National Assessment Governing Board

Committee on Standards, Design and Methodology

May 18, 2012

Joint Session of the Committee on Standards, Design and Methodology (COSDAM) and the Reporting and Dissemination Committee R&D)

COSDAM Attendees: Lou Fabrizio (Chair), John Q. Easton (*Ex officio* member of the Governing Board and Director of the Institute of Education Sciences), Terry Holliday, Jim Popham, Leticia Van de Putte, and Fielding Rolston.
R&D Attendees: Eileen Weiser (Chair), Tom Luna (Vice Chair), Andrés Alonso, David Alukonis, Anitere Flores, Sonny Perdue, and Mary Frances Taymans.
Governing Board Staff: Cornelia Orr, Susan Loomis, Larry Feinberg, Stephaan Harris, Ray Fields, and Michelle Blair.
Other Attendees: NCES: Brenda Wolff. AIR: Cadelle Hemphill. ETS: John Mazzeo. Andreas Oranje and Donnell Butler. HumRRO: Steve Sellman and Lauress Wise. Hager Sharp: Debra Silimeo and Lisa Jacques. MetaMetrics: Heather Koons. Optimal Solutions: Mark Partridge. Reingold: Amy Buckley and Valerie Marrapodi. San Antonio Express-News: Theresa Clift. Westat: Keith Rust and Dianne Walsh. Widmeyer: Jason Smith. West Virginia Department of Education (Policy Task Force Representative) Liza Cordiero.

Lou Fabrizio, Chair of the Committee on Standards, Design and Methodology (COSDAM), called the meeting to order at 10:10 a.m. and welcomed members and guests. Mr. Fabrizio stated that the joint committee session was for the purpose of discussing staff recommendations for reporting results of preparedness research studies that had been underway for over three years. While COSDAM has had updates at each meeting, the R&D Committee members have not had the opportunity for this level of detailed information. There will be a full briefing to the Board later in the day. Eileen Weiser, Chair of the Reporting and Dissemination Committee, had no additional remarks to add before the presentation of the report by Cornelia Orr, Executive Director of the Governing Board.

Ms. Orr noted that the briefing had been made available to members a few days in advance of the meeting. She had a Power Point presentation to show the key points, starting with questions that staff had identified as important questions about the research findings. Staff answers to the questions and the rationale, based on evidence from research findings, were presented for four sets of questions:

1. Can NAEP be used to inform the national discussion about the academic preparedness of U.S. students?
2. Will more than one preparedness reference point on each of the NAEP scales (reading and mathematics) be established?
3. What process was used to determine the recommended reference points and which findings to report?

4. What statements can be made about preparedness for job training?

Ms. Orr noted the following:

- Much more is known about preparedness for college course work than about preparedness for job training.
- The information about college preparedness is very general, referring to "typical" institutions and not to institutions that are differentiated according to admissions criteria.
- A reference point for the "just prepared" or "minimal academic preparedness" level has not yet been identified, but staff feel that the reference points recommended for "likely to succeed in the college freshman year" and "likely to need remediation" are supported.
  - The Board adopted a working definition of prepared to be the minimal level of academic preparedness required for placement in a credit-bearing college course of the sort that fulfills a general education requirement" or the "minimal level of academic preparedness required for entry in a job training program requiring at least three months of training but less than a bachelor's degree. (Other criteria were used to identify the specific occupations for research with the training programs.)
- Charts showing reference points were presented and information regarding each point was provided to explain any caveats regarding the data to be reported.

The recommendation is to report two reference points: (1) likely to be successful in freshman year in college and (2) likely to need remediation. These reference points are largely based on the statistical linking studies for NAEP with the SAT. The staff recommendation was to use the Proficient cut score for each subject as the reference point for preparedness for "college success." The Proficient cut score for mathematics grade 12 NAEP is associated with an 80% probability of scoring 500 on the mathematics SAT assessment, and the College Board has established 500 as the benchmark score having a .67 probability of earning a B- freshman year grade point average. For reading, for which there is a lower correlation between the NAEP and SAT, the probability of scoring 500 on the critical reading SAT assessment is .5 at the NAEP Proficient cut score.

Jim Popham asked about the implications of these recommendations for individuals—do they provide any indication for how to help students be prepared? He wondered what impact these data will have and suggested that the data would have a very short "shelf life" because interest in these results would be short lived. Ms. Orr noted that there is great interest in the question of what is prepared, and this research provides an answer. There will be other answers, and this will contribute to the national conversation in a positive way. For example, this information will be helpful to states that are setting cut scores for high school students to indicate academic preparedness for post-secondary activities.

