Oregon Department of Education

2011–2012 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

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

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 2011-12 school year in grades 3-8 and once in the (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 Oregon Department of Education (ODE) has not developed or adopted extended/expanded grade band expectations for this group of students. The instructional targets for this group are thus established by the grade level standards for all Oregon students that are thoughtfully 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 in the areas of linkage to the Oregon general education content standards, access for students with significant cognitive disabilities, sensitivity, and bias. The items employed as operational have survived a robust and iterative development review, in addition to annual reviews of impact data.

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

2.1 & 2.2

The Oregon Extended assessment, Oregon's 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. Results from the reading and math administrations are included in calculations of participation and performance for Adequate Yearly Progress (AYP). Science participation is also included as part of the Title 1 Assessment System requirements; science is administered in grades 5, 8, & 11:

All academic achievement standards for the Oregon Extended assessment have been submitted for peer review in prior years. Oregon continues to use the same AAS for implementation of our AA-AAS. These AAS were developed by representative samples of 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.

2.3

The alternate achievement standards are composed of four levels (though three are required). In hierarchical order, they are 1) Exceeds, 2) Meets, 3) Nearly Meets, and 4) Does Not Yet Meet. The top two 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

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

2.5

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.

2.6

The *2010-11 Science Technical Report* submitted with Oregon's Peer Review submission in 2010-11 included evidence of the standard setting process conducted in the area of science. 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

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

3.2

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

3.3

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

Oregon has provided documentation of 3.4(a), (b), and (c) in prior submissions. Oregon Extended assessment results continue to be used for AYP calculations in reading and math.

3.5

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

3.6

The Oregon Extended assessment is built upon a variety of items that address a wide range of performance expectations rooted in the Oregon general education content standards. The challenge built into the test design is based first upon the content within each standard in reading, writing, mathematics, and science. That content is reduced in terms of depth, breadth, and complexity in a manner that is approved by Oregon general and special education teachers to develop assessment targets that are appropriate for students with the most significant cognitive disabilities. Within that range of performance, approximately 10% of students statistically hang together at the lower end of the performance continuum, likely because they have yet to develop a formal communication system, and another group of approximately 90% of students who perform toward the middle and upper ranges of the performance continuum, likely because they can communicate effectively using abstract symbols. 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

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

As elaborated by Messick (1989)[1], 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*[2] (AERA et al., 1999) in documenting evidence on content coverage, response processes, internal structure, and relations to other variables. This document follows the essential data requirements of the federal government as needed in the peer review process.[3] 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)

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 aligned with grade level academic content; generate reliable outcomes at the item, task, and 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.

---

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

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

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

4.1(b)

The Oregon Extended 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 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, 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 due to the fact that the State of Oregon adopted new general education content standards in mathematics and science, respectively.

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

4.1(c)

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 test 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)[4].

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).[5]

All items and tasks are linked to grade level standards and a prototype was developed using principles of universal design with traditional item writing techniques. The most important component in these initial steps addressed language complexity and access to students using both receptive, as well as expressive, communication. Additionally, both breadth and depth were addressed. We developed two forms of each grade level test, a

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

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

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 existing panels of teachers who have worked with the Oregon Department of Education in various advising roles on testing content in general education, using the same processes and criteria. While the internal reviews of content were initially conducted within Behavioral Research and Teaching, after the initial draft of prototype items, all reviews involved 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, Oregon has ascertained that its alternate assessment 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)

The primary purpose of the Oregon Extended assessment is to yield technically adequate performance data on grade level state content standards for students with significant cognitive disabilities in reading/writing and mathematics at the test level. All scoring and reporting structures mirror this design and have been shown to be highly reliable measures at the test and task levels (see *Appendix B*).

4.1(e)

Pre-requisite skills assessments are designed to allow teachers to use of various levels of support, as appropriate. Because the Oregon Extended assessment first documents the student's access skill (pre-requisite skill) to assist teachers in presenting the content items, pre-requisite skills were 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 exactly 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)[6] 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 reflects 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).[7]

As mentioned in section 4.1b, our assessments appear to be measuring separate constructs, as intended (see *Appendix D)*, providing Pearson correlational 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 all students are receiving a fair chance to demonstrate performance on the Oregon Extended assessment (see *Appendix A)*.

4.1(f)

As mentioned above in section 4.1a, 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 item, task, and 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)

The state currently utilizes statewide advisory groups to provide consequential validity input. It is also exploring the concept of incorporating a consequential validity survey for stakeholders of the Oregon Extended Assessment system in the spring of 2013.

---

[6] 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.

[7] 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.

4.2(a)

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 test has high reliabilities at both the task and total test levels.

4.2(b)

Oregon has not yet reported the conditional standard error of measurement (SEM) and student classifications that are consistent with each cut score specified in its academic achievement standards. The average SEM associated with each cut score for 2011-12 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**	*2.67*	**103**	*2.04*	**113**	*2.73*
4	**101**	*2.07*	**107**	*2.20*	**116**	*3.24*
5	**105**	*2.09*	**110**	*2.41*	**119**	*4.40*
6	**97**	*2.10*	**103**	*2.10*	**116**	*3.22*
7	**98**	*2.05*	**106**	*2.27*	**117**	*3.40*
8	**102**	*2.16*	**112**	*2.76*	**120**	*3.98**
11	**101**	*2.26*	**109**	*2.53*	**121**	*3.80*

WRITING

Grade	**NM**	*SEM*	**M**	*SEM*	**E**	*SEM*
11	**98**	*2.20*	**103**	*2.40*	**122**	*4.50**

MATHEMATICS

Grade	**NM**	*SEM*	**M**	*SEM*	**E**	*SEM*
3	**97**	*1.89*	**104**	*1.94*	**112**	*2.60**
4	**98**	*1.96*	**106**	*1.93*	**114**	*2.90*
5	**101**	*2.14*	**110**	*1.90*	**118**	*2.90*
6	**99**	*1.74*	**103**	*1.96*	**110**	*2.75*
7	**101**	*1.80*	**102**	*1.86*	**107**	*2.40*
8	**101**	*2.01*	**105**	*1.81*	**110**	*2.01*
11	**99**	*1.97*	**106**	*2.00*	**115**	*3.80**

SCIENCE

Grade	**NM**	*SEM*	**M**	*SEM*	**E**	*SEM*
5	**100**	*2.17*	**107**	*1.96*	**114**	*2.60*
8	**95**	*2.00*	**101**	*1.93*	**113**	*2.50*
11	**98**	*1.80*	**103**	*1.84*	**109**	*2.40*

** = No students achieved this score; therefore, the determination is the average of adjacent scores*

4.2(c)

Oregon has reported evidence of generalizability for all relevant sources, including an analysis of demographic groups, to ensure that items are functioning consistently across demographic groups (see *Appendix A)*. The internal consistency of item responses is analyzed within *Appendix B*. A 2007-08 study 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)

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 and allow for 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 2011-12: 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.

4.3(b)

The Oregon Extended 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. Because of this fact, the Oregon Extended assessment can be translated or interpreted by a qualified administrator in the student's native language, as can all assessments in Oregon. 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 the fact that a systematic prompt and well-defined answers are provided.

4.3(c)

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 alignment, but also for access, sensitivity, and bias. This process makes it more likely that students are receiving instruction that is reflected in the assessment, and also that the items are not biased toward a particular demographic group or sub-group (see *Appendix 1.5*).

Oregon Extended Assessment Field Testing 2011-12

Field testing was conducted in reading, writing, mathematics, and science, in the 2011-12 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 follows a three-year implementation strategy in reading and mathematics to steadily transition toward items and assessments that are aligned with the Common Core State Standards (CCSS). The implementation strategy for science follows an independent course that is defined by the State of Oregon's 2009 Science Standards. Oregon is also involved as a Tier II state in the NCSC GSEG and anticipates utilizing the resources created by the grant in whole or in part.

2011-12 Field Test Development Plan

The table below provides an overview of the field test implementation plan for 2011-2012:

Extended Assessment Subject Area	**Field Test Location on 2011-12 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 Reading and Writing assessment, while eight Mathematics and Science items were developed for each grade band/level.

2011-14 Field Test Development Plan

By the end of the three-year implementation plan, all current items will align to the CCSS in reading/writing and mathematics, while all science items will continue to align with our 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 2011-12

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 2011-12 spring test administration. BRT employed a multi-stage development process to ensure that test items were both linked to relevant content standards, that items were accessible for students with significant cognitive disabilities, and that any perceived item biases were eliminated. This review process included 26 reviewers with an average of 22.3 years of experience in education (see *Appendix 2.4*).

4.3(d)

While accommodations are truly built into the test design to a large degree, as described above, the use of accommodations on the Oregon Extended 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 is researching the possibility of gathering statistical accommodation information for the Oregon Extended through its online data entry process to support relevant analyses.

4.4(a)

The Oregon Extended 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 attentional resources to the task 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)

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

4.5

The Oregon Extended 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 and Appendix 1.1i). The state's training program for test administration is complex; it is described below.

Oregon Extended Assessment Training 2011-12

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

Date	Who/Team	Location
11-3-2012	Team: DC, BL, JT, DF Contact: Kerri Smith kmsmith@nwresd.k12.or.us 503-614-1428 5825 NE Ray Circle, Hillsboro, OR 97124-6436	NWRESD- 5825 NE Ray Cir Hillsboro, OR 97124-6436
11-10-2012	Team: DC, BL, JT, DF Contact: Tom Beach Tom.beach@wesd.org 503-588-5330 Marion Center, 2611 Pringle Rd. SE, Salem, OR, 97302	WESD- 2611 Pringle Rd SE Salem, OR 97302-1533
11-15-2012 (combined webinar)	Team: DC, BL, JT, DF Contact: Catherine Kelly catherine.kelly@hdesd.org 541-693-5600 145 SE Salmon Ave. Redmond, OR 97756 Marian Gerstmar marian_gerstmar@soesd.k12.or.us 541-776-8590 101 North Grape St., Medford, OR 97501	HDESD- 145 SE Salmon Ave Ste A Redmond, OR 97756-8427 SOESD- 101 Grape St. Medford, OR 97501-2793
11-17-2012	Team: BL, DF Contact: Mary Apple mary.apple@umesd.k12.or.us 541-276-6616 2001 SW. Nye Pendleton, OR 97801	UMESD- 2001 SW Nye Ave Pendleton, OR 97801-4416

The Oregon Extended assessment can be administered only by trained assessors, called 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.

An analysis of the results from our required online training assessments for both new and ongoing QAs and QTs is conducted annually to determine if our proficiency assessments for assessors are functioning appropriately. The results of this year's analysis are provided below. The data for this study is gathered from the Oregon Extended Assessment Training and Proficiency Website (http://or.k12test.com/).

New assessors need to pass five proficiencies with a score of 80% or higher. The five proficiencies are in the areas of: Administration, Reading, Math, Writing, and Science. Returning QAs or QTs for the 2011-12 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 total number of assessors documented on the Oregon Extended Assessment Training and Proficiency Website for 2011-12 are provided below.

Assessor in-Training - 449
Qualified Assessors - 1,253
Qualified Trainers - 171

Administration Proficiency (404 Test Participants)

Number of Participants	Percentage of Participants	Attempts to Pass	Average Passing Score
311	77%	1	92%
60	15%	2	88%
22	5%	3	95%
10	2%	4	88%
1	>1%	5	80%

Reading Proficiency (397 Test Participants)

Number of Participants	Percentage of Participants	Attempts to Pass	Average Passing Score
378	95%	1	95%
16	4%	2	92%
3	1%	3	92%

Math Proficiency (395 Test Participants)

Number of Participants	Percentage of Participants	Attempts to Pass	Average Passing Score
389	98%	1	97%
5	1%	2	93%
1	>1%	3	100%

Writing Proficiency (385 Test Participants)

Number of Participants	Percentage of Participants	Attempts to Pass	Average Passing Score
326	85%	1	90%
35	9%	2	86%
19	5%	3	94%
2	>1%	4	90%
3	>1%	5	88%

Science Proficiency (392 Test Participants)

Number of Participants	Percentage of Participants	Attempts to Pass	Average Passing Score
389	99%	1	99%
3	>1%	2	93%

Refresher Proficiency (1,044 Test Participants)

Number of Participants	Percentage of Participants	Attempts to Pass	Average Passing Score
930	89%	1	93%
58	5%	2	86%
49	4%	3	92%
5	1%	4	90%
1	>1%	5	100%
1	>1%	6	100%

An additional analysis was compiled which compared the average score of participants on their first attempt 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	86%	411	91%	404
Reading	94%	397	95%	397
Math	96%	397	97%	395
Writing	87%	385	90%	385
Science	98%	392	99%	392
Refresher	89%	1,062	92%	1,044

The results of these analyses demonstrate that assessors in general had the most difficulty passing the Administration, Writing, and Refresher assessments. However, it is a small number of participants who struggle to pass these assessments by the second attempt, with a range of 8% who did not pass after two attempts on the Administration assessment, to 0% of assessors who did not pass after two attempts on the Science assessment. One assessor required six attempts before s/he was able to pass the Refresher proficiency assessment, but this is an unusual statistic.

Training and Test Appendices

Topic	File Name
Slides for training new qualified assessors, new qualified trainers, and returning assessors	App1.1a_QTTraining2011_12
Slides for orienting assessors to the use of the Training and Proficiency website	App1.1b_ORExtendQTTrng2011_12
A handout which reviews general assessment information and all of the major changes in the Oregon Extended assessment program since the 2010-11 administration	App1.1c_ExtAssessUpdates2011_12
The final test administration calendar for all Oregon assessments	App1.1d_TestSchd2011_12
Sample agenda for training new qualified assessors using the train-the-trainers model	App1.1e_QT_TrainingAgenda2011-12
Provides assessors and trainers instructions regarding how to access the online training and proficiency website	App1.1f_ExtAssessAccess Instr2011_12
Provides qualified trainers with a list of duties associated with their training responsibilities	App1.1g_TrainerResponsibilities2011_12
QT Training Evaluations	App1.1h_QTTrngEvals2011_12
Help Desk log and evaluation report	App1.1i_HelpDeskLog2011_12
General Administration and Scoring Manual	App1.2_ExAssessAdminMan2011-12
Sample test items in RWMS	App1.3_RWMSSampleItems2011_12
Report summarizing the results of the field test item reviews conducted with Oregon teachers	App1.4DIRReport2011_12
Describes how items are reduced in depth, breadth, and complexity, as well as how item bias is reduced/eliminated.	App1.5_ ORExtReduceDepthBreadthComplex2011_12
Mock-up individual student report, demonstrating that the reports contain cutscores and ALDs	App1.6_ORExtend_StudentReport_Mock_Up
Guidance developed by ODE to assist IEP teams in making appropriate assessment decisions	App2.1_AssessDecisionMakingGuidelines2011_12
Guidance for assessors for data entry process	App2.2_DataEntryGuide2011_12
ODE Test Administration Manual (TAM), providing test security and administration requirements for all Oregon assessments,(see *Appendix I*, page I-1)	App2.3_TAM2011_12
Provides ODE's guidance and expectations related to accommodations.	App2.4_ODEAccomMan2011_12

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 trainers (QTs) and qualified assessors (QAs) in four regionally hosted webinar trainings in November 2011. QTs also used the package to train New Qualified Assessors for the 2011-12 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 2011-12
- 2012 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
- 2012 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 Oregon Extended assessment, why, etc.) as well as all changes that have been implemented since the prior year for returning users. The major change this year was not having a writing assessment in grades 4 & 7.

Appendix 1.1d
Appendix 1.1d is the test calendar for the entire Oregon statewide assessment program, including the OAKS, the Oregon Extended, 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 Oregon Extended assessment

Appendix 1.1h
Appendix 1.1h contains the participant evaluation results from the four regionally hosted webinar trainings in November 2011. The results demonstrate a high overall satisfaction with both the policy and procedure training portion, as well as the training and proficiency website components of the program.

Appendix 1.1i
Appendix 1.1i 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 2011-12. The manual establishes ODE's expectations regarding the test window, utilizing the OR Extended 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 Oregon Extended Assessment. 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 be viewing during test administration (Scoring/teacher's Protocol) as well as what the student would be viewing 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 Oregon Extended Assessment 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 Oregon Extended assessment. The mock-up includes cutscores 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 Oregon Extended assessment.

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

4.6(a)

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)

While accommodations are built into the flexibility provided by the Oregon Extended 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 Oregon Extended assessment.

