administered the Standard version of the test, while the remaining 45.2% were administered the Scaffold version of the test.

Reliability

Full reliability statistics for the reading portion of the Oregon Extended Assessment are reported in *Appendix B*. These results demonstrate that the total test reliabilities were quite high, ranging from .88 to .98.

Reading

Elementary. The task reliabilities for the elementary grade-band (3, 4, 5) were moderate to high, ranging from 0.55 for Task 5 to 0.95 for Task 1. The reliability of the total test was quite high, at 0.95.

Reading: Elementary

Task	Cronbach's Alpha
1	0.95
2	0.72
3	0.75
4	0.72
5	0.55
6	0.73
7	0.61
8	0.72
9	0.70
10	0.59
11	0.72
Total Test	0.94

Middle. The task reliabilities for the middle school grade band (grades 6, 7, and 8) were moderate to high, ranging from 0.66 for Task 6 to 0.96 for Task 1. The reliability of the total test was quite high, at 0.96

Reading: Middle

Task	Cronbach's Alpha
1	0.96
2	0.70
3	0.74
4	0.59
5	0.79
6	0.66
7	0.67
8	0.74
9	0.77
10	0.79
11	0.72
Total Test	0.96

High. The task reliabilities for the high school grade band (grade 11) were moderate to high, ranging from 0.66 for Task 7 to 0.97 for Task 1. The reliability of the total test was quite high, at 0.97.

Reading: High

Task	Cronbach's Alpha
1	0.97
2	0.81
3	0.69
4	0.85
5	0.79
6	0.77
7	0.66
8	0.75
9	0.85
10	0.77
11	0.79
Total Test	0.97

Writing

High. Task reliabilities were moderately high to high, ranging from 0.76 for Task 10 to 0.95 for Task 1. The reliability of the total test was quite high, at 0.98.

Writing: High

Task	Cronbach's Alpha
1	0.95
2	0.87
3	0.88
4	0.80
5	0.87
6	0.86
7	0.88
8	0.84
9	0.79
10	0.76
11	0.91
Total Test	0.98

Math

Grade 3. Task reliabilities were moderately low to high, ranging from 0.47 for Task 4 to 0.96 for Task 1. The reliability of the total test was quite high, at 0.91.

Math: Grade 3

Task	Cronbach's Alpha
1	0.96
2	0.74
3	0.55
4	0.47
5	0.57
6	0.58
7	0.48
8	0.63
9	0.67
Total Test	0.91

Grade 4. Task reliabilities were low to high, ranging from 0.40 for Task 8 to 0.96 for Task 1. The reliability of the total test was high, at 0.89.

Math Grade 4

Task	Cronbach's Alpha
1	0.96
2	0.67
3	0.64
4	0.41
5	0.56
6	0.54
7	0.61
8	0.40
9	0.47
Total Test	0.89

Grade 5. Task reliabilities were low to high, ranging from 0.19 for Task 7 to 0.95 for Task 1. The reliability of the total test was high, at 0.88.

Math: Grade 5

Task	Cronbach's Alpha
1	0.95
2	0.50
3	0.55
4	0.54
5	0.57
6	0.46
7	0.19
8	0.37
9	0.60
Total Test	0.88

Grade 6. Task reliabilities were low to high, ranging from 0.41 for Task 8 to 0.96 for Task 1. The reliability of the total test was high, at 0.89.

Math: Grade 6

Task	Cronbach's Alpha
1	0.96
2	0.50
3	0.66
4	0.54
5	0.51
6	0.51
7	0.45
8	0.41
9	0.61
Total Test	0.89

Grade 7. Task reliabilities were moderate to high, ranging from 0.46 for Task 9 to 0.97 for Task 1. The reliability of the total test was quite high, at 0.91.

Math: Grade 7

Task	Cronbach's Alpha
1	0.97
2	0.67
3	0.49
4	0.60
5	0.61
6	0.58
7	0.56
8	0.47
9	0.46
Total Test	0.91

Grade 8. Task reliabilities were low to high, ranging from 0.32 for Task 3 to 0.97 for Task 1. The reliability of the total test was quite high, at 0.90.

Math: Grade 8

Task	Cronbach's Alpha
1	0.97
2	0.54
3	0.32
4	0.48
5	0.66
6	0.57
7	0.69
8	0.42
9	0.41
Total Test	0.90

Grade 11. Task reliabilities were moderate to high, ranging from 0.47 for Task 5 to 0.97 for Task 1. The reliability of the total test was quite high, at 0.93.

Math: Grade 11

Task	Cronbach's Alpha
1	0.97
2	0.61
3	0.55
4	0.61
5	0.47
6	0.70
7	0.56
8	0.81
9	0.71
Total Test	0.93

Science

Grade 5. Task reliabilities were moderate to high, ranging from 0.61 for Task 8 to 0.97 for Task 1. The reliability of the total test was quite high, at 0.94.

Science: Grade 5

Task	Cronbach's Alpha
1	0.97
2	0.73
3	0.67
4	0.64
5	0.73
6	0.75
7	0.63
8	0.61
9	0.64
Total Test	0.94

Grade 8. Task reliabilities were moderately low to high, ranging from 0.48 for Task 4 to 0.98 for Task 1. The reliability of the total test was quite high, at 0.93.

Science: Grade 8

Task	Cronbach's Alpha
1	0.98
2	0.77
3	0.59
4	0.48
5	0.81
6	0.59
7	0.49
8	0.77
9	0.77
Total Test	0.93

Grade 11. Task reliabilities were moderate to high, ranging from 0.66 for Task 7 to 0.98 for Task 1. The reliability of the total test was quite high, at 0.95.

Science: Grade 11

Task	Cronbach's Alpha
1	0.98
2	0.74
3	0.75
4	0.71
5	0.55
6	0.69
7	0.66
8	0.69
9	0.67
Total Test	0.95

Descriptive Statistics

Full descriptive statistics for the reading items of the Oregon Extended Assessment are reported in *Appendix C*. All Task 1 items were scored on a 4-point scale, based upon the level of support needed to bring the student to success (4 = Independent, 3 = Verbal, Gestural, or Visual, 2 = Partial Physical, and 1 pt = Full Physical) . Items on all subsequent tasks were scored on a 2-point scale, where 2 points are awarded for a correct response, 1 point is awarded for a close response (based on scripted scoring expectations or teacher judgment), and 0 points are awarded for an incorrect response. In general, the descriptive statistics suggest that the test has an appropriate range of item difficulties represented, from easy to difficult. Using classical test theory item difficulty indices, the easiest items are located in Task 1, the prerequisite skills items. Item difficulties range from p = .14 (the most difficult item) to p = .98 (the easiest item). Item difficulties are deemed appropriate across all subject areas.

Reading: Elementary

Task 1. Task 1 items were quite easy, with item difficulties (p value) ranging from 0.85 – 0.96. Item 1 had the lowest standard deviation (0.55) and item 10 had the highest (0.85). For Tasks 2-11 all items were scored on a 4-point scale.

Task 2 (field test). Items were relatively easy overall. Item 2 was the most difficult item, p = 0.62, while item 3 were the easiest, p = 0.83. Item 3 had the lowest standard deviation (0.72) while Item 2 had the highest (0.88).

***Task 3 (field test)*.** Items overall were relatively easy. Item 1 was the most difficult item, p = 0.67, while item 4 was the easiest, p = 0.76. Item 4 had the lowest standard deviation (0.80) while Item 5 had the highest (0.91).

***Task 4*.** Item 2 was the most difficult item, p = 0.70, while item 1 was the easiest, p = 0.86. Item 1 had the lowest standard deviation (0.61) while Item 5 had the highest (0.85).

***Task 5*.** Item 2 was the most difficult item, p = 0.56, while item 1 was the easiest, p = 0.82. Item 1 had the lowest standard deviation (0.67) while item 5 had the highest (0.84).

***Task 6*.** Item 2 was the most difficult item, p = 0.69, while item 3 was the easiest, p = 0.90. Item 3 had the lowest standard deviation (0.59) while Item 5 had the highest (0.83).

Task 7. Item 3 was the most difficult item, p = 0.54, while item 1 was the easiest, p = 0.72. Item 5 had the lowest standard deviation (0.75) while item 3 had the highest (0.87).

Task 8. Item 3 was the most difficult item, p = 0.68, while item 5 was the easiest, p = 0.81. Item 3 had the lowest standard deviation (0.66) while Item 1 had the highest (0.89).

Task 9. Item 5 was the most difficult item, p = 0.66, while item 4 was the easiest, p = 0.84. Item 1 had the lowest standard deviation (0.66) while Item 5 had the highest (0.94).

Task 10. Item 5 was the most difficult item, p = 0.51, while item 1 was the easiest, p = 0.69. Item 5 had the lowest standard deviation (0.79) while Item 1 had the highest (0.86).

Task 11. Item 5 was the most difficult item, p = 0.60, while item 1 was the easiest, p = 0.79. Item 5 had the lowest standard deviation (0.66) while Item 3 had the highest (0.74).

Total test. The average total test score was 54.15 with a standard deviation of 19.26. Item difficulties range from p = 0.51 (the most difficult item) to p = 0.96 (the easiest item).

Reading: Middle

Task 1. Task 1 items were quite easy, with item difficulties (*p* value) ranging from 0.86 – 0.96. Item 1 had the lowest standard deviation (0.54) and item 10 had the highest (0.88). For Tasks 2-11 all items were scored on a 4-point scale.

Task 2 (field test). Item 1 was the most difficult item, $p = 0.52$, while item 3 was the easiest, $p = 0.68$. Item 3 had the lowest standard deviation (0.92) while item 1 had the highest (0.99).

Task 3 (field test). Item 3 was the most difficult item, $p = 0.63$, while item 1 was the easiest, $p = 0.77$. Item 1 had the lowest standard deviation (0.76) while Item 3 had the highest (0.87).

Task 4. Item 5 was the most difficult item, $p = 0.52$, while item 1 was the easiest, $p = 0.71$. Items 1 and 3 had the lowest standard deviations (0.74) while item 5 had the highest (1.00).

Task 5. Item 3 was the most difficult item, $p = 0.70$, while item 4 was the easiest, $p = 0.84$. Item 4 had the lowest standard deviation (0.66) while item 2 had the highest (0.75).

Task 6. Item 5 was the most difficult item, $p = 0.61$, while item 3 was the easiest, $p = 0.74$. Item 1 had the lowest standard deviation (0.75) while Item 5 had the highest (0.85).

Task 7. Item 2 was the most difficult item, $p = 0.54$, while item 5 was the easiest, $p = 0.81$. Item 1 had the lowest standard deviation (0.71) while Item 2 had the highest (0.99).

Task 8. Item 5 was the most difficult item, $p = 0.67$, while item 3 was the easiest, $p = 0.88$. Item 3 had the lowest standard deviation (0.61) while Item 5 had the highest (0.78).

Task 9. Item 5 was the most difficult item, $p = 0.72$, while item 1 was the easiest, $p = 0.85$. Item 1 had the lowest standard deviation (0.63) while Item 5 had the highest (0.78).

Task 10. Item 3 was the most difficult item, $p = 0.66$, while item 2 was the easiest, $p = 0.80$. Item 3 had the lowest standard deviation (0.68) while Item 4 had the highest (0.78).

Task 11. Item 2 was the most difficult item, $p = 0.54$, while item 3 was the easiest, $p = 0.79$. Item 2 had the lowest standard deviation (0.62) while Item 4 had the highest (0.77).

Total test. The average total test score was 55.07 with a standard deviation of 20.32. Item difficulties range from $p = 0.52$ (the most difficult item) to $p = 0.96$ (the easiest item).

Reading: High

Task 1. Task 1 items were quite easy, with item difficulties (*p* value) ranging from 0.83– 0.95. Item 1 had the lowest standard deviation (0.75) and item 10 had the highest (1.07). For Tasks 2-11 all items were scored on a 4-point scale.

