

To: Michael Muenks
From: Andrew Porter, Chair, DESE Assessment Technical Advisory Committee
Subject: Advice to DESE based on meeting of 2/8/08

The 27th meeting of the DESE Assessment Technical Advisory Committee took place on February 8, 2008 from 8:30 AM to 4:00 PM at the Hilton St. Louis airport hotel. Members of the committee in attendance were Andy Porter, Chair, University of Pennsylvania; Robert Linn, Professor Emeritus, University of Colorado, Boulder; Ron Mertz, Consultant Emeritus, St. Louis Public Schools; Barbara Plake, Professor Emeritus, University of Nebraska; Ed Roeber, Professor, Michigan State University; Phoebe Winter, Independent Consultant. In attendance from DESE were Michael Muenks, Stan Johnson, Becky Odneal, Andrea Wood, Margie Vandeven and Susan Newbold. Persons in attendance from CTB were Anita Benson, Lindy Weinland, Karla Egen, Thakur Karkee, Dennis Bullard and Alan Sheinker. Persons in attendance from ARC were Tim Parshalls, Lisa Sireno and Rebecca Bryant-Fritz. Persons in attendance from Riverside were Meredith Durgin, Molly Zebrowski, Bill Insko and Sara Gaddis-Brazzle.

Assessment Update

Michael Muenks sketched the Missouri Assessment Program, stating that mathematics is tested in grades 3-8 and 10, and communications arts is tested in grades 3-8 and 11. The high school testing program is being replaced by end of course exams. End of course tests can be taken by students in lieu of taking the course, and, if passed, the student will receive course credit. Science is tested in grades 5, 8, and 11 beginning in Spring 2008, with the standard setting in summer 2008. There are alternate assessments in mathematics, communication arts, and science that parallel the regular assessment.

DESE has entered into contractual agreement with Riverside for the end of course testing program. End of course tests are being developed for Algebra I, English II, Biology, and American government. Item writing took place in September 2007 followed by bias review and field testing in May 2008. The first operational form will be administered in October 2009. End of course exams have yet to be developed in Algebra II, Integrated Math II & III, English I, American government, and American history. End of course exams for science will begin with item writing in Fall 2009 and operational form administered first in October 2010.

DESE is recommending that end of course testing be weighted no more than 10-25% in determining a student's final grade in the course. The tests are not to be used for high school graduation. DESE is still deliberating on how end of course exams will be used for NCLB purposes.

Growth Modeling

Becky Odneal indicated that Missouri will be submitting a proposal to the US Department of Education on or before February 15, 2008 to use "value-added" as a part of the Missouri district and school accountability program for NCLB purposes. No final decision has been made on the growth model to be used, though Odneal thought that the model would give students three or possibly four years to become proficient, with each student having his or her own trajectory for

determining academic yearly progress. The trajectories will be stated in terms of scale scores on the vertical scale for grades 3 to 8.

The TAC anticipated the use of the growth model under the restrictions set by the US Department of Education would likely make little difference in what schools would be identified as meeting or failing their AYP targets. This is because the US Department of Education requires that the growth models result in all students reaching proficiency by 2014. Andy indicated that a student here at Penn has written a paper investigating the properties of several growth models being used by various states, and that he would send the paper to Michael for distribution within DESE as he sees fit.

The TAC recommends that DESE not extend growth modeling into the high school years. Growth modeling is feasible for grades [3-?]4-8 because there is every-year testing in mathematics and reading and results are reported on a vertical scale. For high school, however, DESE is moving to end of course exams, as was just noted, and the TAC sees no way to put those end of course exams on the same vertical scale as the scales for grades 3-8. First, end of course exam content taught will be different than the content assessed in grades 3-8. Second, the time period between end of course testing and grade 8 testing will vary, but is likely to be more than a single year, making development of a vertical scale difficult, if not impossible. Neither did the TAC know how to make end of course exams comparable across courses.

Odneal clarified that growth modeling would be yet a third way that a school might meet their AYP target. Thus, a school could meet AYP based on status (e.g. the percent of students proficient judged against an AYP target), safe harbor, or through growth. In growth the question would be whether sufficient percentages of students are on target to be proficient in three (or four) years, or by 8th grade. The TAC asked whether student specific trajectories would be calculated once or recalculated each year. No decision has been made. Calculated once would be more consistent with the status model, but recalculated each year might make the targets a bit easier to reach and offer a student more time to be proficient.

