National Assessment Governing Board

Committee on Standards, Design and Methodology

May 13, 2016
10:00 am – 12:15 pm

AGENDA

10:00 – 10:05 am	Welcome and Review of Agenda *Andrew Ho, COSDAM Chair*	
10:05 – 10:50 am	Computer Access and Familiarity Study *George Bohrnstedt, American Institutes for Research*	Attachment A
10:50 – 11:40 am	NAEP Validity Framework *Fran Stancavage, American Institutes for Research*	Attachment B
11:40 am – 12:10 pm	Key Findings and Actions from NAEP Linking Studies *Sharyn Rosenberg, Assistant Director for Psychometrics* *William Tirre, NCES*	Attachment C
12:10 – 12:15 pm	Information Items • Update on Evaluation of NAEP Achievement Levels • Student Engagement in NAEP: Critical Review and Synthesis of Research • 2017 Writing Grade 4 Achievement Levels Setting Procurement	Attachment D Attachment E Attachment F

Developing New Indices to Measure Computer/Technology Access and Familiarity

As NAEP moves to becoming a fully digitally-based assessment (DBA), one concern is the degree to which all children are ready for a move to a DBA. In particular, NAEP needs to consider the extent to which all students have the same access to, and familiarity with, the tablet [or digital] technology being used to collect the data, as well as the extent to which access and familiarity with digital technology is correlated with performance on NAEP DBA assessments in reading, mathematics, and science at grades 4, 8 and 12. Assuming that there is a measurable relationship to performance, a second but equally important question is whether access and familiarity differ for disadvantaged students (e.g., Black students, Hispanic students, and students eligible for the National School Lunch Program (NSLP)) compared to non-disadvantaged students.

The current computer access and familiarity study (CAFS) has been or will be investigating this concern by:

1. Developing and administering new student-level items to be used in creating indices of digital technology access and familiarity,
2. Assessing the reliability of these indices, and
3. Analyzing the distribution of these indices across NAEP's subpopulations and the relationship between the indices with achievement on NAEP administered as a paper based assessment (PBA) and as a DBA.

The specific research questions to be addressed by these activities are:

1. Do the access and familiarity items cluster together in ways that suggest that reliable indices of each can be constructed?
2. For those who take the DBA version of the assessment, what is the relationship between access/familiarity and performance on NAEP reading, mathematics, and science? Are these relationships constant across gender and race/ethnic groups?
3. Do the observed relationships between access and familiarity and NAEP performance persist when controlling for SES?
4. What is the differential validity of the two measures in predicting NAEP performance as a function of mode of administration?
5. Are access and familiarity differentially distributed across gender, race/ethnicity and/or SES? If so, and if there is a relationship between access and familiarity and NAEP performance for those taking the DBA version, does this raise equity issues about the use of a technology-based NAEP assessment?

The CAFS surveys were administered to samples of 4th (N=5247), 8th (N=6233) and 12th (N=5628) graders who took the 2015 NAEP reading, mathematics or science assessment. The samples were also split between those in the PBA and DBA conditions. (The CAFS survey will also be re-administered as part of the 2017 assessment.) Data cleaning has been completed and the structure of covariation among the items is currently being examined at grades 4 and 8 for the DBA and PBA samples in reading and mathematics. (The NAEP science performance data are not yet available to merge with the CAFS data and analysis of the grade 12 data have been delayed slightly because of some analysis difficulties).

Initial analyses suggest that access is best represented as having two subdomains—access at home and access at school. Familiarity appears to have four subdomains (three at grade 4) – familiarity based on instruction at school, familiarity with tablets, familiarity with laptops or desktops, and familiarity with digital technology concepts (grades 8 and 12 only).

The goal of follow-up analyses is to construct a common set of indices across grades 4, 8 12 that will measure the various subdomains of access and familiarity. The indices will then be used to examine the research questions noted above. All analyses will be done separately for grades 4, 8 and 12 and by gender, NSLP status and race/ethnicity.

The NVS NAEP Validity Framework

Since its inception in 1995, the NAEP Validity Study (NVS) Panel has been engaged in research on various aspects of the validity of the NAEP assessment program. The choice of topics was informed by the judgments of both panel members and the National Center for Education Statistics (NCES) regarding the most pressing validity research needs at any given point in time. In October 2002, NCES asked the panel to put together a framework for their work and also asked the panel to be more forward looking in generating possible research topics to be studied. As a result of this request, in 2002, the panel developed a research agenda that was based on a framework defined by categories:

1. The constructs measured within each of NAEP's subject domains
2. The manner in which these constructs are measured
3. The representation of the population to be assessed
4. The analysis of data
5. The reporting and use of NAEP results
6. The assessment of trends

This framework, which was published as an NVS report, continued to be used as an organizing tool for the panel for several subsequent annual updates to the validity research agenda until the recent past.

However, by the start of the current five-year contract (2013-2018), it was time to update the NVS framework in light of more recent developments. The most notable of these was criticism from a Congressionally-mandated evaluation of the NAEP program that was completed in 2009 by scholars from the Buros Center for Testing at the University of Nebraska–Lincoln and the Center for Educational Assessment at the University of Massachusetts–Amherst. The evaluators argued that the then-current approach to NAEP validity research seemed to imply that the validity of NAEP was in the instrument rather than in the uses to which NAEP has been put. Instead, validity must be established for each purpose or use. More specifically the evaluation said: "Validation is an ongoing process because it is the interpretation or use of assessment results that are supported (validated), not the assessment instrument itself." (Buckendahl, Davis, Plake, Sireci, Hambleton, Zenisky and Wells, 2009, p.xvii). They also noted that, in their view, much of the validity research that NCES had done to this point in time was piecemeal and without the benefit of a comprehensive framework. The specific language the evaluators used is: "NAEP has not had the benefit of a comprehensive framework to guide the *systematic* accumulation of evidence in order to substantiate the ways in which its assessment results may be reasonably interpreted and applied." (Buckendahl et al., p .xi). Finally, they argued that "there is a need for an ongoing, systematic appraisal of the

validity of the interpretations and uses being built on the NAEP assessments." (Buckendahl et al., p.14).

In response to the criticism of Buckendahl et al. (2009), NCES requested AIR's NAEP Education Statistical Services Institute (NESSI) to construct a comprehensive NAEP validity framework based on the uses to which NAEP is put. In order to keep the task a manageable one, the NESSI team decided to focus only on uses designated by the federal government. That is, the framework does not include the various non-official uses to which researchers might employ NAEP.

