National Assessment Governing Board
Committee on Standards, Design and Methodology

March 1, 2013
10:00 am–12:15 pm

AGENDA

Joint Session with Reporting and Dissemination Committee *Plaza Room [6th Floor]*		
10:00 – 10:40am	Welcome and Introductions *Lou Fabrizio, COSDAM Chair* *Andrés Alonso, R & D Committee Chair* NAEP Participation Issues and Options ▪ Implementation of Board Policy on Students with Disabilities and English Language Learners *Grady Wilburn, NCES*	Attachment A
COSDAM Meeting *Attaché Room [6th Floor]*		
10:45 – 11:10am	NAEP Participation Issues and Options *(continued)* ▪ State Participation in Voluntary NAEP National Assessments *Keith Rust, Westat*	Attachment B
11:10 – 11:50 am	NAEP 12th Grade Academic Preparedness Research ▪ Update on Course Content Analysis Research □ College □ Job Training *Michelle Blair, NAGB Senior Research Associate* *David Conley and Mary Seburn, Education Policy Improvement Center (EPIC)* ▪ Release of Technical Report on Phase 1 Research ▪ Phase 2 Research Plans *Cornelia Orr, NAGB Executive Director*	Attachment C
11:50 am – 12:05 pm	Preliminary Discussion on Setting NAEP Technology and Engineering Literacy (TEL) Achievement Levels *COSDAM Members*	Attachment D
12:05 – 12:10 pm	Update on Evaluation of NAEP Achievement Levels Procurement *NCES Staff*	
12:10 – 12:15 pm	Other Issues or Questions *COSDAM Members*	

Attachment A

Implementation of Board Policy on Students with Disabilities and English Language Learners

Objective To review and discuss options for analysis and reporting of various forms of student non-participation in NAEP.

Background

While recent data show that large numbers of students with disabilities and English language learners are included in NAEP, variations remain in exclusion and accommodation rates among the states and large urban districts. Building on previous inclusion efforts in 2008, the Board formed the Ad Hoc Committee on NAEP Testing and Reporting on Students with Disabilities and English Language Learners to provide another careful review of the issue. After gathering public comments on recommendations and considering the feedback received from various groups, the Board developed a draft policy statement. At the March 2010 Board meeting, the Board unanimously adopted the policy. The Board's Reporting and Dissemination Committee has also been monitoring progress on implementation of the policy, with several briefings on this topic from NCES staff. This agenda item will include the Reporting and Dissemination Committee as part of a joint session.

Attachments A-1 Issue Summary from NCES

A-2 Board Policy Statement: NAEP Testing and Reporting on Students with Disabilities and English Language Learners

ies NATIONAL CENTER FOR EDUCATION STATISTICS
Institute of Education Sciences

Issues in Implementing the Governing Board's 2010 Inclusion Policy

In 2010, the Governing Board adopted the *NAEP Testing and Reporting on Students with Disabilities and English Language Learners* policy. This policy called for changes in how NAEP would both collect and report data on these two student groups. Through this policy, the Governing Board hoped to make NAEP a more inclusive assessment, to make inclusion and accommodation practices more consistent across the states. Further, the policy called for NCES to report which states meet, and do not meet the Board's inclusion targets of assessing at least 95 percent of all students as well as at least 85 percent of students with disabilities (SD) and English language learners (ELL).

Even before the scheduled full implementation of this policy in the 2013 data collection, many states complied with the spirit of this effort to make NAEP more inclusive. Working with the NAEP state coordinators and field staff, more SD and ELL students participated in the 2011 assessments than in 2009. In the grade 4 reading assessment, for example, in 2009, 17 states did not meet the 95 percent target compared with 9 states in 2011. Nearly all the states (45) included less than 85 of their SD and ELL students in this assessment in 2009, compared with only 18 in 2011. The changes were similar at grade 8. The 2011 NAEP report cards included tables showing which states met these targets, as the Board policy requested.

With one exception, NCES implemented the full policy in the 2013 data collection. The last major component was a new "decision tree," based on the policy, that NAEP administrators are using to assist school personnel in deciding which students should be tested and which accommodations they should receive. The purpose of this new decision tree is to make inclusion practices as uniform as possible across all states.

The one aspect of the decision tree that has proven challenging to implement as stated in the policy pertains to the conversion of certain excluded students to refusals. The policy says that in deciding how a disabled student is to participate in NAEP:

"If the student's IEP or 504 plan specifies an accommodation or modification that is not allowed on NAEP, then the student is encouraged to take NAEP without that accommodation or modification."

Examples of such accommodations are reading aloud the reading test and testing over multiple days.

The Governing Board policy further states:

"Students refusing to take the assessment because a particular accommodation is not allowed should not be classified as exclusions but placed in the category of refusals under NAEP data analysis procedures."