The TAC recommends that DESE conduct simulations with whatever growth model they decide upon to show the likely impact in terms of types and percents of students meeting AYP.

To Bank or not to Bank

Some students will take end of course exam earlier than others. For example, a student might take the Algebra I end of course exam in 7th or 8th grade. The question becomes, how will these results be used, if at all, in determining a high school's AYP and other forms of accountability. The idea of banking is that a student who takes the Algebra I test in 7th grade would be counted as having taken the Algebra I test when the student is in high school (say, 10th grade).

The TAC recommends against banking. The rationale is that high school accountability should be kept as simple as possible. High schools should be held accountable for the students that they are educating at the time that they are educating them. Thus, the TAC recommends against banking because that would hold the high school accountable for student accomplishment before they attended the high school. **The TAC further recommends that DESE monitor**

carefully any resistance or disincentives that might apply to students for taking advanced coursework early and/or end of course tests early, since that would defeat a part of the purpose of those tests. The main purpose of end of course tests is, however, to standardize the content of key high school courses across the state. **The TAC further recommends that information routinely be collected and reported on percent of students taking a particular end of course test, and the grade levels of those students, and that this information be reported alongside the results of the student testing.**

Grade Level Expectations

At the last meeting, the TAC recommended that DESE revisit the grade level expectations in communication arts and mathematics, especially for grades 6, 7, and 8 in communication arts and grade 5, 6, and 7 in mathematics. The TAC's recommendation was based on lack of between-grade separation in cumulative achievement distributions on the vertical scale. The TAC hypothesized that a portion of the problem for a lack of between-grade separation might be due to lack of clear distinctions in grade level expectations from one grade to the next. Michael Muenks reported that Missouri teachers had been involved in revisions of the grade level expectations, and that the revised grade level expectations are now being used to guide item writing. Muenks noted in particular that 8th grade is now placing a clearer and greater emphasis on pre-algebra.

Muenks asked the TAC how much change was possible without having to reset the achievement levels and the performance descriptors.

CTB has conducted content analyses of the old and new GLE's. Results of that analysis led them to conclude that there was not sufficient change in content to create concern about needing to reset achievement levels and performance descriptors. Apparently the test specifications have not been changed. Thus, while the new GLE's are used to guide item writing, selection of items into a form for operational use is guided by the unchanged test specifications.

The TAC recommends that DESE conduct a two-stage study. In stage 1, the items of the old test and the items of the new test are content analyzed and compared one to another to see if the content is sufficiently different between the old test and the new test. If the answer is no, stage 2 is not necessary, and the achievement levels and performance descriptors can stand as they are, unchanged. If the content appears to have changed, then student results should be analyzed for the existing assessment and the simulated new test to see if there has been a shift in level of student performance. Some ambiguities remain in the implementation of this recommendation. First, how big of a difference in content and/or student performance is necessary to cause concern? For this, the TAC had no recommendation, but would be willing to look at the results and advise further at that time. Second, if there is a big enough difference to cause concern, what next step should be taken? Here the TAC recommends that since DESE would like not to change the achievement levels or performance descriptors, the next step might be to return to the old test.

The Consequential Validity Study

The Assessment Resource Center at the University of Missouri is conducting a consequential validity study of the MAP using surveys. The sampling design was described in a document submitted to the TAC in advance of the meeting and was discussed at the meeting. The sample has already been taken in most, but not all cases, following the design described in the document. While the TAC found the sample design to be, in general, quite good, the TAC suggests that the description of the design clarify how students and parents are selected within a school. At the meeting, ARC indicated that the principal is instructed to select one classroom and all of the students and parents connected with those students are then in the survey.

Response rates to the surveys varied by group surveyed, from a low of 36% for teachers to a high of 86% for superintendents. The TAC expressed concern about the variable and sometimes quite low response rates to the survey. A rule of thumb for acceptable response rates is 75% or more. **The TAC recommends that ARC describe the characteristics of their obtained sample and compare it to the characteristics of the state for each respondent group.**

The parent survey has not yet been administered. The TAC wondered about surveying parents for whom English is not their primary language. Language was recognized as a problem, but no solution was identified.

