Oregon Science Assessment Peer Review Resubmission Evidence

Critical Element 4.6:
Comparability Study

June 2023

OREGON DEPARTMENT OF EDUCATION

Oregon achieves . . . together!

Introduction

The Oregon Department of Education's (ODE) Next Generation Science Standards (NGSS) assessment has three test versions: English, Spanish/English toggle, and Braille. The English version is available to all students, while the Spanish/English toggle is a support for students who speak Spanish (e.g., multilingual students with English learner status) and the Braille version is an accommodation for students with visual impairments. The Spanish/English toggle and Braille versions have their own item pools; however, these pools are subsets of the English item pool. The reason the item pools for these versions are subsets of the English item pool is because ODE, partner states, and Cambium Assessment develop and calibrate English items before identifying them for Spanish translation and Braille transcription.

Given the reliance on the English item pool, all three versions share the same initial item development process. This includes the claim structure and underlying principles guiding development, item specifications, selection and training of item writers, internal review, external review, field testing, and post-calibration review. The English, Spanish/English toggle, and Braille versions share the same blueprint, test length, and distribution of item types, and cover the same disciplines and acceptable combination of performance expectations. Lastly, the English, Spanish/English, and Braille versions share the same calibration model (i.e., multigroup Rasch testlet model), measure the same construct, align to the same standards, use the same linking methodology, rely on the same student and institution level reporting structures, and share the same universal tools, supports, and accommodations (with some logical exceptions).

ODE's 2018-19 NGSS assessment used a linear-on-the-fly test (LOFT) assembly design. While the English and Spanish/English toggle versions relied on the LOFT assembly design, the Braille version was a linear fixed form because of the limited availability of items for Braille transcription. We acknowledge that the Braille version in 2018-19 didn't share the same test assembly design as the other versions; however, ODE intends to use a computer adaptive test (CAT) assembly design in 2021-22 and subsequent school years for all three versions.

Please note the following:

- The science items are testlets consisting of one or more assertions, and assertions are the scored student interactions within each item.
- There are two item types: standalone and cluster. Standalone items are testlets with three or fewer assertions. Cluster items are testlets with four or more assertions.

This study examines seven comparability elements. These include performance expectation coverage, blueprint adherence, difficulty parameters, form difficulty, test characteristic curve, measurement error, and differential assertion functioning.

Comparability Elements

1. *Performance expectation coverage*: How comparable is the coverage of performance expectations within the English, Spanish/English toggle, and Braille item pools? Note that coverage refers to the number of performance expectations with at least one item ÷ the total number of adopted NGSS performance expectations. The item pools for each test version increased considerably from 2018-19 to the 2022-23 school year. For instance, the 5th grade English, Spanish/English toggle, and Braille item pools increased from 78 to 511, 26 to 158, and 18 to 78 items, respectively.

Table 1.
Performance expectation coverage by grade, test version, and school year.

Test Version	School Year		
	2018-19	2021-22	2022-23
5th Grade			
English	90.5	100.0	100.0
Spanish/English Toggle	54.8	90.5	97.6
Braille	38.1	64.3	88.1
8th Grade			
English	74.5	100.0	100.0
Spanish/English Toggle	45.5	63.6	96.4
Braille	30.9	54.5	78.2
11th Grade			
English	68.7	97.0	98.5
Spanish/English Toggle	40.3	56.7	89.6
Braille	26.9	43.3	61.2

Summary: Coverage of performance expectations within the respective item pools was inadequate and minimally comparable in the 2018-19 school year. Performance expectation coverage improves considerably for all version item pools in the 2021-22 and 2022-23 school years. We expect complete and uniform performance expectation coverage across versions by the 2024-25 school year as ODE, partner states, and Cambium Assessment develop more items to address the performance expectation gaps in the item pools. ODE will annually monitor the coverage of performance expectations by version

item pool until coverage is complete and uniform. While the versions were minimally comparable in the 2018-19 school year, the coverage of performance expectations in 2021-22, 2022-23, and subsequent school years suggests adequate comparability across versions.

2. *Blueprint adherence:* How comparable is blueprint adherence[1] across the English, Spanish/English toggle, and Braille versions? Note that the Braille version in the 2018-19 school year had one administered test in 5th grade and two administered tests in 8th grade. Although there was one administered test in 11th grade, it did not meet ODE's test validation criteria.