Mr. Fabrizio noted that only 11 states participated in the grade 12 assessments, and only those states will have data regarding the preparedness of students, although the Proficient cut score for the nation is the grade 12 indicator of preparedness. For the 11 pilot states in 2009, these data will be very important.

Terry Holliday said that state policy issues are critical in Kentucky, but Kentucky does not assess at grade 12. Preparedness for college and the opportunity to attend are important issues in the

country, particularly for teachers and parents. He said that states that were not included in the grade 12 assessment will still make comparisons of their own data on student performance on the ACT and SAT to NAEP. Comparisons are the key for sates and these data will be very important as states move to the Common Core State Standards (CCSS). States currently have a large discrepancy between their own "proficient" level and the NAEP Proficient level. The CCSS levels will be much closer to NAEP.

John Easton noted that the controversy about NAEP achievement levels still simmers, although the NAEP achievement levels are designed, implemented, and reviewed fully and carefully. He wondered if the preparedness reference points were developed with the same scrutiny. Given concerns for the integrity of data reported, he recommended that the findings be subjected to external review before release.

Susan Loomis responded that the Technical Advisors had monitored the developments throughout the research process. She then named the technical advisors (Reckase, Campbell, Cohen, Bazemore, and Kolen), and identified the technical expertise of each in relation to this NAEP research for preparedness reporting. Each contract had technical guidance by COSDAM, the Contracting Officer's Representative (COR), and a principal investigator for the contractor; and the judgmental standard setting studies were under the technical guidance of a technical advisory committee with considerable expertise in standard setting, including service on the NAEP Technical Advisory Committee for Standard Setting (Haertel, Forsyth and Hambleton). Finally, experts in standard setting were brought in to observe the judgmental standard setting studies and provide additional technical guidance. The plan has been to vet the statements about preparedness and the evidence in support of those statements widely prior to reporting them. The technical community will be enlisted for this review, as well as a much wider audience of stakeholders.

Ms. Orr reported that the goal is not to set cut scores. Rather, the goal is to report research on academic preparedness. The percentage of higher education institutions that use the SAT for placement is too low for much assurance regarding the "needs remediation" cut score. But, the data from the national survey are very closely replicated by the data from Florida regarding the average NAEP score of students who were placed in remedial courses. The recommendations for discussion are about reporting findings—not about setting a preparedness cut score and not about policy on preparedness, per se.

Leticia van de Putte noted that it is extremely important for students to leave high school prepared for post-secondary activities. She stated that the data currently available on what students need to be prepared are irrefutable. The NAEP findings seem consistent with that information. The amount of training needed is huge; resources for K-12 and higher education have been reduced drastically. It is important for the Governing Board to emphasize that these are *findings* for preparedness and not a sort of "stamp of approval' from the Board regarding a specific score on NAEP that signals preparedness. More research is needed, but these data provide further confirmation of the need for more students to be better prepared. The results for NAEP look pretty similar to the picture for Texas. It is important that the reporting be worded carefully.

Mary Frances Taymans agreed that the language in the public release will be important, especially for the higher education community. We need to make sure that our language for reporting these data based on SAT score linkages can appropriately resonate with the higher education community given that a sizable portion of that community does not use the SAT or ACT to inform their admissions or placement decisions. The decisions for admission and for course placement in most institutions are based on more than a test score.

Andrés Alonso noted that the correlations for NAEP scores with SAT scores are different for reading and math, with a lower correlation for reading. That cannot be changed; it is what it is. But, using one probability for reading and one for mathematics seems to be aimed at using the Proficient cut score as the indicator for preparedness. It "feels" odd to do that—as if we are trying to find support for the Proficient cut score. Why look at 50% and 80% probabilities? What happens if we look at 65% or 75%--or any other way we might slice the data? What are the implications of changing the percentages on the results we have to report?

Mr. Fabrizio reminded the members that the overarching goal of this preparedness research work is to determine if 12th grade NAEP can be used for reporting preparedness of students for post-secondary activities in college or the workplace. The finding is that we can say something about preparedness, but that does not represent an endorsement of the Proficient cut score or any other score point on the NAEP scale.