The NESSI staff identified five such official uses:

1. Monitoring student performance at a given point in time in mathematics, reading and other subjects at grades 4 and 8 (and at grade 12) at the national, state and selected district levels using both scale scores and achievement levels.
2. Monitoring trends in mathematics, reading and other subjects (and at grade 12) at the national, state and district levels and reported both by scale scores and achievement levels.
3. Comparing the performance of achievement across states and districts as well as internationally.
4. Disaggregating and reporting results by race, ethnicity, socioeconomic status, gender, disability and limited English proficiency.
5. Using NAEP results to inform and evaluate federal educational policies.

The team then asked what validity questions would have to be answered to be able to assess the validity of a particular use. The crossing of the various uses of NAEP by its related validity questions resulted in the validity framework.

By agreement with NCES, NVS used the NESSI framework as a starting point for the new framework, which was primarily intended to provide structure for an NVS review of prior research on NAEP validity and to guide the choice of topics for future NVS validity studies. NVS retained the fundamental organization of the framework, and made relatively minor refinements based on the research questions that we saw emerging in our review of extant studies. The full NVS framework is attached.

At present, the 2013 NVS validity framework has not been widely disseminated, although it has appeared in briefing materials for several NVS panel meetings.

Reference:

Buckendahl, C. W., Davis, S. L., Plake, B. S., Sireci, S. G., Hambleton, R. K., Zenisky, A. L., & Wells, C. S. (2009). Evaluation of the National Assessment of Educational Progress. US Department of Education.

NVS Validity Framework

The Intended Use of NAEP Data	Questions NAEP should consider given its intended use
Use I. **A status measure of what students know and can do** *From Legislation* A. Providing a measure of student achievement in mathematics, reading, and other subjects at grades 4, 8, and 12 at the national level. B. Providing a measure of student achievement in mathematics and reading at grades 4 and 8 at the state level. C. Providing achievement levels that are consistent with relevant widely accepted professional assessment standards and based on the appropriate level of subject matter knowledge for grade levels to be assessed (or the age of the students). *From the Governing Board* D. Providing a measure of student achievement in mathematics and reading at grades 4 and 8 for participating urban districts. E. Providing evaluative statements regarding levels of student achievement.	1. To what extent are the NAEP frameworks valid for conceptualizing what is meant by mathematics, reading, or other subject areas given the variation in how they are taught in the United States? 2. To what extent do the item pools adhere to accepted standards of high quality? a. Do the items (collectively) cover the framework objectives for each of the content areas? b. Is the size and composition of the item pool sufficient to both adequately cover the framework and measure the high- and low-performing populations? c. Are the item types used (e.g. multiple-choice, extended response, Hands-on, etc.) sufficient to measure the contents being assessed? 3. To what extent do the individual items adhere to accepted standards of high quality? a. Does each item fit within the framework? b. Are the items free of bias, free of construct irrelevant characteristics, and accessible to all students? c. In the case of translated items, are they valid for inferences for the population being assessed? 4. To what extent are results confounded by student factors that introduce construct irrelevant variance? These include: a. Motivation/engagement b. Other student factors such as test taking strategies 5. To what extent are the psychometric and statistical methods used valid for drawing inferences about student performance? Including: a. Psychometric models – including correct functional form / model specification b. Estimation of error (measurement, sampling and equating) in the scaling of items and the estimation of population parameters c. Conditioning analyses (i.e., the conditioning model used to create the plausible values) d. Equating scales across administrations e. Scoring processes, including the use of machine scoring for extended responses

The Intended Use of NAEP Data	Questions NAEP should consider given its intended use
Use I. **A status measure of what students know and can do (continued)**	f. Imputation procedures (e.g., missing data analysis, treatment of items not reached, conditioning analyses) g. Differential item functioning analyses h. Cross-grade scaling (where used) 6. To what extent do the sampling and weighting procedures allow for drawing valid inferences about student performance? Including: a. Sample design (items and students) b. Sample sizes c. Models used to distribute items (e.g., blocks construction) d. Response rates (state, school and individual) e. Weighting procedures 7. To what extent do the administration procedures used allow for drawing statistically valid inferences about student performance? Do the procedures change the construct being measured? Including: a. Mode of administration (e.g., paper and pencil, computer, tablet) b. Standardization of administration conditions c. Accommodations and exclusions procedures and the standard application of those procedures 8. To what extent are the achievement levels statistically and psychometrically defensible, and meaningful? Including: a. Standard setting methods and processes b. Consequential data (e.g., from external empirical studies) resulting from the cut scores selected 9. To what extent does the reporting of results (e.g., Nation's Report Card) accurately reflect the statistical findings of the assessment? Including: a. Are NAEP reports understandable for the general public and education policymakers? b. Does statistical significance get confounded with substantive significance? c. Is there a shared understanding among target audiences about what the achievement levels mean? 10. To what extent do the data provided to users for secondary data analysis allow for analyses that will yield valid parameter estimates?

The Intended Use of NAEP Data	Questions NAEP should consider given its intended use
Use II. **Comparisons over time (Trends)** *From Legislation* A. Providing a measure of trends in mathematics, reading, and other subjects at grades 4, 8 and 12 at the national level. B. Providing a measure of trends in mathematics and reading at grades 4 and 8 at the state level. C. Providing a measure of academic achievement at ages 9, 13, and 17 for the purpose of maintaining long-term trends in reading and mathematics. *From the Governing Board* D. Providing a measure of trends in mathematics and reading at grades 4 and 8 for participating urban districts.	1. To what extent are comparisons over time valid given changes (or stability) in NAEP frameworks? Including: a. Periodic revisions to the NAEP frameworks, individually or cumulatively, that occur in response to changes in state and district educational practices b. Other changes to the content or structure of the NAEP frameworks (e.g., if reading and writing were combined into one assessment) 2. To what extent is the validity of long-term trend (LTT) affected for the current population of students given changes that have occurred in curricula in the U.S.? 3. To what extent is the validity of NAEP affected by confounding factors that affect the measurement of the constructs over time (e.g., demographic changes)? 4. To what extent is there a valid interpretation for what a unit change in the scale score means? 5. To what extent are comparisons over time affected by changes in SD/ELL populations, exclusion rates, and exclusion policy? 6. To what extent is the NAEP trend data valid given changes over time in the administration or measurement process (e.g., change in mode of administration, use of computer adaptive testing, a new IRT model, change in the length of blocks)? 7. To what extent is the validity of the NAEP scale affected over time by the required release of NAEP items after each administration?
Use III. **Comparisons of entities (States, Districts, Nations)** *Federal Government* A. Providing a measure of student achievement for comparing student achievement across states. B. Providing a measure of student achievement for comparing student achievement across urban districts. C. Providing a measure to compare student performance at national and state level to international students (e.g., international benchmarking using NAEP-TIMSS link).	1. To what extent are comparisons across states and nations valid given the degree of alignment between the NAEP frameworks and states' content standards or international assessment frameworks? 2. To what extent is the validity of comparisons across entities affected by differences in participation and exclusion rates (including differences due to different inclusion and accommodation polices)? 3. To what extent are the validities of cross- and within-district comparisons affected by differing or changing definitions of urban districts in the TUDA (e.g., inclusion or exclusion of charter schools)? 4. To what extent is the validity of comparisons with other nations, affected by different languages, engagement factors, and the compositions of the target populations (e.g., differences in populations attending school)? 5. To what extent are samples large enough to detect meaningful differences between

The Intended Use of NAEP Data	Questions NAEP should consider given its intended use
	entities within a year and across years within entities? 6. To what extent does the reporting of results across entities accurately reflect and convey the findings of the assessments (e.g., accurately reporting statistical significance)? 7. To what extent are comparisons across states and nations affected by the linking methods used?