Some early results from the survey were reported. For example, one question asked, “MAP scores accurately reflect (a) student learning, and (b) effective teaching”. The TAC cautioned ARC about interpretation of the results of this question. A respondent might believe that the MAP accurately reflects the level of student achievement in the subject tested, but they might disagree with the question as asked. For example, “student learning” extends beyond the subject tested and doesn’t make clear that what one is asking about is the level of student achievement. As for effective teaching, many might believe that the MAP does not assess teacher effectiveness because it does not estimate value-added to student achievement associated with individual teachers.

The TAC expressed considerable interest in the results of the survey and asked to receive a copy of the report, if that would be possible, when it is completed.

Technical Manual for MAP

The TAC did not receive a copy of the technical manual in advance of the meeting, so no discussion was possible. Muenks will arrange to have the technical manual sent to members of the TAC, and a conference call will be used to hear the TAC’s reactions.

CTB Strategy for Evaluating Anchor Items

The document was distributed to the TAC in advance of the meeting, describing CTB’s plans for evaluating anchor items. Essentially, CTB will use two concurrent criteria. One criterion asks for the correlation of *a*- and *b*-parameters between estimates and inputs. If *a*-parameters are correlated less than 0.8 or *b*-parameters less than 0.9, CTB will look at outlier items. For any outlier items, CTB will investigate issues of content coverage. Anchor items will only be dropped when no reasonable explanation for difference in performance is available. For the second criterion, CTB will compute differences between the item ability regression curves of the anchor items. Where there are large differences, items will be flagged and, again, content

reviewed. Anchor items will be dropped only when there is not a reasonable alternative explanation for the large differences. **The TAC held these plans to be reasonable.**

Standard Setting

CTB distributed to the TAC a document describing a proposed plan to set achievement levels for science in July 2008. A separate document described the CTB standard setting handbook. In Missouri, there is Senate Bill 1080 requiring that MAP achievement standards meet but not exceed the achievement standards for NAEP. To comply with Senate Bill 1080, CTB has used in the past, and proposes for science, a restricted bookmark approach to standard setting. In essence, standard setting analysts are given a range of cut points for proficiency. The range will be the same as was used for communication arts and mathematics in setting standards in those two content areas in December 2005. A high end of the range allows for 26% of students at or above proficient. A low end of the range allows for 43% of the students at or above proficient. CTB proposes to use 15 panelists for each grade.

The TAC was in general support of CTB's plan, noting that it paralleled the way in which standards had been set by CTB for communication arts and mathematics. The TAC did recommend one change for smoothing across grade levels. Rather than using table leaders who do not vote on standard setting within a specific grade level, CTB should identify panelists in advance and designate those panelists to participate in the smoothing process once achievement levels have been set grade level by grade level.

What Counts as a Student Attempt

The US Department of Education has advised Missouri that they should change their procedure for deciding whether or not a student has had a sufficient assessment attempt. The US Department of Education is pushing DESE to count every student as tested who has completed one or more items on the assessment. **The TAC recommends that if DESE goes in this direction, students should not be assigned a scale score of 0, but rather be given the lowest observable scale score for the assessment they are judged to have taken. A scale score of 0 would throw off the results, since it is impossible, for example, for a 4th grader to get a score of 0, no matter how poor their achievement.**

End of Course Exams

Several documents were circulated in advance of the meeting to members of the TAC describing Riverside's work on developing end of course exams. After reviewing the test blueprints, the TAC asked if performance would be reported at the content strand level as well as a total score for each course. Riverside said the intention was to report at the strand level, and that the metric would be percent of items correct. Scale scores would be used for the total score on each end of course test and achievement levels will be set on the total score. **The TAC expressed concern about reporting at the strand level, wondering if there would be sufficient precision. The TAC recommended against reporting percent of items correct, since that metric would likely result in misinterpretation of results. For example, if the student gets 5 of 7 easy items correct, that might be judged better performance than a student who gets 2 of 7 hard items correct, when in fact that might not be the correct interpretation.**