Task 2 (field test). Item 4 was the most difficult item, $p = 0.63$, while item 2 was the easiest, $p = 0.69$. Item 1 had the lowest standard deviation (0.76) while item 4 had the highest (0.95).

Task 3 (field test). Item 5 was the most difficult item, $p = 0.33$, while item 3 was the easiest, $p = 0.72$. Item 3 had the lowest standard deviation (0.81) while item 2 had the highest (0.86).

Task 4. Item 4 was the most difficult item, $p = 0.66$, while item 3 was the easiest, $p = 0.78$. Item 3 had the lowest standard deviation (0.69) while item 4 had the highest (0.79).

Task 5. Item 5 was the most difficult item, $p = 0.56$, while item 2 was the easiest, $p = 0.81$. Item 3 had the lowest standard deviation (0.71) while item 4 had the highest (0.76).

Task 6. Item 5 was the most difficult item, p = 0.56, while item 4 was the easiest, p = 0.82. Item 4 had the lowest standard deviation (0.68) while item 5 had the highest (0.99).

Task 7. Item 4 was the most difficult item, p = 0.51, while item 1 was the easiest, p = 0.64. Item 1 had the lowest standard deviation (0.73) while item 4 had the highest (0.78).

Task 8. Item 3 was the most difficult item, p = 0.60, while item 1 was the easiest, p = 0.80. Item 5 had the lowest standard deviation (0.74) while item 3 had the highest (0.90).

Task 9. Item 5 was the most difficult item, p = 0.67, while item 1 was the easiest, p = 0.82. Item 4 had the lowest standard deviation (0.67) while item 5 had the highest (0.82).

Task 10. Item 2 was the most difficult item, p = 0.62, while item 4 was the easiest, p = 0.77. Item 1 had the lowest standard deviation (0.68) while item 3 had the highest (0.75).

Task 11. Item 1 was the most difficult item, p = 0.55, while item 3 was the easiest, p = 0.68. Item 1 had the lowest standard deviation (0.66) while Item 4 had the highest (0.71).

Total test. The average total test score was 52.97 with a standard deviation of 21.38. Item difficulties range from p = 0.33 (the most difficult item) to p = 0.95 (the easiest item).

Writing: High School

Task 1. Task 1 items were quite easy, with item difficulties (p value) ranging from 0.88 – 0.95. Item 1 had the lowest standard deviation (0.71) and item 8 had the highest (1.06).

Task 2. Item 1 was the most difficult item, p = 0.59, while item 4 was the easiest, p = 0.68. Item 3 had the lowest standard deviation (0.73) while item 1 had the highest (0.84).

Task 3. Item 2 was the most difficult item, p = 0.52, while item 3 was the easiest, p = 0.65. Item 2 had the lowest standard deviation (0.70) while item 1 had the highest (0.86).

Task 4. Item 5 was the most difficult item, p = 0.63, while item 2 was the easiest, p = 0.78. Item 2 had the lowest standard deviation (0.81) while item 5 had the highest (0.96).

Task 5. Item 5 was the most difficult item, p = 0.70, while item 4 was the easiest, p = 0.84. Item 1 had the lowest standard deviation (0.68) while item 2 had the highest (0.78).

Task 6. Item 5 was the most difficult item, p = 0.54, while item 2 was the easiest, p = 0.76. Item 2 had the lowest standard deviation (0.74) while item 5 had the highest (0.78).

Task 7. Item 5 was the most difficult item, p = 0.65, while item 1 was the easiest, p = 0.84. Item 1 had the lowest standard deviation (0.63) while item 4 had the highest (0.80).

Task 8 (field test). Item 3 was the most difficult item, p = 0.52, while item 1 was the easiest, p = 0.72. Item 4 had the lowest standard deviation (0.72) while item 2 had the highest (0.81).

Task 9 (field test). Item 3 was the most difficult item, p = 0.69, while item 4 was the easiest, p = 0.82. Item 4 had the lowest standard deviation (0.65) while item 2 had the highest (0.69).

Task 10. Item 5 was the most difficult item, p = 0.54, while item 3 was the easiest, p = 0.84. Item 3 had the lowest standard deviation (0.67) while item 5 had the highest (0.89).

Task 11. Item 4 was the most difficult item, p = 0.55, while item 5 was the easiest, p = 0.80. Item 5 had the lowest standard deviation (0.77) while item 2 had the highest (0.86).

Total test. The average total test score was 50.44 with a standard deviation of 25.75. Item difficulties range from p = 0.52 (the most difficult item) to p = 0.95 (the easiest item).

Math: Grade 3

Task 1. Task 1 items were quite easy, with item difficulties (*p* value) ranging from 0.85 – 0.95. Item 1 had the lowest standard deviation (0.67) and item 5 had the highest (1.00). For Tasks 2-9 all items were scored on a 4-point scale.

Task 2. Item 5 was the most difficult item, $p = 0.48$, while item 3 was the easiest, $p = 0.73$. Item 1 had the lowest standard deviation (0.89) while item 5 had the highest (1.00).

Task 3. Item 3 was the most difficult item, $p = 0.31$, while item 6 (field test) was the easiest, $p = 0.65$. Item 3 had the lowest standard deviation (0.92) while item 5 had the highest (0.99).

Task 4. Item 1 was the most difficult item, $p = 0.35$, while item 5 was the easiest, $p = 0.64$. Item 1 had the lowest standard deviation (0.95) while item 4 had the highest (0.98).

Task 5. Item 3 was the most difficult item, $p = 0.37$, while item 6 (field test) was the easiest, $p = 0.77$. Item 6 (field test) had the lowest standard deviation (0.83) while item 1 had the highest (1.00).

Task 6. Item 6 (field test) was the most difficult item, $p = 0.28$, while item 4 was the easiest, $p = 0.67$. Item 4 had the lowest standard deviation (0.94) while item 2 had the highest (1.00).

Task 7. Item 1 was the most difficult item, $p = 0.14$, while item 5 was the easiest, $p = 0.75$. Item 1 had the lowest standard deviation (0.69) while item 6 (field test) had the highest (1.00).

Task 8. Item 4 was the most difficult item, $p = 0.51$, while item 6 (field test) was the easiest, $p = 0.71$. Item 6 (field test) had the lowest standard deviation (0.91) while item 4 had the highest (1.00).

Task 9. Item 5 was the most difficult item, $p = 0.43$, while item 1 was the easiest, $p = 0.88$. Item 1 had the lowest standard deviation (0.64) while item 5 had the highest (1.00).

Total test. The average total test score was 39.07 with a standard deviation of 20.08. Item difficulties range from $p = 0.14$ (the most difficult item) to $p = 0.95$ (the easiest item).

Math: Grade 4

Task 1. Task 1 items were quite easy, with item difficulties (*p* value) ranging from 0.88 – 0.97. Item 1 had the lowest standard deviation (0.50), while item 5 had the highest standard deviation (0.89).

Task 2. Item 6 (field test) was the most difficult item, $p = 0.41$, while item 4 was the easiest, $p = 0.81$. Item 4 had the lowest standard deviation (0.78) while item 3 had the highest (1.00).

Task 3. Item 6 (field test) was the most difficult item, $p = 0.41$, while item 3 was the easiest, $p = 0.81$. Item 3 had the lowest standard deviation (0.78) while item 5 had the highest (1.00).

Task 4. Item 4 was the most difficult item, $p = 0.23$, while item 6 (field test) was the easiest, $p = 0.72$. Item 4 had the lowest standard deviation (0.89) while item 3 had the highest (1.00).

Task 5. Item 3 was the most difficult item, $p = 0.50$, while item 4 was the easiest, $p = 0.73$. Item 6 (field test) had the lowest standard deviation (0.89) while item 1 had the highest (1.00).

Task 6. Item 3 was the most difficult item, p = 0.43, while item 5 was the easiest, p = 0.69. Item 5 had the lowest standard deviation (0.93) while item 4 had the highest (1.00).

Task 7. Item 4 was the most difficult item, p = 0.45, while item 2 was the easiest, p = 0.84. Item 2 had the lowest standard deviation (0.74) while item 5 had the highest (1.00).

Task 8. Item 2 was the most difficult item, p = 0.29, while item 5 was the easiest, p = 0.64. Item 2 had the lowest standard deviation (0.90) while item 1 had the highest (1.00).

Task 9. Item 1 was the most difficult item, p = 0.29, while item 3 was the easiest, p = 0.75. Item 1 had the lowest standard deviation (0.91) while item 4 had the highest (1.00).

Total test. The average total test score was 41.73 with a standard deviation of 19.17. Item difficulties range from p = 0.23 (the most difficult item) to p = 0.97 (the easiest item).

Math: Grade 5

Task 1. Task 1 items were quite easy, with item difficulties (p value) ranging from 0.90 – 0.98. Item 1 had the lowest standard deviation (0.44), while item 5 had the highest standard deviation (0.88).

Task 2. Item 4 was the most difficult item, p = 0.36, while item 3 was the easiest, p = 0.68. Item 3 had the lowest standard deviation (0.93) while item 1 had the highest (1.00).

Task 3. Item 1 was the most difficult item, p = 0.32, while item 3 was the easiest, p = 0.60. Item 1 had the lowest standard deviation (0.93) while item 6 (field test) had the highest (1.00).

Task 4. Item 3 was the most difficult item, p = 0.30, while item 2 was the easiest, p = 0.74. Item 2 had the lowest standard deviation (0.88) while item 4 had the highest (0.97).

Task 5. Item 4 was the most difficult item, p = 0.34, while item 2 was the easiest, p = 0.77. Item 2 had the lowest standard deviation (0.85) while item 3 had the highest (1.00).

Task 6. Item 6 (field test) was the most difficult item, p = 0.29, while item 3 was the easiest, p = 0.81. Item 3 had the lowest standard deviation (0.79) while item 2 had the highest (.97).

Task 7. Item 3 was the most difficult item, p = 0.28, while item 4 was the easiest, p = 0.58. Item 3 had the lowest standard deviation (0.90) while item 1 had the highest (1.00).

Task 8. Item 3 was the most difficult item, p = 0.27, while item 1 was the easiest, p = 0.80. Item 1 had the lowest standard deviation (0.80) while item 4 had the highest (0.99).

Task 9. Item 2 was the most difficult item, p = 0.35, while item 1 was the easiest, p = 0.85. Item 1 had the lowest standard deviation (0.72) while item 3 had the highest (0.99).

Total test. The average total test score was 37.98 with a standard deviation of 18.04. Item difficulties range from p = 0.27 (the most difficult item) to p = 0.98 (the easiest item).

Math: Grade 6

Task 1. Task 1 items were quite easy, with item difficulties (p value) ranging from 0.90 – 0.96. Item 1 had the lowest standard deviation (0.53) while item 5 had the highest standard deviation (0.89).

Task 2. Item 3 was the most difficult item, p = 0.39, while item 2 was the easiest, p = 0.55. Item 3 had the lowest standard deviation (0.98) while item 5 had the highest (1.00).

Task 3. Item 4 was the most difficult item, $p = 0.39$, while item 3 was the easiest, $p = 0.66$. Item 3 had the lowest standard deviation (0.95) while item 6 (field test) had the highest (0.99).

Task 4. Item 4 was the most difficult item, $p = 0.31$, while item 5 was the easiest, $p = 0.76$. Item 5 had the lowest standard deviation (0.85) while item 1 had the highest (1.00).

Task 5. Item 4 was the most difficult item, $p = 0.33$, while item 5 was the easiest, $p = 0.50$. Item 4 had the lowest standard deviation (0.94) while items 5 and 6 (field test) had the highest (1.00).