The Intended Use of NAEP Data	Questions NAEP should consider given its intended use
Use IV. **Disaggregating groups** *From Legislation* A. Providing information on special groups at the national level, including, whenever feasible, information collected, cross tabulated, compared, and reported by race, ethnicity, socioeconomic status, gender, disability and limited English proficiency; *From the Governing Board* B. Monitoring trends and achievement gaps at the state level disaggregated by race, ethnicity, socioeconomic status, gender, disability, and limited English proficiency. C. Monitoring trends and achievement gaps at the urban school district level disaggregated by race, ethnicity, socioeconomic status, gender, disability, and limited English proficiency.	1. To what extent is the validity of analyses of disaggregated groups (including gap analyses) affected by differences in construct equivalence across the groups (e.g., difference in science achievement due to different English language ability, changes in construct being measured due to a provided accommodation)? 2. To what extent is the validity of results affected by the reliability of the reporting variables (.e.g., socioeconomic status, gender, disability, and limited English proficiency)? 3. To what extent is the validity of analyses of disaggregated groups affected by changes in the definitions of reporting variables over time (e.g., changes in the definitions of race categories)? 4. To what extent is the validity of analyses of disaggregated groups (including gap analyses) affected by differences in measurement precision across the groups (e.g., validity of reporting achievement in Puerto Rico due to imprecise measurement at the low end of the achievement scale)? 5. To what extent is the validity of analyses of disaggregated groups (including gap analyses) affected by differences in participation and exclusion rates across the groups? 6. To what extent is the validity of analyses of disaggregated groups (including gap analyses) affected by differential effects of mode of administration across subgroups?

The Intended Use of NAEP Data	Questions NAEP should consider given its intended use
Use V. **Informing policy and evaluating programs** *From NAGB* A. Measuring 12th grade preparedness for college and workplace training *From NCES* B. Providing a secondary source of information that can be used as one criterion for confirming increases in student achievement in grades 4 and 8 reading and mathematics (relative to the goal of all students reaching proficiency) on state assessments under NCLB. C. Mapping state standards onto NAEP. D. Monitoring state progress on their state assessments *Other* E. Identifying states with increased student achievement and decreased achievement gaps due to specific educational policies or reforms in those states. F. Identifying changes in uses of technology in the classroom over time G. Identifying how changes in the economy affects student performance	1. To what extent is the validity of comparisons of state assessments and NAEP results affected by differences in the content coverage of the state tests and NAEP? 2. To what extent is the validity of comparing NAEP to state assessments affected by differences in the tests (e.g., item formats, mode of administration, test difficulty, test reliability, definitions of subgroups)? 3. To what extent is it valid to use NAEP as a common metric for cross-state and within-state-over-time comparisons of proficiency standards? 4. To what extent are the contextual items (e.g., parent education, school resources, school climate, teacher qualifications, and teacher practices) accurately measured so that they can validly be used to evaluate potential factors that impact achievement in order to inform policy? 5. To what extent are comparisons of state assessments and NAEP results valid given differences in student engagement when taking the two assessments? 6. To what extent are comparisons of state assessments and NAEP results valid given differences in participation and exclusion rates? 7. To what extent is NAEP valid for evaluating the impact of changes in policy at the national, state, and district (TUDA) levels? 8. To what extent is NAEP valid as a predictor of postsecondary outcomes? Is there variability in which postsecondary outcomes NAEP can predict (e.g., college attendance versus job performance)? What is the concurrent validity of NAEP with other indicators of postsecondary preparedness? 9. To what extent are NAEP achievement levels valid for policy purposes (e.g., are they meaningful and defensible as standards)?

Key Findings and Actions from NAEP Linking Studies

During the November 2015 and March 2016 Governing Board meetings, the Committee on Standards, Design and Methodology (COSDAM) had brief discussions about various studies that were conducted (by both NCES and the Governing Board) to link NAEP to other assessments or data sources. Linking studies involve comparisons between two assessments allowing one to see where a score point on one of the assessments would fall on the scale of the other assessment. One question raised by COSDAM members was about how the findings from these linking studies are actionable. This presentation is intended to provide an overview of the primary ways in which NAEP linking study results have been used.

Sharyn Rosenberg of the Governing Board staff and William Tirre of NCES will discuss the primary ways in which NAEP linking studies have been used, based on findings from studies conducted during the past ten years:

- To estimate state-level performance on international assessments
- To inform the development of a new measure of socio-economic status
- To compare state performance standards on a common scale
- To compare NAEP achievement levels with external benchmarks
- To estimate the percentage of students academically prepared for college

As background to the presentation, an overview of each study to be discussed is provided. The presentation will focus on the key findings and actions rather than the design and methodology of each study. COSDAM has been briefed on most of these studies during previous Board meetings.

2011 NAEP-Trends In Mathematics and Science Study (TIMSS) Linking Study

Purpose: TIMSS is an international comparison study of student achievement in mathematics and science at grades 4 and 8, administered every four years. The purpose of conducting the 2011 NAEP-TIMSS linking study was two-fold. The study was conducted to see whether it is possible to predict TIMSS scores (in mathematics and science) for the states that did not participate in the TIMSS assessment. Secondly, the study was conducted to identify a method among various methodologies suggested in the literature for linking two assessments. The study was done at grade 8 only.

Sample: The study involved four samples of students at grade 8: the 2011 NAEP operational/national sample, the 2011 TIMSS U.S. operational/national sample, students assessed using 2011 NAEP administration procedures who received braided booklets containing one block of NAEP and one block of TIMSS items; and students assessed using 2011 TIMSS administration procedures who received one block of NAEP items and three blocks of TIMSS items. In addition to these linking study samples, nine states—Alabama, California, Colorado, Connecticut, Indiana, Florida, Massachusetts, Minnesota, and North Carolina—participated in 2011 TIMSS as separate jurisdictions to serve as the "validation sample".

Statistical method used to establish the link: Three types of statistical linking were considered in this study: statistical moderation, statistical projection, and IRT calibration.

