National Assessment Governing Board
Committee on Standards, Design and Methodology

December 6, 2013
10:00 am – 12:45 pm

AGENDA

10:00 – 10:45 am	NAEP Testing and Reporting on Students with Disabilities and English Language Learners *Larry Feinberg, NAGB Staff* *Grady Wilburn, NCES* **[Joint meeting with Reporting and Dissemination]**	See Attachment A under Reporting & Dissemination
10:55 – 11:05 am	Introductions and Welcome to Lucille Davy *Lou Fabrizio, COSDAM Chair*	Attachment B
11:05 – 11:45 am	Discussion on Achievement Level Setting (ALS) on the 2014 NAEP Technology and Engineering Literacy (TEL) Assessment **(Closed Session)** • 2013 TEL Field Trial Scaling Analyses • Implications for Planning the TEL ALS *Sharyn Rosenberg, NAGB Staff* *Andreas Oranje, ETS*	Attachment C
11:45 – 12:20 pm	NAEP 12th Grade Academic Preparedness Research • Reporting Grade 12 Results Using Preparedness Research Findings • National and State Partnerships *Ray Fields, NAGB Staff* *Sharyn Rosenberg, NAGB Staff*	Attachment D
12:20 – 12:35 pm	Board Chairman's Charge to the Committee for 2014 *David Driscoll, NAGB Chair*	
12:35 – 12:45 pm	Other Issues • Instructional Sensitivity and NAEP *W. James Popham, COSDAM Member*	Attachment E
	Information Items: • Update on Evaluation of NAEP Achievement Levels Procurement • Reading for Understanding: A Theory-Based, Developmental Approach • NAEP 12th Grade Academic Preparedness Research: Phase 2 Research Updates o Course Content Analysis Research o Research with Frameworks	Attachment F Attachment G Attachment H

Attachment B

Welcome to Lucille E. Davy,
New Governing Board Member

The Committee on Standards, Design and Methodology welcomes new Governing Board member Lucille Davy. Lucille Davy became a member of the Governing Board on October 1, 2013 in the category of General Public Representative. She is President and CEO of Transformative Education Solutions, LLC.

Abbreviated Professional Biography for Lucille E. Davy

Lucille E. Davy is an education policy consultant through her roles as President and CEO of Transformative Education Solutions, LLC, and senior advisor for the James B. Hunt, Jr. Institute for Educational Leadership and Policy. Ms. Davy started her career as a lawyer and adjunct professor of mathematics, a subject for which she received a bachelor's degree and K-12 teacher certification for New Jersey. She later served as a volunteer and leader for a variety of parent groups and organizations in Westfield Public Schools in New Jersey. Her experience led to service as special counsel for education policy for the New Jersey Governor's office and as an education policy advisor for several entities, including the Committee for Working Families. From 2005-2010, Ms. Davy served as New Jersey's Commissioner of Education, overseeing more than 2,400 schools in 600 districts that served 1.4 million children. Since 2010, she has, via her consulting firm and as a Hunt Institute advisor, focused on high school students' college and career readiness after graduation and effective implementation of the Common Core State Standards in English language arts and mathematics.

Setting Achievement Levels on the NAEP 2014 Technology and Engineering Literacy (TEL) Assessment

Background

At the March 1, 2013 meeting, the Committee began discussion on setting achievement levels for the 2014 NAEP TEL assessment. For the May 17, 2013 meeting, an issues paper was developed to support procurement and project planning for developing recommended achievement levels for TEL. In the Committee's May 2013 discussion, the Committee expressed a need for more information before proceeding with procurement plans, particularly regarding TEL scaling issues that could hinder a strong TEL Achievement Level Setting (ALS) effort. Initial results from the analysis of TEL field trial data were presented during the August 2, 2013 meeting, but extensive scaling analyses had not yet been conducted. **Additional results are now available and will be presented in closed session at the December 2013 meeting. An overview of the presentation can be found on page 5.**

Timeline

The following timeline provides a preliminary list of key dates and activities related to TEL assessment development and achievement level setting.

Date	Activity	Responsibility
2008 - 2010	TEL Framework development	ADC, Board, WestEd (contractor)
2010 - 2012	Assessment development for 2013 pilot test	NCES, NAEP contractors
2010 - 2012	Item review for 2013 pilot test	NCES, NAEP contractors, TEL Standing Committee, ADC
Early 2013	Pilot test – national sample, grade 8	NCES, NAEP contractors
May 2013	TEL ALS issues paper	COSDAM, consultant
Early 2014	ALS procurement and contract award	Board staff, COSDAM
Early 2014	Operational administration – national sample, grade 8	NCES, NAEP contractors
2015	Board action on TEL achievement levels	COSDAM, ALS contractor, Board
2015	Reporting TEL results	Board, NCES, contractors

TEL Assessment Design

The 2014 Technology and Engineering Literacy (TEL) assessment is based on the Board-adopted Framework and Specifications (see www.nagb.org, Publications).

The TEL assessment is composed of three major areas:

- Design and Systems
- Information and Communication Technology
- Technology and Society

Another key dimension of the TEL assessment is the three practices, each of which is applicable to the three major areas noted above:

- Understanding Technological Principles
- Developing Solutions and Achieving Goals
- Communicating and Collaborating

The TEL assessment was developed using an evidence-centered design (ECD) approach. From the beginning, all TEL tasks and items were designed using an evidential chain of reasoning that links what is to be measured, the evidence used to make inferences, and the tasks used to collect the desired evidence. In addition to student responses to complex tasks and discrete items, the computer-based TEL assessment allows NAEP to capture a wide array of data on student performance. For example, NAEP will collect information on how students interact with the TEL simulations and experiments. Such data may include the number of experimental trials run and the number and types of variables controlled. These observable data on "strategies and processes" are intended to be used for reporting purposes but are not expected to contribute to the scoring of student performance.

TEL Reporting

Based on the ECD approach, TEL reporting includes plans to expand beyond the traditional NAEP scores. It is expected that data from complex performance tasks and discrete items will be reported in several ways:

- A composite or univariate scale score on which the achievement levels will be set
- Subscores for the content areas (Design and Systems; Information Communication Technology; Technology and Society)

- Reporting on the practices (Understanding Technological Principles; Developing Solutions and Achieving Goals; Communicating and Collaborating)
- Information on students' processes and strategies, related to the ECD model, captured as observable data from their work on the TEL scenario-based tasks.

Potential Discussion Questions for COSDAM

- Given the field trial results, is there sufficient evidence to warrant achievement level setting on the overall construct?
- What additional information would help to inform the standard setting process?

ies NATIONAL CENTER FOR EDUCATION STATISTICS
Institute of Education Sciences

2013 TEL Field Trial Scaling Analyses
(Closed Session)

At the August Board meeting, an update was provided to COSDAM on the Technology and Engineering Literacy (TEL) field trial analyses with the goal of preparing for standard setting and setting a tentative timeline based on when pertinent (empirical) results would be available to support the standard setting. At that point, analyses had just begun and percent correct and item-block-biserial correlations were shared, indicating a reasonable item pool covering a range of proficiencies. In addition to sharing and discussing results, a discussion ensued about the need for some dimensionality analyses to determine at what level (e.g., overall, by domain) meaningful standards can and ought to be set on this new construct of Technology and Engineering Literacy. Considerable interest was generated for the correlations between subscales, to inform the question of whether it is appropriate to set standards on the overall assessment.

The analyses that can be performed on data from the Technology and Engineering Literacy field trial are more extensive than typical of field trials and more like those on operational assessments. This capability derives from the change in format from paper-based to technology-based assessments—the printing-cost limitation on the block spiral design was removed and a complete spiral design became possible. Of course, important limitations remain. The field trial will provide data for item selection for the operational assessment, so some items will not carry forward and the remaining items will be reconfigured into different blocks that will revise the current position and context effects. Some blocks were found to take too little time. Reconfiguration to create longer assessment units for the operational assessment also has position and context effect implications.

At this point, extensive analyses have been completed with the field trial data and a firmer timeline for the 2014 analysis is available. In this session ETS will:

- Share results, including the correlations between subscales and student performance across scales.
- Provide more detail about what further analyses are planned based on the field trial data and the goals of these analyses.
- Provide a timeline for the 2014 operational analysis and reporting of Technology and Engineering Literacy and an indication of when results and data products will be available.

