To: Shaun Bates	**From DRC: Joanna Tomkowicz**
Dept: Missouri Department of Elementary and Secondary Education	**Dept: Psychometric Services**
Location: Missouri	**Location: Maple Grove, Minnesota**
Subject: MAP ELA and Mathematics scaling options	**Date: November 16, 2015**

Overview

This document outlines the Missouri Assessment Program (MAP) for the English Language Arts (ELA) and Mathematics Spring 2016 scaling plan and options. A historical background of the MAP ELA and Mathematics reporting scales is provided first.

In the 2015 administration year, the ELA and Mathematics test scores were reported on the Smarter Balanced Assessment Consortium (SBAC) vertical scales. These scales ranged from ~2100 to ~2700 scale score points for ELA and from ~2200 to ~2800 scale score points for Mathematics across grades 3 through 8. The reported test scores were 4-digit scale scores.

In prior administration years (2006 to 2014), the ELA and Mathematics MAP scores were reported on the custom MAP vertical scales. These scales were developed after the first operational test administration by linking the MAP assessments to CTB's *TerraNova* (2003) Reading and Mathematics vertical scales via *TerraNova* items embedded in the MAP assessments. Although the *TerraNova* vertical scale properties were used to establish the initial MAP vertical scale properties, the MAP vertical scales were further customized (for example, by adjusting lowest and highest obtainable scores based on Missouri student score distributions). As a result, the MAP reported scores were not directly comparable with *TerraNova* scores for the same content area. The custom MAP scales ranged from 455 to 875 scale score points for ELA and from 450 to 885 scale score points for Mathematics across grades 3 through 8. The reported test scores were 3-digit scale scores.

The Spring 2016 MAP ELA and Mathematics assessments will be developed using the same content standards as the Spring 2015 assessments but will not include any items from the 2015 assessments. Also, DESE has expressed a wish to move away from SBAC scales and revert to reporting total test scores as 3-digit scale scores. Therefore, the 2016 MAP ELA and Mathematics assessments cannot be directly equated to the previous year's scales. In addition, due to the most recent content changes to the ELA and Mathematics assessments, it is not recommended to link the 2016 assessments to the 2014 assessments. Therefore, no items from the prior administration years (2014 or earlier) will be included in the 2016 ELA and Mathematics assessments.

Vertical Linking Design

Given the described above constraints, new ELA and Mathematics vertical scales will be established after the 2016 test administration. To accomplish this goal, sets of off-grade level items (that is, items from the grade above, or from the grade below, or from both grades above and below) will be administered to samples of Missouri students taking on-grade level operational tests to facilitate between-grade assessment linking. For example, a sample of grade 3 students will take a set of grade 4 items in addition to a full grade 3 operational test; a sample of grade 4 students will take a set of grade 3 items in addition to a full grade 4 operational test; and, another sample of grade 4 students will take a set of grade 5 items

in addition to a full grade 4 operational test. The off-grade level items (also called vertical linking items) will not count towards the student test score. The vertical linking items will represent the test content of the grade from which they are selected and the grade in which they are administered. There will be approximately 19 items in each vertical linking set for ELA and approximately 11 items in each vertical linking set for Mathematics.

Vertical Linking Methodology

We propose to use two IRT models to calibrate the operational test items and the off-grade level linking items. The three-parameter logistic model (Lord, 1980) will be used to estimate parameters for the multiple-choice items. The two parameter partial credit model (Muraki, 1992; Yen, 1993) will be used to estimate parameters for the constructed response items. This methodology is consistent with the methodology employed for MAP Science data analysis.

We propose to establish the ELA and Mathematics vertical scales in a process of concurrent calibration of test data for all grades within a content area. Concurrent calibration is a method that allows for the establishment of the common scale in a single step—the calibration phase—by simultaneously estimating the parameters for all items at all grades. The resultant estimated parameters will be on the same scale. In addition, population ability estimates will be obtained for multiple groups. The population mean and standard deviation for the base grade (usually the grade 'in the middle') will then be used to compute the transformation constants to convert the parameter estimates of the base grade and other grades onto the common scale score metric.

An alternative method of a vertical scale development is a separate calibration and chain linking method. In this method, the vertical linking is accomplished in two steps. The first step will be the separate calibration of each grade level's data. The second step will be grade-by-grade chain linking using common items between adjacent grades. The separate calibration will result in the establishment of a unique theta metric scale for each grade. Next, the item parameter estimates for common items between grades will be used to estimate scale transformation constants that would allow placement of item parameters from each adjacent grade onto the base grade scale using the Stocking and Lord (1983) equating technique. This step will then be repeated for each adjacent grade until all grades are placed on the common scale.