Main findings: Selected findings are highlighted below (link to NAEP-TIMSS linking study report).

For Mathematics:

- Average scores for public school students in 36 states were higher than the TIMSS average of 500.
- Scores ranged from 466 for Alabama to 561 for Massachusetts.
- Massachusetts scored higher than 42 of the 47 participating education systems.
- Alabama scored higher than 19 education systems.

For Science:

- Average scores for public school students in 47 states were higher than the TIMSS average of 500.
- Scores ranged from 453 for the District of Columbia to 567 for Massachusetts.
- Massachusetts and Vermont scored higher than 43 participating education systems.
- The District of Columbia scored higher than 14 education systems.

The evaluation of results showed that all three methods of linking yielded essentially the same predicted TIMSS results. In addition, among the three methods, the statistical moderation technique is the simplest method requiring the estimation of the fewest parameters and could be applied to the extant national samples of NAEP and TIMSS. (link to NAEP-TIMSS linking study technical report).

Application to NAEP: The predicted TIMSS scores for states were reported and compared to other countries. This study also helps NCES conduct future NAEP-TIMSS linking studies using statistical moderation without the additional resources needed for the braided-booklet samples.

2015 NAEP-TIMSS Linking Study

Purpose: The purpose of conducting the 2015 NAEP-TIMSS linking study is to predict TIMSS scores for the states that did not participate in the 2015 TIMSS assessment.

Sample: The study design involves two samples of students: (a) students assessed in NAEP paper-based mathematics or science during the winter (January–March) 2015 NAEP administration (**NAEP operational/national sample**) and (b) students in the United States assessed in TIMSS (mathematics and science) during the spring (April–June) 2015 TIMSS administration (**TIMSS U.S. operational/national sample**).
Florida is the only state that participated in 2015 operational TIMSS as a separate jurisdiction. Its actual TIMSS results can be used to validate the predicted state TIMSS results.

Statistical method used to establish the link: Statistical moderation will be used in this study.

Main findings: Analysis will start in early 2017, following the release of the 2015 TIMSS results at the end of 2016. A decision is pending on whether to conduct the NAEP-TIMSS linking study at both grades 4 and 8 or one grade only.

Application to NAEP: As an outcome of the study, the predicted TIMSS scores for states will be evaluated for possible reporting including comparisons to countries participating in TIMSS.

2011 NAEP- Progress in International Reading Literacy Study (PIRLS) Linking Study

Purpose: PIRLS is an international comparison study of reading literacy at grade 4, administered every five years. The purpose of this study was to obtain a statistical comparison between NAEP and PIRLS. The results of the 2011 NAEP grade 4 reading assessment were expressed in terms of the metric of the 2011 PIRLS assessment thereby providing international benchmarks for the NAEP grade 4 reading achievement levels.

Sample: Separate operational national samples of 2011 NAEP and 2011 PIRLS (the design did not include administering both assessments to a common sample of students). Florida did participate in 2011 PIRLS at the state level and was used to validate the linking results.

Statistical method to establish the link: Statistical moderation was used.

Main findings: At each level, the linking shows that the NAEP grade 4 reading achievement levels are higher than the PIRLS international benchmarks. The study report can be found at: http://files.eric.ed.gov/fulltext/ED545246.pdf

When the actual PIRLS results for Florida were compared to the projected PIRLS results, the mean difference was not statistically significant. The only significant difference between the two sets of results for Florida was for the percentage of Advanced students (which varied by only one percentage point).

Application to NAEP: The fact that NAEP reading achievement levels are higher than similar PIRLS international benchmarks may help explain why NAEP has historically reported lower rates of reading proficiency for the United States, whereas PIRLS has historically reported higher levels of reading proficiency. For example, in 2011, NAEP reported that 34 percent of fourth graders were reading at the proficient level, while PIRLS reported that 56 percent were reading at the high international benchmark.

2007 NAEP-Early Childhood Longitudinal Study–Kindergarten Cohort of 1998-1999

Purpose: ECLS-K is a longitudinal study conducted by NCES to follow a cohort of students who entered kindergarten during the 1998-1999 school year through their eighth grade year in 2006-2007. The study includes data collected from students, parents, teachers, and schools. The linking study served at least two purposes. One research study investigated the relationship between ECLS-K reading proficiency levels and 8th-grade NAEP achievement levels and explored the relationship between reading performance at earlier grades and performance on the 8th-grade NAEP reading assessment. Another research study investigated the concordance of student-reported parental education on the NAEP student background questionnaire with parent reports on the same variable from the ECLS-K questionnaire.

Sample: Data came from a common sample of public school students (n=1,290) who took both NAEP and ECLS-K grade 8 reading assessments in spring of 2007.

Statistical method to establish the link: Projection by regression was used in this study.

Main findings: The correlation between NAEP Reading and ECLS Reading at grade 8 was estimated at r = .83.

Reading Analysis: The link allowed a comparison between NAEP grade 8 achievement levels in reading and the finer grain and developmentally descriptive ECLS reading proficiency levels. Reading skills students need to master in earlier grades to later reach NAEP's *Proficient* level at grade 8 were identified.

> Dogan, E., Ogut, B., & Kim, Y. (2015). Early childhood reading skills and proficiency in NAEP eighth-grade reading assessment. *Applied Measurement in Education, 28(3),* 187-201.

Parental Education Analysis: With few exceptions, the higher the parent's education, the more accurate the student estimates are of what their parent's education is as reported by one of the parents. Consistent with this result, the higher the parent's education, the lower the percentage of students who report "I don't know". The high polychoric correlations computed with the "don't knows" eliminated and the relatively small bias in analyses using student-reported parental education instead of parent-reported suggest that in spite of the inaccuracies in student reports of parental education, valuable information is nonetheless contained in students' reports of parental education.

> Ogut, B. and Bohrnstedt, G. W. (2012). Reliability of student-reported parental education at NAEP grade 8 mathematics assessment. Paper presented at the annual meeting of the American Educational Research Association, Vancouver.

Application to NAEP: Information from this study on SES is being considered among other pieces of information in the formulation of a new SES measure.

2015 NAEP-ECLS Kindergarten Cohort of 2010-2011

Purpose: ECLS-K is a longitudinal study conducted by NCES to follow a cohort of students who entered kindergarten during the 2010-2011 school year through their fifth grade year in 2015-2016. The study includes data collection from students, parents, teachers, schools, and care providers. The parent interviews include information about income and parental education. The aim of the NAEP/ECLS-K special study is to evaluate the accuracy of grade 4 student reported parental occupation and education (the piloted NAEP SES-related questions), using the ECLS-K parent reported occupation and education as a reference. The results will be useful to inform development and interpretation of SES measures.

