**Annual Item Selection** (page 37 before section on Point Measure Correlations)

Based on the initial development of approximately 5,000+ items in 2014-2015, BRT researchers developed an initial operational test for vertical scaling across grades 3-8 in English language arts (ELA) and Mathematics. Because Science did not have contiguous grades, no such vertical scaling was possible. As noted in the previous sections on Validity (3.1), item selection was concurrently based on a number of dimensions: alignment with standards (academic and content), cognitive processes, and internal structure (of domains within subject areas).

In the first year, these dimensions were the primary considerations until sufficient numbers of students had taken the test and item characteristics calculated. Oregon monitors the quality of its system in several ways in order to support continuous improvement. In terms of the assessment quality, item statistics are reviewed each year and items that are not functioning as intended are removed and replaced by better functioning field-test items. In 2014-15, items were reviewed in two phases, first using classical test theory (CTT) and second using Rasch analyses. All items flagged as a result of the statistical reviews were analyzed, item-by-item, by a team of measurement and content experts at BRT. Not all flagged items were removed, as several did not have apparent design flaws. Considerations regarding domain representation as well as item difficulty range also were considered during the review process. We also employed different decision rules for unique items versus horizontally- or vertically-scaled anchor items. It was important in many cases to maintain anchor items. Items with clear design flaws were removed from subsequent analyses and reporting. The following flagging criteria were employed:

• CTT : A unique item was flagged if it had a p-value of .10 or lower, .90 or higher, or a point biserial < .15. Anchor items were flagged if they had a p-value of .10 or lower or .95 and higher on all forms or a point biserial < .45 on any form.

• Rasch : Unique items were flagged if their outfit mean square values were between 0 and .25 or > 1.5. Anchor items were flagged if their outfit mean square values were < .5, > 1.8 for horizontal items, or > 2.0 for vertical anchor items.

Out of a total of 5,929 items developed in 2014-15, 166 were removed (2.8%).

In subsequent years, not only item characteristic curves (ICCs) but also test characteristics curves (TCCs) were calculated and used to incorporate field test items as operational items and move new field test items into the test. To ensure the vertical scale was functioning properly, with successive values increasing over grades, the item selection process not only considered item fit but also item difficulty. When a grade level was more difficult than a later grade, we drew less difficult items from the previous grade and/or more difficult items from the later grade.



