Oregon Extended Assessment

Technical Report on Standard Setting ORExt Science

ORExt Mathematics

ORExt English Language Arts

Submitted to the Oregon Department of Education

June 2015

by

DCE Educational Communications LLC

Executive Summary

In June of 2015, 53 content area and special education experts, representing three subject areas met over the course of three days and were guided through a judgmental decision-making workshop to set the cut scores for Oregon's 2015 Extended Assessments. Oregon's Extended Assessments are designed for students with the most significant cognitive disabilities and were recently revised to align to the state's Essentialized Assessment Frameworks. The Essentialized Assessment Frameworks provide a direct link to the English Language Arts and Mathematics Common Core State Standards, and Oregon's Science Standards as well as the Next Generation Science Standards in Science, for this population. A third-party, neutral observer was present to document and evaluate the proceedings to determine the validity of the resulting cut scores. The documentation that follows, details the logistical and statistical procedures undertaken in preparation for the workshop, describes the procedures followed during the workshop, and documents steps taken after the workshop toward finalizing the cut scores for use by this population. The results of the workshop are included in this document and the validity of the process is affirmed.

Table of Contents

Setting Achievement Standards for Oregon's Extended Assessments – 2015

Overview

In reference to the process of setting assessment cut scores, the American Educational Research Association (AERA, APA, & NCME, 2014) Standards for Educational and Psychological Testing suggests that

> "if a judgmental standard setting process is followed, the method employed should be described clearly, and the precise nature and reliability of the judgments called for, should be presented… Documentation should also include the selections and qualifications of standard setting panel participants, training provided, any feedback to participants concerning the implications of their provisional judgments, and any opportunities for participants to confer with one another. Where applicable, variability over participants should be reported." (p. 108).

In June of 2015, Behavioral Research and Teaching, developers of Oregon's alternate assessments, in collaboration with the Oregon Department of Education conducted a standard setting workshop in Eugene, Oregon, to determine the cut scores that would delineate the achievement categories for the population of students that takes Oregon's Extended Assessments. In addition to proposing cut scores, participants also reviewed and edited the associated Achievement Level Descriptors that provide qualitative descriptions of proficiency in each category. The workshop was conducted using the Bookmarking method of standard setting and was accomplished over the course of three days. Workshop participants recommended cut scores for the Oregon Extended Assessments in three subject areas: Science, Mathematics, and English Language Arts (ELA).

This document summarizes the main components of the standard setting process, and provides information related to the validity of the process in four areas: procedural consistency, internal consistency, panel membership, panel confidence.

The evaluation of procedural consistency examined whether a formal model of standard setting was implemented with integrity to an established procedure. The evaluation of internal consistency examined the function of the test items and the relationship between test items and the content standards (upon which achievement would be based). Panel membership and diversity was reviewed to ensure that the qualifications and perspective of the standard setting panel aligned with those necessary for the judgments required for standard setting. Finally, panelists were surveyed to determine their support of the process and their confidence in the outcomes -- including projected student impact. The cut scores generated from the standard

setting as well as the projected student impact of the cut scores (in terms of percentages of students falling into each of four achievement categories) are included in this review.

The complete document will be submitted to the Oregon Department of Education as part of a body of evidence documenting the validity of the Oregon Extended Assessment achievement standards.

Oregon's Extended Assessments

Oregon's alternate assessment, referred to as Oregon's Extended Assessment (OExt), is designed to ensure that students in Oregon who have significant cognitive disabilities are exposed to critical, and appropriately stimulating academic content and are included in Oregon's educational accountability system. Oregon's Extended Assessments assess student performance in three subject areas via dichotomously-scored, selected response items that are administered by trained individuals. The assessments were originally developed in 2000 and have undergone at least 4 major revisions (as well as annual refinements) over their 15 years of use by the state of Oregon. The most recent assessments were revised in 2014 and field tested in 2015.

The three subject areas assessed by OExt are as follows: (1) English Language Arts (ELA) which assesses both Reading and Writing and is taken in grades 3, 4, 5, 6, 7, 8, and 11. OExt ELA assesses reading standards for literature, informational text, foundational skills, writing, and language, but excludes the assessment of speaking, listening, or literacy in history, social studies, science, and technical subjects. (2) OExt Mathematics, which is taken in grades 3, 4, 5, 6, 7, 8, and 11 and assesses operations and algebraic thinking, number and operations in base ten, number and operations – fractions, measurement and data, and geometry in grades 3 – 5; ratios and proportional relationships, the number system, expressions and equations, geometry, and statistics and probability in grades 6 – 8, and number and quantity, algebra, functions, modeling, geometry, and statistics and probability in high school. (3) OExt Science, which is taken in grades 5, 8, and 11 and assesses matter and its interactions, motion and stability: forces and interactions, energy, structure and processes of molecules and organisms, interaction, energy, and dynamics of ecosystems, Earth's place in the universe, Earth's systems, Earth and human activity, and engineering design (ODE, 2015).