NAEP 12th Grade Academic Preparedness Research

Phase 1 Research

The first phase of the Governing Board's research on academic preparedness is now complete; results from more than 30 studies are available at: http://www.nagb.org/what-we-do/preparedness-research.html. During the August 2013 meeting, the Board voted on a motion to use the phase 1 research on academic preparedness for college in the reporting of the 2013 grade 12 national results for reading and mathematics. The approved motion and supporting validity argument also appear on the aforementioned website.

During the December 2013 meeting, COSDAM will hear a brief update on plans for reporting the 2013 grade 12 results for reading and mathematics in terms of academic preparedness for college (scheduled for release in April 2014).

Phase 2 Research

The second phase of the Governing Board's research on academic preparedness currently consists of the following studies that are planned or underway:

Study name	Sample	December 2013 Update
Statistical linking of NAEP and ACT	National; FL, IL, MA, MI, TN	See pp. 7-10 for overview and draft research questions
Longitudinal statistical relationships: Grade 12 NAEP	FL, IL, MA, MI, TN	
Statistical linking of NAEP and EXPLORE	KY, NC, TN	
Longitudinal statistical relationships: Grade 8 NAEP	KY, NC, TN	
Content alignment of NAEP and COMPASS		See pp. 11-12 for overview
Content alignment of NAEP and EXPLORE		
College Course Content Analysis		See pp. 27-36 for informational update
Evaluating Reading and Mathematics Frameworks and Item Pools as Measures of Academic Preparedness for College and Job Training (Research with Frameworks)		See pp. 37-38 for informational update

During the December 2013 meeting, COSDAM will receive an update on the status of the national and state partnerships and will discuss draft research questions for the statistical relationship studies.

Brief overviews and status updates on the College Course Content Analysis and Research with Frameworks are provided as information items in Attachment H.

Overarching Research Questions for Statistical Relationship Studies:

1. What scores on the 2013 grade 8 and 12 NAEP Reading and Mathematics assessments predict academic preparedness for college?
2. Is it feasible to use NAEP to make state-level inferences about academic preparedness for college?

National and State Statistical Linking Studies with the ACT

In 2013, the Governing Board is planning to partner with ACT, Inc. to conduct a statistical linking study at the national level between NAEP and the ACT in Reading and Mathematics. Through a procedure that protects student confidentiality, the ACT records of 12th grade NAEP test takers in 2013 will be matched, and through this match, the linking will be performed. A similar study at the national level was performed with the SAT in 2009. There will not be a statistical linking study performed for NAEP and the SAT in 2013.

In addition, the state-level studies, begun in 2009 with Florida, will be expanded in 2013. Again using a procedure that protects student confidentiality, ACT scores of NAEP 12th grade test takers in the state samples in partner states will be linked to NAEP scores. We are in the planning stages with five states to be partners in these studies at grade 12: Florida, Illinois, Massachusetts, Michigan, and Tennessee. In three of these states (IL, MI, TN), the ACT is administered to all students state-wide, regardless of students' intentions for postsecondary activities.

Draft Research Questions for National and State Statistical Linking Studies with the ACT:

1. What are the correlations between the grade 12 NAEP and ACT student score distributions in Reading and Math?
2. What scores on the grade 12 NAEP Reading and Math scales correspond to the ACT college readiness benchmarks? (concordance and/or projection)
3. What are the average grade 12 NAEP Reading and Math scores and interquartile ranges (IQR) for students below, at, and at or above the ACT college readiness benchmarks?
4. Do the results differ by race/ethnicity or gender?

Longitudinal Statistical Relationships: Grade 12 NAEP

In addition to the linking of ACT scores to NAEP 12th grade test scores in partner states, the postsecondary activities of NAEP 12th grade test takers will be followed for up to six years using the state longitudinal databases in Florida, Illinois, Massachusetts, Michigan, and Tennessee. These studies will examine the relationship between 12th grade NAEP scores and scores on placement tests, placement into remedial versus credit-bearing courses, GPA, and persistence. Data sharing agreements are in development for each state partner.

Draft Research Questions for Longitudinal Statistical Relationships, Grade 12 NAEP:

1. What is the relationship between grade 12 NAEP Reading and Math scores and grade 8 state test scores?
2. What are the average grade 12 NAEP Reading and Math scores and interquartile ranges (IQR) for students with placement in remedial and non-remedial courses?
3. What are the average grade 12 NAEP Reading and Math scores (and the IQR) for students with a first-year GPA of B- or above?
4. What are the average grade 12 NAEP Reading and Math scores (and the IQR) for students who remain in college after each year?
5. What are the average grade 12 NAEP Reading and Math scores (and the IQR) for students who graduate from college within 6 years?

State Statistical Linking Studies with EXPLORE

In 2013, linking studies between 8th grade NAEP in Reading and Mathematics and 8th grade EXPLORE, a test developed by ACT, Inc. that is linked to performance on the ACT, are planned with partners in three states: Kentucky, North Carolina, and Tennessee. In all three of these states, EXPLORE is administered to all students state-wide during grade 8.

Draft Research Questions for State Statistical Linking Studies with EXPLORE:

1. What are the correlations between the grade 8 NAEP and EXPLORE scores in Reading and Math?
2. What scores on the grade 8 NAEP Reading and Math scales correspond to the EXPLORE college readiness benchmarks (concordance and/or projection)?
3. What are the average grade 8 NAEP Reading and Math scores and the interquartile ranges (IQR) for students below, at, and at or above the EXPLORE college readiness benchmarks?

Longitudinal Statistical Relationships: Grade 8 NAEP

In 2013, the Governing Board will also expand the state-level studies by partnering with a few states at grade 8. Again using a procedure that protects student confidentiality, secondary and postsecondary data for NAEP 8th grade test takers in the state samples in partner states will be linked to NAEP scores. These studies will examine the relationship between 8th grade NAEP scores and scores on state tests, future ACT scores, placement into remedial versus credit-bearing courses, and first-year college GPA.

Three states will be partners in these studies at grade 8: Kentucky, North Carolina, and Tennessee. Data sharing agreements are in development for each state partner.

Draft Research Questions for Longitudinal Statistical Relationships, Grade 8 NAEP:

1. What is the relationship between NAEP Reading and Math scores at grade 8 and state test scores at grade 4?
2. What are the average NAEP Reading and Math scores and the interquartile ranges (IQR) at grade 8 for students below the ACT benchmarks at grade 11/12? At or above the ACT benchmarks?
3. What are the average NAEP Reading and Math scores and the interquartile ranges (IQR) at grade 8 for students who are placed in remedial and non-remedial courses in college?
4. What are the average NAEP Reading and Math scores (and the IQR) at grade 8 for students who obtain a first-year college GPA of B- or above?
5. What is the relationship between grade 8 NAEP Reading and Math scores and grade 12 NAEP Reading and Math scores? (contingent on feasibility of sampling the same students in TN, NC, and KY)

Content Alignment Study of Grade 12 NAEP Reading and Mathematics and COMPASS

Content alignment studies are a foundation for the trail of evidence needed for establishing the validity of preparedness reporting, and are, therefore, considered a high priority in the Governing Board's Program of Preparedness Research. The alignment studies will inform the interpretations of preparedness research findings from statistical relationship studies and help to shape the statements that can be made about preparedness. Content alignment studies were recommended to evaluate the extent to which NAEP content overlaps with that of the other assessments to be used as indicators of preparedness in the research.

We plan to conduct an alignment study of grade 12 NAEP Reading and Mathematics and ACT COMPASS. At this point in time, details of our agreement with ACT are still being worked out. Detailed plans for conducting this study will be presented at a future meeting.

Content Alignment Study of Grade 8 NAEP Reading and Mathematics and EXPLORE

Content alignment studies are a foundation for the trail of evidence needed for establishing the validity of preparedness reporting, and are, therefore, considered a high priority in the Governing Board's Program of Preparedness Research. The alignment studies will inform the interpretations of preparedness research findings from statistical relationship studies and help to shape the statements that can be made about preparedness. Content alignment studies were recommended to evaluate the extent to which NAEP content overlaps with that of the other assessments to be used as indicators of preparedness in the research.