Table 2.
Blueprint adherence by grade, test version, and school year.

Test Version	School Year	
	2018-19	2021-22
5th Grade		
English	99.9	100.0
Spanish/English Toggle	99.8	100.0
Braille	0.0	100.0
8th Grade		
English	100.0	100.0
Spanish/English Toggle	100.0	100.0
Braille	0.0	100.0
11th Grade		
English	100.0	100.0
Spanish/English Toggle	100.0	100.0
Braille	0.0	100.0

Summary: Blueprint adherence for the English and Spanish/English toggle versions was approximately 100 percent in the 2018-19 school year; however, the adherence was zero percent for the Braille version. Blueprint adherence in the 2021-22 school year for all versions was 100 percent. This suggests adequate comparability across test versions.

3. *Difficulty parameters:* How comparable are the distributions of assertion difficulty parameters within the 2018-19 English, Spanish/English toggle, and Braille item pools? Note that items consist of multiple assertions, and assertions are the smallest unit of

[1] Blueprint adherence refers to the percentage of tests that meet blue print requirements (e.g., specific number of stand-alone and cluster items per NGSS domain, no redundant performance expectations, etc.).

analysis in the item response function. Assertions have a difficulty parameter and items have a variance representing the dependency between assertions within an item.

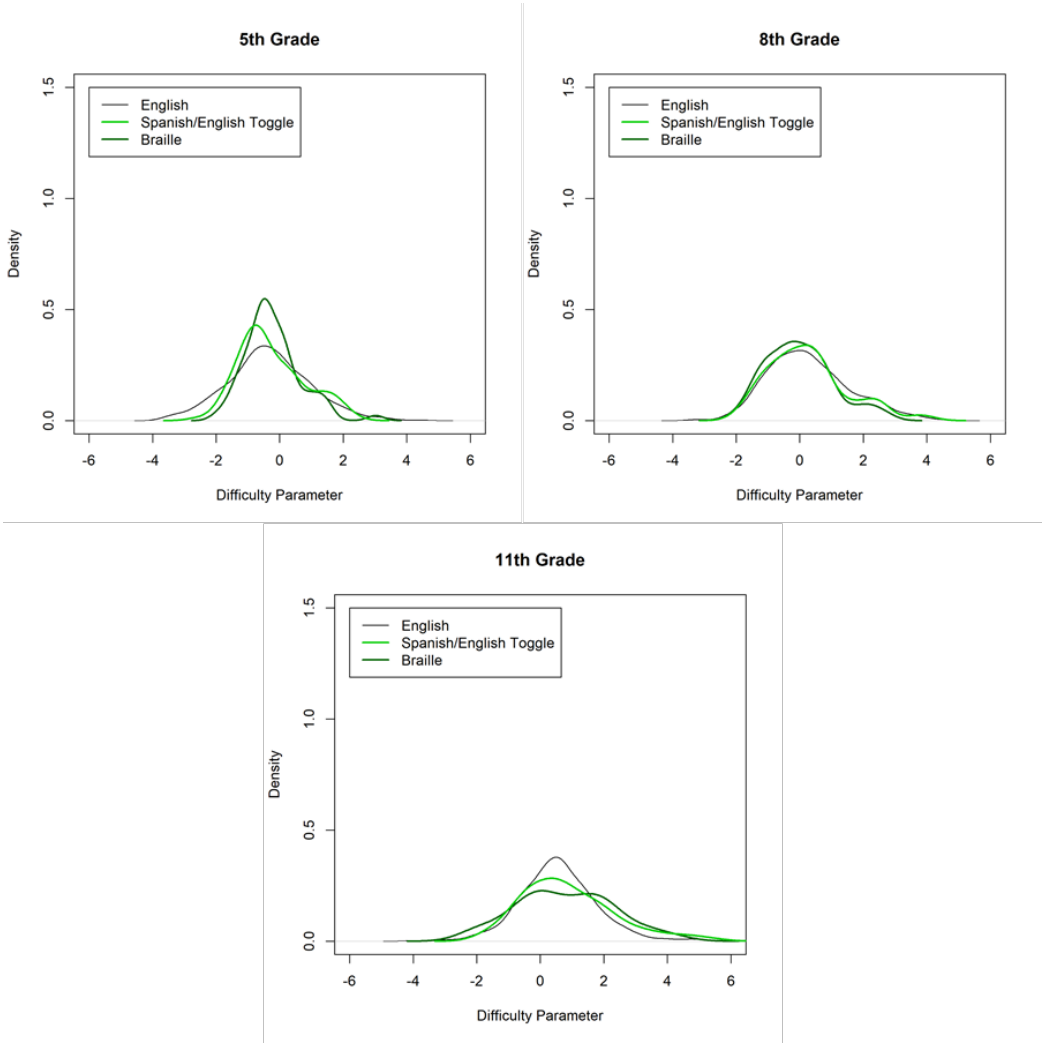

Figure 1. Distribution of difficulty parameters by grade and test version.

Summary: The distribution of assertion difficulty parameters within the English, Spanish/English toggle, and Braille item pools was reasonably similar in the 2018-19 school year. This suggests adequate comparability with respect to this comparability element. The distributions will be approximately equivalent in future school years as ODE, partner states, and Cambium Assessment develop, field test, and add more items to the respective item pools.

4. *Form difficulty:* How comparable is the difficulty of the average form across the 2018-19 English, Spanish/English toggle, and Braille versions? Note that the LOFT assembly design

creates a unique form for each student. This comparability element compares the difficulty[2] of the average form for each test version after ordering the assertions from easiest to most difficult. The average form for the Spanish/English toggle or Braille version will meet comparability expectations if it falls within the English lower and upper bounds (i.e., one standard deviation below and above the difficulty of the average English form).