Both OExt ELA and OExt Mathematics are linked to the Common Core Standards (CCSS) using the Essentialized Assessment Frameworks (EAFs). (The process of "essentializing" standards for students with the most significant cognitive disabilities will be described later in this document.) OExt Science is linked to Next Generation Science Standards using the EAF. Currently in Oregon, a student with a significant cognitive disability may take the general assessment (with appropriate accessibility supports), the alternate assessment, or a combination of the two. Student eligibility for an alternate assessment is based on the IEP team's decision.

Method

Selection of standard setting method: Bookmarking. The Oregon Department of Education (ODE) in conjunction with Behavioral Research and Teaching (BRT) selected the Bookmarking method of standard setting to set standards for the newly revised ORExts. The Bookmarking Method of standard setting is consistent with the method used for the state's general assessment, and is the method previously used with the state's alternate assessment. The Bookmarking method of standard setting, though based on rigorous statistical procedures necessary to develop the Ordered Item Booklets, is a relatively simple procedure to implement with a large-scale state assessment, and is well-accepted among many states (Cizek, 2007). The bookmarking method is typically used with mixed responses items and vertically scaled items similar to those used in Oregon's tests.

Though there are certain variations to the Bookmarking process, the central process as described by Cizek in 2007 is as follows:

> The task presented to participants in a Bookmark standard-setting procedure is straightforward. Using the [Ordered Item Booklet] assembled with one item (or score point) on each page, [panelists] are instructed to indicate the point at which they judge that the borderline or minimally qualified examinee's chances of answering the item correctly (or obtaining the score point) fall below the specified response probability or decision rule. For example, if a 2/3 decision rule is used, participants beginning to work through the OIB would ordinarily judge that the minimally qualified examinee would have better than a 2/3 likelihood of answering items at the beginning of the OIB (i.e., the easiest items) correctly. At some point in the OIB, however, participants would begin to discern that the chances of the minimally qualified examinee answering correctly approach and begin to drop below 2/3. Participants are instructed to indicate the point in the OIB at which the chances of the minimally qualified examinee answering correctly drop below 2/3. They indicate this judgment by placing a page marker—often a self-adhesive note or similar indicator—on the first page in the OIB at which the chance drops below the criterion. That is, the participants are indicating that the items prior to the marker represent content that the minimally qualified examinee would be expected to master at the [Response Probability] or decision rule specified." (p.175).

Instructions for the full Bookmarking procedure that was followed by BRT and ODE in the June standard setting, are documented in Appendices A and B.

Initial Procedures

The newly developed ORExt in Science, Mathematics, and English Language Arts were developed in 2014 and field tested with students in the Spring of 2015. The revised assessments were updated to: assess students on the Essentialized Assessment Frameworks of the CCSS/ORSci/NGSS, support longitudinal growth models, improve administration, remove

administration functions that had become obsolete (such as the administration of the levels of support assessment), and improve general item functioning. A complete summary of the most recent changes to the assessment is included in Appendix C Summary of changes.

Oregon's Essentialized Assessment Frameworks. As part of the development of the assessment, Oregon developed a set of alternate content standards based on the essential components of the Common Core State Standards, Oregon Science Standards, and Next Generation Science Standards. These alternate standards were developed to ensure that Oregon's alternate assessment links to academic content. Almost 200 standards were distilled to under 50 essentialized standards. Each standard was analyzed and reduced to its essential core using a standardized process that is described in Oregon's Extended Assessment administration manual as follows:

> The standards have been "essentialized" by analyzing the content, the intellectual operation being requested, and the delimiters to the content. Structurally, this can be seen in the manner in which standards are written with the content identified by nouns, the intellectual operation by verbs, and the delimiters by either conditional phrases or as placed as the object of the sentence. The essentialization system uses the following conventions: (a) content (nouns) is boxed, (b) intellectual operations (verbs) are underlined (with complex verbs bold), and (c) delimiters (of content or intellectual operations) are italicized. Once the portions of the standard have been appropriately identified, the reduction in depth, breadth, and complexity (RDBC), which is explained below, follows.
>
> The essentialization process involves [the reduction in depth, breadth, and complexity] of the Common Core State Standards (CCSS), Oregon's Science Standards, and the Next Generation Science Standards (NGSS) in order to establish a performance expectation that is relevant and accessible for students who participate in the OExt, while maintaining the highest possible standards of rigor (the science tests will thus be dual-aligned to both the Oregon Science Standards and the NGSS). Complexity is reduced by: 1) focusing on essential content; 2) simplifying the process verb; and, 3) eliminating inappropriate delimiters. For the OExt, all essentialized standards were written at three levels of complexity, which feeds the population of the Low, Medium, and High difficulty forms. The essentialized standards that will be assessed on the OExt are called Essentialized Assessment Frameworks (EAFs) (ODE, 2015).

A flowchart of the standardized process of essentializing Oregon's content standards is included in Appendix D.

Field testing. Items were operationally field tested with Oregon's population of students with the most significant cognitive disabilities. Field testing was conducted in all three subject areas: Science (2,011 students), Mathematics (6,364 students), and English Language Arts (6,627

students). Almost six thousand (6,000) items were developed. Any items that failed to function as anticipated after scoring were eliminated from the item pool.

Ordered item booklet (OIB) development. Following field testing, item difficulty and student ability scores were calculated, using Item Response Theory procedures, in preparation for developing the ordered item booklets (OIBs). Student ability level on Oregon's alternate assessment differed by subject area. ORExt ELA student ability ranged from 1.91 (3rd grade) – 2.65 (7th grade) in consecutive grades, whereas ORExt Mathematics student ability ranged from .13 (3rd grade) to .78 (8th grade) in consecutive grades. In consecutive grades, mean item difficulty also varied from test to test. Mathematics mean item difficulty ranged from 0.7 (3rd grade) to 2.22 (8th grade). ELA mean item difficulty ranged from .93 (5th grade) to 2.14 (8th grade).

To develop the OIBs, items representing the full range of assessed items per grade were identified and then placed into booklets in their order of difficulty. The operational test taken by students was 48 items long in each subject area, however, Ordered Item Booklets constructed for the standard setting workshops ranged in length from 50 to 56 items. Appendix E includes more detailed information on OIB length and item difficulty across tests.

Selection of panelists. Each panelist was recruited by the Oregon Department of Education to play a specialized role as part of a subject-area group. Participants were recruited from among Oregon's licensed teachers throughout the year as well as from Oregon's Qualified Assessors (QAs) and Qualified Trainers (QTs) who are individuals trained in Oregon's Extended Assessments. Individuals were also recruited from among Oregon's Content Specialists who are educators who teach in Oregon and also serve the state in the development of educational materials. Panelists were asked to provide information on their: affiliation, degree, licensure, any certifications, and years of experience working with students with significant cognitive disabilities. Panelists were also asked to share their ethnicity and race.

Workshop

Panel Participants. A total of 53 panelists participated in the event. Eleven panelists in ORExt-Science, and 21 panelists were present each day for both ORExt-Mathematics and ORExt ELA.

The panel was highly educated. Over 90% of the panel possessed a Master's degree or higher. Fifty-seven (57%) percent of the panelists had over 11 years of teaching experience. Seventy-six percent (76%) of the panelists had some experience working with students with significant cognitive disabilities with 64% licensed as Special Educators. The panel was overwhelmingly female (87%), overwhelmingly from the Northwest of the state (87%), and overwhelmingly White (83%). No panel member self-identified with Oregon's major minority population (Hispanic). Panelist demographics collected at the workshop are compiled in Appendix F.

Structure of workshop. On each of the three workshop days a group of panelists met representing their specific subject-area. Each day's group of panelists had the same agenda and sequence of activities. One of the primary procedural differences among the three meetings was related to the number of grade levels assessed in that subject. Participants sat at tables in groups by their grade-level of expertise. On day one (ORExt Science), the three tables represented grades 5, 8, and 11. On days two and three (ORExt Mathematics, and ORExt ELA), the seven tables represented each of grades 3 - 8, and 11.

Across all three days/subject areas, each table had a similar configuration consisting of four to five individuals -- a table facilitator and three or four standard setting participants. The table facilitator was assigned by BRT to manage time and materials, keep the discussions focused, and to complete the rating sheets that captured the results from each of the decision-making rounds.

