Oregon Science Assessment Peer Review Resubmission Evidence

Critical Element 4.1:
Reliability Study

June 2023

OREGON
DEPARTMENT OF EDUCATION
Oregon achieves . . . together!

Introduction

The Oregon Department of Education's (ODE) Next Generation Science Standards (NGSS) assessment used a linear-on-the-fly test (LOFT) assembly design to deliver items to students during the 2018-19 school year. The LOFT assembly design randomly selected stand-alone and cluster items[1] for each student in order to meet the science content blueprint requirements (e.g., a specific number of stand-alone and cluster items for each domain—earth and space science, life science, and physical science). ODE acknowledged in the June 2020 science peer review submission to the U.S. Department of Education that the extreme ends of each grade's science scale score distribution had larger than desirable standard errors. Furthermore, ODE stated that item development and the transition to a computer adaptive test (CAT) assembly design would improve the precision at the extremes of the scale score distribution.

This study examines the impact of item development and the transition to a CAT assembly design on the precision at the extremes of the 5th, 8th, and 11th grade science scale score distributions in the 2021-22 school year.

Item Development

The 5th, 8th, and 11th grade item pools increased considerably from the 2018-19 school year to the 2021-22 school year. We anticipate the size of the item pools will continue to increase as ODE, partner states, and Cambium Assessment develop items to meet pool size targets and address known gaps within the item pools (e.g., performance expectation coverage). Table 1 below describes each grade's item pool by school year and performance level. Because each item consists of at least one assertion, the performance level represents the classification of the average assertion location on the reporting scale. We view items corresponding to performance levels 1 and 4 as representing the extreme ends of the science scale score distribution.

It is clear from Table 1 that the extreme ends were lacking items in the 2018-19 school year. This was particularly evident for performance level 4 in 5th and 8th grades and performance level 1 in 11th grade. While the bulk of the item pool size expansion occurred within performance levels 2 and 3 (which makes sense given that the level 3 cut score signifies proficiency), the number of items within performance levels 1 and 4 increased noticeably in the 2021-22 school year. We anticipate a further increase in the number of performance level 1 and 4 items during the 2022-23 and 2023-24 school years. Nonetheless, we believe a continual increase in the item pool size

[1] The science items are testlets consisting of one or more assertions, and assertions are the scored student interactions within each item. Standalone items are testlets with three or fewer assertions. Cluster items are testlets with four or more assertions.

(both general and targeted) along with the transition to a CAT assembly design will have a meaningful impact on the precision of scale scores at the extreme ends of the distribution.

Table 1.
Item pool by grade, school year, and performance level.

Grade	Pool size in terms of items	Number of items corresponding to each performance level			
		1	2	3	4
5th					
2018-19	78	26	38	10	4
2021-22	358	79	177	91	11
8th					
2018-19	97	15	41	40	1
2021-22	285	46	110	115	14
11th					
2018-19	87	7	13	48	19
2021-22	286	14	51	132	89

CAT assembly design

ODE, partner states, and Cambium Assessment transitioned from a LOFT assembly design to a CAT assembly design in the 2021-22 school year. While the LOFT assembly design randomly selects subsequent items only to meet the science content blueprint requirements, the CAT assembly design selects subsequent items (i.e., items providing optimal information) to match the student's provisional ability estimate as well as to meet the science content blueprint requirements. In addition to improving the student testing experience, we believe ODE's adoption of a CAT assembly design will improve the overall reliability of the test as well as the precision at the extreme ends of the science scale score distribution.

Table 2.
Marginal reliability by grade and test assembly design (i.e., LOFT vs. CAT).

Grade	Marginal Reliability	
	LOFT	CAT
5th	0.874	0.885
8th	0.885	0.901
11th	0.866	0.892

Note. ODE used the LOFT assembly design in the 2018-19 school year, and used the CAT assembly design in the 2021-22 school year.

Table 2 presents the marginal reliabilities for each test assembly design in the 5th, 8th, and 11th grades. The marginal reliability increased for each grade after the implementation of the CAT assembly design in the 2021-22 school year (with 11th grade experiencing the largest increase—0.026 or 2.6 percent of the total variance). We acknowledge that item development and the CAT assembly design likely have a combined influence on the marginal reliability given their intersection and dependency (i.e., a deeper item pool supports item selection and the estimation of student ability). We anticipate a gradual improvement in the marginal reliability as the item pools increase and as ODE, partner states, and Cambium Assessment make adjustments and enhancements to the CAT assembly design.