We plan to conduct an alignment study of grade 8 NAEP Reading and Mathematics and ACT EXPLORE. Results from this content alignment study will be particularly important for interpreting the findings from the NAEP-EXPLORE statistical linking studies. At this point in time, details of our agreement with ACT are still being worked out. Detailed plans for conducting this study will be presented at a future meeting.

OVERVIEW OF REFERENCED ASSESSMENTS

For additional background information, the following list presents a brief description of the assessments referenced in the phase 2 academic preparedness research studies. In each case, only the mathematics and reading portions of the assessments are the targets for analysis, although analyses with the composite scores may be conducted.

- ACT – The ACT assessment is a college admissions test used by colleges and universities to determine the level of knowledge and skills in applicant pools, including Reading, English, and Mathematics tests. ACT has *College Readiness Standards* that connect reading or mathematics knowledge and skills and probabilities of a college course grade of "C" or higher (75%) or "B" or higher (50%) with particular score ranges on the ACT assessment.
- ACT EXPLORE – ACT EXPLORE assesses academic progress of eighth and ninth grade students. It is a component of the ACT College and Career Readiness System and includes assessments of English, Mathematics, Reading, and Science. ACT EXPLORE has *College Readiness Standards* that connect reading or mathematics knowledge and skills and probabilities of a college course grade of "C" or higher (75%) or "B" or higher (50%) by the time students graduate high school with particular score ranges on the EXPLORE assessment.
- COMPASS – ACT Compass is a computer-adaptive college placement test. It is produced by ACT and includes assessments of Reading, Writing Skills, Writing Essay, Math, and English as a Second Language.
- SAT – The SAT reasoning test is a college admissions test produced by the College Board. It is used by colleges and universities to evaluate the knowledge and skills of applicant pools in critical reading, mathematics, and writing. The SAT has calculated preparedness benchmarks are defined as the SAT scores corresponding to a 65% probability of earning a first-year college grade-point average of 2.67 (B-) or better.

THE EVALUATIVE MISUSES OF COMPARATIVELY FOCUSED TESTS[1]

W. James Popham

University of California, Los Angeles

For almost a full century, the mission of U.S. educational measurement has been to elicit test-takers' scores so that those scores can be compared with one another. This is a good and useful thing to do. It is particularly good and useful thing to do in situations where the numbers of applicants exceeds the numbers of openings. To make a flock of important educational decisions, we need to identify those students who are our strongest or weakest performers. I am an enthusiastic supporter of tests that yield comparative score-interpretations.

The legitimacy of such test-based comparisons was firmly established way back in World War One, almost 100 years ago, when a group-administered intelligence test, the *Army Alpha,* was administered to about 1,750,000 U.S. Army recruits in an effort to identify men who would be the most suitable candidates for officer training programs. This use of the *Alpha* to provide comparative score-interpretations was regarded as a smashing success and, although the test was clearly a measure of a test-taker's *aptitude*, the *Alpha's* focus on comparative score-interpretations was soon emulated by the makers of educational *achievement* tests. Indeed, a number of the test-construction and test-refinement tactics used for today's U.S. achievement tests can be traced back to the comparative assessment procedures associated with the *Army Alpha.*

But tests capable of providing comparative score-interpretations are not necessarily tests that should be used to evaluate schools or teachers. Such *evaluative* applications of educational assessment, although similar in some ways to *comparative* applications of educational assessment, are fundamentally different. Increasingly, however, America's educators are being evaluated on the basis of their students' performances on tests that were created to yield comparative score-interpretations rather than to measure instructional quality. This is a terrible mistake.

[1] A written accompaniment to oral remarks, *A Trip to Intolerability,* presented at the first International Instructional Sensitivity Conference hosted by the Achievement and Assessment Institute of the University of Kansas, Lawrence, Kansas, November 13-15, 2013.

This mistake is being made because of a pervasive but erroneous belief by Americans that students' test-measured achievement levels, namely, the knowledge and skills students display when responding to achievement tests, can be attributed to what those students have learned in school. In some instances, this is a warranted belief. Certain skills and bodies of knowledge measured by today's achievement tests have definitely been learned by students because of instruction those student received in school.

Yet, what if the tests we traditionally employ to measure students' achievement, because of those tests' preoccupation with providing comparative score-interpretations, *also* measure many things other than what students were taught in school? What if our traditional achievement tests, in an effort to provide the necessary variance in total-test scores that are so vital for comparative score-interpretations, also measure test-takers' status with respect to such variance-inducing factors as students' socioeconomic status and inherited academic aptitudes? Clearly, such a confounding of causality would make such traditional achievement tests less appropriate for evaluating how well students have been taught. To what extent is a student's performance on a traditional achievement test attributable to what was taught in school rather than what was brought to school? Realistically, for many of today's achievement tests, we just can't tell.

I contend that the traditional way we build and burnish our educational achievement tests *may* lead to those tests' being inappropriate for use in the evaluating of schools and teachers. The italicized *may* is intended to emphasize my conviction that, to date, the suitability of today's traditional achievement tests for evaluative use has not been rigorously scrutinized. But it should be.

Clearly, if one wishes to evaluate the performance of a school's instructional staff, or the performance of a particular teacher, then it would be better to have evidence on hand from students' performances on almost *any* sort of achievement test rather than relying on no achievement evidence at all. Thus, I'd certainly rather use students' scores from the tests we now employ for such evaluative purposes than have access to no data whatsoever regarding students' achievement. But the choice before us is not whether we should try to carry out evaluations using flawed tests instead of using no tests at all. Instead, our challenge is to carry out today's increasingly high-stakes evaluations using the most appropriate tests we can employ. I am certain we can do a better job of evaluating our schools and teachers than we do by using today's achievement tests.

The Cornerstone of Our Assessment Castle

If you were to ask today's educators—irrespective of how much they actually knew about educational testing—what is the single, most important concept in educational measurement, the most frequent response to your query would surely be "validity." That response, happily, turns out to be the correct answer. Educational measurement is predicated on the conviction that by getting students to make *overt* responses to stimuli such as a test's items, educators can arrive at valid inferences about students' *covert* knowledge and skills. Determination of the covert based on the overt, indeed, lies at the heart of all educational assessment.

It is not a *test,* however, that is valid or invalid. Instead, it is the score-based *inference*—an inference based on students' test scores—that is valid or invalid. Validity thus represents the accuracy of test-based inferences (or, if you prefer, test-based interpretations). Increasingly these days, assessment validity is regarded not only as the accuracy of a test-based inference, but also as the appropriateness of the use to which a test's score-based inferences are put (Kane, 2013). Optimally, therefore, not only would a test-based inference be accurate, but then that accurate inference would be employed to accomplish a suitable consequence such as subsequently making sound educational decisions about students.

The validity of a score-based inference, therefore, gets our test-usage ball rolling. If we can't establish that test-takers' performances lead to an accurate inference about what test-takers' scores signify, then the likelihood of then making a sensible inference-based decision is definitely diminished. And this is where we currently are with respect to the tests we use to evaluate U.S. schools and teachers. Although educators have been urged (or, in some instances, been statutorily required) to evaluate schools and teachers using students' performances on educational tests, *we have no meaningful evidence at hand indicating that these tests can accurately distinguish between well taught and badly taught students.* This state of affairs is truly astonishing.

Instructional Sensitivity

Yes, our nation increasingly relies on students' scores on tests, typically using standardized achievement tests, to arrive at inferences about the quality of instruction provided to those students. Yet, the evidence to support the accuracy of such score-based inferences about instructional quality is essentially nonexistent. Today's educators are being asked to sidestep the most important tenet of educational measurement, namely, the obligation to supply validity evidence regarding the interpretations and significant uses of an educational test's results. Putting it differently, no evidence currently exists about these evaluative tests' *instructional sensitivity.*

What is this "instructional sensitivity," and how is it determined? Actually, the concept is quite a straightforward one, and it simply refers to how well a test can accurately distinguish between test-takers who have been taught well and test-takers who have been taught badly. Although a certain amount of definitional disagreement about instructional sensitivity can be found in the measurement community, the following definition reflects what most writers on this topic understand when they refer to a test's instructional sensitivity:

> *Instructional sensitivity is the degree to which students' performances on a test accurately reflect the quality of instruction specifically provided to promote students' mastery of what is being assessed* (Popham, 2006).