Task 6. Item 1 was the most difficult item, $p = 0.30$, while item 4 was the easiest, $p = 0.65$. Item 1 had the lowest standard deviation (0.92) while item 6 (field test) had the highest (1.00).

Task 7. Item 3 was the most difficult item, $p = 0.31$, while item 5 was the easiest, $p = 0.60$. Item 3 had the lowest standard deviation (0.92) while item 4 had the highest (1.00).

Task 8. Item 2 was the most difficult item, $p = 0.32$, while item 6 (field test) was the easiest, $p = 0.73$. Item 6 (field test) had the lowest standard deviation (0.89) while item 5 had the highest (1.00).

Task 9. Item 5 was the most difficult item, $p = 0.36$, while item 4 was the easiest, $p = 0.69$. Item 4 had the lowest standard deviation (0.92) while item 1 had the highest (1.00).

Total test. The average total test score was 35.18 with a standard deviation of 18.33. Item difficulties range from $p = 0.30$ (the most difficult item) to $p = 0.96$ (the easiest item).

Math: Grade 7

Task 1. Task 1 items were quite easy, with item difficulties (p value) ranging from 0.90 – 0.96. Item 1 had the lowest standard deviation (0.54), while item 10 had the highest standard deviation (0.82).

Task 2. Item 6 (field test) was the most difficult item, $p = 0.50$, while item 5 was the easiest, $p = 0.66$. Item 5 had the lowest standard deviation (0.95) while items 2 and 6 (field test) had the highest (1.00).

Task 3. Item 3 was the most difficult item, $p = 0.31$, while item 2 was the easiest, $p = 0.60$. Item 3 had the lowest standard deviation (0.92) while item 5 had the highest (1.00).

Task 4. Item 1 was the most difficult item, $p = 0.41$, while item 3 was the easiest, $p = 0.74$. Item 3 had the lowest standard deviation (0.88) while item 4 had the highest (1.00).

Task 5. Item 3 was the most difficult item, $p = 0.35$, while item 2 was the easiest, $p = 0.62$. Item 2 had the lowest standard deviation (0.97) while items 4 and 5 had the highest (1.00).

Task 6. Item 4 was the most difficult item, $p = 0.35$, while item 5 was the easiest, $p = 0.76$. Item 2 had the lowest standard deviation (0.87) while item 1 had the highest (1.00).

Task 7. Item 2 was the most difficult item, $p = 0.25$, while item 1 was the easiest, $p = 0.62$. Item 2 had the lowest standard deviation (0.87) while item 5 had the highest (1.00).

Task 8. Item 6 (field test) was the most difficult item, $p = 0.34$, while item 3 was the easiest, $p = 0.78$. Item 3 had the lowest standard deviation (0.82) while item 5 had the highest (1.00).

Task 9. Item 4 was the most difficult item, $p = 0.31$, while item 1 was the easiest, $p = 0.65$. Item 4 had the lowest standard deviation (0.92) while item 2 had the highest (1.00).

Total test. The average total test score was 36.83 with a standard deviation of 20.15. Item difficulties range from $p = 0.25$ (the most difficult item) to $p = 0.96$ (the easiest item).

Math: Grade 8

Task 1. Task 1 items were quite easy, with item difficulties (p value) ranging from 0.90 – 0.96. Item 1 had the lowest standard deviation (0.60), while item 5 had the highest standard deviation (0.86).

Task 2. Item 3 was the most difficult item, $p = 0.34$, while item 1 was the easiest, $p = 0.70$. Item 1 had the lowest standard deviation (0.91) while item 2 had the highest (1.00).

Task 3. Item 6 (field test) was the most difficult item, $p = 0.25$, while item 2 was the easiest, $p = 0.46$. Item 5 had the lowest standard deviation (0.86) while item 4 had the highest (1.00).

Task 4. Item 5 was the most difficult item, $p = 0.22$, while item 4 was the easiest, $p = 0.74$. Item 5 had the lowest standard deviation (0.83) while item 6 (field test) had the highest (1.00).

Task 5. Item 2 was the most difficult item, $p = 0.37$, while item 1 was the easiest, $p = 0.71$. Item 1 had the lowest standard deviation (0.91) while item 5 had the highest (1.00).

Task 6. Item 1 was the most difficult item, $p = 0.26$, while item 3 was the easiest, $p = 0.76$. Item 3 had the lowest standard deviation (0.85) while item 5 had the highest (1.00).

Task 7. Item 4 was the most difficult items, $p = 0.32$, while item 6 (field test) was the easiest, $p = 0.75$. Item 6 (field test) had the lowest standard deviation (0.86) while item 2 had the highest (0.98).

Task 8. Item 5 was the most difficult item, $p = 0.29$, while item 3 was the easiest, $p = 0.75$. Item 3 had the lowest standard deviation (0.86) while item 6 (field test) had the highest (1.00).

Task 9. Item 3 was the most difficult item, $p = 0.20$, while item 6 (field test) was the easiest, $p = 0.71$. Item 3 had the lowest standard deviation (0.80) while item 5 had the highest (0.99).

Total test. The average total test score was 35.03 with a standard deviation of 18.00. Item difficulties range from $p = 0.20$ (the most difficult item) to $p = 0.96$ (the easiest item).

Math: Grade 11

Task 1. Task 1 items were quite easy, with item difficulties (p value) ranging from 0.87– 0.95. Item 3 had the lowest standard deviation (0.84), while item 5 had the highest standard deviation (1.00).

Task 2. Item 3 was the most difficult item, $p = 0.35$, while item 6 (field test) was the easiest, $p = 0.66$. Item 3 had the lowest standard deviation (0.94) while item 5 had the highest (1.00).

Task 3. Item 2 were the most difficult items, $p = 0.25$, while item 6 (field test) was the easiest, $p = 0.69$. Item 2 had the lowest standard deviation (0.86) while item 3 had the highest (1.00).

Task 4. Item 6 (field test) was the most difficult items, p = 0.20, while item 4 was the easiest, p = 0.70. Item 6 (field test) had the lowest standard deviation (0.79) while item 1 had the highest (1.00).

Task 5. Item 3 was the most difficult item, p = 0.26, while item 6 (field test) was the easiest, p = 0.63. Item 3 had the lowest standard deviation (0.87) while item 2 had the highest (0.98).

Task 6. Item 1was the most difficult item, p = 0.38, while item 5 was the easiest, p = 0.75. Item 5 had the lowest standard deviation (0.80) while item 3 had the highest (0.99).

Task 7. Item 2 was the most difficult items, p = 0.23, while item 5 was the easiest, p = 0.50. Item 2 had the lowest standard deviation (0.84) while item 4 and 5 had the highest (1.00).

Task 8. Item 1 was the most difficult item, p = 0.59, while item 5 was the easiest, p = 0.79. Item 5 had the lowest standard deviation (0.82) while item 2 had the highest (0.94).

Task 9. Item 3 was the most difficult item, p = 0.50, while item 2 was the easiest, p = 0.83. Item 2 had the lowest standard deviation (0.74) while item 3 had the highest (1.00).

Total test. The average total test score was 35.76 with a standard deviation of 20.37. Item difficulties range from p = 0.20 (the most difficult item) to p = 0.95 (the easiest item).

Science: Grade 5

Task 1. Task 1 items were quite easy, with item difficulties (p value) ranging from 0.93 – 0.96. Item 1 had the lowest standard deviation (0.55), while item 9 had the highest standard deviation (0.75).

Task 2. Item 6 (field test) was the most difficult item, p = 0.53, while item 1 was the easiest, p = 0.88. Item 1 had the lowest standard deviation (0.64) while item 6 (field test) had the highest (1.00).

Task 3. Item 4 was the most difficult item, p = 0.43, while item 1 was the easiest, p = 0.86. Item 1 had the lowest standard deviation (0.69) while item 4 had the highest (0.99).

Task 4. Item 3was the most difficult item, p = 0.88, while item 6 (field test) was the easiest, p = 0.56. Item 3 had the lowest standard deviation (0.66) while item 6 (field test) was the highest (0.99).

Task 5. Item 6 (field test) was the most difficult item, p = 0.43, while item 5 was the easiest, p = 0.87. Item 5 had the lowest standard deviation (0.66) while item 6 (field test) had the highest (0.99).

Task 6. Item 3 was the most difficult item, p = 0.71, while item 4 was the easiest, p = 0.85. Item 4 had the lowest standard deviation (0.72) while item 3 had the highest (0.91).

Task 7. Item 5 was the most difficult item, p = 0.48, while item 1 was the easiest, p = 0.89. Item 1 had the lowest standard deviation (0.63) while items 5 and 6 (field test) had the highest (1.00).

Task 8. Item 1 was the most difficult item, p = 0.47, while item 5 was the easiest, p = 0.75. Item 5 had the lowest standard deviation (0.86) while item 4 had the highest (1.00).

Task 9. Item 4 was the most difficult item, p = 0.41, while item 2 was the easiest, p = 0.81. Item 1 had the lowest standard deviation (0.80) while item 6 (field test) had the highest (0.99).

Total test. The average total test score for the operational items was 53.67 with a standard deviation of 22.92. Item difficulties range from $p = 0.41$ (the most difficult item) to $p = 0.96$ (the easiest item).

Science: Grade 8

Task 1. Task 1 items were quite easy, with item difficulties (p value) ranging from 0.93 – 0.96. Item 1 had the lowest standard deviation (0.59), while item 10 had the highest standard deviation (0.75).

Task 2. Item 6 (field test) was the most difficult item, $p = 0.63$, while item 2 was the easiest, $p = 0.80$. Item 2 had the lowest standard deviation (0.81) while item 6 (field test) had the highest (0.96).

Task 3. Item 6 (field test) was the most difficult item, $p = 0.27$, while item 3 was the easiest, $p = 0.68$. Item 3 had the lowest standard deviation (0.93) while item 5 had the highest (1.00).

Task 4. Item 3 was the most difficult item, $p = 0.28$, while item 6 (field test) was the easiest, $p = 0.64$. Item 3 had the lowest standard deviation (0.90) while item 4 had the highest (1.00).

Task 5. Item 1 was the most difficult item, $p = 0.69$, while item 5 was the easiest, $p = 0.86$. Item 5 had the lowest standard deviation (0.69) while 1 had the highest (0.92).

Task 6. Item 6 (field test) was the most difficult item, $p = 0.42$, while item 1 was the easiest, $p = 0.73$. Item 1 had the lowest standard deviation (0.89) while item 6 (field test) had the highest (0.99).

Task 7. Item 6 (field test) was the most difficult item, $p = 0.27$, while item 1 was the easiest, $p = 0.85$. Item 1 had the lowest standard deviation (0.71) while item 2 had the highest (0.98).

Task 8. Item 1 was the most difficult item, $p = 0.55$, while item 5 was the easiest, $p = 0.81$. Item 5 had the lowest standard deviation (0.79) while item 1 had the highest (1.00).

Task 9. Item 3 was the most difficult items, $p = 0.55$, while items 4 and 6 (field test) were the easiest, $p = 0.75$. Item 6 (field test) had the lowest standard deviation (0.87) while item 3 had the highest (0.99).

Total test. The average total test score for the operational items was 47.98 with a standard deviation of 22.45. Item difficulties range from $p = 0.27$ (the most difficult item) to $p = 0.96$ (the easiest item).

Science: High School

Task 1. Task 1 items were quite easy, with item difficulties (p value) ranging from 0.91 – 0.95. Item 1 had the lowest standard deviation (0.73), while item 9 had the highest standard deviation (0.92).

Task 2. Item 6 (field test) was the most difficult item, $p = 0.49$, while item 5 was the easiest, $p = 0.75$. Item 5 had the lowest standard deviation (0.86) while item 3 had the highest (1.00).

Task 3. Item 2 was the most difficult item, $p = 0.53$, while item 1 was the easiest, $p = 0.82$. Item 1 had the lowest standard deviation (0.77) while item 2 had the highest (0.99).

