Missouri EOC Equating Plan

The equating plan described below will serve to maintain consistency of the MAP End-of-Course Assessments score scales over time and ensure that the performance standards remain at the same level of difficulty until such time they are revised. Riverside Publishing's psychometric team will perform all analyses necessary to scale the End-of-Course item pools and to ensure that each new operational form is equated to all current and previous forms.

As originally proposed, we plan on using a Pre-Equating model (Kolen & Brennan, 2004[1]) to produce equated forms for each End-of-Course Assessment. Once the system is up and running, a clear advantage of this model is that raw-to-scale score conversion tables can be produced *before* a new form is administered operationally. Consequently, it is then possible to report student scores without waiting for post-equating to occur after the administration of the test. The steps we plan to follow for Pre-Equating are described below.

Step 1

The following table shows the number of field-test forms, and their composition, for the spring 2008 stand-alone field test.

2008 Stand-Alone Field Test

Assessment	Session I		Session II	# of Forms
	Selected-Response Items		Number of Performance Events/Writing Prompts	
	Number of Items	Number of Passages		
Algebra I	18	N/A	1	10
English II	30	5	1	10
Biology	18	N/A	1	10

[1] Kolen, M. J. and Brennan, R. L. (2004). *Test Equating, Scaling, and Linking: Methods and Practices (2nd Edition)*. New York: Springer-Verlag.

For the spring 2008 stand-alone field test, for each assessed content area, 10 forms containing multiple-choice items and 10 forms containing a performance event or writing prompt will be spiraled within each classroom. Assuming random equivalent groups, the complete pool of items for each content area will be simultaneously calibrated (put on a common IRT scale). This effort will be accomplished in spring 2008 for Algebra I, English II, and Biology, and in spring 2009 for the nine additional assessments.

Because the characteristics of selected-response and written-response items are different, two item response theory models will be used in the analysis of field test forms. More specifically, the three-parameter logistic (3PL) model (Lord & Novick, 1968[2], Lord, 1980[3]) will be used for the selected-response items and the generalized partial credit (GPC) model (Muraki, 1992[4]; Yen, 1993[5]) for performance event/writing prompt items.

Step 2

Four alternate forms will be constructed, consistent with the test blueprint, and using IRT statistics from the initial, simultaneous calibration. Each operation form will also contain new items embedded for field testing. The following table shows the composition for the operational tests.

Operational Test Design for Core Assessments (Four Forms)

	Session I		**Session II**
	Selected-Response		**Performance Event/Writing Prompt**
	OP	**EFT**	**OP**
Algebra I	35	12	1
Biology	35	12	1
English II	35	12	1

OP = operational items
EFT = embedded field-test items

[2] Lord, F. M. and Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
[3] Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
[4] Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16(2)*, 159-176.
[5] Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30(3)*, 187-213.

Due to scheduling issues in 2008-2009, only the operational form administered in the spring of 2009 will contain "live" field test items.

Step 3

While the simultaneous calibrations following the stand-alone field test are sufficient for building alternate forms (Step 2), the first spring operational administration of a given test form will serve as the base form. The IRT parameter estimates for the items in the base test form will be used to re-center the item pool. This step will scale all items in the pool to the new base (operational) scale. This is a one-time only activity after the first operational administration.

Step 4

For the spring operational forms in subsequent years, new items will be embedded for field testing. The IRT parameters for the already scaled operational items will be fixed and the estimates for field-test items will be allowed to float. This process will allow all new items to be placed on the same MAP End-of-Course Assessments scale as the other items in the pool.

As outlined above, not only can the pre-equating model be used to annually build alternate test forms, but by using an embedded field-testing approach, DESE will also be able to maintain its item pools.