Missouri Assessment Program

Proposed Anchor Evaluation for Communication Arts, Mathematics, and Science

The evaluation of anchor items is a routine part of the linking process. The reliability of the anchor set is critical to the longitudinal stability of a testing program because performance on the anchor helps us evaluate growth. The evaluation of the anchor items is meant to ensure that the anchor set is stable over time and that misbehaving anchors do not adversely influence results.

It should be noted that the removal of anchor items is a serious concern. In general, CTB is reluctant to remove an item from the anchor set unless the item is clearly performing differently than expected. An anchor item should only be removed if it is adversely affecting the quality and accuracy of the scaling. An anchor item should not be removed for the sole purpose of raising or lowering test performance. Also, the removal of an anchor item cannot adversely affect the content structure of the anchor set such that a different construct is being measured.

This document first overviews MAP's linking design. It then outlines the methods that will be used to evaluate the MAP anchors. Finally, it summarizes the steps that will be followed to determine if items should be removed from the anchor set.

Linking Design

The MAP item parameters will be linked to the *TerraNova* scale through the Stocking and Lord (1983) procedure using an embedded form of the *TerraNova* assessment. In 2008, the MAP scale for all grade/content areas (except Grade 11 Communication Arts and Grade 11 Science) will be linked to the *TerraNova* scale using *TerraNova* Survey Form D. The Grade 11 Communication Arts MAP tests will be linked to the *TerraNova* scale using *TerraNova* Survey Form A. The Grade 11 Sciences MAP tests will be linked to the *TerraNova* scale using *TerraNova* Survey Form C. Table 1 summarizes the total number of items and the number of anchor items for each grade/content area of the 2008 MAP.

After the initial IRT item calibration, item parameters will be linked to the *TerraNova* scale using the test characteristic curve (TCC) method proposed by Stocking & Lord (1983), and implemented using PARDUX (Burket, 1995). This approach takes advantage of the vertical properties of the *TerraNova* scale and allows for comparison of the results from the 2008 administration to results from prior and future administrations. Again, the intact *TerraNova* Survey test items will serve as anchors.

Table 1. Number of Points for the Anchor Set and Total Test, 2008 and 2009 Forms

		Total Number of Points		
Content Area	**Grade**	**Anchor Set**	**2008**	**2009**
Mathematics	3	30	67	67
	4	32	77	77
	5	32	69	69
	6	31	68	68
	7	32	69	69
	8	31	76	76
	11	25	70	70
Communication Arts	3	30	62	61
	4	35	63	63
	5	32	63	62
	6	31	62	62
	7	33	73	72
	8	34	69	68
	10	34	73	73
Science	5	25	83	83
	8	25	87	90
	11	25	90	91

Anchor Set Evaluation

The anchor items will be evaluated immediately following the linking of the MAP to the *TerraNova* scale. This paper outlines the methods to be used to evaluate anchor items for the MAP. First, the quality of the overall equating will be analyzed using the test characteristic curve (TCC) method. Following this, individual anchor items will be evaluated to determine if performance on those items has changed between administrations.

Methods Used to Evaluate Anchor Items

For the Missouri assessments, two statistical methods will be used to evaluate anchor items: (1) iterative linking (Candell & Drasgow, 1988) using Stocking and Lord's (SL) (1983) test characteristic curve method; and (2) differences between the item-ability regression curves.

Test Characteristic Curve Method

The Stocking and Lord (1983) procedure, also called the test characteristic curve (TCC) method, minimizes the mean squared difference between the two TCCs, one based on inputs from the previous calibration and the other on transformed estimates from the current calibration. The Stocking and Lord method is used both to link the items to the *TerraNova* scale, and to evaluate the anchor items themselves.