As you can see, this definition revolves around the "quality of instruction" insofar as it specifically contributes to "students' mastery" of whatever the test is measuring. A test, then, can vary in the degree to which it is instructionally sensitive. We need not, therefore, distinguish between a test that is totally sensitive to instruction or totally insensitive to instruction. Instructional sensitivity is a continuous rather than a dichotomous variable. Our quest, therefore, should be to determine a minimum threshold of instructional-sensitivity acceptability for any test being used to evaluate the caliber of instruction. The more significant the stakes are that are associated with a test's use, the higher should be our acceptability-threshold.

The instructional sensitivity of education tests is not a brand new concept. More than 30 years ago, when the high-stakes accountability movement began to capture the attention of American educators, Haladyna and Roid (1981) described the role of instructional sensitivity when judging the merits of accountability tests.

Much earlier, when the initial proponents of criterion-referenced measurement were attempting to sort out how to create and improve tests leading to criterion-referenced inferences, Cox (1971) and other measurement specialists tried to devise ways to maximize a test item's sensitivity to instruction. But those early deliberations among advocates of criterion-referencing were focused almost exclusively on *measurement* challenges, that is, how to build tests capable of yielding more valid criterion-referenced inferences. As the years tumbled by, however, the *evaluative* use of students' test performances has become more significant. During the next several years, for instance, it is almost certain that many American teachers will lose their jobs primarily because of their students' poor performances on tests. The high-stakes decisions riding on students' test scores have, without argument, become higher and higher and higher.

Nonetheless, despite the increased importance now attached to evaluative test-based consequences, the attention given to the instructional sensitivity of the tests being used to arrive at those consequences still ranges from trifling to nonexistent. Perhaps, one might think, today's inattention to tests' instructional sensitivity simply stems from our not knowing how to go about determining the degree of a test's sensitivity to instructional quality. Yet, we already have on hand a demonstrably successful strategy drawn from our experiences in reducing the *assessment bias* found in our important educational tests. Let's look at the chief elements of that strategy.

Serious Problems Demand Serious Responses

Rarely today is a significant educational test created for which considerable attention has not been devoted to the reduction of assessment bias. That is, we currently regard the diminishment of assessment bias as a canon of good test-building. But it was not always thus.

Go back to the 1960s and 1970s, and you will find that if any attention whatsoever was given to the reduction of tests' assessment bias, it was apt to be perfunctory. Usually, it was completely absent. This skimpy attention to assessment bias was quite understandable. That's because in those days we rarely analyzed test results in such a way as to reveal differences in performances among test-taker groups associated with their gender, race, or ethnicity. However, the rules of the educational testing game changed dramatically in the late seventies when a substantial number of states—dismayed by what they perceived to be the poor quality of their state's public schools—began to link high-school graduation to a students' passing "minimum competency tests" demonstrating that students possessed at least rudimentary skills in reading, mathematics, and sometimes writing.

Because those minimum competency tests were administered to *all* students in a state's public schools, and those students' scores were typically made public, we soon began to see astonishing disparities between the performances of racial groups as well as students drawn from different socioeconomic strata. Indeed, it was the difference in the racial pass rates on Florida's diploma-denial tests that triggered a class-action lawsuit in the precedent-setting *Debra P. v. Turlington* case (Popham and Lindheim, 1981). In that case which, even now, remains the operative case law in such litigation, it was affirmed by a federal appellate court that a violation of the U.S. Constitution occurs when students are denied a property right (such as a high-school diploma) if they are tested using a test whose content had not been taught. In the Florida case, the precipitating circumstance was that far more African-American students were failing the state's basic skills test than were white students. The *Debra P.* litigation, and similar disparities in racial pass rates elsewhere, presented a serious problem to America's educational-measurement specialists. They quickly grasped the significance of the situation—and they set out to fix it.

A Two-Pronged Bias-Reduction Strategy

Having recognized the legitimacy of complaints that the nation's tests were biased against certain subgroups, members of the measurement community soon devised a two-tactic strategy to minimize such bias. The first of these two tactics was a during-development *judgmental review* of each test item in an effort to identify and eliminate any items thought to offend or unfairly penalize test-takers because of test-takers' personal characteristics such as their gender or ethnicity. Second, an *empirical analysis* of students' actual test performances was undertaken, usually during field-testing of new items, so that items potentially contributing to a test's assessment bias could be spotted. The typical analytic approach that evolved, after several years of exploratory analyses, was to employ "differential item functioning" (DIF) techniques in which items were isolated that were being answered differently by different subgroups of test-takers. Items identified by DIF as possibly biased were then modified or jettisoned before being used in an operational test. As a consequence of employing this two-tactic strategy, over many years, we have witnessed a substantial reduction in the number of items on high-stakes tests that are biased against particular groups of test-takers.

The actual procedures for these two approaches to the reduction of assessment bias are now well known among measurement specialists. While their use may not have completely *eliminated* assessment bias from the nation's high-states assessments, the marked impact of these procedures on the reduction of assessment bias is undisputed.

Benign Borrowing

The methodological strategy we could employ in reducing the instructional insensitivity of today's evaluatively oriented achievement tests might be nothing more than a straight-out lift from what has been used in the reduction of assessment bias, that is, to employ a blend of judgmental and empirical procedures.

Although we currently do not have a definite, well-honed set of procedures for dealing with the instructional sensitivity of our tests, the essential elements of an attack on this problem could be directly derivative from previous work in minimizing assessment bias. For example, the charge to be issued when asking a group of seasoned educators to scrutinize a set of test items for instructional *insensitivity* could be quite similar to the language employed when we ask a committee of bias reviewers to look for biased elements in test items. To illustrate, a review committee composed of experienced teachers (who are thoroughly familiar with the content and age-levels of the students to be tested) could be oriented to their item-review responsibilities by learning about the most likely ways an item might be instructionally insensitive. After such an orientation, reviewers could then be given the following charge and asked to render a per-item judgment regarding each item intended for inclusion in a high-stakes evaluative test:

> Attention reviewers: Please note the specific curricular aim which, according to the test's developers, this item is assessing. Only then, answer the following

> question: *If a teacher has provided reasonably effective instruction to promote students' mastery of the specific curricular aim being assessed, it is likely that the bulk of the teacher's students will answer this item correctly?* (Choose one: YES, NO, NOT SURE). (Popham, 2014, 397)

Items for which one or more reviewers have supplied a specified proportion of negative and/or not-sure responses would then be scrutinized to discern if the items embody elements apt to render them instructionally insensitive. Such items would, as is true when acting on the judgments of bias-review committees, be revised or removed.

Similarly, procedural elements for carrying out empirical DIF-like studies for instructional sensitivity must surely be generated and refined. The overriding thrust of such DIF analyses is to identify two groups of teachers who, for item-analysis purposes, are decisively different in their demonstrated effectiveness in bringing about improvements in students' assessed achievement levels. Having identified two extreme groups of teachers on the basis of, for instance, their students' performances for several previous years' worth of annual assessments, we can then see if those teachers' current students' responses to a new set of items are consonant with what would be predicted. For example, if students taught by lower-effectiveness teachers actually perform better on particular items than students taught by higher-effectiveness teachers, then those items should certainly be subjected to serious scrutiny to discern what seems to be rendering them instructionally insensitive. Although Joseph Ryan and I (Popham and Ryan, 2012) have proposed one use of DIF procedures using student-growth-percentiles to carry out item-sensitivity analyses, much more exploratory work on this problem should be undertaken.

As with the reduction of assessment bias in high-stakes educational tests, the implementation of the previously described two-tactic strategy for dealing with instructional sensitivity will not transform instructionally insensitive tests, overnight, into assessment that reek of instructional sensitivity. But our colleagues who coped with assessment bias have given us a set of MapQuest.com directions for making our evaluative tests *more* instructionally sensitive. And progress in that direction, of course, will increase not only the validity of test-based inferences about instructional quality, but also the subsequent decisions we make about the teachers or schools being evaluated.