To ensure sufficient expert knowledge of the population, the subject area, the assessment and accountability in all decision-making groups, each grade level group in each subject area was required to be comprised of at least two special educators, and at least one subject-area specialist. The two Oregon special education teachers were present to ensure the panel's judgments included knowledge of the subject area, the population, and the scope and content of the assessment. The Oregon general education teacher(s) at each grade in each subject area was present to ensure the panel's judgments included subject area expertise, familiarity with the general education achievement expectations as they relate to Oregon's educational standards, the CCSS/ORSci/NGSS.

Training and process. Each day's session began with an overall training to ensure that participants understood their role in determining the state's alternate achievement standards, and the rationale for the day's activities. The training provided information on the development of the assessment, its framework, purpose and uses the training materials are included in Appendix G. Participants were provided with the appropriate subject-level materials and instructed on the standard setting procedure. Panelists were trained on the four levels of achievement. Proficiency Levels are referred to as Levels 1-4. Table 1 provides a general description of each performance category as it is currently used in Oregon.

Table 1.
Oregon Alternate Assessment Achievement Categories

Level	Description
1	Students demonstrate **limited to no mastery of knowledge and skills** related to essentialized standards that **do not meet proficiency**.
2	Students demonstrate **inconsistent or partial mastery of knowledge and skills** related to essentialized standards that **do not meet proficiency**.

3	Students demonstrate **adept knowledge and skills** related to essentialized standards that **meet proficiency**.
4	Students demonstrate **adept mastery of knowledge and skills** related to essentialized standards that **exceed the requirements for proficiency**.

During training, panelists were instructed to place their bookmarks (sticky notes) on the first item of each category starting with the determination for Level 3 (the level in which a student is deemed minimally proficient). Panelists were next instructed to work on Level 4, the level in which a student is deemed to have exceeded the expectations of the population for the assessment. Finally, panelists were instructed to place their third marker on Level 2 to delineate the point at which a student who is not meeting even the minimal expectations, begins to demonstrate some understanding of the material.

Panelists were guided to place their marker on the first item that a minimally proficient student in that given level would have an 80% chance of getting right in the category. Panelists were asked to jot notes about what made the item they selected more difficult than the previous item. Following the standard setting, these (jotted) notes were used by BRT psychometricians when it was necessary to make articulation adjustments (the full process of post-standard-setting articulation is described later in this document).

Judgment Rounds

Judgment rounds 1 and 2. Discussions occurred in three rounds: an independent round, a consensus round, and a post-impact adjustment round. During the first (independent) round, individuals were asked to review their OIBs independently and to set all three level markers according to their knowledge of the population and the content of the items. During the second round, individuals discussed their round 1 findings with their grade and subject level colleagues at their table and discussed their findings and values to come to a shared conclusion about the placement of the cut points. In these discussions, individuals were required to support their judgments by providing content-driven explanations as to why the particular placement marked a delineation not only between two items, but between two categories. A sample of the types of discourse the individual engaged in is included in Table 2. Additional discussion points are included in Appendix H.

Table 2.

Panelist content considerations during judgment rounds

Subject	Discussion
Science	**Complexity of academic concepts.** "Abstract concepts for this population are anything that they cannot experience through physical means, even a term like oxygen may be considered an abstract concept. Gravity, orbit, are all abstract concepts for this population [and render an item more difficult as a result]".
Mathematics	**Level of skill (academic verb) required by the item.** "Up until this point there's just a lot of point and matching and so on".
Mathematics	**Complexity of academic concepts:** "Concepts change here. Now they have to know the concepts *same, more,* and *less*".
Mathematics	**Level of skill (academic verb) required by the item.** "Even with manipulatives, this item still requires a lot of accurate counting".
Mathematics	**Complexity of academic concepts.** "At this point we are starting to talk about a student who could be taking the General Assessment, for example this item is about a clock, whereas this item requires in depth knowledge about fractions."
ELA	**Experience with the item type or content.** Individual A: "The length of item is very different from the previous, lots of extra information is provided. My students don't know most of this information." Individual B: "Yes, but look, the information that the question is based on is literally provided immediately before the question is asked."

Judgment round 3. Following the second round, BRT psychometricians calculated impact data for each of the groups to demonstrate the percentages of students that would fall into each of the four levels of achievement based on the cut points. For round 3, groups used this data to make any final adjustments to their cut points in the event that the percentages of students deemed proficient or not proficient were inconsistent across levels or indefensible. Groups were encouraged to maintain a content- and skill-driven discussion (similar to the discussion after round 2) to see whether their cut points would change. Panelists were warned not to use the impact data to simply place students into levels by percentages. Once panels made final (post-impact) changes they were shown a final round of impact data, however, no changes were made after round 3.