4.6(c)

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 Oregon Extended assessment 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)

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.

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 that demonstrate that student performance is dependent primarily upon the difficulty of the content task and not pre-requisite skills or disability:

Data Analyses Appendices Table

Topic	File Name
Demographics for participants in 2011-2012 alternate assessment	AppA_Dems
Reliability of items, tasks, and tests in reading, writing, mathematics and science for all grade levels	AppB_Rel
Descriptive statistics for all tasks in reading, writing, mathematics and science for all grade levels	AppC_Dscrpt
Correlations across subject areas	AppD_Corr
Simultaneous regression model using pre-requisite skills as a predictor of both the scale and raw scores	AppE_Model_1
Simultaneous regression model using pre-requisite skills *and* type of administration as a predictor of both the scale and raw scores	AppF_Model_2
Sequential regression model using demographic variables, test administration type, and disability variables to predict pre-requisite skills total	AppG_Model_3
Sequential regression model using demographic variables, test administration type, and disability variables to predict both the scale and raw scores	AppH_Model_4

Demographics

The full demographics for students taking the Oregon Extended Assessment 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. For grade 3, approximately 67.9% were male, 55.3% were White, and 29.3% were Hispanic. Approximately 72.3% of all students were administered the Standard version of the test, while the remaining 27.7% were administered the Scaffold version of the test. For grade 4, approximately 67.7% were male, 54.6% were White, and 30.4% were Hispanic. Approximately 71.2% of all students were administered the Standard version of the test, while the remaining 28.8% were administered the Scaffold version of the test. For grade 5, 66.5% were male, 54.7% were White, and 30.9% were Hispanic. Approximately 72.4% of all students were administered the Standard version of the test, while the remaining 27.6% were administered the Scaffold version of the test.

Middle. For grade 6, approximately 63.9% were male, 59.3% were White, and 24.5% were Hispanic. Approximately 64.8% of all students were administered the Standard version of the test, while the remaining 35.2% were administered the Scaffold version of the test. For grade 7, approximately 66.4% were male, 61.1% were White, and 23.3% were Hispanic. Approximately 60.9% of all students were administered the Standard version of the test, while the remaining 39.1% were administered the Scaffold version of the test. For grade 8, 64.6% were male, 63.7% were White, and 25.2% were Hispanic. Approximately 58.1% of all students were administered the Standard version of the test, while the remaining 41.9% were administered the Scaffold version of the test.

High. Approximately 62.5% were male, 67.4% were White, and 17.8% were Hispanic. Approximately 53.5% of all students were administered the Standard version of the test, while the remaining 46.5% were administered the Scaffold version of the test.

Writing

Grade 11. Approximately 62.5% were male, 67.2% were White, and 18.1% were Hispanic. Approximately 54% of all students were administered the Standard version of the test, while the remaining 46% were administered the Scaffold version of the test.

Math

Grade 3. Approximately 65% of students taking the mathematics portion of the Oregon Extended Assessment were male, 54.6% were White, and 29.2% were Hispanic. Approximately 66.3% of all students were administered the Standard version of the test, while the remaining 33.7% were administered the Scaffold version of the test.

Grade 4. Approximately 64.6% were male, 55% were White, and 29.1% were Hispanic. Approximately 67.2% of all students were administered the Standard version of the test, while the remaining 32.8% were administered the Scaffold version of the test.

Grade 5. Approximately 64.4% were male, 55.9% were White, and 28.8% were Hispanic. Approximately 68.1% of all students were administered the Standard version of the test, while the remaining 31.9% were administered the Scaffold version of the test.

Grade 6. Approximately 61.6% were male, 59.1% were White, and 24.7% were Hispanic. Approximately 63.4% of all students were administered the Standard version of the test, while the remaining 36.6% were administered the Scaffold version of the test.

Grade 7. Approximately 63.9% were male, 62.2% were White, and 21.7% were Hispanic. Approximately 59% of all students were administered the Standard version of the test, while the remaining 41% were administered the Scaffold version of the test.

Grade 8. Approximately 62% were male, 63.5% were White, and 24.9% were Hispanic. Approximately 56.9% of all students were administered the Standard version of the test, while the remaining 43.1% were administered the Scaffold version of the test.

Grade 11. Approximately 62.4% were male, 68.3% were White, and 17.9% were Hispanic. Approximately 53.5% of all students were administered the Standard version of the test, while the remaining 46.5% were administered the Scaffold version of the test.

Science

Grade 5. Approximately 65.1% of students taking the science portion of the Oregon Extended Assessment were male, 59.9% were White, and 25.8% were Hispanic. Approximately 59% of all students were administered the Standard version of the test, while the remaining 41% were administered the Scaffold version of the test.

Grade 8. Approximately 64% of students taking the science portion of the Oregon Extended Assessment were male, 64.5% were White, and 24.8% were Hispanic. Approximately 54% of all students were administered the Standard version of the test, while the remaining 46% were administered the Scaffold version of the test.

Grade 11. Approximately 62.1% of students taking the science portion of the Oregon Extended Assessment were male, 68% were White, and 18.3% were Hispanic. Approximately 53.3% of all students were administered the Standard version of the test, while the remaining 46.7% 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.91 to .97.

Reading

Elementary. The task reliability for the elementary grade-band (3, 4, 5) was moderate to high, ranging from 0.65 for Task 2 to 0.96 for Task 1. The reliability of the total test was quite high, at 0.96.

Reading: Elementary

Task	Cronbach's Alpha
1	0.96
2	0.65
3	0.74
4	0.80
5	0.71
6	0.77
7	0.73
8	0.75
9	0.75
10	0.70
11	0.73
Total Test	0.96

Middle. The task reliability for the middle school grade band (grades 6, 7, and 8) was moderate to high, ranging from 0.64 for Task 2 to 0.97 for Task 1. The reliability of the total test was quite high, at 0.97

Reading: Middle

Task	Cronbach's Alpha
1	0.97
2	0.64
3	0.83
4	0.83
5	0.78
6	0.78
7	0.72
8	0.75
9	0.80
10	0.80
11	0.75
Total Test	0.97

High. The task reliability for the high school grade band (grade 11) was moderately high to high, ranging from 0.64 for Task 3 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	0.77
3	0.64
4	0.84
5	0.78
6	0.85
7	0.77
8	0.82
9	0.86
10	0.74
11	0.78
Total Test	0.97

Writing

High. Task reliability was moderately high to high, ranging from 0.73 for Task 9 to 0.97 for Tasks 1 and 2. The reliability of the total test was quite high, at 0.97.

Writing: High

Task	Cronbach's Alpha
1	0.97
2	0.97
3	0.87
4	0.77
5	0.87
6	0.83
7	0.86
8	0.85
9	0.73
10	0.85
11	0.90
Total Test	0.97

Math

Grade 3. Task reliability was moderate to high, ranging from 0.46 for Task 4 to 0.96 for Task 1. The reliability of the total test was quite high, at 0.92.

Math: Grade 3

Task	Cronbach's Alpha
1	0.96
2	0.75
3	0.54
4	0.46
5	0.62
6	0.53
7	0.51
8	0.65
9	0.70
Total Test	0.92

Grade 4. Task reliability was again moderate to high, ranging from 0.55 for Task 3 to 0.96 for Task 1. The reliability of the total test was quite high, at 0.93.

Math Grade 4

Task	Cronbach's Alpha
1	0.96
2	0.73
3	0.55
4	0.62
5	0.70
6	0.59
7	0.71
8	0.63
9	0.60
Total Test	0.93

Grade 5. Task reliability was moderate to high, ranging from 0.50 for Tasks 4 and 7 to 0.97 for Task 1. The reliability of the total test was quite high, at 0.92.

Math: Grade 5

Task	Cronbach's Alpha
1	0.97
2	0.52
3	0.60
4	0.50
5	0.59
6	0.54
7	0.50
8	0.63
9	0.64
Total Test	0.92

Grade 6. Task reliability was low to high, ranging from 0.39 for Task 6 to 0.97 for Task 1. The reliability of the total test was high, at 0.91.

Math: Grade 6

Task	Cronbach's Alpha
1	0.97
2	0.54
3	0.75
4	0.40
5	0.52
6	0.39
7	0.45
8	0.49
9	0.60
Total Test	0.91

Grade 7. Task reliability was moderate to high, ranging from 0.42 for Task 7 to 0.97 for Task 1. The reliability of the total test was quite high, at 0.92.

Math: Grade 7

Task	Cronbach's Alpha
1	0.97
2	0.60
3	0.50
4	0.64
5	0.65
6	0.61
7	0.42
8	0.50
9	0.54
Total Test	0.92

Grade 8. Task reliability was low to high, ranging from 0.42 for Tasks 2 and 9 to 0.97 for Task 1. The reliability of the total test was quite high, at 0.92.

Math: Grade 8

Task	Cronbach's Alpha
1	0.97
2	0.42
3	0.52
4	0.52
5	0.65
6	0.73
7	0.50
8	0.52
9	0.42
Total Test	0.92

Grade 11. Task reliability was moderate to high, ranging from 0.50 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	0.61
4	0.63
5	0.50
6	0.72
7	0.52
8	0.79
9	0.64
Total Test	0.93

Science

Grade 5. Task reliability for the operational items was moderate to high, ranging from 0.63 for Task 8 to 0.98 for Task 1. The reliability of the total test was quite high, at 0.96.

Science Operational: Grade 5

Task	Cronbach's Alpha
1	0.98
2	0.74
3	0.75
4	0.80
5	0.77
6	0.71
7	0.70
8	0.63
9	0.78
Total Test	0.96

Grade 8. Task reliability for the operational items was moderately low to high, ranging from 0.42 for Task 4 to 0.98 for Task 1. The reliability of the total test was quite high, at 0.94.

Science Operational: Grade 8

Task	Cronbach's Alpha
1	0.98
2	0.66
3	0.60
4	0.42
5	0.78
6	0.71
7	0.59
8	0.79
9	0.70
Total Test	0.94

Grade 11. Task reliability for the operational items was moderate to high, ranging from 0.56 for Task 7 to 0.98 for Task 1. The reliability of the total test was quite high, at 0.94.

Science Operational: Grade 11

Task	Cronbach's Alpha
1	0.98
2	0.64
3	0.74
4	0.58
5	0.59
6	0.58
7	0.56
8	0.67
9	0.61
Total Test	0.94

Descriptive Statistics

The Oregon Extended Assessments are part of a large-scale assessment system that is developed, administered, scored, and reported in concert with the professional expectations established by the *Standards*[8] (AERA et al., 1999) and best professional practices. Items are developed in an iterative manner that includes evaluation by Oregon teachers and education professionals for bias, accessibility, and alignment to the appropriate Oregon standards.

The assessments evaluate a level of student performance that has been reduced in terms of depth, breadth, and complexity in comparison to Oregon's content standards. These assessments reflect an appropriate range of performance demands (easy to difficult) to assess students with significant cognitive disabilities who exhibit a wide variety of achievement levels.

Full descriptive statistics for the reading items of the Oregon Extended Assessment are reported in *Appendix C*. All Tasks 1 were scored on a 4-point scale. All subsequent Tasks were scored on a 2-point scale. In general, the test has an appropriate range of item difficulties represented, from easy to difficult. The easiest items are located in Task 1, the prerequisite skills items. Item difficulties range from $p = .15$ (the most difficult item) to $p = .97$ (the easiest item). Item difficulties are deemed appropriate across all subject areas.

Reading: Elementary

Task 1. Items were scored on a 4-point scale. Task 1 items were quite easy, with item difficulties (p value) ranging from 0.86 – 0.97. Generally, the less difficult items had a lower standard deviation. For Tasks 2-11 all items were scored on a 2-point scale.

Task 2 (field test). Items were relatively easy overall. Item 2 was the most difficult item, $p = 0.54$, while items 1 and 3 were the easiest, $p = 0.79$. Item 1 had the lowest standard deviation (0.72) while Item 4 had the highest (0.87). Students averaged a total score of 6.53 with a standard deviation of 2.61.

Task 3 (field test). Items overall were relatively easy. Item 3 was the most difficult item, $p = 0.55$, while item 1 was the easiest, $p = 0.70$. Item 2 had the lowest standard deviation (0.80) while Item 3 had the highest (0.89). Students averaged a total score of 6.20 with a standard deviation of 2.78.

Task 4. Item 2 was the most difficult item, $p = 0.72$, while item 1 was the easiest, $p = 0.85$. Item 1 had the lowest standard deviation (0.64) while Item 5 had the highest (0.82). Students averaged a total score of 7.89 with a standard deviation of 2.49.

Task 5. Item 1 was the most difficult item, $p = 0.69$, while item 4 was the easiest, $p = 0.81$. Item 5 had the lowest standard deviation (0.73) while Item 1 had the highest (0.93). Students averaged a total score of 7.67 with a standard deviation of 2.61.

Task 6. Item 5 was the most difficult item, $p = 0.74$, while item 3 was the easiest, $p = 0.91$. Item 3 had the lowest standard deviation (0.54) while Item 5 had the highest (0.71). Students averaged a total score of 8.21 with a standard deviation of 2.31.

---

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

Task 7. Item 5 was the most difficult item, p = 0.66, while item 1 was the easiest, p = 0.90. Item 1 had the lowest standard deviation (0.55) while Item 5 had the highest (0.79). Students averaged a total score of 8.06 with a standard deviation of 2.39.

Task 8. Item 2 was the most difficult item, p = 0.70, while item 5 was the easiest, p = 0.84. Item 4 had the lowest standard deviation (0.66) while Item 1 had the highest (0.84). Students averaged a total score of 7.50 with a standard deviation of 2.59.

Task 9. Item 5 was the most difficult item, p = 0.68, while item 4 was the easiest, p = 0.87. Item 4 had the lowest standard deviation (0.64) while Item 5 had the highest (0.93). Students averaged a total score of 7.74 with a standard deviation of 2.55.

Task 10. Item 5 was the most difficult item, p = 0.58, while item 1 was the easiest, p = 0.75. Item 1 had the lowest standard deviation (0.71) while Item 5 had the highest (0.88). Students averaged a total score of 6.92 with a standard deviation of 2.63.

Task 11. Item 2 was the most difficult item, p = 0.66, while item 1 was the easiest, p = 0.68. Items 2 and 5 had the lowest standard deviation (0.68) while Item 1 had the highest (0.69). Students averaged a total score of 7.18 with a standard deviation of 2.40.

Total test. The average total test score was 58.99 with a standard deviation of 19.46. Item difficulties range from p=.54 (the most difficult item) to p=.97 (the easiest item).

Reading: Middle

Task 1. Items were scored on a 4-point scale. Task 1 items were quite easy, with item difficulties (p value) ranging from 0.86 – 0.96. Item 1 had the lowest standard deviation (0.59) and item 10 had the highest (0.89). For Tasks 2-11 all items were scored on a 2-point scale.

Task 2 (field test). Items in task 2 were more difficult in general, compared to Task 1. Item 5 was the most difficult item, p = 0.51, while items 1, 3, and 4 were the easiest, p = 0.69. Items 1 and 3 had the lowest standard deviation (0.77) while Item 5 had the highest (0.99). Students averaged a total score of 6.31 with a standard deviation of 2.71.

Task 3 (field test). Item 3 was the most difficult item, p = 0.67, while item 4 was the easiest, p = 0.79. Items 4 and 5 had the lowest standard deviation (0.75) while Item 1 had the highest (0.79). Students averaged a total score of 7.45 with a standard deviation of 2.96.

Task 4. Item 4 was the most difficult item, p = 0.49, while item 5 was the easiest, p = 0.76. Item 5 had the lowest standard deviation (0.75) while Item 4 had the highest (0.96). Students averaged a total score of 6.58 with a standard deviation of 3.29.

Task 5. Item 4 was the most difficult item, p = 0.69, while items 3 and 5 were the easiest, p = 0.80. Item 4 had the lowest standard deviation (0.70) while Item 1 had the highest (0.77). Students averaged a total score of 7.49 with a standard deviation of 2.69.

Task 6. Item 3 was the most difficult item, p = 0.77, while items 2 and 4 were the easiest, p = 0.85. Item 2 had the lowest standard deviation (0.65) while Item 1 had the highest (0.72). Students averaged a total score of 8.17 with a standard deviation of 2.55.

Task 7. Item 2 was the most difficult item, p = 0.65, while item 1 was the easiest, p = 0.85. Item 1 had the lowest standard deviation (0.63) while Item 2 had the highest (0.90). Students averaged a total score of 7.44 with a standard deviation of 2.55.