The TAC noted that students will take the end of course exams at the end of the course, and so the results will not be diagnostic for the student. The TAC recommends that reporting be done at the total score level, and not the strand score. **The TAC cautions that strand scores are notoriously unreliable, (a) because of the few number of score points, and (b) because they are highly inter-correlated, making the standard error of their difference quite large.**

The TAC noted that in general there were fewer score points for end of course tests than would be desirable. The TAC wondered if it would be possible to increase the number of score points by decreasing the amount of time spent on performance items to generate more score points. For example, in Algebra I an entire hour is devoted to performance items yielding only 4 score points.

Initially, there will be 4 forms developed for each end of course test, one of which will be released, and the other 3 of which will be used across 3 testing windows within a single year. **The TAC expressed concern about form security, and recommended the DESE work to keep the testing window as narrow as possible. In the future, there will be additional forms created, and that will protect against cheating, but in the short run, cheating may be a problem.**

The TAC recommends that for its next meeting, the purposes of the end of course test be articulated in a document circulated in advance, and that plans for reporting performance on the end of course exams also be circulated in advance of the meeting and discussed at the meeting.

Vertical Scale Linkage

Riverside is considering attempting to put the end of course exams on the same vertical scale as the vertical scale in grades 3-8, one for mathematics and one for communication arts. **The TAC recommends against attempting to put the end of course tests on a vertical scale, connecting performance on the end of course tests to performance in the subject in grades 3-8. The TAC's recommendation is consistent with its recommendation earlier in this report, under growth modeling.**

The TAC understands that there is an interest in being able to say if a student is proficient in mathematics in grade 8, that student is on track to be proficient on the mathematics end of course test that they will take later. The TAC concluded that prediction studies could be conducted to see how performance in earlier grades predicts performance on end of course tests. The TAC further concluded that the achievement level setting for end of course tests could take into account the desire to have proficiency on the end of course test, consistent with proficiency in 8th grade.

Item Bank Scaling/Linking Items

Riverside described that in the field test there would be 20 forms per content area, 10 multiple choice forms and 10 performance item forms. The plan is to spiral these 20 forms within each classroom, and to conduct a joint calibration so that operational forms can be created through pre-equating. In the future, items will be field tested through embedded forms. **The TAC thought the Riverside plans were reasonable, though the TAC recommends that not all**

forms be used with each district; rather, 5 multiple choice and 5 performance forms be used with half the districts, and the other 5 and 5 be used with the other half of districts. This will protect against breaches in security.

Determining the Comparability of Paper and Pencil vs. Online Assessment

Riverside is building both a paper and pencil and an online version of end of course tests. They submitted a plan for studying the comparability of these two modes of assessment, and the TAC discussed their plans. The TAC observed that approximately 40% or more of the students would need to be assessed online before there would be a savings in the overall cost of assessment.

The TAC recommends that the Riverside design be modified. Matching at the district level would not be sufficient for two reasons: first, matching at the district level means that there will be district design effects which will make the effective sample size much less than the 500 proposed. Second, there may well be differences among districts in the percent of students who take a particular end of course test. Thus, matching at the district will not be an effective match as to the type of student tested. The TAC recommends that matching be done at the individual student level, and the sample size be increased to 1000 students minimum for each mode of assessment.

Standard Setting

Riverside described three options for standard setting. All three options set standards on field test data. **The TAC recommended against setting achievement levels on field test data.**

The problem is that there are approximately 40,000 students in Missouri who will not be assessed by the current MAP tests at the high school level, nor will they be assessed with an operational end of course test. At the same time, NCLB requires that all students be tested. **The TAC recommends that students participating in end of course field tests have their data maintained in a bank, and that when the achievement levels are set on the operational data a year later, those achievement levels be applied to the field test data. This allows the achievement levels to be set on operational data and at the same time meet the NCLB requirements that all students be tested and achievement levels for that be reported.**

Future Meetings

The TAC and DESE concluded that in the near term there needed to be more frequent and longer meetings of the TAC. A day and a half meeting will take place May 27 and 28. The TAC will also meet for a day and a half August 21 and 22.