Documentation. Participants used different colored sticky notes for each of the rounds (green sticky notes for round 1, blue for round 2, and pink for round 3). Participants marked each sticky

note with the item number that represented the cut point for the performance category. Round 1 and 2 sticky notes were certified with the participant's initials, round 3 sticky notes were certified by the participant's signature. White sticky notes were also provided for participants to use as markers to indicate any general comments they may have made in the OIBs such as thoughts about items, item difficulty, or their decisions. At the end of the final round (round three), a representative from the group was designated as scribe and captured the group's rationale for each of the placed cut scores. These rationales are included in Appendix I. All judgments from independent reviews in round one, consensus reviews in round 2, and post-impact reviews in round 3 were collected by the table facilitators and are included in this report in Appendices J - L. Examples of data collected at each of the rounds for grade 5 Science is included here in Tables 3 – 5.

Table 3.

ORExt Science Grade 5 Rounds 1 – 3 Judgment Results Item (and Item Difficulty) by person/consensus

Grade 5 Science						
	Round 1				Round 2	Round 3
	Person 1	Person 2	Person 3	Person 4	Consensus	Consensus
Level 2	17 (0.986)	9 (0.556)	16 (0.926)	25 (1.536)	17 (0.986)	9 (0.556)
Level 3	21 (1.176)	17 (0.986)	30 (1.676)	31 (1.776)	29 (1.656)	29 (1.656)
Level 4	29 (1.656)	36 (1.956)	37 (2.006)	40 (2.306)	37 (2.006)	46 (2.956)

Table 4.
Grade 5: Impact following Round 2

Level	Percentage
1	30.9
2	8.6
3	6.5
4	53.9

Table 5.
Grade 5: Impact following Round 3

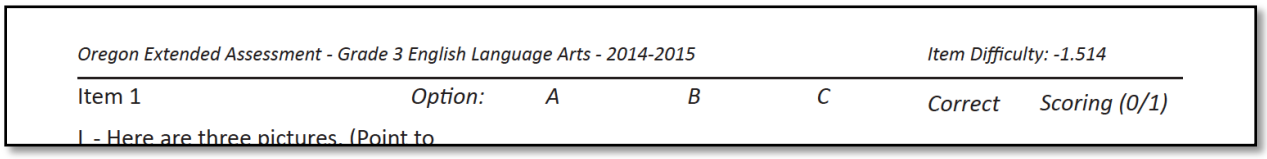
Level	Percentage
1	26.7
2	12.8
3	24.9
4	35.6

Materials

Panelist materials. Each group was provided the following materials:

- A copy of the standard setting procedure Appendices A and B,
- A copy of the training presentation (Appendix G),
- An Ordered Item Booklet (OIB) specific to their grade and subject area, (Individuals were not permitted to remove the OIBs, or the ALDs from the standard setting location.)
- An evaluation survey to share their confidence in the process,
- A background sheet on which they documented their demographic information,
- A copy of the essentialized frameworks, and
- A copy of the Achievement Level Descriptors (ALDs) was provided following the standard setting for the purpose of review and editing.

Ordered Item Booklets consisted of the secure items presented to students in 2014-2015, the language the administrator used to administer the item, the graphics and answer choices that a student was presented in relation to the item, and the correct score associated with the item. An image of the top of an OIB page (with the secure item removed) is shown below.



Achievement Level Descriptor review. Oregon's Achievement Level Descriptors (ALDs) were developed by educators at BRT with a panel of Oregon teachers, and approved by the state board of education, in May of 2015. Following the standard setting, standard setting panelists

were also asked to review the ALDs and to make any edits they deemed necessary. Panelists reviewed according to the following questions:

> Is the language clear enough to communicate to parents?
>
> Does the definition accurately capture a reasonable expectation for this population?
>
> Is the expectation for this population a sufficiently appropriate parallel to expectations for students taking the general benchmark?

No major changes were made as a result of the review. Participants suggested three universal refinements. One such refinement was to alter and reduce the language at level 1 (the does not yet meet) category, to make it clear that Level 1 did not require, expect, or anticipate, any of the skills listed. A brief summary of the panels' suggested changes as shared to the group is included in Appendix M. Specific changes were noted in hardcopy and submitted to BRT and ODE for adjustments and re-submission to the State Board of Education.