Task 8. Item 5 was the most difficult item, p = 0.66, while item 3 was the easiest, p = 0.87. Item 3 had the lowest standard deviation (0.65) while Item 5 had the highest (0.78). Students averaged a total score of 7.57 with a standard deviation of 2.50.

Task 9. Item 5 was the most difficult item, p = 0.72, while item 1 was the easiest, p = 0.87. Item 1 had the lowest standard deviation (0.61) while Item 5 had the highest (0.78). Students averaged a total score of 7.96 with a standard deviation of 2.63.

Task 10. Item 3 was the most difficult item, p = 0.66, while item 2 was the easiest, p = 0.81. Item 2 had the lowest standard deviation (0.68) while Item 5 had the highest (0.72). Students averaged a total score of 7.47 with a standard deviation of 2.65.

Task 11. Item 3 was the most difficult item, p = 0.54, while item 2 was the easiest, p = 0.79. Item 3 had the lowest standard deviation (0.63) while Item 4 had the highest (0.79). Students averaged a total score of 6.89 with a standard deviation of 2.50.

Total test. The average total test score was 56.36 with a standard deviation of 22.19. Item difficulties range from p=.49 (the most difficult item) to p=.96 (the easiest item).

Reading: High

Task 1. Items were scored on a 4-point scale. Task 1 items were quite easy, with item difficulties (p value) ranging from 0.84 – 0.96. Item 1 had the lowest standard deviation (0.57) and item 10 had the highest (1.02). For Tasks 2-11 all items were scored on a 2-point scale.

Task 2 (field test). Items in task 2 were more difficult in general, compared to Task 1. Item 5 was the most difficult item, p = 0.54, while item 3 was the easiest, p = 0.80. Item 3 had the lowest standard deviation (0.71) while Item 5 had the highest (1.00). Students averaged a total score of 6.96 with a standard deviation of 2.94.

Task 3 (field test). Item 2 was the most difficult item, p = 0.52, while items 1 was the easiest, p = 0.66. Item 3 had the lowest standard deviation (0.84) while Item 2 had the highest (0.93). Students averaged a total score of 6.11 with a standard deviation of 2.84.

Task 4. Item 4 was the most difficult item, p = 0.66, while item 3 was the easiest, p = 0.79. Item 3 had the lowest standard deviation (0.70) while Item 4 had the highest (0.82). Students averaged a total score of 7.21 with a standard deviation of 2.85.

Task 5. Item 5 was the most difficult item, p = 0.52, while item 2 was the easiest, p = 0.82. Item 2 had the lowest standard deviation (0.67) while Item 4 had the highest (0.77). Students averaged a total score of 7.26 with a standard deviation of 2.65.

Task 6. Item 4 was the most difficult item, p = 0.79, while item 5 was the easiest, p = 0.87. Item 5 had the lowest standard deviation (0.59) while Items 2 and 4 had the highest (0.69). Students averaged a total score of 8.22 with a standard deviation of 2.58.

Task 7. Item 5 was the most difficult item, p = 0.55, while item 1 was the easiest, p = 0.84. Item 4 had the lowest standard deviation (0.64) while Item 5 had the highest (0.93). Students averaged a total score of 7.34 with a standard deviation of 2.56.

Task 8. Item 5 was the most difficult item, p = 0.72, while item 1 was the easiest, p = 0.86. Item 2 had the lowest standard deviation (0.61) while Item 5 had the highest (0.75). Students averaged a total score of 8.06 with a standard deviation of 2.58.

Task 9. Item 5 was the most difficult item, p = 0.73, while item 1 was the easiest, p = 0.87. Item 1 had the lowest standard deviation (0.60) while Item 5 had the highest (0.79). Students averaged a total score of 7.95 with a standard deviation of 2.65.

Task 10. Item 2 was the most difficult item, p = 0.63, while item 4 was the easiest, p = 0.79. Item 1 had the lowest standard deviation (0.67) while Item 3 had the highest (0.73). Students averaged a total score of 7.12 with a standard deviation of 2.45.

Task 11. Item 1 was the most difficult item, p = 0.56, while item 3 was the easiest, p = 0.71. Item 1 had the lowest standard deviation (0.61) while Item 5 had the highest (0.70). Students averaged a total score of 6.52 with a standard deviation of 2.43.

Total test. The average total test score was 57.13 with a standard deviation of 21.39. Item difficulties range from p=.52 (the most difficult item) to p=.96 (the easiest item).

Writing: High School

Task 1. Items were scored on a 4-point scale. Task 1 items were quite easy, with item difficulties (p value) ranging from 0.90 – 0.97. Item 1 had the lowest standard deviation (0.51) and item 8 had the highest (0.99). For Tasks 2-11 all items were scored on a 2-point scale.

Task 2. Items in task 2 were more difficult in general, compared to Task 1. Item 3 was the most difficult item, p = 0.68, while item 1 was the easiest, p = 0.80. Item 1 had the lowest standard deviation (0.70) while Item 5 had the highest (0.75). Students averaged a total score of 7.55 with a standard deviation of 3.42.

Task 3. Item 2 was the most difficult item, p = 0.56, while items 4 was the easiest, p = 0.67. Item 2 had the lowest standard deviation (0.74) while Item 5 had the highest (0.86). Students averaged a total score of 6.28 with a standard deviation of 3.29.

Task 4. Item 4 was the most difficult item, p = 0.46, while item 1 was the easiest, p = 0.74. Item 5 had the lowest standard deviation (0.65) while Item 4 had the highest (0.84). Students averaged a total score of 5.91 with a standard deviation of 2.77.

Task 5. Item 5 was the most difficult item, p = 0.73, while item 4 was the easiest, p = 0.85. Item 3 had the lowest standard deviation (0.66) while Item 2 had the highest (0.78). Students averaged a total score of 7.92 with a standard deviation of 2.80.

Task 6. Item 5 was the most difficult item, p = 0.54, while item 2 was the easiest, p = 0.77. Items 1 and 4 had the lowest standard deviation (0.72) while Item 5 had the highest (0.80). Students averaged a total score of 6.79 with a standard deviation of 2.88.

***Task 7*.** Item 4 was the most difficult item, p = 0.61, while item 5 was the easiest, p = 0.78. Item 1 had the lowest standard deviation (0.69) while Item 4 had the highest (0.82). Students averaged a total score of 7.02 with a standard deviation of 3.00.

Task 8 (field test). Item 3 was the most difficult item, p = 0.62, while item 5 was the easiest, p = 0.75. Item 3 had the lowest standard deviation (0.64) while Item 1 had the highest (0.76). Students averaged a total score of 6.95 with a standard deviation of 2.71.

Task 9 (field test). Item 5 was the most difficult item, p = 0.54, while item 4 was the easiest, p = 0.88. Item 4 had the lowest standard deviation (0.60) while Item 5 had the highest (0.91). Students averaged a total score of 7.19 with a standard deviation of 2.56.

Task 10. Item 5 was the most difficult item, p = 0.63, while item 3 was the easiest, p = 0.90. Item 3 had the lowest standard deviation (0.56) while Item 2 had the highest (0.79). Students averaged a total score of 7.32 with a standard deviation of 2.93.

Task 11. Item 4 was the most difficult item, p = 0.59, while item 5 was the easiest, p = 0.82. Item 5 had the lowest standard deviation (0.73) while Items 3 and 4 had the highest (0.82). Students averaged a total score of 6.95 with a standard deviation of 3.37.

Total test. The average total test score was 57.78 with a standard deviation of 18.88. Item difficulties range from *p*=.46 (the most difficult item) to *p*=.97 (the easiest item).

Math: Grade 3

Task 1. Items were scored on a 4-point scale. Task 1 items were quite easy, with item difficulties (*p* value) ranging from 0.86 – 0.96. Generally, the more difficult items had a higher standard deviation than the less difficult items. For Tasks 2-9 all items were scored on a 2-point scale.

Task 2. The item range was much larger for task 2 and the items overall were more difficult. Item 2 was the most difficult item, *p* = 0.59, while item 1 was the easiest, *p* = 0.78. Item 1 had the lowest standard deviation (0.82) while item 2 had the highest (0.99). Students averaged a total task score of 7.14 with a standard deviation of 3.10.

Task 3. Item 1 was the most difficult item, *p* = 0.30, while item 6 (field test) was the easiest, *p* = 0.84. Item 1 had the lowest standard deviation (0.78) while item 3 had the highest (0.66). Students averaged a total task score of 4.36 with a standard deviation of 2.71.

Task 4. Item 5 was the most difficult item, *p* = 0.15, while item 4 was the easiest, *p* = 0.61. Item 5 had the lowest standard deviation (0.71) while item 6 (field test) had the highest (0.99). Students averaged a total task score of 3.83 with a standard deviation of 2.48.

Task 5. Item 3 was the most difficult item, *p* = 0.39, while item 6 (field test) was the easiest, *p* = 0.61. Items 3, 4, and 6 (field test) had the lowest standard deviation (0.98) while item 5 had the highest (1.00). Students averaged a total task score of 4.55 with a standard deviation of 3.03.

Task 6. Item 1 was the most difficult item, *p* = 0.41, while item 4 was the easiest, *p* = 0.68. Item 4 had the lowest standard deviation (0.93) while item 2 had the highest (0.99). Students averaged a total task score of 5.67 with a standard deviation of 2.80.

Task 7. Item 1 was the most difficult item, *p* = 0.16, while item 2 was the easiest, *p* = 0.81. Item 1had the lowest standard deviation (0.72) while items 3 and 5 had the highest (0.98). Students averaged a total task score of 5.26 with a standard deviation of 2.46.

Task 8. Item 1 was the most difficult item, *p* = 0.30, while item 5 was the easiest, *p* = 0.85. Item 5 had the lowest standard deviation (0.72) while item 2 had the highest (0.95). Students averaged a total task score of 6.76 with a standard deviation of 2.61.

Task 9. Item 5 was the most difficult item, *p* = 0.41, while item 1 was the easiest, *p* = 0.89. Item 1 had the lowest standard deviation (0.63) while item 6 (field test) had the highest (1.00). Students averaged a total task score of 7.36 with a standard deviation of 2.69.

Total test. The average total test score was 40.91 with a standard deviation of 19.30. Item difficulties range from *p*=.15 (the most difficult item) to *p*=.96 (the easiest item).

Math: Grade 4

Task 1. Items were scored on a 4-point scale. Task 1 items were quite easy, with item difficulties (*p* value) ranging from 0.88 – 0.96. Item 1 had the lowest standard deviation (0.59), while item 5 had the highest standard deviation (.90). For Tasks 2-9 all items were scored on a 2-point scale.

Task 2. The item range was much larger for task 2 and the items overall were more difficult. Item 2 was the most difficult item, $p = 0.30$, while item 6 (field test) was the easiest, $p = 0.80$. Item 6 (field test) had the lowest standard deviation (0.80) while item 3 had the highest (1.00). Students averaged a total task score of 5.91 with a standard deviation of 2.88.

Task 3. Item 3 was the most difficult item, $p = 0.24$, while item 1 was the easiest, $p = 0.68$. Item 3 had the lowest standard deviation (0.85) while item 6 (field test) had the highest (1.00). Students averaged a total task score of 4.72 with a standard deviation of 2.79.

Task 4. Item 5 was the most difficult item, $p = 0.40$, while item 3 was the easiest, $p = 0.69$. Item 3 had the lowest standard deviation (0.77) while item 6 (field test) had the highest (1.00). Students averaged a total task score of 5.50 with a standard deviation of 2.79.

Task 5. Item 5 was the most difficult item, $p = 0.60$, while item 3 was the easiest, $p = 0.82$. Item 3 had the lowest standard deviation (0.77) while item 5 had the highest (0.98). Students averaged a total task score of 6.82 with a standard deviation of 2.97.

Task 6. Items 1 and 3 were the most difficult items, $p = 0.44$, while item 5was the easiest, $p = 0.66$. Item 5 had the lowest standard deviation (0.95) while item 4 had the highest (1.00). Students averaged a total task score of 5.16 with a standard deviation of 2.88.

Task 7. Items 4 and 6 (field test) were the most difficult items, $p = 0.46$, while item 2 was the easiest, $p = 0.86$. Item 2 had the lowest standard deviation (0.71) while items 4 and 6 (field test) had the highest (1.00). Students averaged a total task score of 7.17 with a standard deviation of 2.83.

Task 8. Item 6 (field test) was the most difficult item, $p = 0.34$, while item 1 was the easiest, $p = 0.80$. Item 1 had the lowest standard deviation (0.81) while item 3 had the highest (1.00). Students averaged a total task score of 6.44 with a standard deviation of 2.95.

Task 9. Items 3 and 4 were the most difficult items, $p = 0.47$, while item 1 was the easiest, $p = 0.78$. Item 1 had the lowest standard deviation (0.84) while item 5 had the highest (1.00). Students averaged a total task score of 5.73 with a standard deviation of 2.94.

Total test. The average total test score was 43.76 with a standard deviation of 21.02. Item difficulties range from $p=.24$ (the most difficult item) to $p=.96$ (the easiest item).

Math: Grade 5

Task 1. Items were scored on a 4-point scale. Task 1 items were quite easy, with item difficulties (*p* value) ranging from 0.91 – 0.97. Item 1 had the lowest standard deviation (0.53), while item 5 had the highest standard deviation (.84). For Tasks 2-9 all items were scored on a 2-point scale.

Task 2. The item range was much larger in task 2 and the items overall were more difficult. Item 4 was the most difficult item, $p = 0.15$, while item 1 was the easiest, $p = 0.74$. Item 4 had the lowest standard deviation (0.71) while item 6 (field test) had the highest (1.00). Students averaged a total task score of 4.89 with a standard deviation of 2.51.

Task 3. Item 1 was the most difficult item, $p = 0.31$, while item 5 was the easiest, $p = 0.74$. Item 6 (field test) had the lowest standard deviation (0.91) while item 4 had the highest (0.99). Students averaged a total task score of 4.69 with a standard deviation of 2.73.

Task 4. Item 5 was the most difficult item, $p = 0.24$, while item 2 was the easiest, $p = 0.79$. Item 2 had the lowest standard deviation (0.82) while item 6 (field test) had the highest (1.00). Students averaged a total task score of 5.29 with a standard deviation of 2.64.

Task 5. Item 6 (field test) was the most difficult item, $p = 0.24$, while item 2 was the easiest, $p = 0.74$. Item 6 (field test) had the lowest standard deviation (0.86) while item 3 had the highest (1.00). Students averaged a total task score of 5.83 with a standard deviation of 3.10.

Task 6. Item 6 (field test) was the most difficult item, $p = 0.32$, while item 1 was the easiest, $p = 0.72$. Item 6 (field test) had the lowest standard deviation (0.91) while item 4 had the highest (.99). Students averaged a total task score of 5.07 with a standard deviation of 2.79.

Task 7. Item 1 was the most difficult item, $p = 0.36$, while item 3 was the easiest, $p = 0.80$. Item 3 had the lowest standard deviation (0.80) while item 4 had the highest (1.00). Students averaged a total task score of 5.00 with a standard deviation of 2.54.

Task 8. Item 3 was the most difficult item, $p = 0.29$, while item 1 was the easiest, $p = 0.86$. Item 1 had the lowest standard deviation (0.70) while item 2 had the highest (0.97). Students averaged a total task score of 5.86 with a standard deviation of 2.60.

Task 9. Item 2 was the most difficult item, $p = 0.38$, while item 1 was the easiest, $p = 0.89$. Item 1 had the lowest standard deviation (0.64) while item 3 had the highest (0.99). Students averaged a total task score of 6.64 with a standard deviation of 2.63.

Total test. The average total test score was 40.02 with a standard deviation of 18.87. Item difficulties range from $p=.15$ (the most difficult item) to $p=.97$ (the easiest item).

Math: Grade 6

Task 1. Items were scored on a 4-point scale. Task 1 items were quite easy, with item difficulties (*p* value) ranging from 0.89 – 0.97. Item 1 had the lowest standard deviation (0.47), while item 5 had the highest standard deviation (.89). For Tasks 2-9 all items were scored on a 2-point scale.

Task 2. The item range was much larger in task 2 and the items overall were more difficult. Item 2 was the most difficult item, $p = 0.36$, while item 6 (field test) was the

easiest, p = 0.69. Item 5 had the lowest standard deviation (0.93) while item 1 had the highest (1.00). Students averaged a total task score of 4.66 with a standard deviation of 2.76.