A Discontented Winter

"Now is the winter of our discontent . . ." are the initial seven words of Shakespeare's *Richard the Third.* Well, it is currently winter and I am definitely discontented. I find it altogether intolerable to be a member of a measurement clan that allows hugely important educational decisions to be made on the basis of students' scores on tests not *demonstrated* to be suitable for their evaluative applications. How can we let such misuses continue? How can we, in good conscience, permit our nation's educational leaders and policymakers to rely on test results that may be completely unsuitable for the purposes to which they are being put? How can we allow teachers to be fired because of students' scores on the wrong tests? How can we? And yet we do.

The only way to begin changing an indefensible practice is to set out seriously to alter that practice. It is time, indeed past-time, for those of us who recognize the seriousness of this situation to don our alteration armor and head into battle.

References

Cox, Richard (1971) "Evaluative Aspects of Criterion-Referenced Measures," in *Criterion-Referenced Measurement: An Introduction,* Ed. W. James Popham*,* Englewood Cliffs, NJ: Educational Technology Publications, 67-75.

Haladyna, Tom and Gale Roid (1981) The role of instructional sensitivity in the empirical review of criterion-referenced test items, *Journal of Educational Measurement,* 18(1), 39-53.

Kane, Michael T. (2013) Validating the Interpretations and Uses of Test Scores, *Journal of Educational Measurement,* 50(1), 1-73.

Popham, W. J. (2006) "Determining the Instructional Sensitivity of Accountability Tests," a presentation at the Large-Scale Assessment Conference, Council of Chief State School Officers, San Francisco, California, June 25-28, 2006.

Popham, W. J. (2014) *Classroom Assessment: What Teachers Need to Know,* 7th ed., Pearson: Boston, 396-398.

Popham, W. J. and Lindheim, E. (1981) Implications of a Landmark Ruling on Florida's Minimum Competency Test, *Phi Delta Kappan,* 63(1), 18-22.

Popham, W. J. and Ryan, J. (2012) "Determining a High-Stakes Test's Instructional Sensitivity," a paper presented at the annual meeting of the National Council on Educational Measurement, April 12-16, Vancouver, B.C., Canada.

Update on Evaluation of NAEP Achievement Levels Procurement

Objective To receive a brief informational update from NCES on the current status of the procurement being planned to evaluate NAEP achievement levels. Ongoing updates will be provided at each COSDAM meeting.

Background

The NAEP legislation 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.

In providing further detail, the aforementioned subsection (f) outlines:

> (1) REVIEW-
>
> A. IN GENERAL- The Secretary shall provide for continuing review of any assessment authorized under this section, and student achievement levels, by one or more professional assessment evaluation organizations.
>
> B. ISSUES ADDRESSED- Such continuing review shall address--
>
> (i) whether any authorized assessment is properly administered, produces high quality data that are valid and reliable, is consistent with relevant widely accepted professional assessment standards, and produces data on student achievement that are not otherwise available to the State (other than data comparing participating States to each other and the Nation);
>
> (ii) whether student achievement levels are reasonable, valid, reliable, and informative to the public;-
>
> (iii) whether any authorized assessment is being administered as a random sample and is reporting the trends in academic achievement in a valid and reliable manner in the subject areas being assessed;
>
> (iv) whether any of the test questions are biased, as described in section 302(e)(4); and
>
> (v) whether the appropriate authorized assessments are measuring, consistent with this section, reading ability and mathematical knowledge.
>
> (2) REPORT- The Secretary shall report to the Committee on Education and the Workforce of the House of Representatives and the Committee on Health,

> Education, Labor, and Pensions of the Senate, the President, and the Nation on the findings and recommendations of such reviews.
>
> (3) USE OF FINDINGS AND RECOMMENDATIONS- The Commissioner for Education Statistics and the National Assessment Governing Board shall consider the findings and recommendations of such reviews in designing the competition to select the organization, or organizations, through which the Commissioner for Education Statistics carries out the National Assessment.

Responsively, a procurement has been planned to administer an evaluation of NAEP achievement levels. The last update COSDAM reviewed on this topic was in August 2013.

In this brief written update, NCES provides the Committee with a summary of the status of this procurement.

ies NATIONAL CENTER FOR EDUCATION STATISTICS
Institute of Education Sciences

Evaluation of NAEP Achievement Levels

The National Center for Education Evaluation and Regional Assistance (NCEERA), part of the Institute for Education Sciences (IES), will administer the Evaluation of the NAEP Achievement Levels. NCEERA and the Department of Education's Contracts and Acquisitions Management (CAM) office will begin this procurement during fiscal year 2014. Tentatively, NCEERA will deliver the Request for Comments (RFC) package to CAM in December 2013 and the scheduled award date is June 2014. This will be a full and open competition.

Reading for Understanding: A Theory-Based, Developmental Approach

Objective: To provide a brief overview of an IES grant on assessment innovations.

Background

During the August 2013 COSDAM meeting, Committee members were invited to provide comments on “Other issues or questions.” John Easton noted that there is an IES grant on assessment innovations for parsing out prior knowledge. Committee members expressed interest in hearing more about this project. In this brief overview, NCES provides the Committee with a description of the Reading for Understanding grant.

Reading for Understanding:
A Theory-Based, Developmental Approach

IES Grant Award Number: R305F100005

Principal Investigators: John Sabatini, Tenaha O'Reilly (ETS)

Project Goals: Develop a series of age-appropriate, developmentally-sensitive, and theoretically-based summative reading comprehension assessments.

Population: The assessments are intended for students in grades Pre-Kindergarten through twelfth grade.

Administration: All assessments are computer delivered and take approximately 45-60 minutes to administer.

The Assessments: Our system incorporates two types of assessments that are designed to measure the components of reading as well as higher-level comprehension.

Component assessments: The first type of assessment is designed to measure the components of reading including decoding, phonological awareness, word recognition, morphology, syntax, vocabulary, listening comprehension and spelling. The components assessment is designed to help contextualize and interpret performance on the Global Integrated Scenario-based Assessment (GISA).

GISA: The second type of assessment, the GISA, is designed to measure a set of integrated skills associated with higher-level comprehension. Students are presented with a realistic purpose for reading that requires them to integrate, synthesize, and evaluate a collection of diverse reading materials (e.g., blog, website). Tasks and activities are sequenced to model complex thinking while simultaneously collecting evidence of partial understanding for developing students. Scaffolding techniques coupled with simulated peer interactions are designed to promote the social nature of reading and the structured nature of learning. Particular emphasis was put on measuring and accounting for variables known to affect reading comprehension, but seldom measured in a summative reading assessment. These variables include background knowledge, student motivation, self-regulation/ metacognition, disciplinary reading, learning, and reading strategies.

Technical information: To date, both types of assessments have been piloted in 23 states in a mix of urban, suburban and rural areas. We have tested over 50,000 students on our GISA assessments and over 115,000 students on our component assessments. Preliminary analyses reveal that our GISA assessments are demonstrating good reliability (typically ɑ=.80 or higher). Our component assessments tend to have even higher reliability (typically ɑ=.90 or higher). We have demonstrated validity evidence (e.g., eye tracking data) and answered a number of key research questions. For example, analyses are uncovering key relationships between components and comprehension, dimensionality of the measures, the role of background knowledge and motivation in testing, the ability of students to learn new information during a reading test, the added value of constructed response items and the relationships between local and global comprehension processes.

College Course Content Analysis Study for NAEP Preparedness Research Progress Update

Submitted by Educational Policy Improvement Center (EPIC)

INTRODUCTION AND BACKGROUND

The College Course Content Analysis (CCCA) study is one of a series of studies contributing to the National Assessment of Educational Progress (NAEP) Program of 12th Grade Preparedness Research conducted by the National Assessment Governing Board (NAGB). The purpose of the CCCA study is to identify a comprehensive list of the reading and mathematics knowledge, skills, and abilities (KSAs) that are pre-requisite to entry-level college mathematics courses and courses that require college level reading based on information from a representative sample of U.S. colleges. The Educational Policy Improvement Center (EPIC) is the contractor working for the Board to conduct this study.

Another goal of the CCCA study is to extend the work of the two previous preparedness studies—the Judgmental Standards Setting (JSS)[1] study, implemented in 2011 and the Job Training Program Curriculum (JTPC) study, implemented in 2012. The CCCA study is designed so the results can be compared to the JSS and JTPC studies, reporting on how this new information confirms or extends interpretations of those earlier studies. The design of the CCCA study is based on the JTPC study but with modifications based on the lessons learned.