Panel Confidence

Survey. At the end of each day's workshop, panelists completed a survey to capture their sentiments regarding the day's process and outcomes. Panelists were asked to respond to affirmative statements regarding the process and the outcomes and rate their agreement with the affirmative statements as Strongly Agree, Agree, Disagree, or Strongly Disagree. The 15 affirmative statements are listed in Table 6.

Table 6.
Affirmative Statements to Determine Panelist Confidence

Oregon Extended Assessment Standard Setter Evaluation Form - 2015
1. The orientation provided me with a clear understanding of the purpose of the standard setting meeting.
2. The training helped me understand the bookmark method and how to perform my role as a standard setter.
3. Reviewing the OrExt helped me to understand the assessment.
4. The small and large group discussions aided my understanding of the process.
5. There was an equal opportunity for everyone in my group to contribute his/her ideas and opinions.

6. I was able to follow instructions and complete the rating sheets accurately.
7. The discussions after the first round of ratings were helpful to me.
8. The discussions after the second round of ratings were helpful to me.
9. The information showing the impact of our cut scores on proficiency percentages was helpful to me.
10. I am confident about the defensibility and appropriateness of the final recommended cut scores.
11. The achievement level descriptions were clear and useful.
12. The time provided for discussions was adequate.
13. The workshop leaders helped to answer questions and ensure that all input was respected and valued.
14. The facilities and food service helped create a productive and efficient working environment.
15. Overall, I am confident that the standard setting procedures allowed me to use my experience and expertise to recommend cut scores for the ORExt.

In Science, 100% of participants either strongly agreed or agreed with all 15 of the affirmative statements.

In Mathematics, 95% of participants either strongly agreed, or agreed with all 15 of the affirmative statements, 5% (1 individual) disagreed with statement 11, which read “The achievement level descriptions were clear and useful”.

In ELA 90% of participants either strongly agreed, or agreed with all 15 of the affirmative statements. One individual (5%) disagreed with statement 6, which read: “I was able to follow instructions and complete the rating sheets accurately”. One individual (5%) disagreed with statement 12 that read “The time provided for discussions was adequate.” This participant felt that too much time was provided.

Across all three subject areas, 100% of participants either Strongly agreed, or agreed with statement 10, which read, “I am confident about the defensibility and appropriateness of the final recommended cut scores." Percentages of panel responses by subject area are included in Appendix N.

Articulation Round and Final Results

Articulation. The day following the standard setting workshop, psychometricians met to review the vertical alignment of the proposed cut scores across grades in the assessed subject area. Articulation is reviewed to make sure that, within each subject area of a vertically scaled test, the cut scores set at a given level for one grade do not exceed the cut scores set at the same level for the next grade. A smooth and intuitive progression is anticipated of the item difficulty in a given level as the grades increase. Of the cut scores set, 12 changes were made to maintain integrity across grades. Cut scores were adjusted in consecutive grades 3 – 8 in Mathematics and ELA.

When adjusting to maintain articulation integrity the following rules were followed to ensure that the fewest changes were made following the panelist's input overall:

Articulation Round Guidelines.

1. Identify the fewest number of steps necessary to bring the scores into articulation: Identify the scores that have the least cascading impact on other grades if changed. In reviewing alignment, isolate any scores (at any of the three cut scores levels) that appear to be outliers when compared to scores at other grades.
2. Follow the same order of adjustment as required by panelists: Start at the proficiency (Level 3) cut point, then evaluate Level 4 cut point, followed finally by the does not yet meet (Level 2) cut point.
3. Whenever possible, revert to a score that the panelists had considered previously with particular primacy to round two judgments (prior to their review following impact data): Reverting to round two was based in maintaining panelists' integrity. Panelists came to their round 2 conclusion based on their content review and only changed it in an attempt to influence the impact data if they found the impact data to be skewed.
4. Use booklets to confirm item changes: Whenever possible select the closest item to the panelist's original item selection while maintaining panelist rationale (which was often written in the booklet).
5. Only stray from the "closest item" rule (5 above) if the closest possible item contributes to creating a gap that further compromises the integrity of the articulation.

Articulation Round Summary of Changes

Science

No changes. Grades are not immediately consecutive and the scale was not vertical because of the gap between grades. In addition, the proportions (impact data) were not significantly different from ELA proportions overall.