Task 3. Item 1 was the most difficult item, p = 0.46, while item 6 (field test) was the easiest, p = 0.74. Item 6 (field test) had the lowest standard deviation (0.88) while item 1 had the highest (1.00). Students averaged a total task score of 5.82 with a standard deviation of 3.27.

Task 4. Item 3 was the most difficult item, p = 0.21, while item 6 (field test) was the easiest, p = 0.63. Item 3 had the lowest standard deviation (0.80) while item 1 had the highest (1.00). Students averaged a total task score of 3.40 with a standard deviation of 2.47.

Task 5. Item 4 was the most difficult item, p = 0.35, while item 6 (field test) was the easiest, p = 0.67. Item 6 (field test) had the lowest standard deviation (0.94) while item 5 had the highest (1.00). Students averaged a total task score of 4.26 with a standard deviation of 2.81.

Task 6. Item 2 was the most difficult item, p = 0.28, while item 4 was the easiest, p = 0.68 Item 2 had the lowest standard deviation (0.89) while item 3 had the highest (0.99). Students averaged a total task score of 4.07 with a standard deviation of 2.42.

Task 7. Item 1 was the most difficult item, p = 0.24, while item 5 was the easiest, p = 0.57. Item 1 had the lowest standard deviation (0.85) while items 5 and 6 (field test) had the highest (0.99). Students averaged a total task score of 4.19 with a standard deviation of 2.58.

Task 8. Item 2 was the most difficult item, p = 0.33, while item 3 was the easiest, p = 0.64. Item 2 had the lowest standard deviation (0.94) while item 1 had the highest (1.00). Students averaged a total task score of 4.99 with a standard deviation of 2.67.

Task 9. Item 5 was the most difficult item, p = 0.38, while item 1 was the easiest, p = 0.77. Item 1 had the lowest standard deviation (0.84) while item 6 (field test) had the highest (1.00). Students averaged a total task score of 5.86 with a standard deviation of 2.82.

Total test. The average total test score was 34.27 with a standard deviation of 17.12. Item difficulties range from p=.21 (the most difficult item) to p=.97 (the easiest item).

Math: Grade 7

Task 1. Items in Task 1 were scored on a 4-point scale. Task 1 items were quite easy, with item difficulties (p value) ranging from 0.34 – 0.59. Item 1 had the lowest standard deviation (0.60), while item 5 had the highest standard deviation (.93). For Tasks 2-9 all items were scored on a 2-point scale.

Task 2. The range of items was much larger for task 2, and the items overall were more difficult. Item 5 was the most difficult item, p = 0.32, while item 1 was the easiest, p = 0.58. Item 5 had the lowest standard deviation (0.93) while item 3 had the highest (1.00). Students averaged a total task score of 5.11 with a standard deviation of 3.11.

Task 3. Item 1 was the most difficult item, p = 0.36, while item 2 was the easiest, p = 0.61. Item 1 had the lowest standard deviation (0.96) while item 3 had the highest (1.00). Students averaged a total task score of 4.68 with a standard deviation of 2.68.

Task 4. Item 6 (field test) was the most difficult item, p = 0.35, while item 5 was the easiest, p = 0.72. Item 3 had the lowest standard deviation (0.91) while item 1 had the highest (1.00). Students averaged a total task score of 5.99 with a standard deviation of 3.08.

Task 5. Item 3 was the most difficult item, p = 0.36, while item 1 was the easiest, p = 0.61. Item 2 had the lowest standard deviation (0.96) while item 4 had the highest (1.00). Students averaged a total task score of 5.13 with a standard deviation of 3.04.

Task 6. Item 4 was the most difficult item, p = 0.39, 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). Students averaged a total task score of 6.03 with a standard deviation of 2.86.

Task 7. Item 5 was the most difficult item, p = 0.24, while items 1 and 3 were the easiest, p = 0.49. Item 6 had the lowest standard deviation (0.86) while items 1 and had the highest (1.00). Students averaged a total task score of 3.93 with a standard deviation of 2.56.

Task 8. Item 3 was the most difficult item, p = 0.32, while item 6 (field test) was the easiest, p = 0.71. Item 6 (field test) had the lowest standard deviation (0.92) while item 3 had the highest (1.00). Students averaged a total task score of 4.70 with a standard deviation of 2.65.

Task 9. Item 4 was the most difficult item, p = 0.30, while items 1 was the easiest, p = 0.70. Item 4 had the lowest standard deviation (0.91) while item 5 had the highest (1.00). Students averaged a total task score of 5.51with a standard deviation of 2.81.

Total test. The average total test score was 37.49 with a standard deviation of 19.77. Item difficulties range from p=.39 (the most difficult item) to p=.76 (the easiest item).

Math: Grade 8

Task 1. Items were scored on a 4-point scale. Task 1 items were quite easy, with item difficulties (p value) ranging from 0.88 – 0.96. Item 1 had the lowest standard deviation (0.57), while item 5 had the highest standard deviation (.90). For Tasks 2-9 all items were scored on a 2-point scale.

Task 2. The range of items was much larger for task 2 and the items overall were more difficult. Item 4 was the most difficult item, p = 0.25, while item 1 was the easiest, p = 0.67. Item 4 had the lowest standard deviation (0.87) while item 6 (field test) had the highest (1.00). Students averaged a total task score of 3.57 with a standard deviation of 2.37.

Task 3. Item 5 was the most difficult item, p = 0.21, while item 6 (field test) was the easiest, p = 0.67. Item 5 had the lowest standard deviation (0.81) while items 2 and 4 had the highest (1.00). Students averaged a total task score of 3.88 with a standard deviation of 2.56.

Task 4. Item 5 was the most difficult item, p = 0.24, while item 6 (field test) was the easiest, p = 0.69. Item 5 had the lowest standard deviation (0.85) while item 1 had the highest (1.00). Students averaged a total task score of 3.87 with a standard deviation of 2.59.

Task 5. Item 1 was the most difficult item, p = 0.30, while item 4 was the easiest, p = 0.73. Item 4 had the lowest standard deviation (0.89) while item 6 (field test) had the

highest (1.00). Students averaged a total task score of 4.99 with a standard deviation of 2.90.

Task 6. Item 6 (field test) was the most difficult item, p = 0.56, while item 5 was the easiest, p = 0.78. Item 5 had the lowest standard deviation (0.84) while items 2 and 6 had the highest (0.99). Students averaged a total task score of 6.72 with a standard deviation of 3.15.

Task 7. Item 3 was the most difficult items, p = 0.24, while item 5 was the easiest, p = 0.76. Items 3 and 5 had the lowest standard deviation (0.86) while item 2 had the highest (1.00). Students averaged a total task score of 4.37 with a standard deviation of 2.47.

Task 8. Item 1 was the most difficult item, p = 0.31, while item 6 (field test) was the easiest, p = 0.73. Item 6 (field test) had the lowest standard deviation (0.89) while item 3 had the highest (1.00). Students averaged a total task score of 4.34 with a standard deviation of 2.69.

Task 9. Item 3 was the most difficult item, p = 0.20, while items 6 (field test) was the easiest, p = 0.50. Item 3 had the lowest standard deviation (0.79) while item 6 (field test) had the highest (1.00). Students averaged a total task score of 3.22 with a standard deviation of 2.48.

Total test. The average total test score was 30.98with a standard deviation of 17.78. Item difficulties range from p=.20 (the most difficult item) to p=.96 (the easiest item).

Math: Grade 11

Task 1. Items were scored on a 4-point scale. Task 1 items were quite easy, with item difficulties (p value) ranging from 0.89 – 0.96. Item 1 had the lowest standard deviation (0.59), while item 5 had the highest standard deviation (0.99). For Tasks 2-9 all items were scored on a 2-point scale.

Task 2. The range was much larger, and the items overall were more difficult. Item 1 was the most difficult item, p = 0.28, while item 2 was the easiest, p = 0.60. Item 4 had the lowest standard deviation (0.77) while items 5 and 6 (field test) had the highest (1.00). Students averaged a total task score of 4.57 with a standard deviation of 2.74.

Task 3. Items 3 and 4 were the most difficult items, p = 0.29, while item 2 was the easiest, p = 0.54. Item 1 had the lowest standard deviation (0.87) while item 2 had the highest (0.99). Students averaged a total task score of 4.03 with a standard deviation of 2.81.

Task 4. Item 6 (field test) was the most difficult items, p = 0.33, while item 4 was the easiest, p = 0.62. Item 5 had the lowest standard deviation (0.78) while items 1 and 2 had the highest (1.00). Students averaged a total task score of 5.02 with a standard deviation of 2.87.

Task 5. Item 3 was the most difficult item, p = 0.28, while item 6 (field test) was the easiest, p = 0.73. Item 5 had the lowest standard deviation (0.87) while item 2 had the highest (0.98). Students averaged a total task score of 3.84 with a standard deviation of 2.48.

Task 6. Item 3 was the most difficult item, p = 0.57, while item 5 was the easiest, p = 0.80. Item 5 had the lowest standard deviation (0.73) while item 3 had the highest (0.99). Students averaged a total task score of 6.64 with a standard deviation of 2.94.

Task 7. Item 2 was the most difficult items, p = 0.20, while item 6 (field test) was the easiest, p = 0.54. Item 2 had the lowest standard deviation (0.79) while items 1, 4, 5, and 6

had the highest (1.00). Students averaged a total task score of 4.18 with a standard deviation of 2.70.

Task 8. Item 1 was the most difficult item, p = 0.63, while item 6 (field test) was the easiest, p = 0.84. Item 6 had the lowest standard deviation (0.75) while item 2 had the highest (0.92). Students averaged a total task score of 7.28 with a standard deviation of 2.96.

Task 9. Item 3 was the most difficult item, p = 0.40, while item 2 was the easiest, p = 0.87. Item 5 had the lowest standard deviation (0.66) while item 4 had the highest (0.99). Students averaged a total task score of 5.82 with a standard deviation of 2.50.

Total test. The average total test score was 36.72 with a standard deviation of 19.65. Item difficulties range from p=.20 (the most difficult item) to p=.96 (the easiest item).

Science: Grade 5

Task 1. Items were scored on a 4-point scale. Task 1 items were quite easy, with item difficulties (p value) ranging from 0.94 – 0.96. Generally, the more difficult items had a higher standard deviation than the less difficult items. For Tasks 2-9 all items were scored on a 2-point scale.

Task 2. Item in task 2 had a much larger range and the items overall were more difficult. Item 6 (field test) was the most difficult item, p = 0.48, while item 1 was the easiest, p = 0.88. Item 1 had the lowest standard deviation (0.65) while item 6 had the highest (1.00). Students averaged a total task score of 7.73 with a standard deviation of 2.94.

Task 3. Item 4 was the most difficult item, p = 0.39, while item 1 was the easiest, p = 0.85. Item 1 had the lowest standard deviation (0.71) while item 4 had the highest (0.97). Students averaged a total task score of 6.86 with a standard deviation of 2.88.

Task 4. Item 1 was the most difficult item, p = 0.73, while item 6 (field test) was the easiest, p = 0.86. Item 6 had the lowest standard deviation (0.69) while item 1 was the highest (0.89). Students averaged a total task score of 7.85 with a standard deviation of 2.90.

Task 5. Item 2 was the most difficult item, p = 0.47, while item 5 was the easiest, p = 0.87. Item 5 had the lowest standard deviation (0.68) while items 2 and 6 (field test) had the highest (1.00). Students averaged a total task score of 7.37 with a standard deviation of 2.93.

Task 6. Item 4 was the most difficult item, p = 0.42, while item 5 was the easiest, p = 0.89. Item 5 had the lowest standard deviation (0.64) while item 4 had the highest (0.99). Students averaged a total task score of 7.26 with a standard deviation of 2.59.

Task 7. Item 2 was the most difficult item, p = 0.56, while item 1 was the easiest, p = 0.87. Item 1 had the lowest standard deviation (0.67) while items 2 and 5 had the highest (0.99). Students averaged a total task score of 6.66 with a standard deviation of 2.94.

Task 8. Item 4 was the most difficult item, p = 0.46, while item 6 (field test) was the easiest, p = 0.84. Item 6 had the lowest standard deviation (0.74) while items 1-4 had the highest (1.00). Students averaged a total task score of 5.38 with a standard deviation of 2.91.

Task 9. Item 4 was the most difficult item, p = 0.69, while item 5 was the easiest, p = 0.83. Item 5 had the lowest standard deviation (0.75) while item 4 had the highest (0.92). Students averaged a total task score of 7.77 with a standard deviation of 2.84.

Field test items. Item 6 in task 2 was the most difficult item, p = 0.48, while item 6 in task 6 was the easiest item, p = 0.87. Item 6 in task 6 had the lowest standard deviation (0.67), while items 6 in tasks 2 and 5 had the highest (1.00).

Total test. The average total test score for the operational items was 52.63 with a standard deviation of 23.09. Item difficulties range from p=.39 (the most difficult item) to p=.96 (the easiest item).

Science: Grade 8

Task 1. Items were scored on a 4-point scale. Task 1 items were quite easy, with item difficulties (p value) ranging from 0.92 – 0.96. Item 1 had the lowest standard deviation (0.60), while item 3 had the highest standard deviation (.86). For Tasks 2-9 all items were scored on a 2-point scale.

Task 2. The range of items was much larger for task 2 and the items overall were more difficult. Item 5 was the most difficult item, p = 0.24, while item 1 was the easiest, p = 0.78. Items 1 and 2 had the lowest standard deviation (0.84) while item 3 had the highest (1.00). Students averaged a total task score of 5.94 with a standard deviation of 2.71.

Task 3. Item 6 (field test) was the most difficult item, p = 0.38, while item 4 was the easiest, p = 0.63. Items 4 and 6 had the lowest standard deviation (0.97) while items 1, 2, 3, and 5 had the highest (1.00). Students averaged a total task score of 5.18 with a standard deviation of 3.00.

Task 4. Item 3 was the most difficult item, p = 0.26, while item 1 was the easiest, p = 0.85. Item 1 had the lowest standard deviation (0.71) while item 6 (field test) had the highest (1.00). Students averaged a total task score of 4.65 with a standard deviation of 2.46.

Task 5. Item 1 was the most difficult item, p = 0.37, while item 6 (field test) was the easiest, p = 0.86. Item had the lowest standard deviation (0.70) while item 1 had the highest (0.96). Students averaged a total task score of 7.05 with a standard deviation of 2.81.

Task 6. Item 6 (field test) was the most difficult item, p = 0.61, while item 3 was the easiest, p = 0.71. Item 3 had the lowest standard deviation (0.91) while item 6 had the highest (0.98). Students averaged a total task score of 6.75 with a standard deviation of 3.10.

Task 7. Item 6 (field test) was the most difficult item, p = 0.37, while item 1 was the easiest, p = 0.88. Item 1 had the lowest standard deviation (0.66) while item 3 had the highest (1.00). Students averaged a total task score of 6.66 with a standard deviation of 2.70.

Task 8. Item 1 was the most difficult item, p = 0.54, while item 2 was the easiest, p = 0.83. Item 2 had the lowest standard deviation (0.76) while item 1 had the highest (1.00). Students averaged a total task score of 7.10 with a standard deviation of 3.07.

Task 9. Items 3 and 6 (field test) were the most difficult items, p = 0.59, while item 1 was the easiest, p = 0.78. Item 1 had the lowest standard deviation (0.84) while items 2, 3, and 6 had the highest (0.98). Students averaged a total task score of 6.67 with a standard deviation of 3.06.

Field test items. In grade 8, item 6 in task 7 was the most difficult item, p = 0.37, while item 6 in task 5 was the easiest item, p = 0.86. Item 6 in task 5 had the lowest standard deviation (0.70), while item 6 in task 4 had the highest (1.00).

Total test. The average total test score for the operational items was 45.15 with a standard deviation of 22.10. Item difficulties range from p=.24 (the most difficult item) to p=.96 (the easiest item).

Science: High School

Task 1. Items were scored on a 4-point scale. Task 1 items were quite easy, with item difficulties (p value) ranging from 0.92 – 0.97. Item 1 had the lowest standard deviation (0.50), while item 9 had the highest standard deviation (.87). For Tasks 2-9 all items were scored on a 2-point scale.

Task 2. The range of items was much larger for task 2 and the items overall were more difficult. Item 5 was the most difficult item, p = 0.33, while item 2 was the easiest, p = 0.69. Item 2 had the lowest standard deviation (0.93) while item 3 had the highest (1.00). Students averaged a total task score of 5.21 with a standard deviation of 3.05.

