New York State Regents Examination in Algebra I (Common Core)

2014 Field Test Analysis, Equating Procedure, and Scaling of Operational Test Forms

Technical Report

Prepared for the New York State Education Department

FILE PROCESSING AND DATA CLEANUP

The Regents examinations utilize both multiple-choice (MC) and constructed-response (CR) item types in order to more fully assess student ability. Multiple field test (FT) forms were given

Test Reliability

Reliability is the consistency of the results obtained from a measurement with respect to time or between items or subjects that constitute a test. As such, test reliability can be estimated in a variety of ways. Internal consistency indices are a measure of how consistently examinees respond to items within a test. Two factors influence estimates of internal consistency: (1) test length and (2) homogeneity of the items. In general, the more items on the examination, the higher the reliability; and the more similar the items, the higher the reliability.

Table 3 contains the internal consistency statistics for each of the field test forms under the heading “Test Reliability.” These statistics ranged from 0.44 to 0.70. It should be noted that these FT forms were extremely short (8–10 items); operational tests are composed of more items and can be expected to have higher reliabilities than do these field test forms.

Scoring Reliability

One concern with CR items is the reliability of the scoring process (i.e., consistency of the score assignment). CR items must be read by scorers who assign scores based on a comparison between the rubric and student responses. Consistency between scorers is a critical part of the reliability of the assessment. To track scorer consistency, approximately 10% of the test booklets are scored a second time (these are termed “second read scores”) and compared to the original set of scores (also known as “first read scores”).

As an overall measure of scoring reliability, the Pearson correlation coefficient between the first and second scores for all CR items with second read scores was computed for each form. This statistic is often used as an overall indicator of scoring reliability, and it generally ranges from 0 to 1. Table 3 contains these values in the column headed “Scoring Reliability.” They ranged from 0.70 to 0.95, indicating a high degree of reliability.

Table 3. Test and Scoring Reliability

Form Number	Test Reliability	Scoring Reliability
101		

The anchor form calibration began with all anchor item difficulty parameters fixed to their known values from the previous year. Because it is possible for item parameters to "drift" (shift their difficulties relative to one another), a stability check was integrated into the analysis.

Winsteps provides an item level statistic, termed "displacement." Linacre (2011, p. 545) describes this statistic as:

> …the size of the change in the parameter estimate that would be observed in the next estimation iteration if this parameter was free (unanchored) and all other parameter estimates were anchored at their current values. For a parameter (item or person) that is anchored in the main estimation, (the displacement value) indicates the size of disagreement between an estimate based on the current data and the anchor value.

This statistic was used to identify items with difficulties that had shifted, relative to the difficulties of the other items on the form. After the intial calibration run, the Winsteps displacement values for all anchor form items were examined for absolute values greater than 0.30. If present, the item with the largest absolute displacement value was removed from anchored status, but remained on the test form. Its difficulty value was subsequently reestimated relative to the difficulties of the remaining anchored items. The Winsteps calibration was then rerun with the reduced anchor set, after which the displacement values were again checked for absolute values in excess of 0.30. If another was found, it was also removed from anchored status and the calibration rerun. This iterative procedure continued until all anchored items had displacements of 0.30 or less. Three items were identified as having drifted. In addition, ncho

Form	Item	Score										

Test	Form	Type	Item	Max	N-Count	RID	S1	S2	S3	S4	S5	S6	INFIT
2014_A1CC	125	MC	06	1	566	0.6234							0.95
2014_A1CC	125	MC	07	1	566	0.3131							1.04
2014_A1CC	125	CR	08	2	566	0.9704	í0.1085	0.1085					0.85
2014_A1CC	125	CR	09	2	566	0.2244	0.3528	í0.3528					0.81
2014_A1CC	125	CR	10	4	566	1.2612	í0.2029	í0.4192	í0.3428	0.9649			0.85
2014_A1CC	126	MC	01	1	565	í0.6206							0.96
2014_A1CC	126	MC	02	1	565	í0.3118							

Pos	Item								

2.295
2.441
2.609
2.795
3.010
3.266
3.586
4.025
4.752
50179

Appendi x G: Scoring Tabl es Raw Abili

2.033
2.158

1.807
1.916

Raw Ability