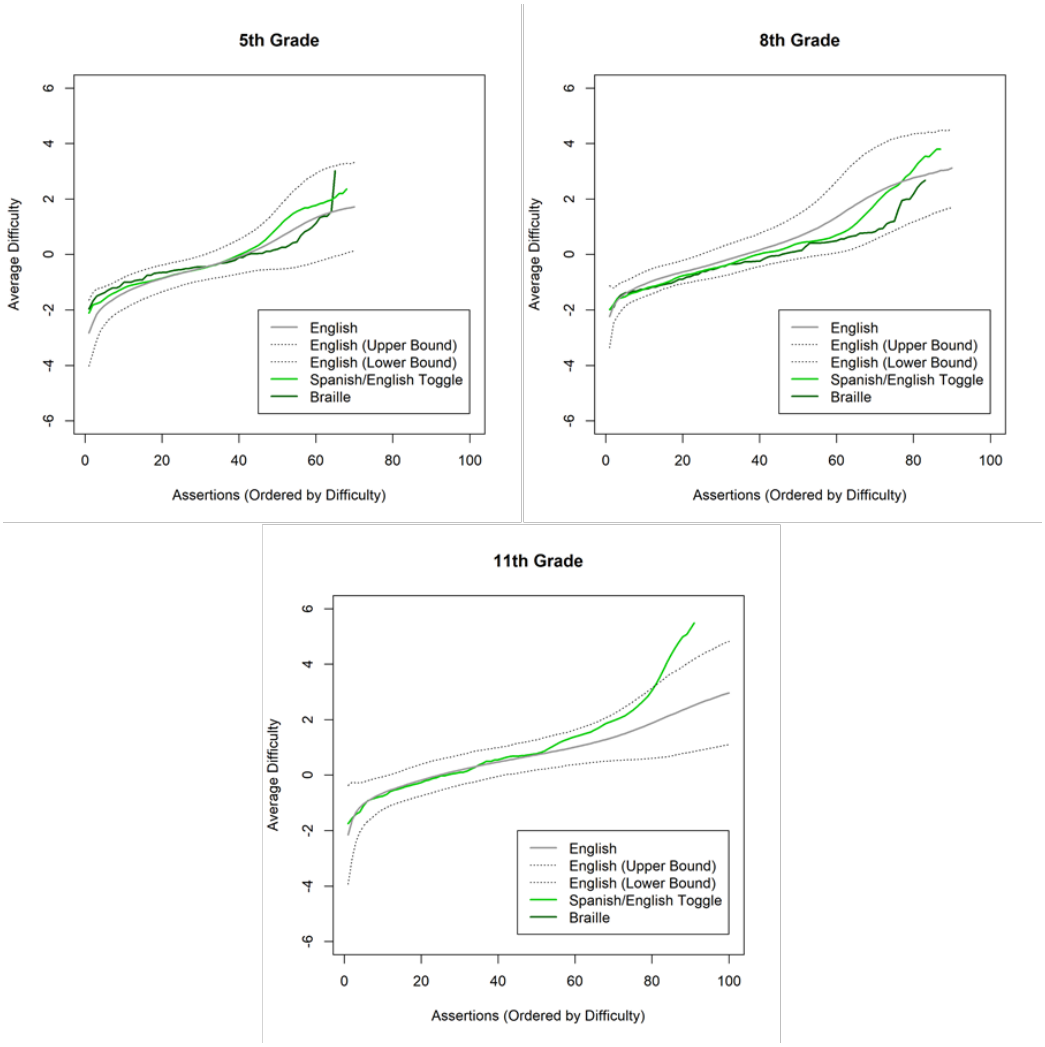

Figure 2. Average difficulty of test assertions by grade and test version.

Summary: The average form difficulty is reasonably similar across the English, Spanish/English toggle, and Braille versions in the 2018-19 school year. The average form for the 11th grade Spanish/English toggle is the exception, however. It exceeded the upper bound of the English version for the most difficult assertions. This reflects less than

[2] Difficulty refers to the assertion difficulty parameter.

desirable comparability. We anticipate reasonably similar form difficulty and adequate comparability in the 2021-22 school year due to the use of the CAT assembly design across test versions.

5. *Test characteristic curve:* How comparable are the test characteristic curves across the 2018-19 English, Spanish/English toggle, and Braille versions? Note that the test characteristic curve is the association between observable science ability (i.e., theta) and the proportion of assertions we would expect students to answer correctly.