Task 3. Item 6 (field test) was the most difficult item, p = 0.52, while item 1 was the easiest, p = 0.82. Item 1 had the lowest standard deviation (0.77) while item 6 had the highest (1.00). Students averaged a total task score of 6.88 with a standard deviation of 3.11.

Task 4. Item 6 (field test) was the most difficult item, p = 0.37, while item 4 was the easiest, p = 0.82. Item 4 had the lowest standard deviation (0.77) while item 2 had the highest (.97). Students averaged a total task score of 6.57 with a standard deviation of 2.79.

Task 5. Item 1 was the most difficult item, p = 0.47, while item 4 was the easiest, p = 0.89. Item 4 had the lowest standard deviation (0.63) while item 1 had the highest (1.00). Students averaged a total task score of 6.42 with a standard deviation of 2.64.

Task 6. Item 4 was the most difficult item, p = 0.39, while item 2 was the easiest, p = 0.87. Item 2 had the lowest standard deviation (0.67) while item 1 had the highest (0.99). Students averaged a total task score of 6.42 with a standard deviation of 2.46.

Task 7. Item 3 was the most difficult item, p = 0.40, while item 1 was the easiest, p = 0.79. Item 1 had the lowest standard deviation (0.82) while items 2 and 6 (field test) had the highest (1.00). Students averaged a total task score of 5.89 with a standard deviation of 2.71.

Task 8. Item 2 was the most difficult item, p = 0.56, while item 4 was the easiest, p = 0.90. Item 4 had the lowest standard deviation (0.61) while item 2 had the highest (0.99). Students averaged a total task score of 7.61 with a standard deviation of 2.62.

Task 9. Item 5 was the most difficult item, p = 0.41, while items 3 and 4 were the easiest, p = 0.91. Items 3 and 4 had the lowest standard deviation (0.59) while items 1 and 6 (field test) had the highest (0.99). Students averaged a total task score of 6.92 with a standard deviation of 2.42.

Field test items. In grade 11, item 6 in task 4 was the most difficult item, p = 0.37, while item 6 in task 6 was the easiest item, p = 0.73. Item 6 in task 6 had the lowest standard deviation (0.89), while item 6 in task 7 had the highest (1.00).

Total test. The average total test score was 47.42 with a standard deviation of 21.74. Item difficulties range from p=.33 (the most difficult item) to p=.97 (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 Oregon Extended Assessment. 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 a students score in one area "went along with" the 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 and Raw Score
2	Administration type Pre-requisite task total	Total Scale and Raw Score
3	Disability category Administration type Race/Ethnicity	Prerequisite Total Score
4	Disability category Administration type Race/Ethnicity	Total Scale and Raw 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? Model 1 tests the strength of this relationship. 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 the 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 55-58. Overall, the model accounted for between 32% - 56% 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 59 - 67. Overall, the full model accounted for between 44% - 79% of the total variance across subjects and grades. In reading and writing test administration type generally accounted for the most unique variance, while in math and science pre-requisite skills task total accounted for the most unique variance.

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. Referent groups included students who were classified with an intellectual disability (formerly known as mental retardation), took the standard administration of the test, and were White. Each referent 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 students' prerequisite skills score, given that the task is intended to assess students' 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, students' 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 12% - 21% of the total variance across subjects and grades. Test administration type accounted for additional variance beyond students' disability (6% - 17%) and was generally the largest predictor. Students' race/ethnicity accounted for minimal variance when added in the third block (0% - 2%), 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 raw and 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 - 78. Overall, students' disability classification accounted for between 10% - 26% of the total variance across subjects and grades. Test administration type accounted for additional variance beyond students' disability (10% - 26%) and was generally the largest predictor. 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 Kolmogorov-Smirnov and Shapiro-Wilk tests of normality were used to examine the assumption of normality for all dependent variables. Frequency distributions and box-plots were also produced to visually interpret the assumption of normality. For all variables, the tests of normality were significant, indicating a non-normal distribution, and a visual examination of the frequency and box-plots confirmed non-normal distributions. However, the central limit theorem protects regression analyses from departures of normality as long as the sample size is reasonably large. Scatterplots were created for each predictor variable and corresponding dependent variable to examine the assumption of linearity. In all cases, the relation between the variables was roughly linear. Finally, the residuals were examined for normality with P-P and Q-Q plots, which revealed roughly normal distributions. The assumption of multicollinearity was investigated with the Tolerance and Variance Inflation Factor (VIF) statistics. All predictor variables were within the acceptable range, and are reported in the appendices for the respective models.

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, $r(877) = .809$, $p < .05$. At grade 4, the correlation was of a similar magnitude $r(909) = .789$, $p < .05$. At grade 5, reading and math had a moderately strong correlation, $r(778) = .762$, $p < .05$. The correlation between reading and science ($n = 627$) and math and science ($n = 672$) were moderate and significant, ranging in the .70's. At grade 6, reading and math had a moderately strong correlation, $r(748) = .754$, $p < .05$. At grade 7, the correlation was of a similar magnitude $r(573) = .784$, $p < .05$. At grade 8, correlation between reading and math ($n = 567$), reading and science ($n = 543$), and math and science ($n = 588$) were all moderately strong and statistically significant, with Pearson's r in the .70's to 80's. Finally, at grade 11, reading was statistically correlated to writing ($n = 428$), math ($n = 429$), and science ($n = 427$), with Pearson's r in the .70's to .80's. Math was statistically correlated to writing ($n = 466$) and science ($n = 474$), with Pearson's r in the .80's. Writing and science had a moderately strong correlation, $r(463) = .854$, $p < .05$.

Model 1 Results: Pre-req on Raw and Scale Scores

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

Reading

Elementary: Scale Score. The regression of pre-requisite skills on students' scale score was statistically significant, $F(1, 3295) = 2588.53$, MSR = 109.05, $p < .05$, $R^2 = 0.44$. Pre-requisite skills task total was a statistically significant predictor of students' scale score, $b = 1.98$, SE = .04, $p < .05$, 95% CI = 1.90 to 2.05. On average, every one-point increase in the pre-requisite skills task total corresponded with a 1.98 increase in students' scale scores.

Elementary: Raw Score. The regression of pre-requisite skills on students' raw score was statistically significant, $F(1, 3295) = 2956.69$, MSR = 198.62, $p < .05$, $R^2 = 0.47$. Pre-requisite skills task total was a statistically significant predictor of students' raw score, $b = 2.85$, SE = .05, $p < .05$, 95% CI = 2.75 to 2.96. On average, every one-point increase in the pre-requisite skills task total corresponded with a 2.85 increase in students' raw scores.

Middle School: Scale Score. The regression of pre-requisite skills on students' scale score was statistically significant, $F(1, 2116) = 1717.71$, MSR =140.81, $p < .05$, $R^2 = 0.45$. Pre-requisite skills task total was a statistically significant predictor of students' scale score, $b = 1.97$, SE = .05, $p < .05$, 95% CI = 1.87 to 2.06. On average, every one-point increase in the pre-requisite skills task total corresponded with a 1.97 increase in students' scale scores.

Middle School: Raw Score. The regression of pre-requisite skills on students' scale score was statistically significant, $F(1, 2116) = 1711.04$, MSR = 272.40, $p < .05$, $R^2 = 0.45$. Pre-requisite skills task total was a statistically significant predictor of students' scale score, $b = 2.73$, SE = .07, $p < .05$, 95% CI = 2.60 to 2.86. On average, every one-point increase in the pre-requisite skills task total corresponded with a 2.73 increase in students' raw scores.

High School: Scale Score. The regression of pre-requisite skills on students' scale score was statistically significant, $F(1, 434) = 386.07$, MSR = 193.37, $p < .05$, $R^2 = 0.47$. Pre-requisite skills task total was a statistically significant predictor of students' scale score, $b = 2.24$, SE = .11, $p < .05$, 95% CI = 2.02 to 2.47. On average, every one-point increase in the pre-requisite skills task total corresponded with a 2.24 increase in students' scale scores.

High School: Raw Score. The regression of pre-requisite skills on students' scale score was statistically significant, $F(1, 434) = 532.64$, MSR = 206.34, $p < .05$, $R^2 = 0.55$. Pre-requisite skills task total was a statistically significant predictor of students' scale score, $b = 2.72$, SE = .12, $p < .05$, 95% CI = 2.49 to 2.95. On average, every one-point increase in the pre-requisite skills task total corresponded with a 2.72 increase in students' raw scores.

Writing

Grade 11: Scale Score. The regression of pre-requisite skills on students' scale score was statistically significant, $F(1, 476) = 600.32$, MSR = 207.71, $p < .05$, $R^2 = 0.56$. Pre-requisite skills task total was a statistically significant predictor of students' scale score, $b = 2.38$, SE = .10, $p < .05$, 95% CI = 2.19 to 2.37. On average, every one-point increase in the pre-requisite skills task total corresponded with a 2.38 increase in students' scale scores.

Grade 11: Raw Score. The regression of pre-requisite skills on students' scale score was statistically significant, $F(1, 476) = 444.01$, MSR = 347.42, $p < .05$, $R^2 = 0.48$. Pre-requisite skills task total was a statistically significant predictor of students' scale score, $b = 2.64$, SE = .13, $p < .05$, 95% CI = 2.40 to 2.89. On average, every one-point increase in the pre-requisite skills task total corresponded with a 2.64 increase in students' scale scores.

Math

Grade 3: Scale Score. The regression of pre-requisite skills on students' scale score was statistically significant, $F(1, 971) = 1076.95$, MSR = 48.91, $p < .05$, $R^2 = 0.53$. Pre-requisite skills task total was a statistically significant predictor of students' scale score, $b = 1.17$, SE = .04, $p < .05$, 95% CI = 1.10 to 1.24. On average, every one-point increase in the pre-requisite skills task total corresponded with a 1.17 increase in students' scale scores.

Grade 3: Raw Score. The regression of pre-requisite skills on students' raw score was statistically significant, $F(1, 971) = 788.74$, MSR = 205.91, $p < .05$, $R^2 = 0.45$. Pre-requisite skills task total was a statistically significant predictor of students' raw score, $b = 2.07$, SE = .07, $p < .05$, 95% CI = 1.92 to 2.21. On average, every one-point increase in the pre-requisite skills task total corresponded with a 2.07 increase in students' raw scores.

Grade 4: Scale Score. The regression of pre-requisite skills on students' scale score was statistically significant, $F(1, 1049) = 811.58$, MSR = 77.90, $p < .05$, $R^2 = 0.44$. Pre-requisite skills task total was a statistically significant predictor of students' scale score, $b = 1.23$, SE = .04, $p < .05$, 95% CI = 1.15 to 1.32. On average, every one-point increase in the pre-requisite skills task total corresponded with a 1.23 increase in students' scale scores.

Grade 4 Raw Score. The regression of pre-requisite skills on students' raw score was statistically significant, $F(1, 1049) = 695.74$, MSR = 265.98, $p < .05$, $R^2 = 0.39$. Pre-requisite skills task total was a statistically significant predictor of students' raw score, $b = 2.11$, SE = .08, $p < .05$, 95% CI = 1.95 to 2.27. On average, every one-point increase in the pre-requisite skills task total corresponded with a 2.11 increase in students' raw scores.

Grade 5: Scale Score. The regression of pre-requisite skills on students' scale score was statistically significant, $F(1, 908) = 671.88$, MSR = 48.09, $p < .05$, $R^2 = 0.43$. Pre-requisite skills task total was a statistically significant predictor of students' scale score, $b = 1.07$, SE = .04, $p < .05$, 95% CI =.99 to 1.15. On average, every one-point increase in the pre-requisite skills task total corresponded with a 1.07 increase in students' scale scores.

Grade 5: Raw Score. The regression of pre-requisite skills on students' raw score was statistically significant, $F(1, 908) = 474.20$, MSR = 234.31, $p < .05$, $R^2 = 0.34$. Pre-requisite skills task total was a statistically significant predictor of students' raw score, $b = 1.98$, SE = .09, $p < .05$, 95% CI = 1.80 to 2.15. On average, every one-point increase in the pre-requisite skills task total corresponded with a 1.98 increase in students' raw scores.

Grade 6: Scale Score. The regression of pre-requisite skills on students' scale score was statistically significant, $F(1, 889) = 853.60$, MSR = 38.12, $p < .05$, $R^2 = 0.49$. Pre-requisite skills task total was a statistically significant predictor of students' scale score, $b = 0.96$, SE = .03, $p < .05$, 95% CI = 0.90 to 1.02. On average, every one-point increase in the pre-requisite skills task total corresponded with a 0.96 increase in students' scale scores.

Grade 6: Raw Score. The regression of pre-requisite skills on students' raw score was statistically significant, $F(1, 889) = 498.27$, MSR = 187.92, $p < .05$, $R^2 = 0.36$. Pre-requisite skills task total was a statistically significant predictor of students' raw score, $b = 1.63$, SE = .08, $p < .07$, 95% CI = 1.49 to 1.77. On average, every one-point increase in the pre-requisite skills task total corresponded with a 1.63 increase in students' raw scores.

Grade 7: Scale Score. The regression of pre-requisite skills on students' scale score was statistically significant, $F(1, 673) = 700.51$, MSR = 43.46, $p < .05$, $R^2 = 0.51$. Pre-requisite skills task total was a statistically significant predictor of students' scale score, $b = 1.03$, SE = .04, $p < .05$, 95% CI = 0.95 to 1.10. On average, every one-point increase in the pre-requisite skills task total corresponded with a 1.03 increase in students' scale scores.

Grade 7: Raw Score. The regression of pre-requisite skills on students' raw score was statistically significant, $F(1, 673) = 392.14$, MSR = 247.39, $p < .05$, $R^2 = 0.35$. Pre-requisite skills task total was a statistically significant predictor of students' raw score, $b = 1.83$, SE = .10, $p < .05$, 95% CI = 1.65 to 2.01. On average, every one-point increase in the pre-requisite skills task total corresponded with a 1.83 increase in students' raw scores.

Grade 8: Scale Score. The regression of pre-requisite skills on students' scale score was statistically significant, $F(1, 660) = 542.62$, MSR = 38.65, $p < .05$, $R^2 = 0.37$. Pre-requisite skills task total was a statistically significant predictor of students' scale score, $b = 0.88$, SE = .04, $p < .05$, 95% CI = 0.80 to 0.95. On average, every one-point increase in the pre-requisite skills task total corresponded with a 0.88 increase in students' scale scores.

Grade 8: Raw Score. The regression of pre-requisite skills on students' raw score was statistically significant, $F(1, 660) = 316.83$, MSR = 214.00, $p < .05$, $R^2 = 0.32$. Pre-requisite skills task total was a statistically significant predictor of students' raw score, $b = 1.57$, SE = .09, $p < .05$, 95% CI = 1.40 to 1.75. On average, every one-point increase in the pre-requisite skills task total corresponded with a 1.57 increase in students' raw scores.

Grade 11: Scale Score. The regression of pre-requisite skills on students' scale score was statistically significant, $F(1, 496) = 488.44$, MSR = 62.18, $p < .05$, $R^2 = 0.50$. Pre-requisite skills task total was a statistically significant predictor of students' scale score, $b = 1.12$, SE

= .05, $p < .05$, 95% CI = 1.02 to 1.22. On average, every one-point increase in the pre-requisite skills task total corresponded with a 1.12 increase in students' scale scores.

Grade 11: Raw Score. The regression of pre-requisite skills on students' raw score was statistically significant, $F(1, 491) = 322.75$, MSR = 233.92, $p < .05$, $R^2 = 0.39$. Pre-requisite skills task total was a statistically significant predictor of students' raw score, $b = 1.77$, SE = .10, $p < .05$, 95% CI = 1.57 to 1.96. On average, every one-point increase in the pre-requisite skills task total corresponded with a 1.77 increase in students' raw scores.

Science

Grade 5: Scale Score. The regression of pre-requisite skills on students' scale score was statistically significant, $F(1, 685) = 524.28$, MSR = 70.83, $p < .05$, $R^2 = 0.43$. Pre-requisite skills task total was a statistically significant predictor of students' scale score, $b = 1.19$, SE = .05, $p < .05$, 95% CI = 1.09 to 1.29. 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: Raw Score. The regression of pre-requisite skills on students' raw score was statistically significant, $F(1, 685) = 438.54$, MSR = 325.56, $p < .05$, $R^2 = 0.39$. Pre-requisite skills task total was a statistically significant predictor of students' raw score, $b = 2.33$, SE = .11, $p < .05$, 95% CI = 2.11 to 2.55. On average, every one-point increase in the pre-requisite skills task total corresponded with a 2.33 increase in students' raw scores.