The CCCA study will answer four core research questions.

1. What are the prerequisite KSAs in reading and mathematics to qualify for entry-level, credit-bearing courses that satisfy general education requirements?
2. How do these prerequisite KSAs compare with the 2009 and 2013 NAEP reading and mathematics frameworks and item pools?
3. How do these prerequisite KSAs compare with previous NAEP preparedness research (i.e., the descriptions of minimal academic preparedness requirements produced in the JSS research)?
4. How can these prerequisites inform future NAEP preparedness research?

The final report is due May 2014, and until then COSDAM will receive detailed reports at each Board meeting.

[1] National Assessment Governing Board. (2010). *Work Statement for Judgmental Standard Setting Workshops for the 2009 Grade 12 Reading and Mathematics National Assessment of Educational Progress to Reference Academic Preparedness for College Course Placement.* (Higher Education Solicitation number ED-R-10-0005).

METHODOLOGY

The Design Document for the CCCA study is complete. It provides guidance for the study by describing:

- Criteria for collecting courses and artifacts;
- A sampling plan to comprise a representative sample of institutions;
- Review and rating processes, including a training plan and process for ensuring reviewer effectiveness and consistency; and
- The process for ensuring reliability across reviewers providing artifact analysis.

This study comprises three primary phases:

1. Identification and collection of course artifacts,
2. Review of course artifacts by Review Teams, and
3. Analysis and reporting.

OVERVIEW OF ACTIVITIES BY PHASE

Phase 1: Identification and collection of course artifacts

In the CCCA study, a *course artifact* is defined as a syllabus, a non-textbook based assignment or assessment, and textbook excerpt. In mathematics, there are some instances where the only specifically identified assignments were listed in the syllabus and were from the textbook. In those cases, a textbook based assignment or assessment was allowed. The CCCA sample of artifacts is derived from extant artifacts and combined with newly gathered course artifacts. Extant artifacts contributing to the CCCA sample were extracted from EPIC's repository of artifacts compiled during previous research on entry-level curricula at postsecondary educational institutions. Project staff solicited new course artifacts as needed to create a complete and nationally representative sample.

EPIC identified a set of inclusion criteria that courses must meet to be included in the CCCA study as well as a set of institutional characteristics of which the final CCCA Artifact Bank must be representative. The final CCCA Artifact Bank comprises a set of courses and artifacts that are to be used as the basis for the content reviews to be conducted by mathematics and reading content review teams in the second phase of the study.

Phase 1 preparatory work also included the convening of NAEP advisory panels, for reading and mathematics respectively, to obtain content-based guidance and recommendations. In these meetings, preliminary coding schemas, training materials and decision rules were reviewed. NAEP advisors also reviewed all of the course packets to be used in validation data analyses, training sessions, and determining sufficient reviewer competence (qualifying). Guidance from these NAEP advisory panels was integrated into the implementation of the study.

Phase 2: Review of course artifacts by Review Teams

In Phase 2, content reviewers are recruited and training materials are developed in preparation for the review of course artifacts and the content reviews are conducted. Content reviewers are first trained to review the course packets from a "holistic" perspective and identify prerequisite mathematics and reading KSAs. In the second independent review training, the NAEP frameworks for grade 12 reading and mathematics are used as a basis for coding the packets. All additional KSAs beyond the NAEP frameworks are documented and included in all successive reviews, comparisons and data analyses. The overarching goal of the CCCA study is to identify all prerequisite KSAs, not just those KSAs associated with the NAEP frameworks.

Subsets of the course artifact packets were set aside to serve as training packets and qualifying packets. These packets are annotated by the NAEP advisory panel members for use as exemplars of expert coding. After the holistic reviews, content reviewers are trained with respect to the NAEP frameworks, and as part of the training process, the reviewers code the training packets with respect to the NAEP frameworks in small groups. Then, the reviewers code qualifying packets independently. EPIC project staff then compares and scores this coding with respect to the exemplars provided by the NAEP advisory panel. If a reviewer scores below a certain threshold, retraining is provided. Reviewers who receive a second low score are not invited to participate in the study. Qualified reviewers proceed to the next stage: coding 28 course artifact packets independently. Group review meetings are then held to discuss discrepancies identified in independent reviews.

NAEP experts attend the group review meetings as on-site assistance, answering questions about the NAEP framework as they arise. Validity checks are also embedded in the group review process. Validity packets are annotated by the NAEP experts at the advisory panel meetings to be used as reference coding. Those packets are reviewed by all content reviewers without the knowledge that the packets were for validity purposes. This provides the opportunity for evaluating the reliability of the review team coding. The percent agreement between the four review teams' group consensus coding on the validation packets and the NAEP reference coding as reliability evidence will be calculated within each course title and across course titles.

In summary, the CCCA Study's Phase 2 combines independent individual judgments with panel processes. The primary goal of the second, or group, review is to adjudicate differences where possible in coding of the packets completed during the independent review and to produce group-level coding of the additional prerequisite KSAs that were not found in the NAEP frameworks. The final result of this two-part review process is a comprehensive list of prerequisite KSAs, answering the Board's first research question: what are the prerequisite KSAs in reading and mathematics to qualify for entry-level, credit-bearing courses that satisfy general education requirements?

The final step in Phase 2 is for the NAEP experts to review the results of the KSA prerequisite data collected from the content reviewers, which will be summarized in content maps. The NAEP experts' primary task is to compare these data with the 12th grade NAEP 2009 and 2013 items, achievement level descriptions, and minimal academic preparedness descriptions (from the JSS studies) in both mathematics and reading.

Phase 3: Analysis and reporting

Phase 3 includes processing and analyzing the judgments collected during the review of course artifacts by review teams, and preparing the data to be reported in ways that are directly responsive to research questions in accordance with the analysis plan specified within the Design Document. Standard statistical methods and metrics necessary will provide evidence of validity and reliability, and both conceptual (information processing/document analysis) and technical (quantitative) analyses will be conducted. The CCCA study is structured to provide a fully crossed, three factor design to ensure that results can be reviewed in statistical generalizability analyses, which will allow us to evaluate the reliability of the study design.

Final results will include narrative summaries of the prerequisite knowledge, skills, and abilities in mathematics and reading. Summary analyses will also address all aspects of the CCCA study design (see Illustration 1). As project elements are completed, the appropriate sections of Illustration 1 are shaded in dark gray. Project elements that have begun and are in progress are shaded in a lighter gray. Those project elements that have just begun have no shading in the diagram.