Let $\hat{\psi}_j$ be the test characteristic curve based on estimates from the previous calibration and $\hat{\psi}_j^*$ be the test characteristic curve based on transformed estimates from the current calibration.

$$\hat{\psi}_j = \hat{\psi}(\theta_j) = \sum_{i=1}^{n} P_i(\theta_j; a_i, b_i, c_i)$$

$$\hat{\psi}_j^* = \hat{\psi}(\theta_j) = \sum_{i=1}^{n} P_i(\theta_j; \frac{a_i}{M_1}, M_1 b_i + M_2, c_i)$$

The TCC method determines the scaling constants (M_1 and M_2) by minimizing the following quadratic loss function (F):

$$F = \frac{1}{N} \sum_{a=1}^{N} (\hat{\psi}_j - \hat{\psi}_j^*)^2$$

With this approach, the overall quality of the linking is evaluated through the correlation of the estimate and input anchor item parameters. As a rule of thumb the *b*-parameter correlation value between the input and estimate difficulty parameters should be greater than .90. For the *a*-parameter, the correlation value of the relationship of the input and estimate discrimination parameter should be greater than .80.

When the correlation values associated with the overall linking fall below the specified values, then outlying items are closely examined. For evaluating individual outlying anchor items, the root mean square difference (RMSD) of the anchor items in the anchor set will be calculated. The individual anchor item with an absolute difference of parameters (after putting them on a common scale) greater than two times the RMSD will be flagged as an outlier.

IRT Item-Ability Regression Curves

We will also compute differences between the item-ability regression curves of the anchor items for the TerraNova and 2008 calibrations. The differences between the curves will be evaluated using the following statistics:

- UnWtd Mean = Average signed difference in estimated probability.
- UnWtd Mean Abs Dif = Average Absolute (unsigned) difference in estimated probability.
- UnWtd RMSD = Root mean squared difference.

Both un-weighted and weighted versions of these statistics will be calculated. Un-weighted differences give equal weight to differences across the ability spectrum. Weighted differences assign weights according to the number of test-takers that are impacted.

The weighted versions of these differences are:

- Wtd Mean = Weighted average signed difference in estimated probability.
- Wtd Mean Abs = Weighted average Absolute (unsigned) difference in estimated probability.
- WtdRMSD = Weighted Root mean squared difference.

For the six statistics listed above, differences greater than ±.10 are considered large and differences between ±.07 and .10 are considered moderate.

Additionally, the Maximum Absolute difference (MaxAbsDifPC) will be identified. For MaxAbsDIFPC, large differences are those greater than ±.15, and moderate differences are all differences between ±.125 and .15.

Implementation

For equating and parameter comparisons, Stocking and Lord's TCC method will be implemented in PARDUX and the differences between the IRT regression curves will be implemented in a SAS routine.

Removing Anchor Items

While dropping an anchor item flagged based solely on statistical criteria has its simplicity, this option may change the content coverage and equating constants, shift scale score distributions, and affect the classification of students by moving them into different proficiency levels. Before an anchor item may be dropped from an anchor set, the adequacy of the content coverage must be evaluated.

As stated above, an item is only removed from the anchor set if it adversely affects quality of scaling, not desirability of results. As such, CTB will not consider how the removal of an item affects the overall mean scale score or the impact data (percent of students in each achievement level) when recommending items for removal.

Items removed from the anchor set are still scored as part of the whole test. Anchor items are considered for exclusion from the MAP under the following conditions:

1.) Items flagged using the TCC method are considered for exclusion when the correlation between the input and equated item parameters is below .90 for the *b*-parameter or below .80 for the *a*-parameter. If the exclusion of an outlying anchor item increases the correlation to above .90 for the *b*-parameter or above .80 for the *a*-parameter, then the anchor is a candidate for removal.
2.) An item is a candidate for removal when it is flagged for large differences on four of the seven statistics considered when examining the differences between the IRT regression curves.
3.) Removal of the item will only be considered after alternative explanations have been considered that may explain shifts in performance. For example, performance on the anchor item may improve because of a statewide initiative emphasizing instruction on a particular set of skills. In this case, improved performance on the item represents true growth in that area. Removing the anchor item may artificially lower test scores.
4.) Removal of the item may not significantly alter the content distribution of the anchor set. The distribution of the anchor items across the content standards must remain within 10% of the 2008 test blueprint.
5.) The number of remaining items will remain at an acceptable level of anchor set reliability. Operationally, this means the anchor set will still be representative of the total test blueprint and that the anchor may not be less than 20% of the total test length.