Task 4. Item 6 (field test) was the most difficult item, $p = 0.58$, while item 4 was the easiest, $p = 0.80$. Item 4 had the lowest standard deviation (0.79) while item 1 had the highest (0.98).

Task 5. Item 6 (field test) was the most difficult item, $p = 0.23$, while item 4 was the easiest, $p = 0.85$. Item 4 had the lowest standard deviation (0.72) while item 5 had the highest (1.00).

Task 6. Item 6 (field test) was the most difficult item, $p = 0.52$, while item 4 was the easiest, $p = 0.78$. Item 4 had the lowest standard deviation (0.82) while item 6 (field test) had the highest (1.00).

Task 7. Item 2 was the most difficult item, $p = 0.37$, while item 1 was the easiest, $p = 0.72$. Item 1 had the lowest standard deviation (0.89) while item 5 had the highest (0.98).

Task 8. Item 6 (field test) was the most difficult item, $p = 0.47$, while item 4 was the easiest, $p = 0.84$. Item 4 had the lowest standard deviation (0.72) while item 6 (field test) had the highest (1.00).

Task 9. Item 5 was the most difficult item, $p = 0.42$, while item 3 was the easiest, $p = 0.89$. Item 5 had the lowest standard deviation (0.75) while item 1 had the highest (1.00).

Total test. The average total test score was 47.48 with a standard deviation of 23.68. Item difficulties range from $p = 0.23$ (the most difficult item) to $p = 0.95$ (the easiest item).

Analyses Within and Across Subject Areas

We conducted one correlational analysis and a series of four regression models to explore the validity of the OREXt. In this section, we describe the purpose of each analysis, as well as our anticipated results. We then discuss our observed results before concluding with an overall evaluative judgment of the validity of the test. Each regression model is briefly introduced below, and discussed in more depth later in the report.

In **Correlation Analysis 1**, we explore the correlations among students' total scores across subject areas. The purpose of the analysis was to investigate how strongly students' scores in one area "went along with" students' scores in other subject areas. If the correlations were exceedingly high (e.g., above .90), it would indicate that the score a student receives in an individual subject has less to do with the intended construct (i.e., reading) than with factors idiosyncratic to the student. For example, if all subject areas correlated at .95, then it would provide strong evidence that the tests would be measuring a global student-specific construct (i.e., intelligence), and not the individual subject constructs. We would expect, however, that the tests would correlate quite strongly given that the same students were assessed multiple times. Therefore, we would expect moderately strong correlations (e.g., 0.7) simply because of the within-subject design. Idiosyncratic variance associated with the individual student is thus captured.

Regression models

Four regression models were run to examine the functioning of the Oregon Alternate Assessment. Each model was run by grade-level for each subject, with the exception of reading, which was conducted by grade-band. These analyses provide information supporting the validity of inferences as a function of performance in a content subject area rather than pre-requisite skills, administration type, disability categories, or race/ethnicity.

Regression Models

Model	Predictors	Dependent Variable
1	Pre-requisite task total	Total Scale Score
2	Administration type Pre-requisite task total	Total Scale Score
3	Disability category Administration type Race/Ethnicity	Prerequisite Total Score
4	Disability category Administration type Race/Ethnicity	Total Scale Score

In regression **Model 1**, we test the extent to which the pre-requisite skills task moderates students' total test score. In other words, did students scoring high on the pre-requisite skills task generally score high on the content tasks? The pre-requisite skills task assesses students' level of independence, while the total scale score assesses students' content knowledge. A strong relation between the pre-requisite skills task and the content tasks would indicate that students' level of independence plays a large role in the content score they receive. Similarly, a low relation would indicate that students level of independence has very little to do with the score they receive. It is important to note that the score the student received on the pre-requisite skills task also determines the level of support the student receives on the content tasks. Thus, we would expect the relation between the pre-requisite skills task total and the content task total to be quite low given that: (a) the tasks assess distinctly different constructs, and (b) students with lower levels of independence were supported during the content task administration to reduce the effect of any impeding factors that would preclude them from demonstrating their content knowledge. The full results are described on pages 61 - 62. Overall, the model accounted for between 34% - 58% of the total variance across subjects and grades.

In regression **Model 2**, we test the influence of the type of test students' were administered (scaffold versus standard) on their total test score, while controlling for their pre-requisite skills total. Students taking the standard administration of the test were entered as the referent group. The scaffold administration has built in supports not available in the standard administration (i.e., auditory prompts by the Assessor). The extra supports are intended to minimize the effect of factors that would preclude students from demonstrating their content knowledge. However, it is also important to note that students taking the scaffold version of the test are generally lower performing students compared to those taking the standard version of the test. The type of administration a student receives is determined prior to the student taking the test by the student's IEP team. Thus, although the scaffold version helps students access the test and display their content knowledge, the observed effects cannot be attributed fully to the differences in test design. Rather, the observed effect represents the combined effects of the test design differences and the student group differences. Model 2 provides an indication of the magnitude of the differences between groups in terms of performance. The observed effects are generally quite large. Because students receiving the scaffold administration receive additional support, we would logically predict that the test would be easier than the standard administration. However, when inspecting the unstandardized regression weights (with standard administration as the referent group) it is apparent that students receiving the scaffold administration scored lower than students receiving the standard administration. Thus, the observed differences are likely due more to the student groups taking each version of the test than to the test itself. The full results are described on pages 63 - 67. Overall, the full model accounted for between 25% - 55% of the total variance across subjects and grades.

In regression **Model 3**, we conducted a sequential regression model to examine the influence of students' disability type, test administration type, and race/ethnicity on their pre-requisite skills total. Administration type was entered primarily as a control variable, but was also used to examine the proportion of students with each disability type in each

administration type. Reference groups included students who were classified with an intellectual disability, took the standard administration of the test, and were White. Each reference group was chosen based on the subgroup with the largest proportion of students. It was necessary to control for the variance associated with different administration types (scaffold versus standard), given that different student groups are represented in each (see results of Model 2). Holding administration constant, an examination of how students performed on the pre-requisite skills by the type of disability and race/ethnicity was performed. Hypothetically, the student's disability should play a role in the student's prerequisite skills score, given that the task is intended to assess a student's level of independence. Different disability types could then logically be associated with different levels of independence. However, all students taking the assessment also have a significant cognitive disability and we would therefore not expect disability to play a substantial role. Ideally, a student's race/ethnicity would have essentially nothing to do with the score the student received on any portion of the test, including the prerequisite skills. The full results are described on pages 68 - 71. Overall, students' disability classification accounted for between 13% - 21% of the total variance across subjects and grades. Test administration type accounted for additional variance beyond students' disability (7% - 16%). Students' race/ethnicity accounted for minimal variance when added in the third block (0% - 1%), and was generally not a statistically significant addition.

In regression **Model 4**, we conducted the same analysis as Model 3, but used the variables as a predictor of students' total scale scores, instead of pre-requisite skills task. Again, we would not expect race/ethnicity variables to substantially influence the observed results. The full results for Model 4 are described on pages 72 - 75. Overall, students' disability classification accounted for between 14% to 22% of the total variance across subjects and grades. Test administration type accounted for additional variance beyond students' disability (13% to 21%). Students' race/ethnicity accounted for minimal variance when added in the third block (0% - 2%), and was generally not a statistically significant addition.

Regression Procedures. Simultaneous regression was used for Model 2. Predictor variables were examined relative to the variables' regression weights (b) and unique contribution to the regression equation (semi-partial correlations).

Sequential regression was used for Models 3 and 4, with disability category entered into the first block, test administration type into second block, and race/ethnicity into the third block. Predictor variables were again examined relative to the variables' regression weights and unique contribution to the regression equation. However, blocking variables into steps also allowed for an evaluation of the change in overall model fit between sets of variables.

Assumptions. The central limit theorem protects regression analyses from departures of normality as long as the sample size is reasonably large, so verifications of normal distribution were not conducted.

Correlational Analyses Results

Full results of the correlation analysis are reported in *Appendix D*. At grade 3, reading and math had a moderately strong correlation, $r(881) = .86$, $p < .01$. At grade 4, the correlation was of a similar magnitude $r(851) = .83$, $p < .01$. At grade 5, reading and math had a moderately strong correlation, $r(858) = .82$, $p < .01$. The correlation between reading and science ($n = 735$) and math and science ($n = 737$) were moderately strong, ranging in the .80s. At grade 6, reading and math had a moderately strong correlation, $r(786) = .80$, $p < .01$. At grade 7, the correlation was of a similar magnitude $r(697) = .82$, $p < .01$. At grade 8, correlation between reading and math ($n = 611$), reading and science ($n = 576$), and math and science ($n = 612$) were moderately strong and statistically significant, with Pearson's *r* in the .70s to 80s. Finally, at grade 11, reading was statistically correlated to writing ($n = 493$), math ($n = 490$), and science ($n = 484$), with Pearson's *r* in the .80s and .90s. Math was statistically correlated to writing ($n = 495$) and science ($n = 489$), with Pearson's *r* in the .80s. Writing and science had a moderately strong correlation, $r(486) = .87$, $p < .01$.

Model 1 Results: Pre-req on Scale Scores

The full regression model, including correlations and descriptive statistics, are reported in *Appendix E*.

Reading

Elementary: The regression of scale score on pre-requisite skills was statistically significant, $F(1, 3249) = 3{,}605.34$, MSR = 141.161, $p < .01$, $R^2 = 0.53$. Pre-requisite skills task total was a statistically significant predictor of students' scale score, $b = 1.99$, SE = .03, $p < .02$, 95% CI = 1.96 to 2.02 On average, every one-point increase in the pre-requisite skills task total corresponded with a 1.99 increase in students' scale scores.

Middle School: The regression of scale score on pre-requisite skills was statistically significant, $F(1, 2268) = 2082.66$, MSR =243.01, $p < .01$, $R^2 = 0.48$. Pre-requisite skills task total was a statistically significant predictor of students' scale score, $b = 2.26$, SE = .05, $p < .01$, 95% CI = 2.21 to 2.31. On average, every one-point increase in the pre-requisite skills task total corresponded with a 2.26 increase in students' scale scores.

High School: The regression of scale score on pre-requisite skills was statistically significant, $F(1, 495) = 678.09$, MSR = 283.90, $p < .01$, $R^2 = 0.58$. Pre-requisite skills task total was a statistically significant predictor of students' scale score, $b = 2.21$, SE = .09, $p < .01$, 95% CI = 2.12 to 2.29. On average, every one-point increase in the pre-requisite skills task total corresponded with a 2.21 increase in students' scale scores.

Writing

Grade 11: The regression of scale score on pre-requisite skills was statistically significant, $F(1, 504) = 696.86$, MSR = 231.01, $p < .01$, $R^2 = 0.58$. Pre-requisite skills task total was a statistically significant predictor of students' scale score, $b = 2.05$, SE = .08, $p < .01$, 95% CI = 1.97 to 2.12. On average, every one-point increase in the pre-requisite skills task total corresponded with a 2.05 increase in students' scale scores.

Math

Grade 3: The regression of scale score on pre-requisite skills was statistically significant, $F(1, 926) = 1235.02$, MSR = 44.59, $p < .01$, $R^2 = 0.57$. Pre-requisite skills task total was a statistically significant predictor of students' scale score, $b = 1.05$, SE = .03, $p < .01$, 95% CI = 1.02 to 1.08. On average, every one-point increase in the pre-requisite skills task total corresponded with a 1.05 increase in students' scale scores.

Grade 4: The regression of scale score on pre-requisite skills was statistically significant, $F(1, 937) = 1{,}011.57$, MSR = 55.68, $p < .01$, $R^2 = 0.52$. Pre-requisite skills task total was a statistically significant predictor of students' scale score, $b = 1.19$, SE = .04, $p < .01$, 95% CI = 1.15 to 1.23. On average, every one-point increase in the pre-requisite skills task total corresponded with a 1.19 increase in students' scale scores.