Sample: About 1,500 grade 4 students were assessed for both NAEP and ECLS-K in 2015 and were given an extended NAEP student questionnaire. The extended student questionnaire included a set of SES questions on parental occupation and education which are also being administered as part of the 2016 NAEP pilots, and were tested in cognitive interviews prior to administration in the special study.

Statistical method used to establish the link: Data from the ECLS-K and NAEP datasets will be merged by matching students based on common identification. Where available, one or both parents were interviewed as part of the 2015 ECLS-K grade 4 data collection, including SES-related questions of occupation and education. For households with two parents, the mother and father were interviewed separately.

Main findings: N/a. Analyses are currently underway.

Application to NAEP: The goal of this study is to define an SES measure for use in reporting 2017 results. Results of this analysis will inform the selection of SES items for operational administration in 2017.

2009 Preparedness Research: Statistical Linking of NAEP and the SAT

Purpose: This study was conducted as part of the Governing Board's research program on using NAEP as an indicator of academic preparedness for college. The purpose of this study was to identify a reference point or range on the NAEP 12th grade reading and mathematics scales that might be associated with the College Board's SAT preparedness benchmarks. The NAEP and SAT scores for 12th grade students who had taken NAEP in 2009 and had also taken the SAT were the basis for this linking (via an agreement with the College Board).

Sample: The overall NAEP sample size for 2009 12th grade was 49,000 (reading) and 46,000 (math). Students who also took the SAT were matched to NAEP resulting in 16,200 students (reading) and 15,300 students (mathematics), or approximately 33% of students. Note this was conducted for public-school students only. This match rate compares favorably to the national SAT participation rate of approximately 36% of public school students.

Statistical method used to establish the link: Two types of statistical linking were considered in this study: concordance and projection. Projection was preferred primarily due to the moderate correlation of 0.74 for NAEP reading and SAT-reading. (The correlation for math was 0.91.)

Main findings: Based on the College Board's designation of 500 as the preparedness benchmark for each subject at the time the study was conducted, using statistical projection defined the preparedness cut-point for NAEP at 302 (reading) and 164 (math). Note that 302 is the reading proficient cut score and 176 is the math proficient cut score. A report of the results is available on the Governing Board website at (link to NAEP/SAT Report).

Application to NAEP: Findings from this study and others were used to report estimates of the percentage of students academically prepared for college in the 2013 and 2015 NAEP grade 12 report cards. A similar methodology will be applied in a planned linking study of 2013 12th grade NAEP and ACT data at the national level (via a data sharing agreement with ACT) and for a few states (via data sharing agreements with states). In addition, 2013 12th grade NAEP and SAT scores will be linked for students in one state via a data sharing agreement with Massachusetts.

2013 State Mapping Study

Purpose: Since 2003, NCES has conducted studies, which compare each state's academic performance levels in reading and mathematics in grades 4 and 8 by placing the state standards onto the NAEP scale, which is a common metric for all states. These studies, also known as "state mapping" studies, allow states to examine (a) how stringent their state's academic proficiency criteria compare to other states, and (b) whether the rigor of its own standards has changed over time.

Data sources: The study involved two sets of data:

a. The NAEP data from the 50 states and the District of Columbia that participated in the 2011 and 2013 reading and mathematics assessments.
b. State assessment school-level achievement data from the 2010-2011 and 2012-2013 school years provided by each state. The state alternate and modified assessments were excluded from the state mapping studies.

Statistical method to establish the link: By comparing the percentages of students in each NAEP school who achieve each of a state's performance standards with the distribution of NAEP performance by the random sample of students participating in NAEP in the school, we can approximately estimate the position of each of the state standards on a common scale. The method employed to map the state standards and the NAEP scores is known as equipercentile equating. Detailed information on the estimation methods is available at http://nces.ed.gov/nationsreportcard/pdf/studies/2010456.pdf .

Main findings: Results discussed here are from the most current state mapping study available to the public, which was conducted using NAEP and public school data from 2011 and 2013.

1. State proficiency standards for grade 4 reading and mathematics classified into NAEP achievement levels: 2013
    - In reading: The range of the states' NAEP equivalent scores for the "proficient" level, as defined by each state, was 76 points on the 0-500 NAEP scale (twice the size of the standard deviation of the of the NAEP grade 4 reading assessment)
    - In mathematics: The range of the states' NAEP equivalent scores for the "proficient" level, as defined by each state, was 49 points on NAEP 0-500 scales (1.5 times the size of the standard deviation of the NAEP grade 4 mathematics assessment)

2. State proficiency standards for grade 8 reading and mathematics classified into NAEP achievement levels: 2013
    - In reading: The range of the states' NAEP equivalent scores for the "proficient" level, as defined by each state, was 83 points on NAEP 0-500 scales (twice the size of the standard deviation of the grade 8 reading assessment)
    - In mathematics: The range of the states' NAEP equivalent scores for the "proficient" level, as defined by each state, was 60 points on NAEP 0-500 scales (1.5 times the size of the standard deviation of the NAEP grade 8 mathematics assessment)

Application to NAEP: Findings from this study can help states to examine the rigor of their academic standards compared to other states as well as against the NAEP standards.

2009 Preparedness Research: Longitudinal Analyses of Performance on NAEP Related to Performance in College and Other Outcomes of Florida Students:

Purpose: The purpose of this study was to relate 2009 grade 12 NAEP scores to ACT and SAT scores, college performance and other outcomes. Working with Florida state officials and their K-20 Education Data Warehouse (a longitudinal database) scores for students who had participated in the 2009 NAEP 12th-grade assessments and were subsequently enrolled in Florida's public colleges in 2010 were linked to a variety of outcome indicators.

Sample: The overall NAEP sample size for 2009 Florida 12th grade was 3,400 (reading) and 3,200 (math). Sample size for students attending Florida public colleges in 2010 was 1,800 (math) and 1,900 (reading), or about 55% of the NAEP-sampled students. Approximately one-third of these students attended 4-year colleges and about two-thirds attended community colleges.

Statistical method: Average 2009 grade 12 NAEP scores (and interquartile ranges) were reported for seven variables related to postsecondary performance: SAT preparedness benchmarks; ACT preparedness benchmarks; Accuplacer performance; students' self-reported program of study in high school; college enrollment; first year college coursetaking; and first year grade point average.

Main findings: Based on the College Board's designation of 500 as the preparedness benchmark for each subject, 53% of Florida's 12th graders were deemed college ready for mathematics and 54% were for critical reading. Based on the ACT benchmarks of 22 for mathematics and 21 for reading, 34% of Florida's 12th graders were college-ready for mathematics and 46% were college-ready in critical reading. Finally, first year of college results showed a greater percentage of students achieving GPA of B- or better during their first year of college scored at or above the potential NAEP preparedness reference points from the NAEP-SAT linking study compared to students whose GPA was less than a B- during their first year of college. The limitations of the Florida data, namely the availability of data only for students enrolled in Florida public postsecondary institutions, must be taken into consideration when interpreting these results. The report can be found on the Governing Board website: (link to Florida report).