Illustration 1: Project Design

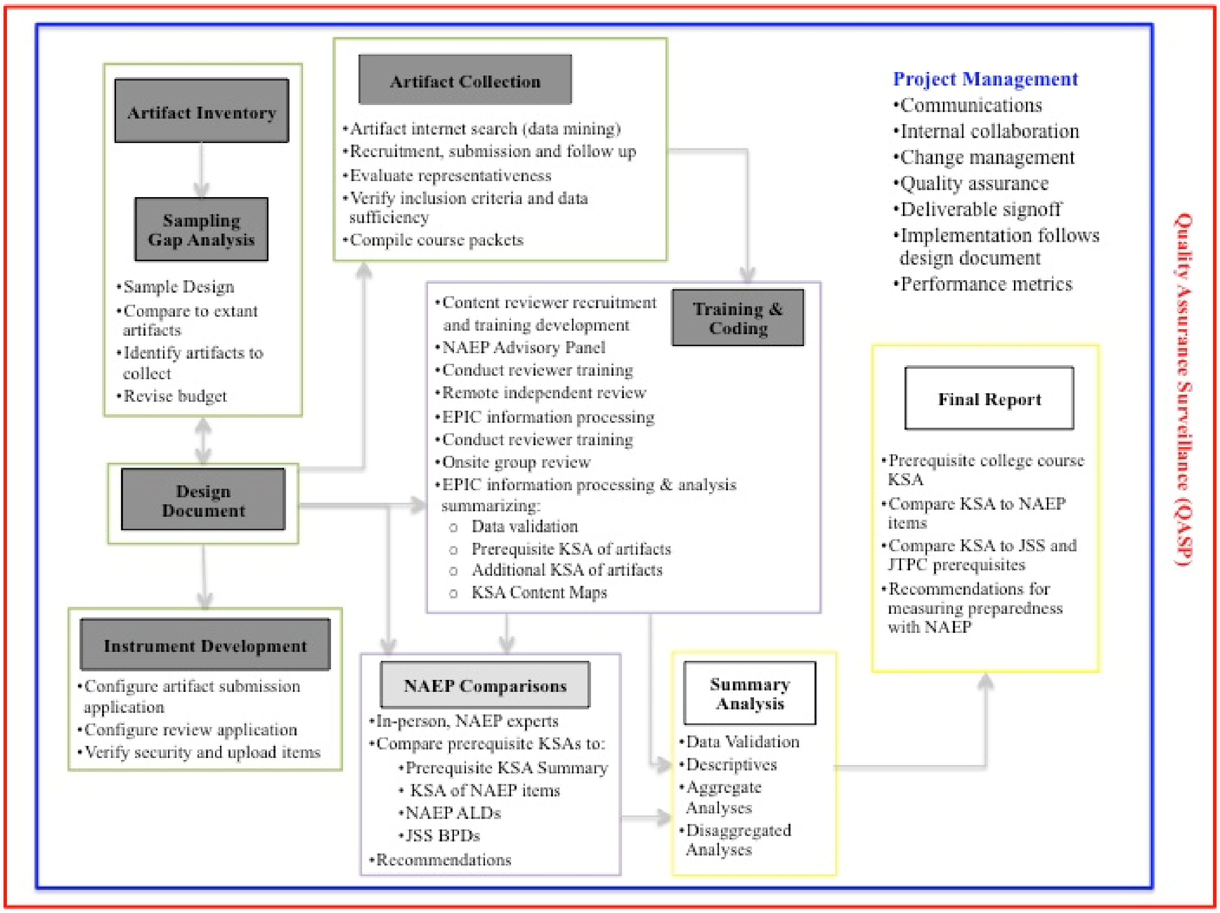
Artifact Inventory
Sampling Gap Analysis
•Sample Design
•Compare to extant artifacts
•Identify artifacts to collect
•Revise budget
Artifact Collection
•Artifact internet search (data mining)
•Recruitment, submission and follow up
•Evaluate representativeness
•Verify inclusion criteria and data sufficiency
•Compile course packets
Project Management
•Communications
•Internal collaboration
•Change management
•Quality assurance
•Deliverable signoff
•Implementation follows design document
•Performance metrics
Quality Assurance Surveillance (QASP)
Training & Coding
•Content reviewer recruitment and training development
•NAEP Advisory Panel
•Conduct reviewer training
•Remote independent review
•EPIC information processing
•Conduct reviewer training
•Onsite group review
•EPIC information processing & analysis summarizing:
○ Data validation
○ Prerequisite KSA of artifacts
○ Additional KSA of artifacts
○ KSA Content Maps
Design Document
Final Report
•Prerequisite college course KSA
•Compare KSA to NAEP items
•Compare KSA to JSS and JTPC prerequisites
•Recommendations for measuring preparedness with NAEP
Instrument Development
•Configure artifact submission application
•Configure review application
•Verify security and upload items
NAEP Comparisons
•In-person, NAEP experts
•Compare prerequisite KSAs to:
•Prerequisite KSA Summary
• KSA of NAEP items
•NAEP ALDs
•JSS BPDs
•Recommendations
Summary Analysis
•Data Validation
•Descriptives
•Aggregate Analyses
•Disaggregated Analyses

Illustration 2 displays a schedule of the CCCA study. As meetings or events are completed, they are noted and shaded in dark gray.

Illustration 2: CCCA Study Gantt chart

MEETING OR EVENT	Start Date	End Date	Duration	Quarter 2			Quarter 3			Quarter 4			Quarter 1			FINAL
				APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	JAN	FEB	MAR	APR
NAEP Technical Panel Meeting	21-Jun	23-Jun	COMPLETE			■										
NAEP Technical Panel Meeting	5-Jun	9-Jun	COMPLETE			■										
Facilitator Training	8-Jul	12-Jul	COMPLETE				■									
Math Content Holistic Training	8-Jul	12-Jul	COMPLETE				■									
Reading Content Holistic Training	8-Jul	12-Jul	COMPLETE				■									
Math Content NAEP Training	22-Jul	24-Jul	COMPLETE				■									
Reading Content NAEP Training	22-Jul	24-Jul	COMPLETE				■									
Independent Content Reviews	9-Jul	16-Aug	COMPLETE					■	■							
EPIC Data Analysis Period 1	19-Aug	6-Sep	COMPLETE					■	■							
Content Review Meeting 1	26-Sep	29-Sep	COMPLETE						■							
Content Review Meeting 2	3-Oct	9-Oct	COMPLETE							■						
EPIC Data Analysis Period 2	29-Sep	15-Nov	COMPLETE						■	■	■					
NAEP Math Expert Review Meeting	3-Jan	5-Jan	3 days										▨			
NAEP Reading Expert Review Meeting	15-Nov	17-Nov	3 days								▨					
EPIC Data Analysis Period 3	18-Nov	31-Jan	10.5 weeks								▨	▨	▨			
Final Report Writing and Review	3-Feb	28-Mar	7.5 weeks											▨	▨	
Board Review and Presentation	1-Apr	30-Apr	4+ weeks													▨
Final Report Deliverable Due	30-Apr	30-Apr	FINAL DELIVERY													▨

PROGRESS UPDATE

Phase 1, *Identification and collection of course artifacts*, is complete. For Phase 2, *Review of course artifacts by Review Teams*, the work of the content reviews, independent and group reviews, is complete. The outcome of the content review is a list of prerequisite KSAs, including additional KSAs and those associated with the NAEP frameworks. This list of prerequisite KSAs, in the form of content maps, is the basis of the upcoming NAEP content expert review. Phase 3, *Analysis and Reporting*, is ongoing. Data are being compiled from the group reviews, the NAEP expert reviews and generalizability analyses.

Independent Content Review (Phase 2)

Content reviews included training, the independent review and group review by mathematics and reading content reviewers. Content reviewers, separated into four mathematics and four reading groups, conducted reviews of 20 course packets and 8 validity packets. The initial holistic review was conducted to elicit additional KSA for each course packet without the influence of the NAEP framework, and familiarized the content reviewers with the review process and course packets. The independent review generated the following data that is the basis of the group content review:

- Applicability and importance coding on all NAEP framework objective level statements
- KSA exclusions on relevant NAEP framework objective level statements (i.e., phrases in the NAEP framework objectives that were not applicable)
- List of additional KSAs evident within the course packets

Thirty-two mathematics and reading content experts participated in training for independent review. Training consisted of the following:

- Attend Holistic Review training webinar
- Complete Holistic Review of 28 course packets
- Attend Independent (NAEP) Review training webinar
- Complete Training Packet #1
- Review feedback and attend re-training (as necessary)
- Complete Training Packet #2 (optional)
- Complete Qualifying Packet #1
- Review Scoring and attend re-training (as necessary)
- Complete Qualifying Packet #2 (as necessary)

Content Reviewers were required to reach a level of coding consistency with NAEP advisory panel coding on a set of qualifying course packets in order to proceed to independent (NAEP framework) review. One content reviewer did not reach the level required and was released from further work on the CCCA project.

Twenty-nine content reviewers completed the independent (NAEP framework) review. Twenty-four of content reviewers were asked to attend the group review meetings. Consistency of content reviewer validity packet coding with the NAEP advisory panel coding was the main factor to determine whether a reviewer was selected to attend the group review meetings.

Group Review (Phase 2)

Two group review sessions were scheduled upon completion of the independent (NAEP framework) review in the Portland, Oregon metropolitan area. Two mathematics and two reading groups attended each session.

Illustration 3: Group Meeting

Group Review Meeting 2 September 26-29, 2013	**Group Review Meeting 2 October 3-6, 2013**
Mathematics Group 1	Mathematics Group 3
Mathematics Group 2	Mathematics Group 4
Reading Group 2	Reading Group 1
Reading Group 4	Reading Group 3

A technical working group, familiar with the CCCA project, considered the multiple content area configurations of the group review meetings and potential threats to the procedural validity of the study. It was determined that no threats to validity existed as all decision rules were finalized prior to the group review sessions. Also, content-area calibration in both mathematics and reading primarily occurred during training and through the review process.