Grade 5: The regression of scale score on pre-requisite skills was statistically significant, $F(1, 949) = 847.76$, MSR = 46.71, $p < .01$, $R^2 = 0.47$. Pre-requisite skills task total was a

statistically significant predictor of students' scale score, b = 1.09, SE = .04, p < .01, 95% CI = 1.05 to 1.12. On average, every one-point increase in the pre-requisite skills task total corresponded with a 1.09 increase in students' scale scores.

Grade 6: The regression of scale score on pre-requisite skills was statistically significant, $F(1, 875) = 804.80$, MSR = 46.70, $p < .01$, $R^2 = 0.48$. Pre-requisite skills task total was a statistically significant predictor of students' scale score, b = 1.02, SE = .04, p < .01, 95% CI = 0.98 to 1.06. On average, every one-point increase in the pre-requisite skills task total corresponded with a 1.02 increase in students' scale scores.

Grade 7: The regression of scale score on pre-requisite skills was statistically significant, $F(1,782) = 530.73$, MSR = 54.87, $p < .01$, $R^2 = 0.41$. Pre-requisite skills task total was a statistically significant predictor of students' scale score, b = .97, SE = .04, p < .01, 95% CI = 0.92 to 1.01. On average, every one-point increase in the pre-requisite skills task total corresponded with a .97 increase in students' scale scores.

Grade 8: The regression of scale score on pre-requisite skills was statistically significant, $F(1, 723) = 366.77$, MSR = 60.13, $p < .01$, $R^2 = 0.34$. Pre-requisite skills task total was a statistically significant predictor of students' scale score, b = .83, SE = .04, p < .01, 95% CI = .79 to .87. On average, every one-point increase in the pre-requisite skills task total corresponded with a .83 increase in students' scale scores.

Grade 11: The regression of scale score on pre-requisite skills was statistically significant, $F(1, 509) = 606.50$, MSR = 76.33, $p < .01$, $R^2 = 0.54$. Pre-requisite skills task total was a statistically significant predictor of students' scale score, b = 1.08, SE = .05, p < .01, 95% CI = 1.06 to 1.15. On average, every one-point increase in the pre-requisite skills task total corresponded with a 1.08 increase in students' scale scores.

Science

Grade 5: The regression of scale score on pre-requisite skills was statistically significant, $F(1, 752) = 660.60$, MSR = 76.77, $p < .01$, $R^2 = 0.47$. Pre-requisite skills task total was a statistically significant predictor of students' scale score, b = 1.27, SE = .05, p < .01, 95% CI = 1.22 to 1.32. On average, every one-point increase in the pre-requisite skills task total corresponded with a 1.27 increase in students' scale scores.

Grade 8: The regression of scale score on pre-requisite skills was statistically significant, $F(1, 627) = 443.81$, MSR = 65.97, $p < .01$, $R^2 = 0.42$. Pre-requisite skills task total was a statistically significant predictor of students' scale score, b = 0.98, SE = .05, p < .01, 95% CI = 0.93 to 1.03. On average, every one-point increase in the pre-requisite skills task total corresponded with a 0.98 increase in students' scale scores.

Grade 11: The regression of scale score on pre-requisite skills was statistically significant, $F(1, 494) = 557.72$, MSR = 84.18, $p < .01$, $R^2 = 0.53$. Pre-requisite skills task total was a statistically significant predictor of students' scale score, b = 1.10, SE = .05, p < .01, 95% CI = 1.15 to 1.06. On average, every one-point increase in the pre-requisite skills task total corresponded with a 1.10 increase in students' scale scores.

Model 2 Results (Simultaneous): Admin Type and Pre-req on Scale Scores

The full regression model, including correlations and descriptive statistics, are reported in *Appendix F*.

Reading

Elementary: The regression of scale score on pre-requisite skills and test administration type was statistically significant, $F(2, 3249) = 2{,}253.62$, MSR = 124.76, $p < .01$, $R^2 = .58$. Test administration type was a statistically significant predictor of students' scale score, $b = -10.33$, SE = .50, $p < .05$, 95% CI = -11.21 to -9.35. On average, students taking the scaffold administration scored 10.33 scale score points lower than students taking the standard administration. Students' pre-requisite task total was also a statistically significant predictor of students' scale score, $b = 1.60$, SE = .04, $p < .05$, 95% CI = 1.53 to 1.68. On average, every one-point increase in the pre-requisite skills task total corresponded with a 1.06 increase in students' scale scores. Examination of the squared semipartial correlations revealed that approximately 5.52% of the total scale score variance was uniquely accounted for by test administration type, while 25.20% was uniquely accounted for by the pre-requisite task total. Together test administration type and Pre-requisite skills accounted for 31% of the total variability in scale scores.

Middle School: The regression of scale score on pre-requisite skills and test administration type was statistically significant, $F(2, 2268) = 1{,}410.68$, MSR = 207.77, $p < .01$, $R^2 = .56$. Test administration type was a statistically significant predictor of students' scale score, $b = -14.09$ SE = .72, $p < .01$, 95% CI = -15.50 to -12.69. On average, students taking the scaffold administration scored 14.09 scale score points lower than students taking the standard administration. Students' pre-requisite task total was also a statistically significant predictor of students' scale score, $b = 1.77$, SE = .05, $p < .01$, 95% CI = 1.67 to -1.87. On average, every one-point increase in the pre-requisite skills task total corresponded with a 1.77 increase in students' scale scores. Examination of the squared semipartial correlations revealed that approximately 7.56% of the total scale score variance was uniquely accounted for by test administration type, while 47.89% was uniquely accounted for by the pre-requisite task total. Together test administration type and Pre-requisite skills accounted for 55% of the total variability in scale scores.

High School: The regression of scale score on pre-requisite skills and test administration type was statistically significant, $F(2, 495) = 408.61$, MSR = 253.97, $p < .01$, $R^2 = .62$. Test administration type was a statistically significant predictor of students' scale score, $b = -12.26$, SE = 1.59, $p < .01$, 95% CI = -15.39 to -9.13. On average, students taking the scaffold administration scored 12.26 scale score points lower than students taking the standard administration. Students' pre-requisite task total was also a statistically significant predictor of students' scale score, $b = 1.91$, SE = .89, $p < .01$, 95% CI = 1.74 to 2.09. On average, every one-point increase in the pre-requisite skills task total corresponded with a 1.91 increase in students' scale scores. Examination of the squared semipartial correlations revealed that approximately 2.54% of the total scale score variance was uniquely accounted for by test administration type, while 35.52% was uniquely accounted for by the

pre-requisite task total. Together test administration type and Pre-requisite skills accounted for 38% of the total variability in scale scores.

Writing

Grade 11: The regression of scale score on pre-requisite skills and test administration type was statistically significant, $F(2, 504) = 438.88$, MSR = 200.89, $p < .01$, $R^2 = .64$. Test administration type was a statistically significant predictor of students' scale score, $b = -12.10$, SE = 1.34, $p < .01$, 95% CI = -14.82 to -9.38. On average, students taking the scaffold administration scored 12.10 scale score points lower than students taking the standard administration. Students' pre-requisite task total was also a statistically significant predictor of students' scale score, $b = 1.77$, SE = .08, $p < .01$, 95% CI = 1.62 to 1.93. On average, every one-point increase in the pre-requisite skills task total corresponded with a 1.77 increase in students' scale scores. Examination of the squared semipartial correlations revealed that approximately 5.52% of the total scale score variance was uniquely accounted for by test administration type, while 36.60% was uniquely accounted for by the pre-requisite task total. Together test administration type and Pre-requisite skills accounted for 42% of the total variability in scale scores.

Math

Grade 3: The regression of scale score on pre-requisite skills and test administration type was statistically significant, $F(2, 926) = 731.68$, MSR = 40.35, $p < .01$, $R^2 = .61$. Test administration type was a statistically significant predictor of students' scale score, $b = -4.96$, SE = .50, $p < .01$, 95% CI = -5.94 to -3.98. On average, students taking the scaffold administration scored 4.96 scale score points lower than students taking the standard administration. Students' pre-requisite task total was also a statistically significant predictor of students' scale score, $b = 0.88$, SE = .03, $p < .01$, 95% CI = .82 to .95. On average, every one-point increase in the pre-requisite skills task total corresponded with a 0.88 increase in students' scale scores. Examination of the squared semipartial correlations revealed that approximately 4.12% of the total scale score variance was uniquely accounted for by test administration type, while 29.70% was uniquely accounted for by the pre-requisite task total. Together test administration type and Pre-requisite skills accounted for 34% of the total variability in scale scores.

Grade 4: The regression of scale score on pre-requisite skills and test administration type was statistically significant, $F(2, 937) = 657.30$, MSR = 48.21, $p < .01$, $R^2 = .58$. Test administration type was a statistically significant predictor of students' scale score, $b = -6.52$, SE = .54, $p < .01$, 95% CI = -7.57 to -5.49. On average, students taking the scaffold administration scored 6.52 scale score points lower than students taking the standard administration. Students' pre-requisite task total was also a statistically significant predictor of students' scale score, $b = 0.97$, SE = .04, $p < .01$, 95% CI = 0.89 to 1.05. On average, every one-point increase in the pre-requisite skills task total corresponded with a 0.97 increase in students' scale scores. Examination of the squared semipartial correlations revealed that approximately 6.50% of the total scale score variance was uniquely accounted for by test administration type, while 27.46% was uniquely accounted for by the

pre-requisite task total. Together test administration type and Pre-requisite skills accounted for 34% of the total variability in scale scores.

Grade 5: The regression of scale score on pre-requisite skills and test administration type was statistically significant, $F(2, 949) = 549.62$, MSR = 40.99, $p < .01$, $R^2 = .54$. Test administration type was a statistically significant predictor of students' scale score, $b = -5.91$, SE = .51, $p < .01$, 95% CI = -6.91 to -4.90. On average, students taking the scaffold administration scored 5.91 scale score points lower than students taking the standard administration. Students' pre-requisite task total was also a statistically significant predictor of students' scale score, $b = .87$, SE = .04, $p < .01$, 95% CI = .79 to .95. On average, every one-point increase in the pre-requisite skills task total corresponded with a .87 increase in students' scale scores. Examination of the squared semipartial correlations revealed that approximately 6.50% of the total scale score variance was uniquely accounted for by test administration type, while 23.52% was uniquely accounted for by the pre-requisite task total. Together test administration type and Pre-requisite skills accounted for 30% of the total variability in scale scores.

Grade 6: The regression of scale score on pre-requisite skills and test administration type was statistically significant, $F(2, 875) = 479.03$, MSR = 42.82 $p < .01$, $R^2 = .52$. Test administration type was a statistically significant predictor of students' scale score, $b = -4.76$, SE = .53, $p < .01$, 95% CI = -5.80 to -3.72. On average, students taking the scaffold administration scored 4.76 scale score points lower than students taking the standard administration. Students' pre-requisite task total was also a statistically significant predictor of students' scale score, $b = .84$, SE = .04, $p < .01$, 95% CI = .77 to .92. On average, every one-point increase in the pre-requisite skills task total corresponded with a .84 increase in students' scale scores. Examination of the squared semipartial correlations revealed that approximately 4.37% of the total scale score variance was uniquely accounted for by test administration type, while 24.80% was uniquely accounted for by the pre-requisite task total. Together test administration type and Pre-requisite skills accounted for 29% of the total variability in scale scores.

