National Assessment Governing Board

Committee on Standards, Design and Methodology

May 13, 2011

COSDAM Attendees: Lou Fabrizio (Chair), Tonya Miles (Vice Chair), Steven Paine, James Popham, and Blair Taylor.
Other Governing Board Members: John Q. Easton (*Ex officio*), Director of the Institute of Education Sciences.
Governing Board Staff: Susan Loomis, Michelle Blair, and Ray Fields.
Other Attendees: NCES: Commissioner Jack Buckley, Associate Commissioner Peggy Carr and Andy Kolstad. AIR: George Bohrnstedt. California Department of Education: Deborah Sigman. ETS: Andreas Oranje. HumRRO: Lauress Wise. Measured Progress: Luz Bay. MetaMetrics: Heather Koons. Northwest Evaluation Association: Steven Wise. Oregon Department of Education: Tony Alpert. Pearson: Connie Smith. SMARTER Balanced Assessment Consortium: Joe Willhoft. Westat: Marcie Hickman and Dianne Walsh.

Lou Fabrizio, Chair of the Committee on Standards, Design and Methodology (COSDAM), called the meeting to order at 10:00 a.m. and welcomed members and guests.

1. Developing Writing Achievement Levels Descriptions

Susan Loomis provided an overview of the process underway to develop writing achievement levels descriptions (ALDs) to be used for setting achievement levels for the 2011 NAEP in writing and for reporting results of the writing NAEP. Ms. Loomis noted three major stages for the process of developing the descriptions, each involving reviewing and editing the descriptions by a core group of content experts who participated in development of the new NAEP framework. Draft ALDs are evaluated for alignment to the policy definitions of Basic, Proficient and Advanced, for the calibration across achievement levels within each grade, and for the calibration within achievement levels across grades.

1. ALDs drafted by a core group of content experts who participated in development of NAEP writing framework
    - Draft ALDs reviewed and evaluated by a larger group of content experts who participated in development of NAEP writing framework
    - Modifications by a core group of content experts

2. Delphi Process implemented with state curriculum specialists who reached agreement on modifications to be recommended in order to increase alignment of ALDs to policy definitions and calibration across levels within grades and within levels across grades. A

total of 39 recommendations were agreed upon and addressed by the core group of content experts who implemented the recommendations, as they deemed appropriate.

3. Focus groups with curriculum experts, teachers, school and district administrators, parents, business leaders, and students
   - Modifications by core content experts
   - Review by larger group of content experts

The achievement levels descriptions are to be finalized and prepared for presentation to the Governing Board by June 2, 2011. COSDAM will evaluate the achievement levels descriptions at the August 2011 meeting and determine whether to give provisional approval for their use in the achievement levels-setting panel studies. Final approval of the descriptions will be decided in May 2012 as part of the full set of achievement levels-setting outcomes. Jim Popham had questions about the Delphi Method. Specifically, he asked if any "false" descriptors were planted in the ALDs to check for the attentiveness of reviewers. Ms. Loomis stated that there was no plan to do that and that the nature of the Delphi Method would largely eliminate the potential for negative effect by a small number of panel members.

2. Writing Achievement Levels Setting Update and the Body of Work Standard Setting Method

Luz Bay, of Measured Progress, presented an overview of the Body of Work method to be implemented for the writing NAEP. The Body of Work (BoW) method of standard setting was developed by Measured Progress specifically for use with performance assessments, and it is considered to be most appropriate for an assessment such as the writing NAEP that includes only two measures of performance for each student. Ms. Bay noted that the achievement levels-setting (ALS) process to be implemented for the 2011 writing NAEP includes all of the aspects of the ALS process typically included, and only the method of collecting panelists' ratings and the feedback provided will be different.

This is a "booklet classification" type process to determine the maximum discrimination between each achievement level to represent the cut score. A sample of student test booklets will be presented in order from lowest to highest scores. Panelists will review booklets relative to the achievement levels descriptions for Basic, Proficient and Advanced performance and classify each booklet into an achievement level category. Several questions were asked about the ordering of booklets—whether panelists would be instructed that booklets may be classified without regard to the ordering presented. Ms. Bay affirmed that panelists are allowed to classify booklets without regard to ordering and they will be so instructed. John Easton asked about the computation of cut scores with the BoW methodology, and Ms. Bay gave a brief description of how logistic regression would be used for the computation to find the score that maximally differentiates classifications between two levels. Mr. Popham expressed concern that individual panelists highly motivated to promote a specific outcome might have a large impact on the ALS process during discussions. He also expressed concern about the choice of "consequences" as the term and suggested that another term be adopted for this feedback of information on the percentage of students scoring at or above each level of achievement. This term has been used for providing feedback in NAEP achievement levels-setting procedures since the mid-1990s. Mr. Popham also noted that the exemplars for writing would not be "items," per se, but responses. Ms. Bay said that this term will be changed to "responses" in all instances for the writing ALS process.

Ms. Bay showed a graphical representation of the consequences data to display the ways the data can be shown with the Body of Work Technological Integration and Enhancements (BoWTIE) computerized version of the process. She concluded her presentation by discussing an impressive array of advantages available in the computerized version of the Body of Work methodology.

3. Update on 12th Grade Preparedness Research Program

Ms. Loomis noted that there are two important updates to discuss at this meeting: (1) the statistical relationship between NAEP and the SAT for reading and mathematics and (2) the judgmental standard setting studies.

Andreas Oranje of ETS reported on the analyses of statistical relationships for NAEP and the SAT in reading and mathematics. Two primary methods were examined: concordance and projection. The results of the two statistical methods were quite similar. The correlation between NAEP performance and SAT scores for mathematics was high and strong enough to warrant reporting concordance scores, but that would not be warranted for reading because of the lower correlation between scores.