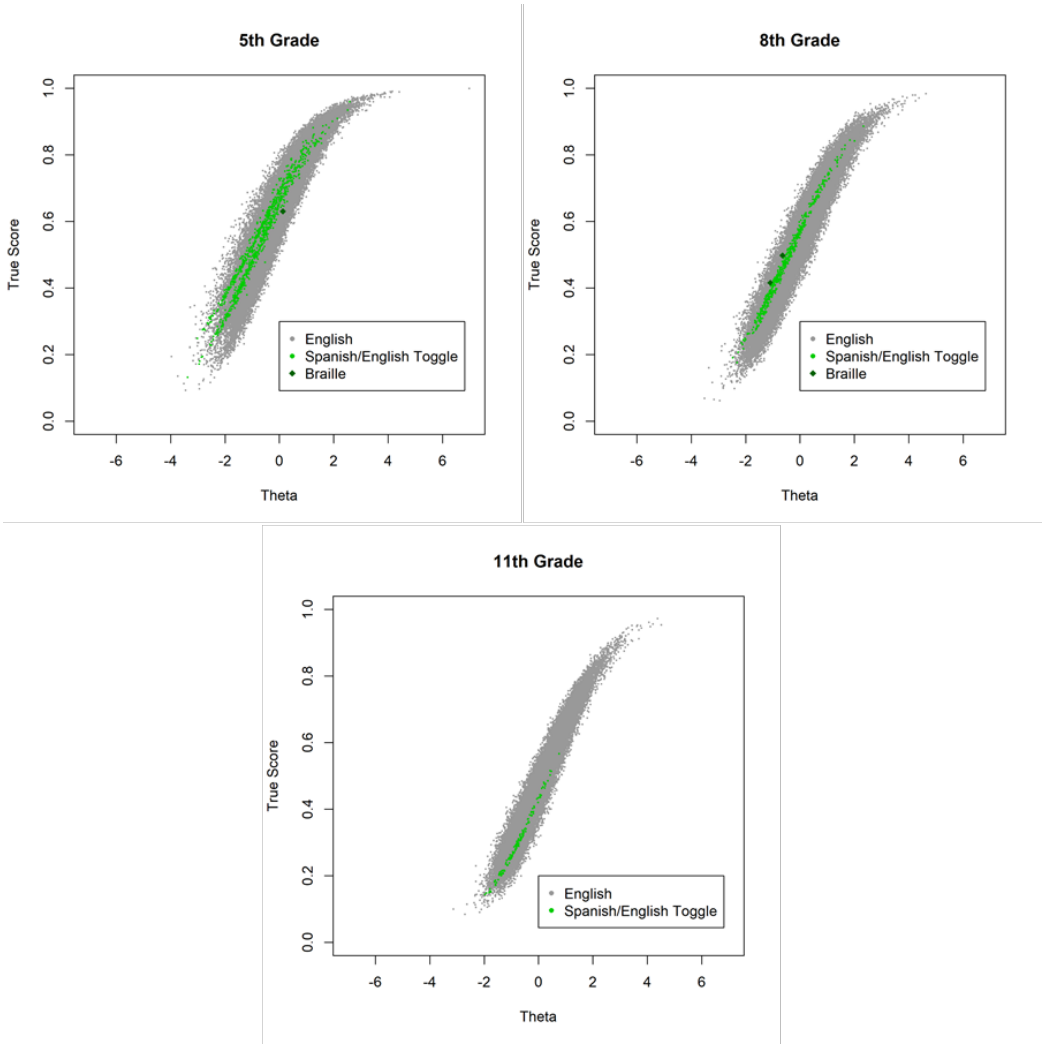

Figure 3. Test characteristic curve by grade and test version.

Summary: The test characteristic curves were reasonably similar across the English, Spanish/English toggle, and Braille versions in the 2018-19 school year. This suggests adequate comparability. It is important to note the narrow performance of students using the Spanish/English toggle in the 11th grade. ODE believes the homogenous performance

of multilingual students with English learner status in high school is largely due to lower reclassification rates (i.e., lower numbers of students exiting from Title III English language programs); however, the average form difficulty may also influence the homogeneous performance of multilingual students with English learner status. ODE intends to monitor the homogenous performance of multilingual students with English learner status in high school in subsequent years, and explore the impact of average form difficulty on homogeneous test performance.

6. *Measurement error:* How comparable is the measurement error across the 2018-19 English, Spanish/English toggle, and Braille versions? Note that this comparability element compares the association between scale scores and their respective conditional standard error of measurement (CSEM) across test versions.