Grade 7: The regression of scale score on pre-requisite skills and test administration type was statistically significant, $F(2, 782) = 370.99$, MSR = 47.29, $p < .01$, $R^2 = .49$. Test administration type was a statistically significant predictor of students' scale score, $b = -6.55$, SE = .58, $p < .01$, 95% CI = -7.70 to -5.41. On average, students taking the scaffold administration scored 6.55 scale score points lower than students taking the standard administration. Students' pre-requisite task total was also a statistically significant predictor of students' scale score, $b = .73$, SE = .04, $p < .01$, 95% CI = .65 to .82. On average, every one-point increase in the pre-requisite skills task total corresponded with a .73 increase in students' scale scores. Examination of the squared semipartial correlations revealed that approximately 8.29% of the total scale score variance was uniquely accounted for by test administration type, while 18.06% was uniquely accounted for by the pre-requisite task total. Together test administration type and Pre-requisite skills accounted for 26% of the total variability in scale scores.

Grade 8: The regression of scale score on pre-requisite skills and test administration type was statistically significant, $F(2, 723) = 238.10$, MSR = 54.68, $p < .01$, $R^2 = .40$. Test administration type was a statistically significant predictor of students' scale score, $b = -5.24$, SE = .61, $p < .01$, 95% CI = -6.45 to -4.04. On average, students taking the scaffold administration scored 5.24 scale score points lower than students taking the standard administration. Students' pre-requisite task total was also a statistically significant predictor of students' scale score, $b = .67$, SE = .05, $p < .01$, 95% CI = .59 to .76. On average, every one-point increase in the pre-requisite skills task total corresponded with a .67 increase in students' scale scores. Examination of the squared semipartial correlations revealed that approximately 6.10% of the total scale score variance was uniquely accounted for by test administration type, while 18.66% was uniquely accounted for by the pre-requisite task total. Together test administration type and Pre-requisite skills accounted for 25% of the total variability in scale scores.

Grade 11: The regression of scale score on pre-requisite skills and test administration type was statistically significant, $F(2, 509) = 374.88$, MSR = 67.69, $p < .01$, $R^2 = .60$. Test administration type was a statistically significant predictor of students' scale score, $b = -6.44$, SE = .79, $p < .01$, 95% CI = -8.01 to -4.89. On average, students taking the scaffold administration scored 6.44 scale score points lower than students taking the standard administration. Students' pre-requisite task total was also a statistically significant predictor of students' scale score, $b = .97$, SE = .05, $p < .01$, 95% CI = .88 to 1.06. On average, every one-point increase in the pre-requisite skills task total corresponded with a .97 increase in students' scale scores. Examination of the squared semipartial correlations revealed that approximately 5.24% of the total scale score variance was uniquely accounted for by test administration type, while 35.16% was uniquely accounted for by the pre-requisite task total. Together test administration type and Pre-requisite skills accounted for 40% of the total variability in scale scores.

Science

Grade 5: The regression of scale score on pre-requisite skills and test administration type was statistically significant, $F(2, 753) = 484.53$, MSR = 63.05, $p < .01$, $R^2 = .56$. Test administration type was a statistically significant predictor of students' scale score, $b = -8.44$, SE = .66, $p < .01$, 95% CI = -9.73 to -7.15. On average, students taking the scaffold administration scored 8.44 scale score points lower than students taking the standard administration. Students' pre-requisite task total was also a statistically significant predictor of students' scale score, $b = 1.01$, SE = .075 $p < .01$, 95% CI = 0.91 to 1.10. On average, every one-point increase in the pre-requisite skills task total corresponded with a 1.01 increase in students' scale scores. Examination of the squared semipartial correlations revealed that approximately 9.55% of the total scale score variance was uniquely accounted for by test administration type, while 24.21% was uniquely accounted for by the pre-requisite task total. Together test administration type and Pre-requisite skills accounted for 34% of the total variability in scale scores.

Grade 8: The regression of scale score on pre-requisite skills and test administration was statistically significant, $F(2, 627) = 333.70$, MSR = 54.61, $p < .01$, $R^2 = .52$. Test

administration type was a statistically significant predictor of students' scale score, b = -7.22, SE = .63, p < .01, 95% CI = -8.46 to -5.98. On average, students taking the scaffold administration scored 7.22 scale score points lower than students taking the standard administration. Students' pre-requisite task total was also a statistically significant predictor of students' scale score, b = .81, SE = .05, p < .01, 95% CI = .72 to .90. On average, every one-point increase in the pre-requisite skills task total corresponded with a .81 increase in students' scale scores. Examination of the squared semipartial correlations revealed that approximately 10.18% of the total scale score variance was uniquely accounted for by test administration type, while 25.40% was uniquely accounted for by the pre-requisite task total. Together test administration type and Pre-requisite skills accounted for 36% of the total variability in scale scores.

Grade 11: The regression of scale score on pre-requisite skills and test administration type was statistically significant, $F(2, 494) = 317.53$, MSR = 71.59, p < .01, R^2 = .60. Test administration type was a statistically significant predictor of students' scale score, b = -7.60, SE = .81, p < .01, 95% CI = -9.19 to -6.00. On average, students taking the scaffold administration scored 7.60 scale score points lower than students taking the standard administration. Students' pre-requisite task total was also a statistically significant predictor of students' scale score, b = .96, SE = .05, p < .01, 95% CI = .87 to 1.05. On average, every one-point increase in the pre-requisite skills task total corresponded with a .96 increase in students' scale scores. Examination of the squared semipartial correlations revealed that approximately 7.08% of the total scale score variance was uniquely accounted for by test administration type, while 35.88% was uniquely accounted for by the pre-requisite task total. Together test administration type and Pre-requisite skills accounted for 43% of the total variability in scale scores.

Model 3 Results (Sequential): Dis, Admin, & Race/Ethnicity on Pre-Requisite Skills

The full regression model, including correlations and descriptive statistics, are reported in *Appendix G*.

Reading

Elementary: The first block of the regression model included only students' disability category, and was statistically significant, $F(9, 3273) = 94.07$, MSR = 32.06, $p < .01$, $R^2 = .21$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*$(1, 3273) = 166.07$, $p < .01$, R^2 *Change* = .13. For the third block, students race/ethnicity was added to the model, which did not result in a significant change in model fit, *F Change*$(8, 3273) = 92.38$, $p = .84$, R^2 *Change* = .00. For the final model, Test Administration Type had the largest standardized regression weight, β = --.41, $p < .01$, and accounted for the most variance, uniquely accounting for approximately 11.42% of the total variance in students Pre-requisite skills total.

Middle School: The first block of the regression model included only students' disability category, and was statistically significant, $F(9, 2348) = 48.07$, MSR = 41.48, $p < .01$, $R^2 = .16$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*$(1, 2348) = 101.40$, $p < .01$, R^2 *Change* = .15. For the third block, students race/ethnicity was added to the model, which did not result in a significant change in model fit, *F Change*$(8, 2348) = 56.95$, $p = .26$, R^2 *Change* = .00. For the final model, Test Administration Type had the largest standardized regression weight, β = -.41, $p < .01$, and accounted for the most variance, uniquely accounting for approximately 9.36% of the total variance in students Pre-requisite skills total.

High School: The first block of the regression model included only students' disability category, and was statistically significant, $F(9, 506) = 10.76$, MSR = 73.38, $p < .01$, $R^2 = .16$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*$(1, 506) = 19.98$, $p < .01$, R^2 *Change* = .12. For the third block, students race/ethnicity was added to the model, which did not result in a significant change in model fit, *F Change*$(8, 506) = 11.23$, $p = .86$, R^2 *Change* = .00. For the final model, Test Administration Type had the largest standardized regression weight, β = -.371, $p < .01$, and accounted for the most variance, uniquely accounting for approximately 8.59% of the total variance in students Pre-requisite skills total.

Writing

Grade 11: The first block of the regression model included only students' disability category, and was statistically significant, $F(9, 512) = 10.49$, MSR = 68.69, $p < .01$, $R^2 = .16$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*$(10, 512) = 18.43$, $p < .01$, R^2 *Change* = .11. For the third block, students race/ethnicity was added to the model, which did not result in a significant change in model fit, *F Change*$(18, 512) = 10.40$, $p = .83$, R^2 *Change* = .01. For the final model, Test Administration Type had the largest standardized regression weight, β = -.35, $p < .01$, and accounted for the most variance, uniquely accounting for approximately 7.56% of the total variance in students Pre-requisite skills.

Math

Grade 3: The first block of the regression model included only students' disability category, and was statistically significant, $F(9, 923) = 25.11$, MSR = 43.88, $p < .01$, $R^2 = .20$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*(10, 923) = 47.13, $p < .01$, R^2 *Change* = .14. For the third block, students race/ethnicity was added to the model, which did not result in a significant change in model fit, *F Change*(18, 923) = 26.55, $p = .54$, R^2 *Change* = .01. For the final model, Test Administration Type had the largest standardized regression weight, $\beta = -.42$, $p < .01$, and accounted for the most variance, uniquely accounting for 11.76% of the total variance in students' scale score.

Grade 4: The first block of the regression model included only students' disability category, and was statistically significant, $F(9, 945) = 21.74$, MSR = 36.61 $p < .01$, $R^2 = .17$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*(10, 945) = 36.97, $p < .01$, R^2 *Change* = .11. For the third block, students race/ethnicity was added to the model, which did not result in a significant change in model fit, *F Change*(18, 945) = 20.88, $p = .58$, R^2 *Change* = .01. For the final model, Test Administration Type had the largest standardized regression weight, $\beta = -.37$, $p < .01$, and accounted for the most variance, uniquely accounting for approximately 8.29% of the total variance in students Pre-requisite skills.

Grade 5: The first block of the regression model included only students' disability category, and was statistically significant, $F(9, 954) = 21.04$, MSR = 31.80, $p < .01$, $R^2 = .17$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*(10, 954) = 41.08, $p < .01$, R^2 *Change* = .14. For the third block, students race/ethnicity was added to the model, did not result in a significant change in model fit, *F Change*(18, 954) = 22.90, $p = .91$, R^2 *Change* = .00. For the final model, Test Administration Type had the largest standardized regression weight, $\beta = -.41$, $p < .01$, and accounted for the most variance, uniquely accounting for approximately 9.36% of the total variance in students Pre-requisite skills.

Grade 6: The first block of the regression model included only students' disability category, and was statistically significant, $F(9, 890) = 18.30$, MSR = 37.05, $p < .01$, $R^2 = .16$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*(10, 890) = 50.53, $p < .01$, R^2 *Change* = .16. For the third block, students race/ethnicity was added to the model, which did not result in a significant change in model fit, *F Change*(18, 890) = 23.37, $p = .11$, R^2 *Change* = .01. For the final model, Test Administration Type had the largest standardized regression weight, $\beta = -.44$, $p < .01$, and accounted for the most variance, uniquely accounting for approximately 10.56% of the total variance in students Pre-requisite skills.

Grade 7: The first block of the regression model included only students' disability category, and was statistically significant, $F(9, 798) = 17.93$, MSR = 35.27, $p < .01$, $R^2 = .17$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*(10, 798) = 35.73, $p < .01$, R^2 *Change* = .14. For the third block, students race/ethnicity was added to the model, which did not result in a

significant change in model fit, *F Change*(18, 798) = 20.40, $p = .32$, R^2 *Change* = .01. For the final model, Test Administration Type had the largest standardized regression weight, $\beta = -.41$, $p < .01$, and accounted for the most variance, uniquely accounting for approximately 10.24% of the total variance in students Pre-requisite skills.

Grade 8: The first block of the regression model included only students' disability category, and was statistically significant, $F(9, 743) = 13.59$, MSR = 41.52, $p < .01$, $R^2 = .14$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*(10, 743) = 23.76, $p < .01$, R^2 *Change* = .10. For the third block, students race/ethnicity was added to the model, which did not result in a significant change in model fit, *F Change*(18, 743) = 13.22, $p = .97$, R^2 *Change* = .00. For the final model, Test Administration Type had the largest standardized regression weight, $\beta = -.35$, $p < .01$, and accounted for the most variance, uniquely accounting for approximately 6.10% of the total variance in students Pre-requisite skills.