The "college readiness" benchmark for the SAT in each subject test is 500. That score corresponds to a score around the Proficient cut score for the reading grade 12 NAEP (302) and a score a little below (165) the Proficient cut score for the mathematics grade 12 NAEP. In general, Mr. Oranje noted that the results of the two methods converge reasonably, especially the concordance results and the projections based on a 50% criterion, and he also noted that the results for each of the subgroups are generally consistent with expectations from other studies in NAEP. When estimated for the full population, as opposed to the matched sample of NAEP and SAT examinees, the percentage of students scoring at or above the SAT "college ready benchmark" on NAEP would be nearly 40% for both mathematics and reading.

John Easton asked about the focus on test results for NAEP preparedness reporting, given the advice of the College Board and ACT to supplement test performance scores to predict readiness or for decisions on college admission. Steven Paine voiced his agreement that this is an important issue. Ms. Loomis suggested that the statistical relationships will produce one set of data to evaluate, relative to other sources of information. She also suggested that the Board may use data from the 2009 High School Transcript Study to add more information for analysis of the indicators of preparedness. Ray Fields added that analysis of data for Florida students, available through special agreement with Florida, provides a source of these data for analysis with Florida NAEP data and performance on both the ACT and SAT.

Mr. Fields elaborated on plans for the future and the potential for adding partners as sources of additional data to inform our research on 12th grade NAEP preparedness. Further, he noted that the 2015 High School Transcript Study provides the potential for collecting information about course rigor, high school grade point average, and other information to evaluate in conjunction with the NAEP data and test data from other sources, such as the SAT, to have a more complete set of predictors of college preparedness.

Blair Taylor asked for more information about how the preparedness research data will be used, and Ms. Loomis described plans for reporting the preparedness results by early 2012. Mr. Fabrizio noted that the chart on pages 18-22 of the COSDAM briefing materials provides the timelines for each set of studies.

Ms. Loomis next discussed the judgmental standard setting studies (JSS) that are now underway. She reported that the pilot studies were conducted two weeks ago and that the operational studies will begin in less than two weeks, May 24, 2011.

The design of the JSS calls for replicate panels, and both panelists and the item pool are divided to be as similar as possible and assigned to each paired group. A total of eight panels are convened concurrently for each JSS session: replicate panels for each subject, mathematics and reading, for each of two post-secondary activities at each session. JSS studies will be implemented for job training programs in five different occupations (automotive master technicians; computer support specialists; heating, ventilation, and air conditioning technicians; licensed practical nurses; and pharmacy technicians) and for placement of college level credit bearing courses that fulfill general education requirements.

Ms. Loomis presented results for the pilot study across rounds and across groups, and showed results relative to NAEP achievement levels. The pilot cut scores were quite high—higher than the NAEP proficient cut score for both reading and mathematics. And, relative to the ACT “college and career readiness” benchmark and the SAT “college success” benchmark, the performance requirements for NAEP preparedness appeared to be more demanding.

A list of ten recommendations for improving the JSS process developed with the technical advisors was shared with COSDAM.

Mr. Easton noted that he had observed several NAEP ALS procedures while a member of the Governing Board, and he urged members of COSDAM to take the opportunity to observe these studies. Ms. Loomis thanked him for this encouragement and again invited all members of COSDAM to observe either the JSS studies or the writing ALS.

4. Motivation Research: Measuring Test-Taking Behavior

Steven Wise of the Northwest Evaluation Association in Portland presented information from his research on detecting motivation, or the lack thereof, through response time analysis based on student behaviors in computer-based assessments. This is important information as the Board moves toward computer-based testing in NAEP. Mr. Wise's research shows that rapid guessing behaviors are associated with low motivation. Several correlates of rapid guessing behavior were presented, including gender, age, academic ability, item position in the assessment, reading load of the item, and time of day. Rapid guessing behavior becomes more prevalent as grade level increases, and the difference in guess behavior between grade 8 and grade 9 students was greater than for any other adjacent grade levels. The impact of time of day was especially evident for males, compared to females; and the impact for males increased across the day so that motivation seems to decline rather sharply for assessments administered in the afternoon. However, performance score differences are greater at lower grades by time of day than for higher grade levels.

Several solutions to rapid guessing behaviors were suggested by Mr. Wise. For example, research by Mr. Wise shows that students typically engage in non-effortful behavior for only a relatively small portion of the test, so scores may be based on only the portion of the test for which student behavior is focused on solutions to test questions. Other research focuses on identifying rapid-guessing behavior in computer based tests. The computer can detect if non-effortful behavior is detected and send a message to encourage the student to re-focus on the test. Mr. Wise noted that computer based testing provides better opportunities for detecting behaviors associated with low-motivation and for correcting these behaviors during the assessment. Some of these solutions are being considered by the NAEP program and it may be possible to initiate evaluation of test-taking behaviors with the computer-based mathematics field trial administered in 2011. Mr. Wise acknowledged, however, that it is hard to know how much these strategies can impact NAEP.

5. Recommendations for Future Agenda Topics

Mr. Fabrizio asked the committee for suggestions of topics to be discussed in future sessions of COSDAM. Mr. Popham suggested that Mr. Fabrizio should contact the committee at the midway point between meetings to ask for suggestions. Mr. Fabrizio said that he would try that strategy.

The meeting was adjourned at 12:30 p.m.

I certify the accuracy of this report.

Louis M. Fabrizio

5-31-11

Lou Fabrizio, Chair

Date