EPIC staff, trained in facilitation and data collection, was engaged to facilitate the group review process. Two required training sessions were convened for five facilitators and five scribes. The first training session focused on the project overview and independent review materials and data. The second session instructed facilitators and scribes on roles, data collection process, and survey instrument instruction.

The objective of the group review meeting sessions was to determine if consensus could be reached on discrepant coding across reviewers in how well the KSAs described in the NAEP framework align with the KSAs evident in course materials. All three members of each review group coded course packets for evidence of mathematics or reading KSAs during independent review by applying the decision rules and using their expert judgment based on evidence in the course packet. At the group review, the same groups discussed and reached consensus on discrepant applicability, importance coding and related KSA exclusions.

Decision points are the number of decisions a group of reviewers was asked to make over the course of the independent review and to come to consensus on in the group review. The number of points is calculated by multiplying the number of packets (28) times the sum of the decisions applicability/importance coding (130 for mathematics or 37 for reading) and KSA exclusions times the number of reviewers.

Illustration 4: Distribution of Coding Decisions

Group	NAEP KSA Coding Decision Points	Points of Discussion/ KSA Coding Discrepancies
Mathematics Group 1	21,840	1,467
Mathematics Group 2	21,840	771
Mathematics Group 3	21,840	959
Mathematics Group 4	21,840	971
Reading Group 1	6,216	965
Reading Group 2	6,216	827
Reading Group 3	6,216	883
Reading Group 4	6,216	938

In the group reviews, coding for applicability and importance of KSAs were reviewed. Reviewers also examined their areas of agreement and as well as phrases in the NAEP framework objectives that were not applicable, i.e., partial matches to NAEP framework objectives (termed "KSA exclusions"). All of the coding is the basis for, and summarized in, content maps to be used for at the NAEP expert review meetings, which are being held in November 2013 and January 2014.

Analysis Conducted to Date (Phase 3)

EPIC conducted a fully-crossed generalizability study on the independent review data in order to determine the inter-rater reliability (i.e., consistency of the reviewers' ratings) on the NAEP standards and framework objectives (KSA) for both reading and mathematics. Generalizability analyses allow analysts to disentangle the contributions made to measurement error by different facets. EPIC analyzed three facets for their contributors to the variance in coding: individual reviewer, NAEP standard, and packet for reading and mathematics. As the G and Phi coefficients approach 1.0, consistency increases; coefficients between .70 and 1.0 are in the acceptable range. Because there are more NAEP standards and framework objectives in mathematics than in reading, EPIC anticipated that the G and Phi coefficients for mathematics would be smaller than for reading, however preliminary results indicate that raters consistently rated the packets at both the standard and objective level for both mathematics and reading.

Content maps have been prepared from the group review data for both mathematics and reading to show the coding provided for each KSA across course packet and course title. Content maps will be generated in spreadsheet form and will be incorporated into a narrative document during the NAEP review material preparation.

Preparation for analysis and the final reporting have begun with the majority of the effort in data management. Staff are working with sample data and testing to ensure that accurate data collection protocols and routines of effective quality control, data cleaning procedures and data storage/security protocols are in place and use.

The final report is also underway. The table of contents has been established and preliminary table shells have been drafted.

STATUS SUMMARY

The first phase of the study is complete. The course artifacts have been identified, all artifacts have been collected, review packets have been created from those artifacts, and the course packets were reviewed by content reviewers independently and then again in group review meetings.

The second phase of the study is nearing completion. Based on guidance from NAEP advisory panels, in both reading and mathematics, feedback was integrated into the content review training and coding schemes and the overall approach to training. Content reviewers were trained in two sessions and required to obtain an acceptable score on training and qualifying packets prior to beginning the process of remote independent content reviews. Next, trained facilitators managed a process to determine and record group level coding of the course packets at onsite group review meetings in September. Project staff will facilitate the comparison work of the NAEP experts at onsite meetings in November 2013 and January 2014. Process evaluations were conducted after training, after independent review and after the group review meetings. Evaluations were largely positive.

The third phase of the study has begun. The data from the independent reviews was compiled for presentation at the onsite group review meeting using online recording tools. A generalizability analysis was conducted on the independent coding data to quantify the variance that certain factors contribute to the dataset.

The data from the group reviews is being compiled for presentation at meetings of the NAEP mathematics and reading content experts. These data are also being used in a generalizability analysis on the group coding data. Preparation of the final report is ongoing.

Evaluating Reading and Mathematics Frameworks and Item Pools as Measures of Academic Preparedness for College and Job Training

Project Status Update
Contract ED-NAG-13-C-0001

The National Assessment Governing Board contracted with the Human Resources Research Organization (HumRRO) in June 2013 to conduct three tasks related to research on 12th grade preparedness:

1. **Evaluation of the Alignment of Grade 8 and Grade 12 NAEP to an Established Measure of Job Preparedness:** This study will extend prior analysis of the relation of NAEP to measures such as WorkKeys by including the NAEP grade 8 assessments and by expanding the method for assessing content alignment. The study method will follow the Governing Board content alignment design document for preparedness research studies, with some modifications. The two-pronged approach includes alignment of: (a) the training preparedness measure to the NAEP frameworks; and (b) NAEP items to the framework from which the training preparedness measure was developed.

2. **O*NET Linkage Study:** This study is a content validity investigation. Major duties (MDs) for the five target occupations will be identified. The occupations are automotive master technician, computer support specialist, HVAC technician, licensed practical nurse, and pharmacy technician. Expert raters will link NAEP content to MDs; NAEP content to O*NET knowledge, skills, and abilities (KSAs); and O*NET KSAs to MDs. This study will identify any disconnects between the level of constructs measured by NAEP and the level of those constructs required for entry into job training programs.

3. **Technical Advisory Panel (TAP) Symposium:** As part of the current contract, HumRRO assembled a technical advisory panel (TAP) of five experts in educational measurement and five experts in industrial-organizational (I-O) psychology to review extant research and to generate ideas for commissioned papers on preparedness. Each panelist is being asked to propose a paper that he/she could develop. Governing Board staff and members will review the proposals and commission up to 10 papers. Panelists will have several months to develop the papers, after which the TAP will reconvene in a late 2014 symposium. Authors will present their papers and the entire panel will discuss implications for preparedness research. HumRRO will produce a proceedings document summarizing the commissioned papers and discussion. (A list of TAP members is included on the next page.)

In addition, HumRRO will produce a comprehensive report at the conclusion of the contract in December 2014.

Work completed to date:

O*NET Linkage: HumRRO has developed lists of major duties (MDs) for each occupation based on O*NET task lists and course objectives from training curricula. Content experts in each occupation have reviewed and vetted the MDs. We also have obtained NAEP items for the O*NET Linkage Study. We are currently assembling materials for conducting the linking exercise.

TAP Symposium: The initial Brainstorming Meeting of the TAP was convened on October 25, 2013 in Crystal City, VA. Panelists will submit paper proposals in late 2013.

Technical Advisory Panel (TAP) Members

John Campbell
Professor of Psychology
University of Minnesota
(Member, NAGB Technical Panel on 12th Grade Preparedness Research, 2007-2008)

Michael Campion
Herman C. Krannert
Professor of Management
Purdue University

Gregory Cizek
Professor of Educational Measurement and Evaluation
University of North Carolina at Chapel Hill

Brian Gong
Executive Director of Center for Assessment
National Center for the Improvement of Educational Assessment, Inc.

Ronald Hambleton
Distinguished University Professor, Educational
Policy, Research, & Administration
Executive Director, Center for Educational Assessment
University of Massachusetts at Amherst

Suzanne Lane
Professor, Research Methodology
University of Pittsburgh School of Education

Kenneth Pearlman
Independent Consultant in Industrial-Organizational Psychology
Sarasota, FL

Barbara Plake
University Distinguished Professor, Emeritus
University of Nebraska-Lincoln

Ann Marie Ryan
Professor of Psychology
Michigan State University

Nancy Tippins
Senior Vice President
CEB Valtera