Grade 11: The first block of the regression model included only students' disability category, and was statistically significant, $F(9, 522) = 11.91$, MSR = 67.66, $p < .01$, $R^2 = .17$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*(10, 522) = 19.54, $p < .01$, R^2 *Change* = .10. For the third block, students race/ethnicity was added to the model, which did not result in a significant change in model fit, *F Change*(18, 522) = 10.88, $p = .96$, R^2 *Change* = .00. For the final model, Test Administration Type had the largest standardized regression weight, $\beta = -.34$, $p < .01$, and accounted for the most variance, uniquely accounting for approximately 7.84% of the total variance in students Pre-requisite skills.

Science

Grade 5: The first block of the regression model included only students' disability category, and was statistically significant, $F(9, 742) = 14.61$, MSR = 36.52, $p < .01$, $R^2 = .14$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*(10, 741) = 25.69, $p < .01$, R^2 *Change* = .12. For the third block, students race/ethnicity was added to the model, which did not result in a significant change in model fit, *F Change*(18, 733) = 14.24, $p = .99$, R^2 *Change* = .00. For the final model, Test Administration Type had the largest standardized regression weight, $\beta = -.35$, $p < .01$, and accounted for the most variance, uniquely accounting for approximately 6.60% of the total variance in students Pre-requisite skills.

Grade 8: The first block of the regression model included only students' disability category, and was statistically significant, $F(9, 647) = 10.51$, MSR = 43.93, $p < .01$, $R^2 = .13$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*(10, 647) = 15.78, $p < .01$, R^2 *Change* = .07. For the third block, students race/ethnicity was added to the model, which resulted in a significant change in model fit, *F Change*(18, 647) = 8.87, $p = .93$, R^2 *Change* = .00. For the final model, Test Administration Type had the largest standardized regression weight, $\beta = -.29$, $p < .01$, and accounted for the most variance, uniquely accounting for approximately 4.08% of the total variance in students Pre-requisite skills.

Grade 11: The first block of the regression model included only students' disability category, and was statistically significant, $F(9, 503) = 11.25$, MSR = 69.58, $p < .01$, $R^2 = .17$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*(10, 503) = 15.86, $p < .01$, R^2 *Change* = .07. For the third block, students race/ethnicity was added to the model, which did not result in a significant change in model fit, *F Change*(18, 503) = 8.82, $p = .98$, R^2 *Change* = .00. For the final model, Test Administration Type had the largest standardized regression weight, $\beta = -.28$, $p < .01$, and accounted for the most variance, uniquely accounting for approximately 6.10% of the total variance in students Pre-requisite skills.

Model 4 Results (Sequential): Dis, Admin, & Race/Ethnicity on Scale Score

The full regression model, including correlations and descriptive statistics, are reported in *Appendix H*.

Reading

Elementary: The first block of the regression model included only students' disability category, and was statistically significant, $F(9, 3254) = 93.19$, MSR = 239.91, $p < .01$, $R^2 = .21$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*$(1, 3254) = 198.05$, $p < .01$, R^2 *Change* = .17. For the third block, students race/ethnicity was added to the model, which did not result in a significant change in model fit, *F Change*$(9, 3254) = 110.99$, $p = .09$, R^2 *Change* = .00. For the final model, Test Administration Type had the largest standardized regression weight, $\beta = -.47$, $p < .01$, and accounted for the most variance, uniquely accounting for 14.59% of the total variance in students' scale score.

Middle School: The first block of the regression model included only students' disability category, and was statistically significant, $F(9, 2270) = 42.58$, MSR = 401.98, $p < .01$, $R^2 = .14$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*$(1, 2270) = 125.34$, $p < .01$, R^2 *Change* = .21. For the third block, students race/ethnicity was added to the model, which resulted in a significant change in model fit, *F Change*$(9, 2270) = 71.99$, $p < .01$, R^2 *Change* = .01. For the final model, Test Administration Type had the largest standardized regression weight, $\beta = -.50$, $p < .01$, and accounted for the most variance, uniquely accounting for 13.32% of the total variance in students' scale score.

High School: The first block of the regression model included only students' disability category, and was statistically significant, $F(9, 498) = 8.71$, MSR = 602.21, $p < .01$, $R^2 = .14$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*$(1, 498) = 23.23$, $p < .01$, R^2 *Change* = .18. For the third block, students race/ethnicity was added to the model, which did not result in a significant change in model fit, *F Change*$(9, 498) = 13.04$, $p = .84$, R^2 *Change* = .01. For the final model, Test Administration Type had the largest standardized regression weight, $\beta = -.45$, $p < .01$, and accounted for the most variance, uniquely accounting for 10.76% of the total variance in students' scale score.

Writing

Grade 11:. The first block of the regression model included only students' disability category, and was statistically significant, $F(9, 507) = 10.60$, MSR = 479.60, $p < .01$, $R^2 = .16$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*$(1, 507) = 26.98$, $p < .01$, R^2 *Change* = .19. For the third block, students race/ethnicity was added to the model, which did not result in a significant change in model fit, *F Change*$(9, 507) = 15.07$, $p = .88$, R^2 *Change* = .01. For the final model, Test Administration Type had the largest standardized regression weight, $\beta = -.45$, $p < .01$, and accounted for the most variance, uniquely accounting for 12.74% of the total variance in students' scale score.

Math

Grade 3: The first block of the regression model included only students' disability category, and was statistically significant, $F(9, 928) = 27.68$, MSR = 83.55, $p < .01$, $R^2 = .21$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*$(1, 928) = 57.48$, $p < .01$, R^2 *Change* = .17. For the third block, students race/ethnicity was added to the model, which resulted in a significant change in model fit, *F Change*$(9, 928) = 33.13$, $p < .05$, R^2 *Change* = .01. For the final model, Test Administration Type had the largest standardized regression weight, $\beta = -.47$, $p < .01$, and accounted for the most variance, uniquely accounting for 15.68% of the total variance in students' scale score.

Grade 4: The first block of the regression model included only students' disability category, and was statistically significant, $F(9, 937) = 19.49$, MSR = 98.28, $p < .01$, $R^2 = .16$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*$(1, 937) = 48.51$ $p < .01$, R^2 *Change* = .19. For the third block, students race/ethnicity was added to the model, which did not result in a significant change in model fit, *F Change*$(9, 937) = 27.26$, $p = .60$, R^2 *Change* = .01. For the final model, Test Administration Type had the largest standardized regression weight, $\beta = -.48$, $p < .01$, and accounted for the most variance, uniquely accounting for 12.11% of the total variance in students' scale score.

Grade 5: The first block of the regression model included only students' disability category, and was statistically significant, $F(9, 952) = 23.43$, MSR = 74.76, $p < .01$, $R^2 = .18$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*$(1, 952) = 51.55$, $p < .01$, R^2 *Change* = .17. For the third block, students race/ethnicity was added to the model, which resulted in a significant change in model fit, *F Change*$(9, 952) = 29.83$, $p = .035$, R^2 *Change* = .01. For the final model, Test Administration Type had the largest standardized regression weight, $\beta = -.46$, $p < .01$, and accounted for the most variance, uniquely accounting for 13.32% of the total variance in students' scale score.

Grade 6: The first block of the regression model included only students' disability category, and was statistically significant, $F(9, 876) = 19.57$, MSR = 75.78, $p < .01$, $R^2 = .17$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*$(1, 876) = 41.71$, $p < .01$, R^2 *Change* = .16. For the third block, students race/ethnicity was added to the model, which did not result in a significant change in model fit, *F Change*$(9, 876) = 23.32$, $p = .81$, R^2 *Change* = .00. For the final model, Test Administration Type had the largest standardized regression weight, $\beta = -.44$, $p < .01$, and accounted for the most variance, uniquely accounting for 10.82% of the total variance in students' scale score.

Grade 7: The first block of the regression model included only students' disability category, and was statistically significant, $F(9, 782) = 17.89$, MSR = 77.06, $p < .01$, $R^2 = .17$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*$(1, 782) = 43.72$, $p < .01$, R^2 *Change* = .19. For the third block, students race/ethnicity was added to the model, which did not result in a

significant change in model fit, *F Change*(9, 782) = 23.35, $p = .06$, R^2 *Change* = .01. For the final model, Test Administration Type had the largest standardized regression weight, $\beta = -.47$, $p < .01$, and accounted for the most variance, uniquely accounting for 13.99% of the total variance in students' scale score.

Grade 8: The first block of the regression model included only students' disability category, and was statistically significant, *F*(9, 724) = 13.23, MSR = 79.25, $p < .01$, $R^2 = .14$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*(1, 724) = 26.22, $p < .01$, R^2 *Change* = .13. For the third block, students race/ethnicity was added to the model, which did not result in a significant change in model fit, *F Change*(9, 724) = 15.02, $p = .43$, R^2 *Change* = .01. For the final model, Test Administration Type had the largest standardized regression weight, $\beta = -.38$, $p < .01$, and accounted for the most variance, uniquely accounting for 7.67% of the total variance in students' scale score.

Grade 11: The first block of the regression model included only students' disability category, and was statistically significant, *F*(9, 512) = 10.329, MSR = 146.30, $p < .01$, $R^2 = .16$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*(1, 512) = 23.36, $p < .01$, R^2 *Change* = .16. For the third block, students race/ethnicity was added to the model, which did not result in a significant change in model fit, *F Change*(9, 512) = 13.41, $p = .46$, R^2 *Change* = .01. For the final model, Test Administration Type had the largest standardized regression weight, $\beta = -.42$, $p < .01$, and accounted for the most variance, uniquely accounting for 10.76% of the total variance in students' scale score.

Science

Grade 5: The first block of the regression model included only students' disability category, and was statistically significant, *F*(9, 754) = 23.34, MSR = 114.75, $p < .01$, $R^2 = .22$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*(1, 754) = 50.21, $p < .01$, R^2 *Change* = .18. For the third block, students race/ethnicity was added to the model, which resulted in a significant change in model fit, *F Change*(9, 754) = 30.40, $p < .01$, R^2 *Change* = .02. For the final model, Test Administration Type had the largest standardized regression weight, $\beta = -.48$, $p < .01$, and accounted for the most variance, uniquely accounting for 18.15% of the total variance in students' scale score.

Grade 8: The first block of the regression model included only students' disability category, and was statistically significant, *F*(9, 627) = 11.63, MSR = 97.65, $p < .01$, $R^2 = .15$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*(1, 627) = 24.49, $p < .01$, R^2 *Change* = .17. For the third block, students race/ethnicity was added to the model, which did not result in a significant change in model fit, *F Change*(9, 627) = 16.57, $p = .17$, R^2 *Change* = .01. For the final model, Test Administration Type had the largest standardized regression weight, $\beta = -.44$, $p < .01$, and accounted for the most variance, uniquely accounting for 10.82% of the total variance in students' scale score.

Grade 11: The first block of the regression model included only students' disability category, and was statistically significant, $F(9, 497) = 9.68$, MSR = 158.85, $p < .01$, $R^2 = .15$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*$(1, 497) = 22.20$, $p < .01$, R^2 *Change* = .16. For the third block, students race/ethnicity was added to the model, which did not result in a significant change in model fit, *F Change*$(9, 497) = 12.85$, $p = .35$, R^2 *Change* = .01. For the final model, Test Administration Type had the largest standardized regression weight, $\beta = -.42$, $p < .01$, and accounted for the most variance, uniquely accounting for 10.63% of the total variance in students' scale score.