David Alukonis stated that he was very concerned about the lack of data to report on career preparedness. This seems to indicate that career preparedness is a "dead end" issue for now. He wanted to know about next steps that would add information for reporting on career preparedness.

Ms. Weiser agreed with the concerns expressed by Mr. Alukonis regarding the lack of data for reporting on career preparedness, but she asked that Ms. Orr complete her presentation by moving to findings for reading next.

Ms. Orr reiterated that a 50% probability of scoring 500 on the SAT was chosen for representing "college success" as a point on the NAEP scale for reading. She explained that the decision was to be more conservative with reporting for reading, due to the lower correlation between SAT and NAEP reading scores. Staff recommended the 50% probability to reduce the likelihood of an under estimate of students prepared for college success and a corresponding over estimate of the need for remediation. The greater uncertainty in the relationship between NAEP and SAT leads to a greater difference in the scale scores for 50% and 80% probabilities.

Mr Popham address his next question to Ms. Weiser. He noted that Ms. Orr had referred to these as "staff recommendations," but he needed clarification regarding the purpose of the recommendations. It seemed a "funny game" to make recommendations to the Board about reference points while stating that the Board is not being asked to set standards.

Ms. Orr responded that COSDAM had asked staff to provide recommendations for preparedness research reporting. Staff had asked for COSDAM advice regarding the statements to make and their judgment regarding the extent to which findings provide compelling evidence in support of the statements. In response, COSDAM had asked staff to provide recommendations for their

review. This presentation of recommendations was developed in direct response to that request by COSDAM.
Mr. Alonso reiterated his concerns that using different probabilities to reference college preparedness for math and reading is problematic. It seems to elicit doubt and suspicion regarding the role of the Proficient cut score in reporting preparedness.

Ms. Orr agreed that this may seem that we are trying to support the Proficient cut score. She added that the Technical Advisors saw no problem with using different probabilities for reporting preparedness in math and reading.

Mr. Fabrizio reminded the group that the big question was "What can we say about preparedness?" and this report is presenting what we can say. We know full well that there is a need to do more research and to be able to say more about preparedness. He stated that he felt neither worried nor surprised to find that the level of preparedness for reading and math were different.

Ms. Orr then asked the members if there were any points that they would consider to be "show stoppers"—issues that would mean we should not report findings for the 2009 NAEP.

Mr. Fabrizio noted that while not a "show stopper," he was concerned about the plan to release the technical report a month later than the public report. He recommended that the two reports be released at the same time to provide the technical information that would be needed for the level of scrutiny that would ensue regarding the findings. It would be prudent to have the technical report available at the same time as the public report. (There was general agreement with this point.)

Jack Buckley urged the committees to think about two points. First, the Governing Board really needs to share the findings from the research on preparedness for job training programs. The difficulty of producing conclusive information on the job training preparedness is an important finding that needs to be shared with the research community. In addition, he recommended that more attention be given to describing and explaining the "indeterminate" region on the scale between "likely to need remediation" and "likely to succeed in college." People will think of these reference points as cut scores. The indeterminate zone is the zone of minimal academic preparedness that the Board aims to identify, and more needs to be said about that portion of the NAEP scale. There is implied approbation on the part of the Board regarding these findings. The Board guards zealously against error in NAEP achievement levels setting. So, the Board must make it very clear that the technical and public reports are being accepted by the Board but that the findings reported as reference points do not represent cut scores in the same sense as achievement levels cut scores.

Ms. Weiser noted that she is very interested in knowing more about the "soft skills" that NAEP cannot measure. And, she would like to know more about the actual role of algebra II as a pre-requisite for career preparedness. There is current controversy in the state of Michigan regarding this requirement for students in high school.

Mr. Alonso summarized his concerns by contrasting reporting *judgments* versus reporting *correlations*. If we are reporting judgments, then he feels that the Governing Board needs much

more deliberation on the issues. If we are reporting correlations only, then there is much less for the Board to discuss. It seems to him that the Board is reporting a judgment because of the recommendation to use different probabilities for reading and mathematics. This concerns him. It does not seem reasonable to report that more 12th graders are prepared for success in college based on their reading score than based on their math score.

Mr. Holliday affirmed the importance of soft skills to employers. He noted that Kentucky uses WorkKeys and industry certification data for career preparedness indicators. But, there are still no good measures of soft skills. He also noted that when he talked about the NAEP results to the other Chief State School Officers, not many seemed interested in having another measure of academic preparedness.