Application to NAEP: Findings from this study and others were used to report estimates of the percentage of students academically prepared for college in the 2013 and 2015 NAEP grade 12 report cards. Longitudinal research is ongoing and also includes a few additional state partners for 2013 NAEP.

2013 NAEP-High School Longitudinal Study (HSLS)

Purpose: HSLS is a longitudinal study conducted by NCES to follow a cohort of students who were in ninth grade during the 2009-2010 school year throughout their secondary years and into their postsecondary years. Data for students who had participated in both the 2013 NAEP 12th grade assessments and the HSLS were linked so that information from the HSLS student and parent questionnaires could provide a broader context for understanding NAEP results. In addition, the study explored using the relationship between the HSLS questionnaire variables and NAEP scores to predict NAEP mathematics scale scores for the full HSLS sample. The results from this research study are under review by NCES.

Sample: Students in the HSLS study who were also tested in NAEP in the 12th grade. N = 3,471 NAEP 2013 Math; 717 NAEP 2013 Reading.

Statistical method to establish the link: Imputation by multiple regression.

Main findings: The results from regression analyses and validation tests show that it is feasible to impute NAEP scale scores with acceptable accuracy for the full ~20,000 HSLS sample using data from the NAEP-HSLS overlap sample (N=3,471). Specifically, models that use HSLS algebra performance in grades 11 and 9 combined with student student-level covariates including race/ethnicity, gender, SD status, ELL status, and parental education proved to work best in recovering actual mean scores of student subgroups from the HSLS-NAEP overlap sample. The pseudo R-squared of the best fitting model with the least bias was 0.744 (R = .863).

Application to NAEP: There are multiple applications. For example, the study that investigated SES in the NAEP overlap sample and follow-on research resulting from this study (as well as additional similar efforts proposed for the NAEP-ECLS-K overlap sample of 2015) could inform the development of a simple and effective SES index based on student level SES items (existing one and/or newly piloted ones). Also possible with the HSLS is the derivation of preparedness benchmarks for college attendance and graduation (eventually).

2013 NAEP-EXPLORE (KY, NC, TN) and Longitudinal Analyses (NC, TN) – Grade 8

Purpose: The ACT Explore assessments were designed to assess a specific student's academic progress at the 8th or 9th grade levels, especially with respect to college and career readiness. As part of the Governing Board's research on using NAEP to estimate the percentage of students academically prepared for college, the NAEP-EXPLORE linking studies tried to identify reasonable points on the grade 8 NAEP reading and mathematics scales that indicate being on track for academic preparedness for college by the end of high school. Longitudinal analyses will follow this cohort of students in two states through high school and into the first year of postsecondary pursuits.

Sample:

- 3,700 and 3,800 for reading and math respectively in KY (including TUDA sample), and overall matching rates are 96% for both subjects.
- 4,000 and 3,900 for reading and math respectively in NC (including TUDA sample), and overall matching rates are 96% for both subjects.
- 2,700 each for reading and math in TN, and overall matching rates are 93% and 94% respectively.

Statistical method: Given that the correlation between NAEP and EXPLORE was not strong enough to support concordance, it was decided a statistical projection was a more appropriate choice. The correlations ranged from 0.72 to 0.74 for reading and from 0.81 to 0.82 for mathematics.

Main findings: In general, the relationship between NAEP and EXPLORE is moderate. Based on the Explore benchmarks of 16 for reading and 17 for mathematics, the NAEP *Proficient* achievement levels for reading and mathematics at grade 8 correspond well with the EXPLORE benchmarks and could possibly be used to form reasonable basis for reporting 'on track for preparedness'. The reports can be found on the Governing Board website at (link to Explore reports). Longitudinal analyses are not yet available.

Application to NAEP: Results have not been applied to operational NAEP but could potentially be used to explore the feasibility of reporting estimates of the percentage of students on track to be academically prepared for college by the end of high school. The Governing Board has not decided whether to pursue a program of research to support this goal.

GLOSSARY

Depending on how the link is established (common items, common test takers, or randomly equivalent groups), how closely comparable the contents of the two tests are, and other considerations (e.g., the reliabilities of the compared tests or the correlation between them), one can use one of four linking procedures: equating, calibration, projection and moderation[1].

In ***equating***, both tests, *X* and *Y*, have been designed and developed to be equally reliable and each measures the same content. Equating is most often used when the goal is to relate two alternate forms of the same test, such as alternate forms of the ACT or the SAT. In equating the distributions of test *X* and *Y* are aligned or matched up directly. The matching can be done with equipercentile equating or linear equating, and the distributions can be either observed score distributions or estimates of unobserved true score distributions. Sometimes IRT scaling is applied and the resulting relationship is invariant across different populations.

In ***calibration*** (e.g., with item-response theory), two tests are assumed to measure the same content, but they are not equally reliable. For example, one test *X* might be a long test whereas the other test *Y* is short. The two versions of the test are not equated, but they are indirectly comparable because they have been calibrated to a common scale θ. This type of linking is done across years in NAEP, TIMSS, PISA, PIRLS, most state criterion-referenced tests, as well as most nationally standardized norm-referenced tests. Calibration procedures provide unbiased estimates for individual students and means (average scores), but additional statistical machinery is needed to accurately estimate group characteristics such as the variance or the percent at and above achievement levels. In the 2011 NAEP/TIMSS linking study, calibration was accomplished by scaling in the same analysis the NAEP and TIMSS items that were administered within braided (one block NAEP paired with one block TIMSS) test booklets.

In ***projection***, a regression equation uses the correlation between the two tests to predict the scores on one test *Y* from those of another test *X*. There is no assumption that the two tests measure the same content or that they are equally reliable. However, there is an assumption that the tests are highly correlated. With projection, there is no longer a symmetric relationship between one test and the other. The conversion table for predicting the first test from the second is different from the table predicting the second test from the first. A statistical link was established between the NAEP and ECLS-K grade 8 reading scales using the marginal maximum likelihood (MML) composite regression procedure with the AM software (Cohen, 2005).

In ***statistical moderation***, the scores on the first test *X* are adjusted to have the same distributional characteristics as the scores on the second test *Y*. In this case it is assumed *X* is linked to *Y*. This is typically done by matching the means and standard deviations of *X* and *Y*, or matching their percentile ranks. The usual requirement for statistical moderation is that both *X* and *Y* have been administered to comparable populations of students (e.g., the student populations taking both tests are randomly equivalent). The State Mapping Study estimated the position of each state's standards on a common scale by comparing the percentages of students in each NAEP school who achieved each of a state's performance standards with the distribution of NAEP scores by the random sample of students in the school who took NAEP.