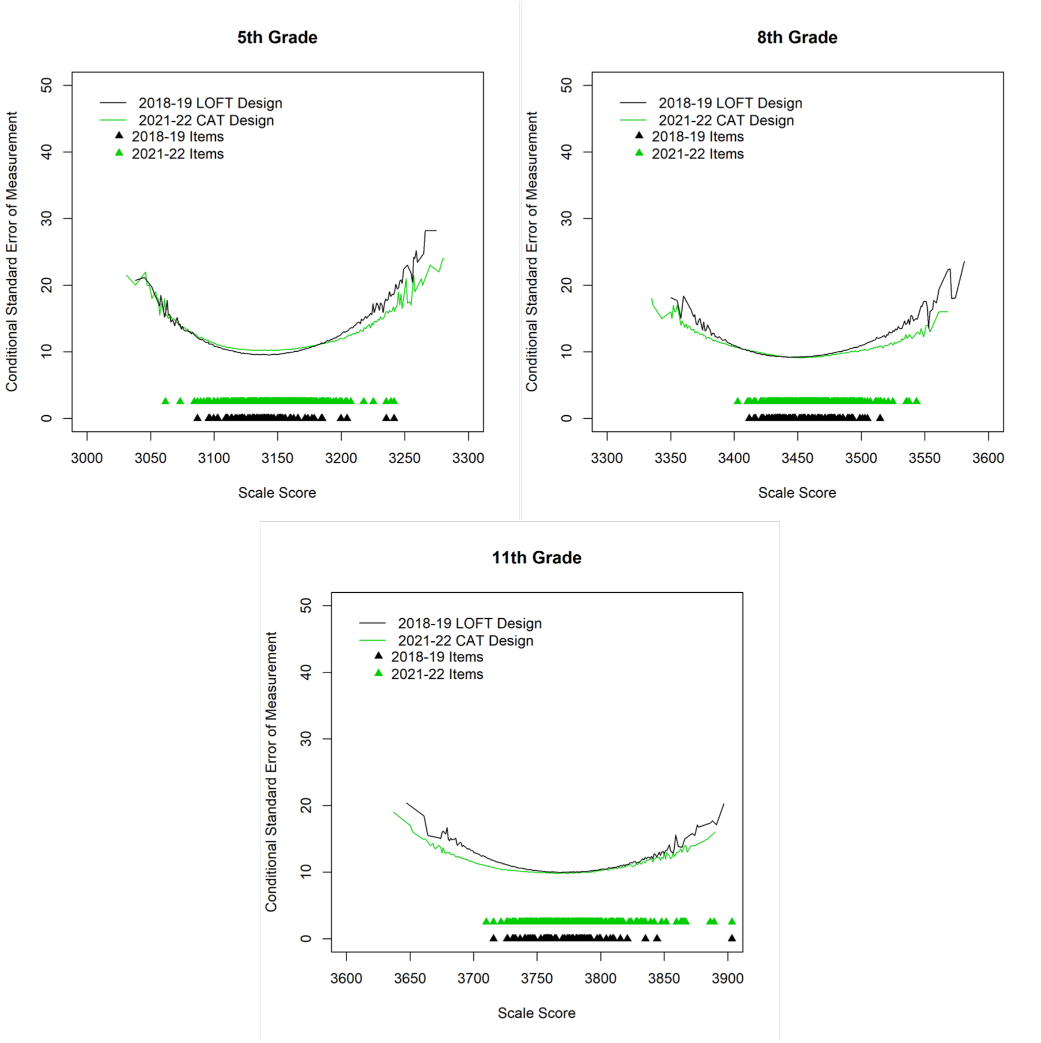

Figure 1. Association between the conditional standard error of measurement and scale score by grade and test assembly design (with item locations on the reporting scale).

Figure 1 displays three plots representing the association between the conditional standard error of measurement (CSEM) and the scale score for each grade. Additionally, each plot shows the average assertion location for each item on the reporting scale. The black lines and triangles symbolize the LOFT assembly design, and the green lines and triangles represent the CAT assembly design. It is clear that item development since the 2018-19 school year and the transition to the CAT assembly design in the 2021-22 school year improved the precision at the extreme ends of the science scale score distribution for each grade. The association between the CSEM and the scale score is flatter at the extremes in comparison to what occurred in the 2018-19 school year. There is one exception, however. The left side of the 5th grade scale score distribution has similar precision between 2018-19 and 2021-22 regardless of item development and test assembly design. ODE is currently investigating this and believes it may be due to item information. That is, if the newer 5th grade items are less informative than the original items, it is possible the CAT algorithm favors the original items as part of the item selection process.

Conclusion

This study examines the impact of item development and the transition to a CAT assembly design on the precision at the extremes of the 5th, 8th, and 11th grade science scale score distributions in the 2021-22 school year. We find that item development since the 2018-19 school year and the transition to a CAT assembly design in the 2021-22 school year had a meaningful impact on precision. We anticipate the precision will improve at the extreme ends of the science scale score distribution as the item pools increase and as ODE, partner states, and Cambium Assessment make adjustments and enhancements to the CAT assembly design.