Each method has its merit. The separate calibration is believed to be more appropriate when multidimensionality may be present in the data (as the data may be more likely to be unidimensional between two adjacent grades of a content area than across a larger grade span (Kolen & Brennan, 2004)). Concurrent calibration is more efficient, as all grades are calibrated in a single run and a common scale is developed as a result of simultaneous calibrations of the common items across grades. Further, the pooling of data across grades under concurrent calibration tends to produce more stable parameter estimates for the common items. Research conducted to date has not provided a definitive answer with regard to the best method for practice. Rather, a review of the literature provides a better understanding of the complexities involved in establishing a vertical scale, and unique problems associated with the different methods. Kim and Cohen (1998) argued that separate calibrations are preferable when the number of common items is small; however, as the number of common items increases different scaling approaches tend to produce similar results. Research also suggests that concurrent estimation results in better performance than separate estimation when the model is correctly specified (Hanson & Béguin, 2002). Additionally, Kim and Cohen (2002) found that parameter recovery using the concurrent estimation method was consistently better than parameter recovery from the separate calibration and the chain-linking for both item and ability estimates. The differences were, however, very small.

Regardless of which method is decided upon for implementation with the MAP assessments, the psychometric properties of linking items will be evaluated in terms of item classical statistics, item fit, and item standard error. Non-convergent items will be removed from linking sets. Items with poor classical statistics or poor item fit may also be removed from the linking sets.

In summary, it has been generally our experience that both methods of scale development tend to provide very similar results; therefore, we recommend the more efficient approach: concurrent calibration of all grades within a content area.

Vertical Scale Properties

The vertical scale properties are established in the year in which the scale is developed. The properties of a 'base' grade student score distribution are typically chosen in an arbitrary manner and the test properties of the remaining grades are aligned with the base grade to create scale continuity across multiple grades. Typically, the mean and standard deviation of ability estimates for the base grade are used to identify transformation constants that would allow transformation of item parameter estimates in a theta metric (from calibration) into a scale score metric and to produce student scale score distribution with a desired mean and standard deviation for the base grade. For example, a scale score mean of 500 and a standard deviation of 50 can be specified for grade 5. The transformation constants will then be developed to transform the grade 5 mean and standard deviation in the theta metric to the scale score mean of 500 and standard deviation of 50. The same transformation constants will then be applied to all other grades and the item parameter estimates in the theta metric for these grades will be transformed to the scale score metric. The item parameters in scale score metric will then be used to score students. The resulting grade level scale score means are expected to increase as the grade level increases.

Note, that if the concurrent calibration method is employed to establish a vertical scale, any grade can be selected as a 'base' grade. If the separate calibration/chain linking method is used for scale development, it is strongly recommended that the grade 'in the middle' is selected as a 'base' grade to reduce cumulative linking error for the lowest and highest grades.

Setting the base grade mean and standard deviation in the year in which the scale is developed is to a large degree arbitrary and does not affect the overall psychometric properties of a vertical scale. However, in case of MAP, it may affect the stakeholders' *perception* of the ELA and Mathematics scales. For example, setting ELA and Mathematics scale properties to be similar to the properties of the 2014 vertical scales may lead to unintended scale score comparison between the 2014 and 2016 administration years. Such score comparison will not be appropriate given that there will be no direct link between the two scales. If new ELA and Mathematics scales are set to be 'similar looking' to the past MAP scales, we recommend that an appropriate message about the past and new scale differences be communicated to the stakeholders. In order to avoid unintended scale and score comparisons, DESE may consider setting the new vertical scale properties to be very different from the properties of the past vertical scales. For example, setting grade 5 mean and standard deviation to 500 and 50, respectively, will result in the overall grades 3-8 scale for either ELA or Mathematics to be different from the past Missouri scales, which had a mean of about 675 and the standard deviation of about 35 for grade 5 ELA and the mean of about 670 and the standard deviation of about 45 for grade 5 Mathematics. Other base grade mean and standard deviation values may also be considered and different base grade psychometric properties may be considered for the ELA and Mathematics assessments.

Vertical scale evaluation

The scale evaluation will include examination of the following scale properties: the ordinality and separation of test characteristic curves, test standard error curves, pattern of grade to-grade growth (grade level scale score means), grade-to grade variability (grade level standard deviations), verticality of grade-to-grade lowest and highest obtainable scores, and separation of grade score distributions. These results will be shared and discussed with DESE and the Missouri TAC prior to finalizing the ELA and Mathematics scales.

References

CTB/McGraw-Hill (2003). *TerraNova, The 2nd Edition.* Monterey. CA: Author.

Hanson, B. A., & Béguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common item equating design. *Applied Psychological Measurement*, Vol. 26, No. 1, 3-24.

Kim, S., & Cohen, A. S. (2002). A comparison of linking and concurrent calibration under the graded response model. *Applied Psychological Measurement*, Vol. 26, No. 1, 25-41.

Kim, S., & Cohen, A. S. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement*, Vol. 22, No. 2, 131- 143.

Kolen, M. J. and R. L. Brennan. (2004). *Test Equating, Scaling, and Linking: Methods and Practices* (2nd ed.). New York: Springer

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems.* Hillsdale, NJ: Lawrence Erlbaum.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement* 16, 159–176.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7, 201–210.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement* 30, 187–213.