Grade 8: Scale Score. The regression of pre-requisite skills on students' scale score was statistically significant, $F(1, 601) = 627.53$, MSR = 42.17, $p < .05$, $R^2 = 0.51$. Pre-requisite skills task total was a statistically significant predictor of students' scale score, $b = 1.06$, SE = .04, $p < .05$, 95% CI = 0.97 to 1.14. On average, every one-point increase in the pre-requisite skills task total corresponded with a 1.06 increase in students' scale scores.

Grade 8: Raw Score. The regression of pre-requisite skills on students' raw score was statistically significant, $F(1, 601) = 350.48$, MSR = 307.32, $p < .05$, $R^2 = 0.37$. Pre-requisite skills task total was a statistically significant predictor of students' raw score, $b = 2.13$, SE = .11, $p < .05$, 95% CI = 1.91 to 2.36. On average, every one-point increase in the pre-requisite skills task total corresponded with a 2.13 increase in students' raw scores.

Grade 11: Scale Score. The regression of pre-requisite skills on students' scale score was statistically significant, $F(1, 480) = 447.73$, MSR = 49.64, $p < .05$, $R^2 = 0.48$. Pre-requisite skills task total was a statistically significant predictor of students' scale score, $b = 1.02$, SE = .05, $p < .05$, 95% CI = 0.93 to 1.12. On average, every one-point increase in the pre-requisite skills task total corresponded with a 1.02 increase in students' scale scores.

Grade 11: Raw Score. The regression of pre-requisite skills on students' raw score was statistically significant, $F(1, 480) = 353.84$, MSR = 271.23, $p < .05$, $R^2 = 0.42$. Pre-requisite skills task total was a statistically significant predictor of students' raw score, $b = 2.12$, SE = .11, $p < .05$, 95% CI = 1.90 to 2.34. On average, every one-point increase in the pre-requisite skills task total corresponded with a 2.12 increase in students' raw score.

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: Scale Score. The regression of pre-requisite skills and test administration type on students' scale score was statistically significant, $F(2, 3294) = 1494.039$, MSR = 102.13, $p < .05$, $R^2 = .48$. Test administration type was a statistically significant predictor of students' scale score, $b = -6.83$, SE = .46, $p < .05$, 95% CI = -7.72 to -5.93. On average, students taking the scaffold administration scored 6.83 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, $b = 1.70$, SE = .04, $p < .05$, 95% CI = 1.61 to 1.78. On average, every one-point increase in the pre-requisite skills task total corresponded with a 1.70 increase in students' scale scores. Examination of the squared semipartial correlations revealed that approximately 4% of the total scale score variance was uniquely accounted for by test administration type, while 26% was uniquely accounted for by the pre-requisite task total. Together test administration type and Pre-requisite skills accounted for 49% of the total variability in scale scores.

Elementary: Raw Score. The regression of pre-requisite skills and test administration type on students' raw score was statistically significant, $F(2, 3294) = 1850.070$, MSR = 177.537, $p < .05$, $R^2 = .53$. Test administration type was a statistically significant predictor of students' raw score, $b = -11.91$, SE = .60, $p < .05$, 95% CI = -13.08 to -10.73. On average, students taking the scaffold administration scored 11.91 raw score points lower than students taking the standard administration. Students' pre-requisite task total was also a statistically significant predictor of students' aw score, $b = 2.36$, SE = .06, $p < .05$, 95% CI = 2. 25 to 2. 47. On average, every one-point increase in the pre-requisite skills task total corresponded with a 2.36 increase in students' raw scores. Examination of the squared semipartial correlations revealed that approximately 6% of the total raw score variance was uniquely accounted for by test administration type, while 26% was uniquely accounted for by the pre-requisite task total. Together test administration type and Pre-requisite skills accounted for 53% of the total variability in raw scores.

Middle School: Scale Score. The regression of pre-requisite skills and test administration type on students' scale score was statistically significant, $F(2, 2115) = 1199.447$, MSR = 119.594, $p < .05$, $R^2 = .53$. Test administration type was a statistically significant predictor of students' scale score, $b = -10.85$, SE = .56, $p < .05$, 95% CI = -11.95 to -9.76. On average, students taking the scaffold administration scored 10.85 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, $b = 1.54$, SE = .05, $p < .05$, 95% CI = 1.44 to 1.63. On average, every one-point increase in the pre-requisite skills task total corresponded with a 1.54 increase in students' scale scores. Examination of the squared semipartial correlations revealed that approximately 8% of the total scale score variance was uniquely accounted for by test administration type, while 22% was uniquely

accounted for by the pre-requisite task total. Together test administration type and Pre-requisite skills accounted for 523% of the total variability in scale scores.

Middle School: Raw Score. The regression of pre-requisite skills and test administration type on students' raw score was statistically significant, $F(2, 2115) = 1314.981$, MSR = 219.702, $p < .05$, $R^2 = .55$. Test administration type was a statistically significant predictor of students' raw score, $b = -17.10$, SE = .76, $p < .05$, 95% CI = -18.59 to -15.61. On average, students taking the scaffold administration scored 17.10 raw score points lower than students taking the standard administration. Students' pre-requisite task total was also a statistically significant predictor of students' aw score, $b = 2.06$, SE = .07, $p < .05$, 95% CI = 1.93 to 2.19. On average, every one-point increase in the pre-requisite skills task total corresponded with a 2.06 increase in students' raw scores. Examination of the squared semipartial correlations revealed that approximately 11% of the total raw score variance was uniquely accounted for by test administration type, while 20% 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 raw scores.

High School: Scale Score. The regression of pre-requisite skills and test administration type on students' scale score was statistically significant, $F(2, 433) = 236.401$, MSR = 175.068, $p < .05$, $R^2 = .52$. Test administration type was a statistically significant predictor of students' scale score, $b = -9.61$, SE = 1.41, $p < .05$, 95% CI = -12.38 to -6.84. On average, students taking the scaffold administration scored 9.61 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 = .12, $p < .05$, 95% CI = 1.68 to 2.15. 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 5% of the total scale score variance was uniquely accounted for by test administration type, while 29% was uniquely accounted for by the pre-requisite task total. Together test administration type and Pre-requisite skills accounted for 52% of the total variability in scale scores.

High School: Raw Score. The regression of pre-requisite skills and test administration type on students' raw score was statistically significant, $F(2, 433) = 347.244$, MSR = 176.900, $p < .05$, $R^2 = .62$. Test administration type was a statistically significant predictor of students' raw score, $b = -12.14$, SE = 1.42, $p < .05$, 95% CI = -14.93 to -9.35. On average, students taking the scaffold administration scored 12.14 raw score points lower than students taking the standard administration. Students' pre-requisite task total was also a statistically significant predictor of students' aw score, $b = 2.30$, SE = .12, $p < .05$, 95% CI = 2.07 to 2.54. On average, every one-point increase in the pre-requisite skills task total corresponded with a 2.30 increase in students' raw scores. Examination of the squared semipartial correlations revealed that approximately 7% of the total raw score variance was uniquely accounted for by test administration type, while 33% was uniquely accounted for by the pre-requisite task total. Together test administration type and Pre-requisite skills accounted for 62% of the total variability in raw scores.

Writing

Grade 11: Scale Score. The regression of pre-requisite skills and test administration type on students' scale score was statistically significant, $F(2, 475) = 394.198$, MSR = 176.948, $p < .05$, $R^2 = .62$. Test administration type was a statistically significant predictor of students' scale score, $b = -12.27$, SE = 1. 34, $p < .05$, 95% CI = -14.90 to -9.63. On average, students taking the scaffold administration scored 12.27 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, $b = 2.01$, SE = .10, $p < .05$, 95% CI = 1.82 to 2.20. On average, every one-point increase in the pre-requisite skills task total corresponded with a 2.01 increase in students' scale scores. Examination of the squared semipartial correlations revealed that approximately 7% of the total scale score variance was uniquely accounted for by test administration type, while 33% was uniquely accounted for by the pre-requisite task total. Together test administration type and Pre-requisite skills accounted for 62% of the total variability in scale scores.

Grade 11: Raw Score. The regression of pre-requisite skills and test administration type on students' raw score was statistically significant, $F(2, 475) = 350.027$, MSR = 272.007, $p < .05$, $R^2 = .60$. Test administration type was a statistically significant predictor of students' raw score, $b = -19.17$, SE = 1.66, $p < .05$, 95% CI = -22.43 to -15.90. On average, students taking the scaffold administration scored 19.17 raw score points lower than students taking the standard administration. Students' pre-requisite task total was also a statistically significant predictor of students' aw score, $b = 2.07$, SE = .12, $p < .05$, 95% CI = 1.83 to 2.31. On average, every one-point increase in the pre-requisite skills task total corresponded with a 2.07 increase in students' raw scores. Examination of the squared semipartial correlations revealed that approximately 11% of the total raw score variance was uniquely accounted for by test administration type, while 25% was uniquely accounted for by the pre-requisite task total. Together test administration type and Pre-requisite skills accounted for 60% of the total variability in raw scores.

Math

Grade 3: Scale Score. The regression of pre-requisite skills and test administration type on students' scale score was statistically significant, $F(2, 970) = 613.01$, MSR = 45.61, $p < .05$, $R^2 = .56$. Test administration type was a statistically significant predictor of students' scale score, $b = -4.34$, SE = .51, $p < .05$, 95% CI = -5.35 to -3.33. On average, students taking the scaffold administration scored 4.34 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, $b = 1.03$, SE = .04, $p < .05$, 95% CI = .95 to 1.10. On average, every one-point increase in the pre-requisite skills task total corresponded with a 1.03 increase in students' scale scores. Examination of the squared semipartial correlations revealed that approximately 3% of the total scale score variance was uniquely accounted for by test administration type, while 32% was uniquely accounted for by the pre-requisite task total. Together test administration type and Pre-requisite skills accounted for 56% of the total variability in scale scores.

Grade 3: Raw Score. The regression of pre-requisite skills and test administration type on students' raw score was statistically significant, $F(2, 970) = 544.47$, MSR = 175.98, $p < .05$, $R^2 = .52$. Test administration type was a statistically significant predictor of students' raw score, $b = -13.02$, SE = 1.01, $p < .05$, 95% CI = -15.00 to -11.04. On average, students taking the scaffold administration scored 13.02 raw score points lower than students taking the standard administration. Students' pre-requisite task total was also a statistically significant predictor of students' aw score, $b = 1.62$, SE = .08, $p < .05$, 95% CI = 1.47 to 1.77. On average, every one-point increase in the pre-requisite skills task total corresponded with a 1.62 increase in students' raw scores. Examination of the squared semipartial correlations revealed that approximately 8% of the total raw score variance was uniquely accounted for by test administration type, while 22% was uniquely accounted for by the pre-requisite task total. Together test administration type and Pre-requisite skills accounted for 52% of the total variability in raw scores.

Grade 4: Scale Score. The regression of pre-requisite skills and test administration type on students' scale score was statistically significant, $F(2, 1048) = 551.72$, MSR = 67.37, $p < .05$, $R^2 = .51$. Test administration type was a statistically significant predictor of students' scale score, $b = -7.97$, SE = .62, $p < .05$, 95% CI = -9.18 to -6.75. On average, students taking the scaffold administration scored 7.97 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, $b = .95$, SE = .05, $p < .05$, 95% CI = .85 to 1.04. On average, every one-point increase in the pre-requisite skills task total corresponded with a .95 increase in students' scale scores. Examination of the squared semipartial correlations revealed that approximately 8% of the total scale score variance was uniquely accounted for by test administration type, while 19% was uniquely accounted for by the pre-requisite task total. Together test administration type and Pre-requisite skills accounted for 51% of the total variability in scale scores.

Grade 4: Raw Score. The regression of pre-requisite skills and test administration type on students' raw score was statistically significant, $F(2, 1048) = 646.19$, MSR = 198.28, $p < .05$, $R^2 = .55$. Test administration type was a statistically significant predictor of students' raw score, $b = -20.16$, SE = 1.07, $p < .05$, 95% CI = -22.25 to -18.07. On average, students taking the scaffold administration scored 20.16 raw score points lower than students taking the standard administration. Students' pre-requisite task total was also a statistically significant predictor of students' aw score, $b = 1.38$, SE = .08, $p < .05$, 95% CI = 1.22 to 1.53. On average, every one-point increase in the pre-requisite skills task total corresponded with a 1.38 increase in students' raw scores. Examination of the squared semipartial correlations revealed that approximately 15% of the total raw score variance was uniquely accounted for by test administration type, while 13% 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 raw scores.

Grade 5: Scale Score. The regression of pre-requisite skills and test administration type on students' scale score was statistically significant, $F(2, 907) = 417.00$, MSR = 43.64, $p < .05$, $R^2 = .48$. Test administration type was a statistically significant predictor of students' scale score, $b = -5.09$, SE = .53, $p < .05$, 95% CI = -6.12 to -4.06. On average, students taking the

scaffold administration scored 5.09 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 = .88, SE = .04, p < .05, 95% CI = .80 to .97. On average, every one-point increase in the pre-requisite skills task total corresponded with a .88 increase in students' scale scores. Examination of the squared semipartial correlations revealed that approximately 5% of the total scale score variance was uniquely accounted for by test administration type, while 24% was uniquely accounted for by the pre-requisite task total. Together test administration type and Pre-requisite skills accounted for 48% of the total variability in scale scores.

Grade 5: Raw Score. The regression of pre-requisite skills and test administration type on students' raw score was statistically significant, $F(2, 907) = 378.53$, MSR = 194.62, $p < .05$, $R^2 = .46$. Test administration type was a statistically significant predictor of students' raw score, b = -15.16, SE = 1.11, p < .05, 95% CI = -17.34 to -12.98. On average, students taking the scaffold administration scored 15.14 raw score points lower than students taking the standard administration. Students' pre-requisite task total was also a statistically significant predictor of students' aw score, b = 1.43, SE = .09, p < .05, 95% CI = 1.25 to 1.61. On average, every one-point increase in the pre-requisite skills task total corresponded with a 1.43 increase in students' raw scores. Examination of the squared semipartial correlations revealed that approximately 11% of the total raw score variance was uniquely accounted for by test administration type, while 14% was uniquely accounted for by the pre-requisite task total. Together test administration type and Pre-requisite skills accounted for 46% of the total variability in raw scores.

Grade 6: Scale Score. The regression of pre-requisite skills and test administration type on students' scale score was statistically significant, $F(2, 888) = 514.863$, MSR = 34.64, $p < .05$, $R^2 = .54$. Test administration type was a statistically significant predictor of students' scale score, b = -4.35, SE = .46, p < .05, 95% CI = -5.25 to -3.45. On average, students taking the scaffold administration scored 4.35 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 = .04, p < .05, 95% CI = .74 to .88. 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 5% of the total scale score variance was uniquely accounted for by test administration type, while 28% was uniquely accounted for by the pre-requisite task total. Together test administration type and Pre-requisite skills accounted for 54% of the total variability in scale scores.

Grade 6: Raw Score. The regression of pre-requisite skills and test administration type on students' raw score was statistically significant, $F(2, 888) = 352.86$, MSR = 163.58, $p < .05$, $R^2 = .44$. Test administration type was a statistically significant predictor of students' raw score, b = -11.48, SE =.99, p < .05, 95% CI = -13.43 to -9.53. On average, students taking the scaffold administration scored 11.48 raw score points lower than students taking the standard administration. Students' pre-requisite task total was also a statistically significant predictor of students' aw score, b = 1.24, SE = .08, p < .05, 95% CI = 1.09 to 1.39. On average, every one-point increase in the pre-requisite skills task total corresponded

with a 1.24 increase in students' raw scores. Examination of the squared semipartial correlations revealed that approximately 8% of the total raw score variance was uniquely accounted for by test administration type, while 17% was uniquely accounted for by the pre-requisite task total. Together test administration type and Pre-requisite skills accounted for 44% of the total variability in raw scores.

Grade 7: Scale Score. The regression of pre-requisite skills and test administration type on students' scale score was statistically significant, $F(2, 672) = 432.029$, MSR = 38.86, $p < .05$, $R^2 = .56$. Test administration type was a statistically significant predictor of students' scale score, $b = -4.85$, SE = .54, $p < .05$, 95% CI = -5.92 to -3.79. On average, students taking the scaffold administration scored 4.85 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, $p < .05$, 95% CI = .80 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 5% of the total scale score variance was uniquely accounted for by test administration type, while 30% was uniquely accounted for by the pre-requisite task total. Together test administration type and Pre-requisite skills accounted for 56% of the total variability in scale scores.