Anitere Flores stated that she did not know if we needed a new measure, but she did find it reassuring to note that the Florida data were consistent with the national data. The NAEP preparedness findings seemed to match reality and she thought it was positive to find such alignment of results.

Mr. Alukonis stated that he would send a report to Ms. Orr to be distributed to all members. The report is from the Federal Reserve Bank of Boston and it is relevant to the issues discussed today.

Mr. Fabrizio noted that the joint session had already lasted longer than scheduled, and he suggested that the joint session be adjourned so that the two committees could resume deliberations separately.

The joint meeting of COSDAM and the R&D Committee ended at 11:10 AM.

COSDAM Session
COSDAM Attendees: Lou Fabrizio (Chair), John Q. Easton (*Ex officio* member of the Governing Board and Director of the Institute of Education Sciences), Terry Holliday, Jim Popham, Leticia Van de Putte, and Fielding Rolston.
Governing Board Staff: Cornelia Orr, Susan Loomis, and Ray Fields
Other Attendees: ETS: Andreas Oranje. HumRRO: Lauress Wise. Measured Progress: Luz Bay. Senator Van de Puttee's Office: Amber Hausenfluck. Westat: Keith Rust.

Trial Urban District Assessment Policy: Ray Fields had prepared a document with suggested changes to the Trial Urban District Assessment (TUDA) policy for clarification of the eligibility criteria and procedures for applying for participation in the program. The recommendations were generally to add more detail to the procedures for districts to apply for participation to represent more accurately the actual procedures being followed. The modification to eligibility requirements adds two districts to the eligible list. Mr. Fields noted that these changes would not impact the assessments at all. And, he noted that the impact of adding districts, should the eligibility for participation be modified, would not impact actual participation until the 2015 NAEP assessment cycle.

John Eason thanked Mr. Fields for bringing this issue to the attention of the Governing Board. He asked about the provision for perpetual inclusion of districts, once they are eligible and opt to

participate in TUDA. What happens if the district enrollments drop to levels that do not support assessments in three subjects? Why is that an important criterion?
Mr. Fields noted that it is more cost efficient to be able to assess students in the three subjects that are administered in a single assessment cycle. Cornelia Orr explained that the booklets are packaged to be distributed to students in a classroom setting such that the three assessments are "spiraled" across students in the class. Having to package assessments differently and administer them differently requires more resources and more costs.

Susan Loomis noted that there was concern for maintaining flexibility when the policies were first developed by COSDAM. The committee wanted to assure that there be flexibility to avoid "expelling" districts due to short-term fluctuations in student enrollments or demographic composition. If the changes are long term, however, then perhaps some other action would be needed.

Mr. Fields noted that if long-term changes were the case, NCES would probably address the issue directly with the district. He also stated that the policy could be modified further to provide some specific rules regarding this potential. Leticia van de Putte stated that the eligibility requirements are quite clear and seem to cover the issue well. She recommended against further specificity regarding potential changes that would lead to ineligibility. Jim Popham agreed with that recommendation.

There was no need for action on the TUDA policy at this meeting. The proposed changes will be brought back to COSDAM and the Governing Board again at the August 2012 meeting for further consideration and action.

Future Topics for COSDAM
Mr. Fabrizio then asked COSDAM for recommendations regarding issues or topics that they would like to have presented for discussion at future meetings.

Mr. Popham again recommended that sensitivity to instruction on the part of the NAEP be an issue for COSDAM discussion.

Ms. van de Putte stated that she hoped the presentation to the Board (later in the day) on demographics and education would be of interest to everyone, and that related issues might be discussed by COSDAM in the future.

CLOSED SESSION 11:25 a.m. – 12:25 p.m.
COSDAM Attendees: Lou Fabrizio (Chair), John Q. Easton (*Ex officio* member of the Governing Board and Director of the Institute of Education Sciences), Terry Holliday, Jim Popham, Leticia Van de Putte, and Fielding Rolston.
Governing Board Staff: Susan Loomis
Other Attendees: AIR: George Bohrnstedt. ETS: Andreas Oranje. HumRRO: Lauress Wise. Measured Progress: Luz Bay. Westat: Keith Rust.