---

[1] Phillips, G. W. (2014). *Linking the 2011 National Assessment of Educational Progress (NAEP) in Reading to the 2011 Progress in International Reading Literacy Study (PIRLS*). Washington, DC: NAEP Validity Studies, American Institutes for Research.

Evaluation of NAEP Achievement Levels

Background

Public Law 107-279 states:

> *The achievement levels shall be used on a trial basis until the Commissioner for Education Statistics determines, as a result of an evaluation under subsection (f), that such levels are reasonable, valid, and informative to the public.*

Even after being in use for about 25 years and undergoing previous evaluations (1993, 1998, 2009), the achievement levels are still considered to be on a trial basis. Jack Buckley initiated a new evaluation during his tenure as NCES Commissioner to determine whether the trial status could be resolved.

About the Evaluation

The National Center for Education Evaluation and Regional Assistance (NCEE), part of the Institute for Education Sciences (IES), is administering the current evaluation of the NAEP achievement levels. On September 29, 2014, NCEE awarded a contract to The National Academy of Sciences to perform this work.

Objectives for the evaluation include the following:

- Determine how "reasonable, valid, reliable and informative to the public" will be operationalized in this study.
- Identify the kinds of objective data and research findings that will be examined.
- Review and analyze extant information related to the study's purpose.
- Gather other objective information from relevant experts and stakeholders, without creating burden for the public through new, large-scale data collection.
- Organize, summarize, and present the findings from the evaluation in a written report, including a summary that is accessible for nontechnical audiences, discussing the strengths/weaknesses and gaps in knowledge in relation to the evaluation criteria.
- Provide, prior to release of the study report, for an independent external review of that report for comprehensiveness, objectivity, and freedom from bias.
- Plan and conduct dissemination events to communicate the conclusions of the final report to different audiences of stakeholders.

Design

This study focuses on the achievement levels used in reporting NAEP results for the reading and mathematics assessments in grades 4, 8, and 12. Specifically, the study is reviewing

developments over the past decade in the ways achievement levels for NAEP are set and used and will evaluate whether the resulting achievement levels are "reasonable, valid, reliable, and informative to the public." The study relies on an independent committee of experts with a broad range of expertise related to assessment, statistics, social science, and education policy. The project receives oversight from the Board on Testing and Assessment (BOTA) and the Committee on National Statistics (CNSTAT) of the National Research Council.

Members of the interdisciplinary review committee were selected in early 2015 (see below):

Name	**Affiliation**
Dr. Christopher F. Edley, Jr. (Chair)	The Opportunity Institute
Dr. Peter Afflerbach	University of Maryland, College Park
Dr. Sybilla Beckmann	University of Georgia
Dr. H. Russell Bernard	University of Florida
Dr. Karla Egan	EdMetric LLC
Dr. David J. Francis	University of Houston
Dr. Margaret E. Goertz	University of Pennsylvania
Dr. Laura Hamilton	The RAND Corporation
Dr. Brian W. Junker	Carnegie Mellon University
Dr. Suzanne Lane	University of Pittsburgh
Ms. Sharon J. Lewis	Retired (formerly with the Council of the Great City Schools)
Dr. Bernard L. Madison	University of Arkansas
Dr. Scott Norton	Council of Chief State School Officers
Dr. Sharon Vaughn	The University of Texas at Austin
Dr. Lauress L. Wise	HumRRO

Additional information about the Committee and project activities is available at: http://www8.nationalacademies.org/cp/projectview.aspx?key=49677. The first Committee meeting took place in Washington, DC on February 19-20, 2015. Governing Board staff attended the open session and made a presentation to the Committee on the history of the NAEP achievement levels setting activities. The second meeting of the Committee took place in Washington, DC on May 27-28, 2015. Governing Board staff attended the open session on the afternoon of May 27th to listen to panel discussions involving representatives of the media, state and local policymakers, advocacy organizations, and the Common Core State Standards assessment consortia, about interpretations and uses of NAEP achievement levels. Several additional meetings were conducted in the latter half of 2015 in closed session. The final report is expected to be released in mid-2016.

Next steps

The final report is expected to be available soon. NCES and Governing Board staff will be briefed on the findings, and we will also arrange a briefing for Board members. The briefing for

Board members will occur either via a webinar or during the August 2016 Board meeting, depending on the timing of when the report will be made available and disseminated to various stakeholder groups.

As stated in the NAEP legislation, the Commissioner of NCES will use the findings from the evaluation to decide whether the achievement levels should continue to be used on a "trial basis" or whether that designation can be removed. In addition, the final report may include conclusions and recommendations that have implications for future Governing Board achievement levels-setting activities. Public Law 107-279 also specifies that the Governing Board must prepare a formal response to the evaluation:

> *Not later than 90 days after an evaluation of the student achievement levels under section 303(e), the Assessment Board shall make a report to the Secretary, the Committee on Education and the Workforce of the House of Representatives, and the Committee on Health, Education, Labor, and Pensions of the Senate describing the steps the Assessment Board is taking to respond to each of the recommendations contained in such evaluation.*

COSDAM will lead the process of responding to the evaluation and considering any potential implications for future achievement levels-setting work, with input from the full Board.

PARTICIPANT ENGAGEMENT IN NAEP:
CRITICAL REVIEW AND SYNTHESIS OF RESEARCH

BACKGROUND

In September 2015, AnLar Incorporated, along with its subcontractors, Abt Associates and Minds Incorporated, were awarded a contract by the Governing Board to conduct a systematic literature review documented via an annotated bibliography and synthesis summary, addressing what the field knows about the extent to which sub-optimal engagement may affect NAEP student performance and NAEP test administration.

The following provides an overview of progress on project milestones since the March 2016 quarterly Governing Board meeting. Updates detailed below include: the completion of operational coding, submission of the List of Relevant Sources and the Systematic Review Table, submission of the draft Annotated Bibliography and Technical review, and commencement of the meta-analysis of eligible studies and the draft Synthesis Report.

PROJECT MILESTONES

OPERATIONAL CODING

Operational coding of studies for Phases 1-3 concluded in March. All sources containing an abstract or full text were processed through *Phase 1: Relevance* (1,026 sources). Sources that remained eligible after Phase 1 were coded through *Phase 2: Methodology*, and sources that maintained the minimal level of rigorous methodology were coded through *Phase 3: Full Coding* (15 sources). All sources were duplicate-coded by two (2) research associates during each phase of review. The Principal Researcher, Dr. Joe Taylor, provided expert guidance and reconciled disagreements between the research associates. The 15 studies that remained eligible through Phase 3 were recorded into the List of Relevant Sources and Systematic Review Table. These studies also comprise the entries of the Annotated Bibliography.