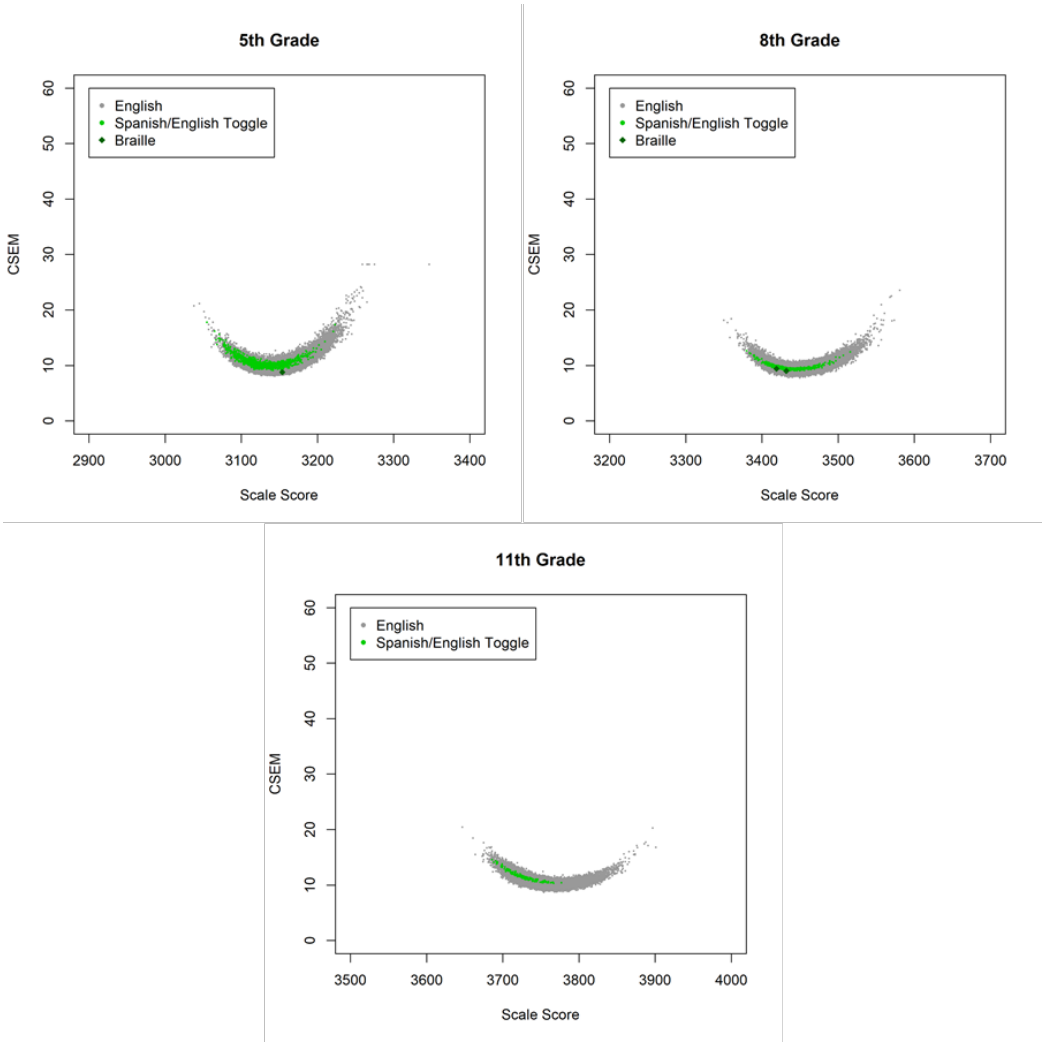

Figure 5. CSEM and Scale score distribution by grade and test version.

Summary: The measurement error by scale score is reasonably similar across the English, Spanish/English toggle, and Braille versions in the 2018-19 school year. This suggests adequate comparability (conditional on student performance). Similar to the previous comparability element, it is important to note the narrow performance of students using the Spanish/English toggle in the 11th grade. In general, the performance of multilingual students with English learner status is more homogenous in high school than in elementary and middle school grades. While this is largely due to lower reclassification rates in high schools, ODE believes the average form difficulty may influence the homogenous performance of multilingual students with English learner status. Moreover, ODE intends to monitor the homogenous performance of multilingual students with English learner status in high school in subsequent years, and explore the impact of average form difficulty on homogeneous test performance.

7. *Differential assertion functioning:* Do assertions function or behave differently across the 2018-19 English and Spanish/English toggle versions? Note that this comparability element does not examine the Braille version because of insufficient operational tests in the 2018-19 school year. We use the Mantel-Haenszel (M-H) approach to examine differential assertion functioning, and exclude assertions that do not meet the minimum n-size of 100 responses for the Spanish/English toggle and 500 responses for the English version. We intend to examine these assertions in the future by combining responses across multiple school years.

Table 3.
Number of Assertions by M-H delta classification, direction, and grade.