Grade 7: Raw Score. The regression of pre-requisite skills and test administration type on students' raw score was statistically significant, $F(2, 672) = 311.94$, MSR = 203.34, $p < .05$, $R^2 = .48$. Test administration type was a statistically significant predictor of students' raw score, $b = -14.98$, SE = 1.24, $p < .05$, 95% CI = -17.41 to -12.55. On average, students taking the scaffold administration scored 14.98 raw score points lower than students taking the standard administration. Students' pre-requisite task total was also a statistically significant predictor of students' aw score, $b = 1.36$, SE = .09, $p < .05$, 95% CI = 1.18 to 1.54. On average, every one-point increase in the pre-requisite skills task total corresponded with a 1.36 increase in students' raw scores. Examination of the squared semipartial correlations revealed that approximately 11% of the total raw score variance was uniquely accounted for by test administration type, while 17% was uniquely accounted for by the pre-requisite task total. Together test administration type and Pre-requisite skills accounted for 48% of the total variability in raw scores.

Grade 8: Scale Score. The regression of pre-requisite skills and test administration type on students' scale score was statistically significant, $F(2, 659) = 347.995$, MSR = 34.30, $p < .05$, $R^2 = .51$. Test administration type was a statistically significant predictor of students' scale score, $b = -4.76$, SE = .52, $p < .05$, 95% CI = -5.78 to -3.74. 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, $b = .71$, SE = .04, $p < .05$, 95% CI = .64 to .80. On average, every one-point increase in the pre-requisite skills task total corresponded with a .71 increase in students' scale scores. Examination of the squared semipartial correlations revealed that approximately 6% of the total scale score variance was uniquely accounted for by test administration type, while 24% was uniquely accounted for by the

pre-requisite task total. Together test administration type and Pre-requisite skills accounted for 51% of the total variability in scale scores.

Grade 8: Raw Score. The regression of pre-requisite skills and test administration type on students' raw score was statistically significant, $F(2, 659) = 269.53$, MSR = 174.48, $p < .05$, $R^2 = .45$. Test administration type was a statistically significant predictor of students' raw score, $b = -14.31$, SE = 1.17, $p < .05$, 95% CI = -16.60 to -12.02. On average, students taking the scaffold administration scored 14.31 raw score points lower than students taking the standard administration. Students' pre-requisite task total was also a statistically significant predictor of students' aw score, $b = 1.08$, SE = .09, $p < .05$, 95% CI = .91 to 1.26. On average, every one-point increase in the pre-requisite skills task total corresponded with a 1.08 increase in students' raw scores. Examination of the squared semipartial correlations revealed that approximately 13% of the total raw score variance was uniquely accounted for by test administration type, while 12% was uniquely accounted for by the pre-requisite task total. Together test administration type and Pre-requisite skills accounted for 45% of the total variability in raw scores.

Grade 11: Scale Score. The regression of pre-requisite skills and test administration type on students' scale score was statistically significant, $F(2, 495) = 276.25$, MSR = 58.44, $p < .05$, $R^2 = .53$. Test administration type was a statistically significant predictor of students' scale score, $b = -4.32$, SE = .75, $p < .05$, 95% CI = -5.80 to -2.83. On average, students taking the scaffold administration scored 4.32 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, $b = .99$, SE = .05, $p < .05$, 95% CI = .89 to 1.10. On average, every one-point increase in the pre-requisite skills task total corresponded with a .99 increase in students' scale scores. Examination of the squared semipartial correlations revealed that approximately 3% of the total scale score variance was uniquely accounted for by test administration type, while 32% was uniquely accounted for by the pre-requisite task total. Together test administration type and Pre-requisite skills accounted for 53% of the total variability in scale scores.

Grade 11: Raw Score. The regression of pre-requisite skills and test administration type on students' raw score was statistically significant, $F(2, 495) = 217.05$, MSR = 206.14, $p < .05$, $R^2 = .47$. Test administration type was a statistically significant predictor of students' raw score, $b = -11.66$, SE = 1.42, $p < .05$, 95% CI = -14.44 to -8.88. On average, students taking the scaffold administration scored 11.66 raw score points lower than students taking the standard administration. Students' pre-requisite task total was also a statistically significant predictor of students' aw score, $b = 1.42$, SE = .11, $p < .05$, 95% CI = 1.23 to 1.62. On average, every one-point increase in the pre-requisite skills task total corresponded with a 1.42 increase in students' raw scores. Examination of the squared semipartial correlations revealed that approximately 7% of the total raw score variance was uniquely accounted for by test administration type, while 21% was uniquely accounted for by the pre-requisite task total. Together test administration type and Pre-requisite skills accounted for 47% of the total variability in raw scores.

Science

Grade 5: Scale Score. The regression of pre-requisite skills and test administration type on students' scale score was statistically significant, $F(2, 684) = 409.92$, MSR = 56.95, $p < .05$, $R^2 = .55$. Test administration type was a statistically significant predictor of students' scale score, $b = -8.01$, SE = .62, $p < .05$, 95% CI = -9.22 to -6.79. On average, students taking the scaffold administration scored 8.01 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, $b = .99$, SE = .05, $p < .05$, 95% CI = .89 to 1.09. On average, every one-point increase in the pre-requisite skills task total corresponded with a.99 increase in students' scale scores. Examination of the squared semipartial correlations revealed that approximately 11% of the total scale score variance was uniquely accounted for by test administration type, while 27% 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.

Grade 5: Raw Score. The regression of pre-requisite skills and test administration type on students' raw score was statistically significant, $F(2, 684) = 369.16$, MSR = 257.17, $p < .05$, $R^2 = .52$. Test administration type was a statistically significant predictor of students' raw score, $b = -17.77$, SE = 1.31, $p < .05$, 95% CI = -20.35 to -15.20. On average, students taking the scaffold administration scored 17.77 raw score points lower than students taking the standard administration. Students' pre-requisite task total was also a statistically significant predictor of students' aw score, $b = 1.89$, SE = .10, $p < .05$, 95% CI = 1.68 to 2.09. On average, every one-point increase in the pre-requisite skills task total corresponded with a 1.89 increase in students' raw scores. Examination of the squared semipartial correlations revealed that approximately 13% of the total raw score variance was uniquely accounted for by test administration type, while 23% was uniquely accounted for by the pre-requisite task total. Together test administration type and Pre-requisite skills accounted for 52% of the total variability in raw scores.

Grade 8: Scale Score. The regression of pre-requisite skills and test administration type on students' scale score was statistically significant, $F(2, 600) = 420.42$, MSR = 35.95, $p < .05$, $R^2 = .58$. Test administration type was a statistically significant predictor of students' scale score, $b = -5.41$, SE = .53, $p < .05$, 95% CI = -6.45 to -4.37. On average, students taking the scaffold administration scored 5.41 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, $b = .90$, SE = .04, $p < .05$, 95% CI = .82 to .98. On average, every one-point increase in the pre-requisite skills task total corresponded with a .90 increase in students' scale scores. Examination of the squared semipartial correlations revealed that approximately 7% of the total scale score variance was uniquely accounted for by test administration type, while 32% was uniquely accounted for by the pre-requisite task total. Together test administration type and Pre-requisite skills accounted for 58% of the total variability in scale scores.

Grade 8: Raw Score. The regression of pre-requisite skills and test administration type on students' raw score was statistically significant, $F(2, 600) = 305.80$, MSR = 241.34, $p < .05$, $R^2 = .51$. Test administration type was a statistically significant predictor of students' raw

score, b = -17.60, SE = 1.37, p < .05, 95% CI = -20.29 to -14.91. On average, students taking the scaffold administration scored 17.60 raw score points lower than students taking the standard administration. Students' pre-requisite task total was also a statistically significant predictor of students' aw score, b = 1.62, SE = .11, p < .05, 95% CI = 1.41 to 1.84. On average, every one-point increase in the pre-requisite skills task total corresponded with a 1.62 increase in students' raw scores. Examination of the squared semipartial correlations revealed that approximately 14% of the total raw score variance was uniquely accounted for by test administration type, while 18% was uniquely accounted for by the pre-requisite task total. Together test administration type and Pre-requisite skills accounted for 51% of the total variability in raw scores.

Grade 11: Scale Score. The regression of pre-requisite skills and test administration type on students' scale score was statistically significant, $F(2, 479) = 305.34$, MSR = 42.86, p < .05, R^2 = .56. Test administration type was a statistically significant predictor of students' scale score, b = -5. 86, SE = .64, p < .05, 95% CI = -7.11 to -4.61. On average, students taking the scaffold administration scored 5.86 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 = .05, p < .05, 95% CI = .78 to .97. 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 8% of the total scale score variance was uniquely accounted for by test administration type, while 31% was uniquely accounted for by the pre-requisite task total. Together test administration type and Pre-requisite skills accounted for 56% of the total variability in scale scores.

Grade 11: Raw Score. The regression of pre-requisite skills and test administration type on students' raw score was statistically significant, $F(2, 479) = 282.41$, MSR = 217.82, p < .05, R^2 = .54. Test administration type was a statistically significant predictor of students' raw score, b = -15.84, SE = 1.44, p < .05, 95% CI = -18.66 to -13.02. On average, students taking the scaffold administration scored 15.84 raw score points lower than students taking the standard administration. Students' pre-requisite task total was also a statistically significant predictor of students' aw score, b = 1.71, SE = .11, p < .05, 95% CI = 1.50 to 1.92. On average, every one-point increase in the pre-requisite skills task total corresponded with a 1.71 increase in students' raw scores. Examination of the squared semipartial correlations revealed that approximately 12% of the total raw score variance was uniquely accounted for by test administration type, while 25% was uniquely accounted for by the pre-requisite task total. Together test administration type and Pre-requisite skills accounted for 54% of the total variability in raw scores.

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

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, $F(9, 3441) = 87.72$, MSR = 30.57, $p < .05$, $R^2 = .19$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*(1, 3441) = 612.50, MSR = 25.96, $p < .05$, 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, 3433) = 0.84, MSR = 25.97, $p = .57$, R^2 *Change* = .00. For the final model, test administration type had the largest standardized regression weight, $\beta = -.39$, $p < .05$, and accounted for the most variance, uniquely accounting for approximately 12% 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, $F(9, 2249) = 46.11$, MSR = 50.26, $p < .05$, $R^2 = .16$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*(1, 2248) = 417.76, MSR = 42.40, $p < .05$, 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, 2240) = 0.50, MSR = 42.48, $p = .86$, R^2 *Change* = .00. For the final model, test administration type had the largest standardized regression weight, $\beta = -.40$, $p < .05$, and accounted for the most variance, uniquely accounting for approximately 13% 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, $F(9, 478) = 11.988$, MSR = 56.25, $p < .05$, $R^2 = .18$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*(1, 477) = 60.99, MSR = 49.98, $p < .05$, R^2 *Change* = .09. 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, 469) = 0.58, MSR = 50.12, $p = .58$, R^2 *Change* = .01. For the final model, test administration type had the largest standardized regression weight, $\beta = -.40$, $p < .05$, and accounted for the most variance, uniquely accounting for approximately 13% 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, $F(9, 468) = 13.06$, MSR = 47.24, $p < .05$, $R^2 = .20$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*(1, 467) = 65.15, MSR = 41.55, $p < .05$, 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*(8, 459) = 1.51, MSR = 41.19, $p = .15$, R^2

Change = .02. For the final model, test administration type had the largest standardized regression weight, β = -.41, $p < .05$, and accounted for the most variance, uniquely accounting for approximately 17% 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, $F(9, 926) = 19.83$, MSR = 40.30, $p < .05$, $R^2 = .16$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*(1, 925) = 129.30, MSR = 35.39, $p < .05$, 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*(8, 917) = 1.77, MSR = 35.16, $p = .08$, R^2 *Change* = .01. For the final model, test administration type had the largest standardized regression weight, β = -.35, $p < .05$, and accounted for the most variance, uniquely accounting for approximately 10% of the total variance in students Pre-requisite skills.

Grade 4. The first block of the regression model included only students' disability category, and was statistically significant, $F(9, 956) = 22.72$, MSR = 37.34, $p < .05$, $R^2 = .18$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*(1, 955) = 187.22, MSR = 31.25, $p < .05$, 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*(8, 947) = 1.77, MSR = 31.39, $p = .87$, R^2 *Change* = .00. For the final model, test administration type had the largest standardized regression weight, β = -.41, $p < .05$, and accounted for the most variance, uniquely accounting for approximately 14% 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, $F(9, 808) = 21.54$, MSR = 34.17, $p < .05$, $R^2 = .19$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*(1, 807) = 122.22, MSR = 29.71, $p < .05$, 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*(8, 799) = 0.37, MSR = 29.90, $p = .94$, R^2 *Change* = .00. For the final model, test administration type had the largest standardized regression weight, β = -.36, $p < .05$, and accounted for the most variance, uniquely accounting for approximately 10% 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, $F(9, 784) = 18.66$, MSR = 43.13, $p < .05$, $R^2 = .18$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*(1, 783) = 113.68, MSR = 37.74, $p < .05$, 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*(8, 775) = 0.93, MSR = 37.74, $p = .49$, R^2 *Change* = .01. For the final model, test administration type had the largest standardized regression weight, β = -.36, $p < .05$, and accounted for the most variance, uniquely accounting for approximately 10% 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, $F(9, 605) = 15.82$, MSR = 54.85, $p < .05$, $R^2 = .19$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*(1, 604) = 93.46, MSR = 47.58, $p < .05$, 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*(8, 596) = 0.83, MSR = 47.69, $p = .58$, R^2 *Change* = .01. For the final model, test administration type had the largest standardized regression weight, $\beta = -.36$, $p < .05$, and accounted for the most variance, uniquely accounting for approximately 11% 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, $F(9, 609) = 9.74$, MSR = 62.24, $p < .05$, $R^2 = .13$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*(1, 608) = 119.81, MSR = 52.08, $p < .05$, 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*(8, 601) = 0.82, MSR = 52.18, $p = .57$, R^2 *Change* = .01. For the final model, test administration type had the largest standardized regression weight, $\beta = -.42$, $p < .05$, and accounted for the most variance, uniquely accounting for approximately 15% 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, $F(9, 471) = 11.40$, MSR = 56.64, $p < .05$, $R^2 = .16$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*(1, 470) = 48.93, MSR = 51.41, $p < .05$, R^2 *Change* = .08. 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, 462) = 0.96, MSR = 51.44, $p = .47$, R^2 *Change* = .01. For the final model, Orthopedic Impairment had the largest standardized regression weight, $\beta = -.32$, $p < .05$, and accounted for the most variance, uniquely accounting for approximately 9% 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, $F(9, 655) = 19.26$, MSR = 37.25, $p < .05$, $R^2 = .21$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*(1, 654) = 53.81, MSR = 34.47, $p < .05$, R^2 *Change* = .06. 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, 646) = 0.57, MSR = 34.66, $p = .81$, R^2 *Change* = .01. For the final model, Visual Impairment had the largest standardized regression weight, $\beta = -.29$, $p < .05$, and accounted for the most variance, uniquely accounting for approximately 8% 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, $F(9, 584) = 8.81$, MSR = 60.93, $p < .05$, $R^2 = .12$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*(1, 583) = 76.40, MSR = 53.97, $p < .05$, 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*(8, 576) = 1.44, MSR = 53.69, $p = .19$, R^2 *Change* = .01. For the final model, Test Administration Type had the largest standardized regression weight, $\beta = -.35$, $p < .05$, and accounted for the most variance, uniquely accounting for approximately 11% 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, $F(9, 466) = 10.11$, MSR = 48.95, $p < .05$, $R^2 = .16$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*(1, 465) = 38.68, MSR = 45.02, $p < .05$, R^2 *Change* = .06. 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, 457) = 1.34, MSR = 45.02, $p = .22$, R^2 *Change* = .02. For the final model, Test Administration Type had the largest standardized regression weight, $\beta = -.27$, $p < .05$, and accounted for the most variance, uniquely accounting for approximately 6% 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: Scale Score. The first block of the regression model included only students' disability category, and was statistically significant, $F(9, 3289) = 79.30$, MSR = 161.25, $p < .05$, $R^2 = .18$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*$(1, 3288) = 471.35$, MSR = 141.08, $p < .05$, 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*$(8, 3280) = 0.31$, MSR = 141.06, $p = .41$, R^2 *Change* = .00. For the final model, Test Administration Type had the largest standardized regression weight, $\beta = -.35$, $p < .05$, and accounted for the most variance, uniquely accounting for 10% of the total variance in students' scale score.