Table 7.
Changes Made to Cut Scores in ORExt Mathematics

Grade	Level Adjusted	Previous Item Difficulty (item)	New Item Difficulty (item)	Shift in number of items
4	Level 1 – 2 (Nearly Meets)	-0.994 (5)	-0.734 (6)	1
4	Level 2 – 3 (Meets)	0.676 (25)	0.606 (21)	-4
4	Level 3 – 4 (Exceeds)	2.326 (48)	1.906 (42)	-6
5	Level 3 – 4 (Exceeds)	1.586 (35)	2.016 (41)	6
7	Level 1 – 2 (Nearly Meets)	-0.244 (6)	0.746 (18)	12
7	Level 3 – 4 (Exceeds)	2.776 (50)	2.276(43)	-7

Table 8.
Changes Made to Cut Scores in ORExt ELA

Grade	Level Adjusted	Previous Item Difficulty (Item)	New Item Difficulty (Item)	Shift in number of items
3	Level 3 – 4 (Exceeds)	3.006 (54)	2.776 (52)	2
4	Level 3 – 4 (Exceeds)	2.746 (45)	2.816 (46)	1
5	Level 1 – 2 (Nearly Meets)	0.516 (12)	0.166 (9)	-3
6	Level 2 – 3 (Meets)	1.666 (25)	2.036 (32)	7
6	Level 3 – 4 (Exceeds)	2.976 (45)	3.266 (49)	4
7	Level 1 – 2 (Nearly Meets)	0.386 (3)	0.776 (6)	6

Post Articulation Cut Scores. Tables 9 - 11 document the final cut scores and associated impact by level following the cross-grade articulation review. (Shaded cells are cells in which cut scores were changed from round 3.)

Table 9.
Science Post Articulation Final Recommended Cut Scores and Impact

	Level 1	Level 2	Level 3	Level 4
Grade 5 cut point (item difficulty)		9 (0.556)	29 (1.656)	46 (2.956)
Grade 5 Impact	26.7%	12.8%	24.9%	35.6%
Grade 8 cut point (item difficulty)		19 (0.956)	36 (2.016)	51 (3.106)
Grade 8 Impact	28.8%	13.7%	15.2%	42.3%
Grade 11 cut point (item difficulty)		5 (0.106)	24 (1.406)	47 (2.856)
Grade 11 Impact	20.8%	10.8%	21.2%	47.2%
Mean Cross Grade Impact	25.43%	12.43%	20.43%	41.7%
SD of Impact	4.15	1.48	4.90	5.8

Table 10.
Mathematics Post Articulation Final Recommended Cut Scores and Impact

	Level 1	Level 2	Level 3	Level 4
Grade 3 cut point (item difficulty)		6 (-0.764)	16 (0.136)	44 (1.816)
Grade 3 Impact	25.9%	13.9%	44.5%	15.7%
Grade 4 cut point (item difficulty)		6 (-0.734)	21 (0.606)	42 (1.906)
Grade 4 Impact	15.4%	30.5%	34.8%	19.3%
Grade 5 cut point (item difficulty)		8 (-0.664)	22 (0.616)	41 (2.016)
Grade 5 Impact	15.5%	25.6%	45%	14%
Grade 6 cut point (item difficulty)		6 (0.406)	13 (0.846)	37 (2.176)
Grade 6 Impact	32.1%	10.7%	39.1%	18.1%
Grade 7 cut point (item difficulty)		18 (0.746)	22 (0.916)	43 (2.276)
Grade 7 Impact	19.5%	25.3%	39.9%	15.4%
Grade 8 cut point (item difficulty)		5 (0.806)	18 (1.236)	35 (2.566)
Grade 8 Impact	41.9%	13%	38.5%	6.7%
Grade 11 cut point (item difficulty)		6 (0.136)	13 (0.656)	43 (2.206)
Grade 11 Impact	38.2%	11.9%	36.2%	13.8%
Mean Cross Grade Impact	26.93%	18.7%	39.71%	14.71%
SD of Impact	10.78	8.13	3.86	4.07

Table 11.
ELA Post Articulation Final Recommended Cut Scores and Impact

	Level 1	Level 2	Level 3	Level 4
Grade 3 cut point (item difficulty)		5 (-0.764)	18 (1.316)	52 (2.776)
Grade 3 Impact	12.1%	23.4%	23%	41.5%
Grade 4 cut point (item difficulty)		8 (0.096)	23 (1.346)	46 (2.816)
Grade 4 Impact	15.2%	13.3%	23.6%	48%
Grade 5 cut point (item difficulty)		9 (0.166)	30 (2.006)	47 (3.246)
Grade 5 Impact	17.5%	16.2%	19.3%	47%
Grade 6 cut point (item difficulty)		5 (0.466)	32 (2.036)	49 (3.266)
Grade 6 Impact	19%	13%	23.1%	44.8%
Grade 7 cut point (item difficulty)		6 (0.776)	30 (2.226)	48 (3.636)
Grade 7 Impact	22.4%	12.8%	21.8%	43%
Grade 8 cut point (item difficulty)		5 (1.266)	18 (2.426)	50 (3.646)
Grade 8 Impact	27.3%	14.2%	24.1%	34.5%
Grade 11 cut point (item difficulty)		3 (-0.124)	35 (1.996)	48 (2.736)
Grade 11 Impact	19.5%	17.3%	11.8%	51.5%
Mean Cross Grade Impact	19%	15.74%	20.96%	44.33%
SD of Impact	4.92	3.78	4.34	5.46