Conclusions

Overall the results were consistent with prior years and with what was expected. The correlations between students' content scores across subjects were not overly strong, implying that each test measures a distinct construct. Model 1 demonstrated that students' pre-requisite skills total was a significant, but not overly large, predictor of students' content score, implying that student's level of independence significantly impacts performance on the test. Model 2 demonstrated that student disability category and test administration type are strong predictors of student performance, likely relating more to the distinct student groups than to the test itself. Finally, Models 3 and 4 demonstrated that, after controlling for students' disability and test administration type, race/ethnicity was rarely a significant predictor of their performance on the pre-requisite skills or content tasks total score.

Section 5: Alignment

5.1 & 5.2 System Alignment and Range

The overall alignment of Oregon's assessments is not addressed here, as that is a systems-level consideration.

The Oregon Extended assessments have been determined to link to grade level academic content, as specified for all tested subject areas in May 2008, as presented in the *2007-08 Technical Report*. Subsequent alignment studies were implemented in mathematics and science due to the fact that the State of Oregon adopted new general education content standards in those two content areas after the 2007-08 school year. Alignment documentation in mathematics was submitted in the *Oregon Alternate Assessment 2011 Alignment Study in Mathematics*, completed on February 12, 2011. In the area of science, alignment has been documented in the *Oregon Alternate Assessment 2011 Alignment Study in Science*, completed on May 4, 2011. These studies were both required when Oregon adopted new general education content standards in mathematics and science. The original assessments are linked to grade level content. However, Oregon continues to look at ongoing linkage to grade level content due to the development of field test items. We are also in the final stages of completing the process of transitioning toward an AA-AAS that is linked to the Common Core State Standards (CCSS) with our 2013-14 field testing efforts.

The Oregon Extended is designed to allow for continuous improvement. Field test items are developed in all content areas on an annual basis, at an average of 20% new items. These items are compared to operational items based on item functioning and test design factors. These data are used to replace items on an annual basis, incorporating the new items that fill a needed gap with regard to categorical concurrence, or provide for a wider range of functioning with regard to DOK. (see *Section 4.1(c)*)

5.3 Content and Process

The Oregon Extended assessments have been determined to link to grade level academic content in terms of content, as reflected in the item development process. Oregon also had each operational item used on the Oregon Extended assessment evaluated for alignment by an independent contractor, Dr. Lindy Crawford, using a structured and credible process. The professional reviewers included both special and general Oregon education experts, with content knowledge and experience in addition to special education expertise. Reviewers were trained by synchronous webinar regarding their alignment tasks, which were conducted online via BRT's Distributed Item Review (DIR) website. Training topics included the concepts of depth, breadth, and complexity. Mock linkage ratings were conducted in order to address questions and ensure appropriate calibration. Reviewers rated each item on a 4-point scale (0 = not at all linked, 1= vaguely linked, 2= somewhat linked, 3= very well linked) as it related to the standard the item developers had defined for that item. Adequate linkage was defined as being rated a 2 or 3 by at least two raters. Additional comment was requested for any item whose linkage was rated 0 or 1. Items that did not meet this standard were not utilized for the operational assessment.

5.4 Degree and Pattern of Emphasis

The Oregon Extended assessments reflect similar degrees and patterns of emphasis when compared to the OAKS. These similarities can be seen in the test specifications documents, which convey the balance of representation both within and across standards (as evaluated by categorical concurrence). The process of addressing any gaps or weaknesses in the system is accomplished via field testing (see *Section 4.3(c)*).

5.5 Scores Reflect Range

The Oregon Extended assessments yield scores that reflect the full range of achievement implied by Oregon's alternate achievement standards. Evidence of this claim is found in the standard setting documentation submitted in prior years. Standards were set for all subject areas, reading, writing, mathematics, and science on May 21, 2007 and June 3-4, 2007. Standards included achievement level descriptors and cut scores, which defined Oregon's alternate achievement standards (AAS) at that time. Since that time, new standard have been set in mathematics and science. A standard setting for mathematics was conducted August 16, 2010. The mathematics AAS were adopted by the State Board of Education in October 2010. The standard setting for science was conducted on August 9, 2011. The State Board of Education officially adopted the science AAS in October 2011. Documentation for all standard settings has been reviewed in prior submissions.

5.6 Results Expressed in Terms of AAS

The mock-up student report template includes the full AAS (cut scores and achievement level descriptors), not only scale scores or percentiles (see *Appendix 1.6*).

5.7 Improving Alignment

The Oregon Extended assessment system uses field testing to improve the alignment of operational assessments each year. Field testing at approximately 20% of operational items in each subject area allows us to remove not only items with weaker alignment statistics, but also items that are no longer functioning as expected, and/or items that are not aligned to the CCSS as Oregon carefully transitions toward full alignment with the CCSS in English language arts and mathematics. Our current field test development plan addresses these continuous improvement strategies in each content area (see *Section 4.3(c)*). This approach is supported by existing alignment documentation (see *Section 5.2*).

Section 6: Inclusion of All Students in the Assessment System

6.1.1 Participation Data

Oregon's participation data indicate that all students in the tested grade levels are included in our assessment system, including students with significant cognitive disabilities. Documentation of this requirement is provided within the Annual Performance Report, Indicator B3, which is submitted to the United States Department of Education's (USED's) Office of Special Education Programs (OSEP).

6.1.2 Separate Reporting

Oregon reports separately the number and percent of students with disabilities assessed on the regular assessment without accommodations, on the regular assessment with accommodations, and the alternate assessment based on alternate achievement standards Documentation of this requirement is provided within the Annual Performance Report, Indicator B3, which is submitted to USED/OSEP.

6.2.1(a) Promoted Use of Accommodations

Oregon has developed, disseminated information on, and promoted the use of appropriate accommodations to increase the number of students with disabilities who are tested against academic achievement standards for the grade levels in which they are enrolled (see *Appendix 2.4*)

6.2.1(b) Assessor Training

Oregon has ensured that general and special education teachers and other appropriate staff know how to administer assessments, including making use of accommodations, for students with disabilities and students covered under Section 504 (see *Section 4.5*).

6.2.2(a) Clear Guidelines for IEP Teams

Oregon has provided IEP teams with guidance and expectations surrounding appropriate participation decisions for the Oregon Extended assessment (see *Appendix 2.1*)

6.2.2(b) Any Disability Category Eligible

The guidance that Oregon has provided to IEP teams both during training (see Section 4.5) and in terms of procedural documentation (see *Appendix 2.1*) makes it clear that students who participate in AA-AAS may be from any disability category.

6.2.2(c) Clear Explanation of Differences

Oregon has made it clear that the performance based on AA-AAS is not comparable to performance from the OAKS, which is based on grade-level academic achievement standards (see *Appendix 1.1a, slide 29 of 46*).

6.2.2(d) Parents Informed

Oregon will add specific language to their Assessment Decision Making Guidelines to ensure that parents are informed of the potential consequences associated with having their child assessed against AAS.

6.2.3 Modified Achievement Standards

Oregon has not developed an AA-MAS, so this section is not addressed.

6.2.4 Involved in General Curriculum

Oregon has documented that students with the most significant cognitive disabilities are, to the extent possible, included in the general education curriculum. Documentation of this requirement can be found within our SPR&I monitoring system.

6.3(a) Assessments Available: Language/Form

Oregon has made available, to the extent practicable, assessments in the language and form most likely to yield accurate and reliable information on what these students know and can do. This effort is based partially on test design using universal design principles (see *Appendix 1.5*), as well as upon the allowable language accommodations (see *Appendix 2.4*).

6.3(b) LEP Student Participation

Oregon requires the participation of all students with limited English proficiency, except for students who are exempt in reading/language arts (see *Appendix 2.6*).

6.3(c) LEP Student Assessment Policies

Oregon has adopted policies requiring students with limited English proficiency to be assessed in reading/language arts in English when they have been enrolled in US schools for three years or more (see *Appendix 2.6*).

6.4 Identification and Inclusion of Migrant Students

Oregon has policies and procedures in place to ensure the identification and inclusion of migrant and other mobile students in the tested grades in our assessment system (see *Appendices 2.5 - 2.6*).

ODE Policy and Procedures Appendices

Topic	File Name
ODE's existing policies regarding which student results are/are not included in all AYP reports	App2.5_AsmtInclusionRules2010_11
ODE's AYP Policy and Technical Manual, a summary document including all AYP procedures and reporting	App2.6_AYPManual2011_12

Appendix 2.5

Appendix 2.5 is the manual defining the state of Oregon's policies and procedures regarding how students are included in AYP reporting.

Appendix 2.6

Appendix 2.6 includes all adequate yearly progress processes, making it clear that all students in the grades tested are to participate in Oregon's statewide assessments, including the OAKS, the Oregon Extended, and the ELPA. The manual also includes official expectations regarding how the 1% reporting cap is handled for the Oregon Extended assessment.

Section 7: Assessment Reports

7.1 Reporting System

Oregon's reporting system facilitates appropriate, credible, and defensible interpretation and use of its assessment data. With regard to the Oregon Extended, the purpose is clearly to provides the state technically adequate student performance data to ascertain proficiency on grade level state content standards for students with significant cognitive disabilities (see *Section 4.1a*). In addition, the state makes it clear that results from the Oregon Extended and not comparable to results from the OAKS (see *Section 6.2.2c*). In addition, the data also meets rigorous reliability expectations (see *Appendices A-H*). Validity is considered here as an overarching summation of the Oregon Extended assessment system, as well as the mechanisms that Oregon uses to continuously improve the Oregon Extended assessment.

7.2 Reporting Requirements

Oregon reports participation and assessment results for all students and for each of the required subgroups in its reports at the school, LEA, and state levels. The state does not report subgroup results when these results would reveal personally identifiable information about an individual student. The calculation rule followed is that the number of students in the subgroup must meet the minimum cell size requirement for each AYP decision: participation, achievement in English language arts and math, attendance, and graduation, where appropriate (see *Appendix 2.6*)

7.3 Individual Reports (IRs)

Oregon develops and disseminates individual student data upon final determination of accuracy. The state provides districts with individual student reports (ISRs) that meet most relevant requirements. The state is in the process of incorporating the Standard Error of Measure (SEM) for each student score into the report templates. However, the SEM associated with each cut score is provided in *Section 4.2b*. Also, see the mock-up ISR in *Appendix 1.6*.

7.3(a) IRs Provide Reliable and Valid Information

Oregon's student reports provide valid and reliable information regarding achievement on the assessments relative to the AAS. The reliability of the data is addressed in *Appendices A-H.* Validity is considered here as an overarching summation of the Oregon Extended assessment system, as well as the mechanisms that Oregon uses to continuously improve the Oregon Extended assessment. The ISRs clearly demonstrate the students' scale score relative the AAS that is relevant for that content area and grade level (see *Section 4.2b* and *Appendix 1.6*).

7.3(b) IRs Provide Information for Stakeholders

The Oregon ISRs provide information for parents, teachers, and administrators to help them understand and address a student's academic needs. These reports are displayed in a simple format that is easy for stakeholders to understand. Results can be translated for

parents by district representatives, as necessary. Guidelines for interpreting individual student reports will be developed (see *Appendix 1.6*).

7.3(c) IRs are Delivered to Stakeholders

The Oregon ISRs are made available via online secure district website upon completion of final AYP analyses. Districts are then expected to deliver the ISRs to schools. Schools are subsequently expected to share results with parents and staff.

7.4 Student Data are Secure

Oregon ensures that student-level assessment data from the Oregon Extended are maintained securely to protect student confidentiality in several manners. First, the data is entered via a secure data entry system. All data sharing is conducted via the state's secure file-sharing system. All servers used for student data storage and analyses are secure, as are the individual PCs and laptops of staff who review and analyze student data via encryption procedures.

7.5 Provided Score Analyses

The results for the Oregon Extended assessment are provided in content area summative scores. They are not provided in disaggregated strand scores, as the information at this level is not always reliable or meaningful (see *Appendix 1.6*).