Elementary: Raw Score. The first block of the regression model included only students' disability category, and was statistically significant, $F(9, 3289) = 90.90$, MSR = 303.98, $p < .05$, $R^2 = .20$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*$(1, 3288) = 671.32$, MSR = 252.51, $p < .05$, 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*$(8, 3280) = 1.39$, MSR = 252.28, $p = .20$, R^2 *Change* = .00. For the final model, Test Administration Type had the largest standardized regression weight, $\beta = -.41$, $p < .05$, and accounted for the most variance, uniquely accounting for 14% of the total variance in students' raw score total.

Middle School: Scale Score. The first block of the regression model included only students' disability category, and was statistically significant, $F(9, 2109) = 44.19$, MSR = 215.50, $p < .05$, $R^2 = .16$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*$(1, 2108) = 628.21$, MSR = 166.10, $p < .05$, R^2 *Change* = .19. For the third block, students race/ethnicity was added to the model, which resulted in a significant change in model fit, *F Change*$(8, 2100) = 2.63$, MSR = 65.13, $p = .01$, R^2 *Change* = .01. For the final model, Test Administration Type had the largest standardized regression weight, $\beta = -.48$, $p < .05$, and accounted for the most variance, uniquely accounting for 19% of the total variance in students' scale score.

Middle School: Raw Score. The first block of the regression model included only students' disability category, and was statistically significant, $F(9, 2109) = 47.75$, MSR = 410.80, $p < .05$, $R^2 = .17$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*$(1, 2108) = 763.24$, MSR = 301.74, $p < .05$, R^2 *Change* = .22. 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*$(8, 2100) = 2.34$, MSR = 300.21, $p = .02$, R^2 *Change* = .01. For the final model, Test Administration Type had the largest standardized regression weight, $\beta = -.51$, $p < .05$, and accounted for the most variance, uniquely accounting for 22% of the total variance in students' raw score total.

High School: Scale Score. The first block of the regression model included only students' disability category, and was statistically significant, $F(9, 427) = 5.93$, MSR = 330.14, $p < .05$, $R^2 = .11$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*(1, 426) = 98.96, MSR = 268.54, $p < .05$, 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*(8, 418) = 1.24, MSR = 267.32, $p = .27$, R^2 *Change* = .02. For the final model, Test Administration Type had the largest standardized regression weight, $\beta = -.43$, $p < .05$, and accounted for the most variance, uniquely accounting for 17% of the total variance in students' scale score.

High School: Raw Score. The first block of the regression model included only students' disability category, and was statistically significant, $F(9, 427) = 6.50$, MSR = 410.85, $p < .05$, $R^2 = .12$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*(1, 426) = 131.80, MSR = 314.51, $p < .05$, R^2 *Change* = .21. 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, 418) = .69, MSR = 316.37, $p = .70$, R^2 *Change* = .01. For the final model, Test Administration Type had the largest standardized regression weight, $\beta = -.48$, $p < .05$, and accounted for the most variance, uniquely accounting for 20% of the total variance in students' raw score total.

Writing

Grade 11: Scale Score. The first block of the regression model included only students' disability category, and was statistically significant, $F(9, 457) = 7.45$, MSR = 416.00, $p < .05$, $R^2 = .13$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*(1, 456) = 141.81, MSR = 318.01, $p < .05$, R^2 *Change* = .21. 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, 448) = 1.67, MSR = 314.29, $p = .10$, R^2 *Change* = .02. For the final model, Test Administration Type had the largest standardized regression weight, $\beta = -.48$, $p < .05$, and accounted for the most variance, uniquely accounting for 20% of the total variance in students' scale score.

High School: Raw Score. The first block of the regression model included only students' disability category, and was statistically significant, $F(9, 388) = 4.59$, MSR = 331.64, $p < .05$, $R^2 = .10$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*(1, 387) = 81.40, MSR = 274.72, $p < .05$, 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*(8, 379) = .88, MSR = 275.42, $p = .54$, R^2 *Change* = .01. For the final model, Test Administration Type had the largest standardized regression weight, $\beta = -.41$, $p < .05$, and accounted for the most variance, uniquely accounting for 15% of the total variance in students' raw score total.

Math

Grade 3: Scale Score. The first block of the regression model included only students' disability category, and was statistically significant, $F(9, 919) = 18.20$, MSR = 90.27, $p < .05$, $R^2 = .15$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*(1, 918) = 170.81, MSR = 76.19, $p < .05$, 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, 910) = 1.84, MSR = 75.64, $p = .07$, R^2 *Change* = .01. For the final model, Test Administration Type had the largest standardized regression weight, $\beta = -.40$, $p < .05$, and accounted for the most variance, uniquely accounting for 14% of the total variance in students' scale score.

Grade 3: Raw Score. The first block of the regression model included only students' disability category, and was statistically significant, $F(9, 919) = 24.14$, MSR = 308.78, $p < .05$, $R^2 = .19$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*(1, 918) = 243.30, MSR = 244.35, $p < .05$, 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*(8, 910) = 1.21, MSR = 243.91, $p = .29$, R^2 *Change* = .01. For the final model, Test Administration Type had the largest standardized regression weight, $\beta = -.45$, $p < .05$, and accounted for the most variance, uniquely accounting for 17% of the total variance in students' raw score total.

Grade 4: Scale Score. The first block of the regression model included only students' disability category, and was statistically significant, $F(9, 951) = 24.07$, MSR = 117.39, $p < .05$, $R^2 = .19$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*(1, 950) = 283.07, MSR = 90.54, $p < .05$, 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*(8, 942) = 0.73, MSR = 90.75, $p = .67$, R^2 *Change* = .00. For the final model, Test Administration Type had the largest standardized regression weight, $\beta = -.49$, $p < .05$, and accounted for the most variance, uniquely accounting for 18% of the total variance in students' scale score.

Grade 4: Raw Score. The first block of the regression model included only students' disability category, and was statistically significant, $F(9, 951) = 34.74$, MSR = 343.91, $p < .05$, $R^2 = .25$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*(1, 950) = 445.64, MSR = 234.34, $p < .05$, R^2 *Change* = .24. 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, 942) = 1.89, MSR = 232.60, $p = .06$, R^2 *Change* = .01. For the final model, Test Administration Type had the largest standardized regression weight, $\beta = -.55$, $p < .05$, and accounted for the most variance, uniquely accounting for 24% of the total variance in students' raw score total.

Grade 5: Scale Score. The first block of the regression model included only students' disability category, and was statistically significant, $F(9, 801) = 21.41$, MSR = 72.27, $p < .05$, $R^2 = .19$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*(1, 800) = 137.76, MSR = 61.73, $p < .05$, 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, 792) = 0.71, MSR = 61.91, p = .68, R^2 *Change* = .01. For the final model, Test Administration Type had the largest standardized regression weight, β = -.38, p < .05, and accounted for the most variance, uniquely accounting for 12% of the total variance in students' scale score.

Grade 5: Raw Score. The first block of the regression model included only students' disability category, and was statistically significant, *F*(9, 801) = 30.79, MSR = 276.33, p < .05, R^2 = .26. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*(1, 800) = 195.74, MSR = 222.29, p < .05, 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, 792) = 1.22, MSR = 221.80, p = .29, R^2 *Change* = .01. For the final model, Test Administration Type had the largest standardized regression weight, β = -.42, p < .05, and accounted for the most variance, uniquely accounting for 14% of the total variance in students' raw score total.

Grade 6: Scale Score. The first block of the regression model included only students' disability category, and was statistically significant, *F*(9, 776) = 15.81, MSR = 68.99, p < .05, 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, 775) = 157.13, MSR = 57.43, p < .05, 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*(8, 767) = 0.63, MSR = 57.65, p = .76, R^2 *Change* = .01. For the final model, Test Administration Type had the largest standardized regression weight, β = -.42, p < .05, and accounted for the most variance, uniquely accounting for 14% of the total variance in students' scale score.

Grade 6: Raw Score. The first block of the regression model included only students' disability category, and was statistically significant, *F*(9, 776) = 19.51, MSR = 250.18, p < .05, R^2 = .19. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*(1, 775) = 165.06, MSR = 206.52, p < .05, 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*(8, 767) = .59, MSR = 207.40, p = .79, R^2 *Change* = .00. For the final model, Test Administration Type had the largest standardized regression weight, β = -.42, p < .05, and accounted for the most variance, uniquely accounting for 14% of the total variance in students' raw score total.

Grade 7: Scale Score. The first block of the regression model included only students' disability category, and was statistically significant, *F*(9, 594) = 12.74, MSR = 77.55, p < .05, 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, 593) = 136.64, MSR = 63.13, p < .05, R^2 *Change* = .16. For the third block, students race/ethnicity was added to the model, which resulted in a significant change in model fit, *F Change*(8, 585) = 2.55, MSR = 61.84, p = .01, R^2 *Change* = .02. For the final model, Test Administration Type had the largest standardized regression weight, β = -.42, p < .05, and accounted for the most variance, uniquely accounting for 15% of the total variance in students' scale score.

Grade 7: Raw Score. The first block of the regression model included only students' disability category, and was statistically significant, $F(9, 594) = 14.57$, MSR = 321.99, $p < .05$, $R^2 = .18$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*(1, 593) = 192.94, MSR = 243.35, $p < .05$, R^2 *Change* = .20. 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, 585) = 1.25, MSR = 242.54, $p = .27$, R^2 *Change* = .01. For the final model, Test Administration Type had the largest standardized regression weight, $\beta = -.48$, $p < .05$, and accounted for the most variance, uniquely accounting for 19% of the total variance in students' raw score total.

Grade 8: Scale Score. The first block of the regression model included only students' disability category, and was statistically significant, $F(9, 592) = 8.24$, MSR = 65.97, $p < .05$, $R^2 = .11$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*(1, 591) = 164.21, MSR = 25.88, $p < .05$, 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*(8, 584) = 1.28, MSR = 51.54, $p = .26$, R^2 *Change* = .01. For the final model, Test Administration Type had the largest standardized regression weight, $\beta = -.48$, $p < .05$, and accounted for the most variance, uniquely accounting for 20% of the total variance in students' scale score.

Grade 8: Raw Score. The first block of the regression model included only students' disability category, and was statistically significant, $F(9, 592) = 11.81$, MSR = 272.65, $p < .05$, $R^2 = .15$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*(1, 591) = 216.02, MSR = 200.01, $p < .05$, R^2 *Change* = .23. 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, 584) = 1.95, MSR = 197.78, $p = .06$, R^2 *Change* = .01. For the final model, Test Administration Type had the largest standardized regression weight, $\beta = -.52$, $p < .05$, and accounted for the most variance, uniquely accounting for 23% of the total variance in students' raw score total.

Grade 11: Scale Score. The first block of the regression model included only students' disability category, and was statistically significant, $F(9, 462) = 6.34$, MSR = 114.50, $p < .05$, $R^2 = .11$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*(1, 461) = 88.09, MSR = 96.34, $p < .05$, 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*(8, 453) = 1.32, MSR = 95.81, $p = .23$, R^2 *Change* = .02. For the final model, Test Administration Type had the largest standardized regression weight, $\beta = -.40$, $p < .05$, and accounted for the most variance, uniquely accounting for 14% of the total variance in students' scale score.

Grade 11: Raw Score. The first block of the regression model included only students' disability category, and was statistically significant, $F(9, 462) = 9.12$, MSR = 334.58, $p < .05$, $R^2 = .15$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*(1, 461) = 115.26, MSR = 268.24, $p < .05$, 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*(8, 453) = 1.15, MSR =

267.55, $p = .33$, R^2 *Change* = .01. For the final model, Test Administration Type had the largest standardized regression weight, $\beta = -.45$, $p < .05$, and accounted for the most variance, uniquely accounting for 17% of the total variance in students' raw score total.

Science

Grade 5: Scale Score. The first block of the regression model included only students' disability category, and was statistically significant, $F(9, 650) = 21.13$, MSR = 99.02, $p < .05$, $R^2 = .23$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*(1, 649) = 174.90, MSR = 78.12, $p < .05$, 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*(8, 641) = 1.75, MSR = 77.40, $p = .23$, R^2 *Change* = .01. For the final model, Test Administration Type had the largest standardized regression weight, $\beta = -.43$, $p < .05$, and accounted for the most variance, uniquely accounting for 16% of the total variance in students' scale score.

Grade 5: Raw Score. The first block of the regression model included only students' disability category, and was statistically significant, $F(9, 650) = 19.47$, MSR = 432.96, $p < .05$, $R^2 = .21$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*(1, 649) = 187.12, MSR = 336.59, $p < .05$, 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*(8, 641) = 1.97, MSR = 332.61, $p = .05$, R^2 *Change* = .02. For the final model, Test Administration Type had the largest standardized regression weight, $\beta = -.45$, $p < .05$, and accounted for the most variance, uniquely accounting for 17% of the total variance in students' raw score total.

Grade 8: Scale Score. The first block of the regression model included only students' disability category, and was statistically significant, $F(9, 569) = 7.58$, MSR = 80.14, $p < .05$, $R^2 = .11$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*(1, 568) = 177.91, MSR = 61.13, $p < .05$, 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*(8, 561) = 2.54, MSR = 60.00, $p = .02$, R^2 *Change* = .02. For the final model, Test Administration Type had the largest standardized regression weight, $\beta = -.43$, $p < .05$, and accounted for the most variance, uniquely accounting for 16% of the total variance in students' scale score.

Grade 8: Raw Score. The first block of the regression model included only students' disability category, and was statistically significant, $F(9, 569) = 9.82$, MSR = 434.42, $p < .05$, $R^2 = .13$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*(1, 568) = 241.98, MSR = 305.17, $p < .05$, R^2 *Change* = .26. For the third block, students race/ethnicity was added to the model, which resulted in a significant change in model fit, *F Change*(8, 561) = 2.85, MSR = 298.39, $p = .01$, R^2 *Change* = .02. For the final model, Test Administration Type had the largest standardized regression weight, $\beta = -.56$, $p < .05$, and accounted for the most variance, uniquely accounting for 27% of the total variance in students' raw score total.

Grade 11: Scale Score. The first block of the regression model included only students' disability category, and was statistically significant, $F(9, 457) = 8.37$, MSR = 85.58, $p < .05$, $R^2 = .14$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*(1, 456) = 102.83, MSR = 69.99, $p < .05$, 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*(8, 448) = 1.95, MSR = 68.84, $p = .05$, R^2 *Change* = .02. For the final model, Test Administration Type had the largest standardized regression weight, $\beta = -.48$, $p < .05$, and accounted for the most variance, uniquely accounting for 20% of the total variance in students' raw score total.

Grade 11: Raw Score. The first block of the regression model included only students' disability category, and was statistically significant, $F(9, 457) = 9.25$, MSR = 410.58, $p < .05$, $R^2 = .15$. Test Administration Type was added to the model for the second block, which resulted in a significant change in model fit, *F Change*(1, 456) = 142.30, MSR = 313.62, $p < .05$, R^2 *Change* = .20. 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, 448) = 2.10, MSR = 307.69, $p = .04$, R^2 *Change* = .02. For the final model, Test Administration Type had the largest standardized regression weight, $\beta = -.48$, $p < .05$, and accounted for the most variance, uniquely accounting for 20% of the total variance in students' raw score total.

Conclusions

Overall the results were consistent with what we 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 test administration type was a strong predictor 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 almost never a significant predictor of their performance on the pre-requisite skills or content tasks total score.

Section 5: Alignment

5.2

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 due to the fact that the State of 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 process of transitioning toward an AA-AAS that is linked to the Common Core State Standards (CCSS).

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 differential 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(c)*)

5.3

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

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

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 cutscores, 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

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

5.7

The Oregon Extended assessment system uses field testing to improve the alignment of our 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

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

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)

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)

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)

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)

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)

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 31 of 43*).

6.2.2(d)

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

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

6.2.4(a)

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)

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)

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)

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

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_AYPManual2010_11

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

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

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

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 cutscore is provided in *Section 4.2b*. Also, see the mock-up ISR in *Appendix 1.6*.

7.3(a)

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)

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)

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

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

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*).