Grade	A		B		C		Zero Score Points
	-	+	-	+	-	+	
5th Grade	39	53	1	5	1	1	0
8th Grade	47	49	8	4	1	0	0
11th Grade	24	34	5	6	4	0	4

Summary: We identified eleven assertions for review (corresponding to eight items). Seven assertions had C+ or C- M-H delta classifications and four assertions had zero student responses with a score point. ODE psychometricians, science content specialists, and staff with Spanish expertise reviewed the eleven assertions for translation issues and construct irrelevant content. After internal review, we sent our findings to Cambium Assessment with the following recommendations: (1) the annual monitoring of item

behavior (5 items), (2) the correction of minor Spanish translation issues (2 items), (3) and the correction of minor errors in both versions (1 item). The minimal amount of differential assertion functioning suggests adequate comparability across the test versions.

Conclusion

This study examines seven comparability elements (i.e., performance expectation coverage, blueprint adherence, difficulty parameters, form difficulty, test characteristic curve, measurement error, and differential assertion functioning). With a few exceptions, the comparability elements suggest adequate comparability across the English, Spanish/English toggle, and Braille versions of ODE's 2018-29 NGSS. We anticipate an improvement in comparability with the transition to the CAT assembly design in 2021-22 and further item development by ODE, partner states, and Cambium Assessment. Where the comparability elements suggest minimal or less than desirable comparability across test versions (i.e., performance expectation coverage and form difficulty), ODE intends to monitor the comparability elements annually and explore solutions to improve the comparability.