Conclusion

Because a Bookmarking standard setting process is, at its heart, based on human judgments, no single piece of information can easily confirm the validity of the standards that result. To determine the validity of the cut scores from Oregon's 2015 standard setting workshop described in this document, a convergence of evidence model was used to evaluate the likelihood of valid outcomes from four perspectives: procedural consistency, internal consistency, panel

membership, and panel confidence in the results. Overall, the process undertaken in Oregon for the ORExt subject area assessments is likely to have resulted in valid outcomes due to soundness in the major procedural areas. Some minor deficits are noted in the summaries below.

Procedural consistency. Procedural consistency was evaluated by a review of: the methods used to set the standards, the integrity to which those responsible for the workshop adhered to the formal procedures, and the rationale used when diversions from formal procedure were necessary. The structure of the workshop, the quality and integrity of the training and materials, as well as the participants' adherence to training guidelines during rounds, contributed to strong procedural consistency of the workshop.

Internal consistency. Internal consistency was evaluated by a review of: the soundness of the initial procedures that went into the essentialization process, the soundness of the OIB development and IRT calculations, the scope of the field testing and associated scoring, and the soundness of the judgments used to guide the post-round articulation. While all internal procedures were carried out with fidelity to the statistical expectations of IRT, the range of item difficulty and student ability did not always fit the expected range of tests on an IRT scale. This likely contributed to some weakness in the internal consistency of the standards. However, the following consideration is an important one: In Oregon, the range of students eligible to take the ORExt is broad. Eligibility criteria currently is provided in the form of broad guidance for IEP team decision-makers and does not require empirical evidence of student ability as eligibility criteria for participation, see Appendix O (ODE, 2015). As a result, the population taking this assessment ranges from students who have difficulty interacting with items in any setting, to students who are close to being (but not quite) able to participate in the general assessment. This range of student skill level has an annual impact on item difficulty scores of Oregon's alternate assessment. Cut scores were made for this year's test with panelist knowledge that the tests (particularly ELA) would require additional, more difficult items in the coming years and that eligibility criteria for the assessment may be more stringent in future test populations.

Panel membership. Panel membership was evaluated by: a review of the diversity and expertise of the panel. As noted, the panels were highly educated with over 90% of the panel possessing a Master's degree or higher. The majority of the panel had had experience working with the population of students with significant disabilities, and while 64% had a special education license, decisions were balanced by the presence of general educators familiar with the expectations of the general population. The panel diversity was low, particularly racial/ethnic diversity, gender diversity and regional diversity. No panel member self-identified with Oregon's major minority population (Hispanic). However, the concentrations of educator gender, and regional representation aligned loosely with proportions of educators in the state. It is not clear how different the cut scores would have been if there had been greater racial diversity in the panel. The educational level of the panel and the quality of the training (with a focus on the specialized needs of the population of students with significant cognitive disabilities) may

mitigate any variance resulting from the panel, however, future panel membership would benefit from greater diversity.

Panel Confidence. Panel confidence was measured via survey following the final round of the decision-making. Panelists had an opportunity to discuss their rationale with colleagues, work toward consensus, and adjust decisions after a review of the impact data. Following the workshop, panelists had full confidence in the standards they had set for the population. Only three of the 53 panelists deviated from agreement to affirmative statements about the process. None of the 3 disagreements impacted the individuals' confidence in the outcome.

References

AERA, APA, & NCME (2014) Standards for Educational and Psychological Testing. Washington, DC: AERA.

Cizek, G. J. (2007). The Bookmark Method. In G. J. Cizek & M.B. Bunch (Eds.), Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests. Thousand Oaks, CA: Sage Publications.

Oregon Department of Education (2015). Oregon Extended Assessment Administration Manual. Retrieved from http://www.ode.state.or.us/teachlearn/testing/admin/alt/ea/updates/orext_adminman_14-15.pdf