In accordance with the provisions of exemption (9)(B) of Section 552b(c) of Title 5 U.S.C., the Committee on Standards, Design and Methodology met in closed session on May 18, 2012 from

11:25 a.m. to 12:25 p.m. in order to review and discuss reports including secure data and results of research conducted to expand the measurement precision of NAEP and research conducted to set achievement levels cut scores for the National Assessment of Educational Progress in writing.

Studies to Expand NAEP Measurement Precision

Andreas Oranje of ETS provided a presentation on two research programs to expand NAEP measurement precision. First, he reported on the Mathematics Computer Based Study (MCBS) which is a multi-stage computer adaptive assessment of grade 8 mathematics. Next he reported on the Knowledge and Skills Appropriate (KASA) study that was developed to address the precision of measurement—especially at the lower ranges of the scale.

The MCBS used an experimental design. The results of the MCBS showed that there was no bias introduced in average scores and that measurement precision was generally increased from a minimum of 10% to over 30% at the individual level. Measurement error is generally lower with adaptive tests, especially at the higher and lower ends of the distribution tails. The adaptive test results in an overall better measure.

Mr. Oranje noted that the items used in this study were taken directly from current NAEP items and not designed for an adaptive test. The results of the study would likely have been even more positive had the items been developed specifically for the purpose.

Jim Popham asked about the number of items in the two stages and how that was determined. He asked for clarification on how the adaptive blocks were made to be representative of the framework. John Easton noted that this issue related not only to the content areas but to the representation of constructed response and multiple choice items.

Mr. Oranje responded that the content was proportionally represented across the blocks administered to individual students. However, in order to have immediate scoring of responses to the router blocks, it was necessary to have only multiple choice items included in that first stage of the adaptive testing. In the future, it will be necessary to use artificial intelligence scoring engines so that constructed response items can be included in the first stage/router blocks.

Leticia van de Putte noted that Texas had brought in gaming experts to advise the state on development of adaptive tests. The experts provided valuable information about the minimal number of items, time, and so forth that would be needed for reliable measures. Mr. Oranje confirmed that gaming expertise is an important part of the design of adaptive tests.

The multi-stage routing seemed to work well. Future research will need to focus on analysis of the student data to gather more information about engagement and performance in the adaptive setting. And, future research will focus on statistical targets, especially those related to performance at the lower end of the NAEP scale.

This discussion provided a perfect segue to the next research report on the KASA study, which was specifically designed to provide reliable measures of lower performance. For the KASA study, the distribution of item complexity was modified. The operational mathematics NAEP includes 25% high complexity, 50% medium complexity, and 25% low complexity items. For

KASA, there were no high complexity items, 30% medium complexity, and 70% low complexity.

Overall, the study results seemed positive. The KASA items yielded higher average performance and lower non-response rates for students; the data fit the scaling mode; and measurement precision was improved for students in the Puerto Rico sample.

Mr. Orange cautioned, however, that the results using KASA items need to be evaluated over time. Replication of the study in 2013 is recommended so that evaluation of data in comparison to 2011 results can be completed before 2011 results are reported.

The Committee was very impressed with the research and asked Mr. Oranje to provide more information about these research studies at a future meeting.

Achievement Levels for 2011 Writing NAEP at Grades 8 and 12

The Committee has been briefed at each meeting since the writing achievement levels setting contract was awarded to Measured Progress in September 2010. Luz Bay had provided a complete review of the results of studies at the March 2012 meeting, and she provided a brief review of the process and results to the Committee at this May 18, 2012 session. Committee members had an opportunity to ask questions in preparation for their action on the achievement levels for writing.

The closed session adjourned at 12:25 PM.

OPEN SESSION 12:25 – 12:30 PM
ACTION

The COSDAM meeting was opened at 12:25 p.m. at which time Mr. Fabrizio asked for a motion to approve the achievement levels cut scores, descriptions of each level, and exemplar performances for reporting the 2011 writing NAEP for grades 8 and 12.

Rolston Fielding moved, and Leticia van de Putte seconded, the following motion:

> The Committee on Standards, Design, and Methodology approves the achievement levels descriptions, cut scores and exemplar performance at each level for reporting the results of the writing NAEP for grades 8 and 12 starting with the Nation's Report Card for 2011.

The Committee unanimously approved the motion, and will recommend approval to the full Board on Saturday, May 19, 2012.

The May 2012 meeting of COSDAM was adjourned at 12:30 PM.

I certify the accuracy of this report.

Louis M. Fabrizio

6/11/2012

Lou Fabrizio, Chair

Date