The 15 studies included in the Annotated Bibliography were also processed through *Phase 4: Comprehensive Critical Analysis* review. During Phase 4 review, Dr. Taylor will code for critiques of methodology, findings, limitations, and recommendations. The details of this comprehensive analysis will be summarized in the Technical Review entries of the Annotated Bibliography and will be completed in May 2016.

LIST OF RELEVANT SOURCES

Upon completion of Phase 1-3 operational coding, AnLar sorted all sources into one of four categories:

- Phase 3 Eligible (NAEP-relevant): Contains sources that were coded all the way through Phase 3. It contains eight (8) correlational studies, three (3) intervention studies, and four (4) descriptive studies.
- Ineligible (Non-NAEP-relevant): Contains sources that were identified as relevant during Phase 1 coding, but are ineligible because they are not specific to NAEP (international assessments or other). These sources were coded through Phase 1.
- Ineligible Sources: Includes sources that were ineligible for a variety of reasons based on Phase 1 coding.
- Un-Coded: Contains sources that were identified by initial search strings, but for which researchers were unable to locate an abstract or full text. These sources were not coded during any phase of this project.

Phase 3 Eligible (NAEP-relevant) studies comprise the List of Relevant Sources, which was completed in March 2016. Each study in the list was also included as an entry in the Systematic Review Table and Annotated Bibliography. Additionally, Dr. Taylor will conduct statistical analyses across similar study types (i.e., correlational, intervention, or descriptive) to inform the findings, limitations, and recommendations sections of the Synthesis Report. These analyses will be completed in early May 2016.

SYSTEMATIC REVIEW TABLE

Concurrent to the completion of the List of Relevant Sources, AnLar entered corresponding data for the 15 eligible sources into a Systematic Review Table (SRT). The SRT contains a subset of pertinent codes for each eligible source that highlight the key illustrative data about each article, providing an accessible at-a-glance presentation. Categories in the SRT include: identifying information (e.g., reference, year published, source of study, and funding entity); descriptive characteristics (e.g., year(s) of data collection, sample size, participant grade(s), assessment type, assessment subject area, administration mode, motivation construct, and number of citations); and study characteristics (e.g., study type, nature of relationship between motivation and achievement on NAEP, direction of treatment effect on motivation, magnitude of relationship between motivation and achievement on NAEP, magnitude of treatment effect/effect size, p-value of relationship, statistical significance, met minimum criteria for either Osborn or WWC Frameworks, attrition, baseline equivalence, and alignment with research question(s)). The final SRT is complete and will be included in the final Synthesis Report.

ANNOTATED BIBLIOGRAPHY AND TECHNICAL REVIEW

AnLar drafted annotated bibliography entries for all 15 sources included in the List of Relevant Sources. Phase 4 review of the 15 studies is currently being conducted by Dr. Taylor. Research associates will use the critiques and data provided during the Phase 4 review to write technical review summaries for each study. Each technical review entry will provide data-specific information on primary findings, significance, limitations, and recommendations. The Annotated Bibliography and Technical Review will be finalized by May 2016.

SYNTHESIS REPORT

Throughout the Operational Coding phase, researchers identified a number of articles that provided relevant context or contributed to the public discourse on motivation and NAEP; however, for a variety of reasons, these articles were found to be ineligible for inclusion in the final List of Relevant Sources during Phase 1 or 2 reviews. Some reasons for exclusion include: focus only on international assessments (PISA, TIMSS, PIRLS) or other various assessments without comparative connection to NAEP; sources, such as technical or literature reviews, that did not include empirical research; or studies that used populations outside of the K-12 scope of research. While no longer eligible for inclusion in the Annotated Bibliography and Technical Review, researchers determined that a subset of these articles likely contained background and context to inform the Synthesis Report.

AnLar obtained citation counts for all sources (when citation counts were available) and calculated the mean and median number of citations, and identified the top five percent as the most cited sources. Researchers also reviewed additional Governing Board-sponsored articles that were not included in the potential source lists provided by NCES, literature search strings, or reference harvesting to account for sources influential to the Governing Board's discussions prompting the work of this project. The two processes yielded 42 articles that were neither in the top five percent of most-cited studies, or directly relevant to the two research questions. Researchers then reviewed the abstract or full-text of these 42 sources to determine relevance to motivation and NAEP, in general. Ultimately, AnLar narrowed this list to seven (7) articles: three (3) are Governing Board-sponsored, and four (4) from the initial search strings. While these sources will not be coded, research associates will consider their content while writing the background, context, and recommendations sections of the Synthesis Report.

All study information captured in Phases 3 and 4 will be presented in a comprehensive report to summarize findings and overall conclusions most relevant to NAEP, while noting and explaining points of agreement and disagreement. Dr. Taylor will complete the meta-analysis of the 15 eligible sources by early May and researchers will incorporate this synthesis into the final report. Study information related to rigor (Phase 2) will be discussed in the report for the subset of ineligible studies selected for background context, as well as the 15 eligible studies included in Phase 4. This Synthesis Report will also present recommendations for future research. The report will be presented to COSDAM during the August 5, 2016 meeting.

2017 Writing Grade 4 Achievement Levels Setting Procurement

The 2017 NAEP writing assessment is the first administration of the grade 4 assessment under the current computer-based Writing Framework (https://www.nagb.org/publications/frameworks/writing/2017-writing-framework.html)[1]. Pursuant to the Governing Board's legislative mandate, achievement levels must be set for the grade 4 writing assessment. In accordance with the Board policy on setting performance levels for NAEP, the achievement levels setting process includes achievement levels descriptions (ALDs), cut scores, and exemplar items. In 2012, the Board formally approved the updated achievement levels descriptions for writing at all three grade levels. A procurement is in process for a contractor to design and implement studies to recommend cut scores and exemplar items.

The 2017 grade 4 writing achievement levels setting will include a field trial (to test logistics associated with any software used to conduct the process), a pilot study, and an operational achievement levels setting study. In addition, the design procedures will require the collection of multiple sources of validity evidence. COSDAM will receive briefings and have the opportunity to provide input on the process throughout the life of the project, with Board action on the grade 4 writing achievement levels planned for the May 2018 Governing Board meeting.

On March 31, 2016, a Request for Proposals (RFP) was issued on www.fbo.gov: https://www.fbo.gov/index?s=opportunity&mode=form&id=40ccabce125cfdff76ca698e7b2c1c13&tab=core&_cview=0 . Proposals are due on May 26, 2016, with an anticipated award date of summer 2016. The contract period of performance is anticipated to be 24 months.

[1] In 2011, NAEP writing assessments were administered at grades 8 and 12 under the current Writing Framework, and achievement levels were set for grades 8 and 12. The grade 4 assessment initially was planned for 2013 administration but was postponed to 2017 due to budgetary constraints.