In NAEP's statistical methodology, however, the category of refusals has been set aside for those students who actually refuse to participate in the assessment or whose parents refuse permission for them to participate. Classifying disabled students who don't take the assessments because their IEP accommodations are not offered in NAEP as "refusals" would result in a technical distortion of the way in which the NAEP sample is adjusted to ensure that it accurately represents the student population as a whole.

NAEP uses a procedure known as "weight-class adjustments" to ensure that results collected from a sample of students accurately reflects the results that would be obtained from testing all students. Weight-class adjustments are defined as follows in the NAEP technical documentation:

"The student nonresponse adjustment procedure inflates the weights of assessed students to account for eligible sampled students who did not participate in the assessment. These inflation factors offset the loss of data associated with absent students. The adjustments are computed within nonresponse cells and are based on the assumption that the assessed and absent students within the same cell are more similar to one another than to students from different cells. Like its counterpart at the school level, the student nonresponse adjustment is intended to reduce the mean square error and thus improve the accuracy of NAEP assessment estimates."

In this procedure, students who refuse participation, as well as absent students, are given weights because presumably they would be able to take the assessments. If students not taking the tests due to unavailability of accommodations were classified as refusals, then refusals would no longer be a random group, since most of these students are relatively low-performing and according to their schools would not be able to take the assessments. Weight-class adjustments made for this group would then constitute an inappropriate use of the statistic, not comparable to the way it is used in other large-scale assessments such as TIMSS and PIRLS.

Classifying these students as "refusals" would result in other, unintended, consequences as well:

- The trend line may not be maintained if the methodology is changed;
- Exclusion rates would be artificially lowered though fewer students were tested;
- Participation rates would decrease as refusals increased; and
- Average scores on the assessments could be lowered in some jurisdictions.

NCES will discuss these issues in more detail at the meeting. We will describe the 2013 data being collected to analyze the impact of classifying these students as other than excluded, and to better understand the barriers still preventing some students from taking the assessments. We will also describe measures being taken to increase participation of students with disabilities and English language learners, and to make inclusion practices in NAEP more consistent across states and school districts. In support of the intent of the Board policy of converting excluded students to refusals, NCES will discuss alternative ways of reporting state exclusion rate data that will show the proportion of excluded students who could not participate in the 2013 assessments because their accommodations were not allowed or provided in NAEP.

Adopted: March 6, 2010

National Assessment Governing Board

NAEP Testing and Reporting on Students with Disabilities and English Language Learners

Policy Statement

INTRODUCTION

To serve as the Nation's Report Card, the National Assessment of Educational Progress (NAEP) must produce valid, comparable data on the academic achievement of American students. Public confidence in NAEP results must be high. But in recent years it has been threatened by continuing, substantial variations in exclusion rates for students with disabilities (SD) and English language learners (ELL) among the states and urban districts taking part.

Student participation in NAEP is voluntary, and the assessment is prohibited by law from providing results for individual children or schools. But NAEP's national, state, and district results are closely scrutinized, and the National Assessment Governing Board (NAGB) believes NAEP must act affirmatively to ensure that the samples reported are truly representative and that public confidence is maintained.

To ensure that NAEP is fully representative, a very high proportion of the students selected must participate in its samples, including students with disabilities and English language learners. Exclusion of such students must be minimized; they should be counted in the Nation's Report Card. Accommodations should be offered to make the assessment accessible, but these changes from standard test administration procedures should not alter the knowledge and skills being assessed.

The following policies and guidelines are based on recommendations by expert panels convened by the Governing Board to propose uniform national rules for NAEP testing of SD and ELL students. The Board has also taken into consideration the views expressed in a wide range of public comment and in detailed analyses provided by the National Center for Education Statistics, which is responsible for conducting the assessment under the policy guidance of the Board**.** The policies are presented not as statistically-derived standards but as policy guidelines intended to maximize student participation, minimize the potential for bias, promote fair comparisons, and maintain trends. They signify the Board's strong belief that NAEP must retain public confidence that it is fair and fully-representative of the jurisdictions and groups on which the assessment reports.

POLICY PRINCIPLES

1. As many students as possible should be encouraged to participate in the National Assessment. Accommodations should be offered, if necessary, to enable students with disabilities and English language learners to participate, but should not alter the constructs assessed, as defined in assessment frameworks approved by the National Assessment Governing Board.

2. To attain comparable inclusion rates across states and districts, special efforts should be made to inform and solicit the cooperation of state and local officials, including school personnel who decide upon the participation of individual students.

3. The proportion of all students excluded from any NAEP sample should not exceed 5 percent. Samples falling below this goal shall be prominently designated in reports as not attaining the desired inclusion rate of 95 percent.

4. Among students classified as either ELL or SD a goal of 85 percent inclusion shall be established. National, state, and district samples falling below this goal shall be identified in NAEP reporting.

5. In assessment frameworks adopted by the Board, the constructs to be tested should be carefully defined, and allowable accommodations should be identified.

6. All items and directions in NAEP assessments should be clearly written and free of linguistic complexity irrelevant to the constructs assessed.

7. Enhanced efforts should be made to provide a short clear description of the purpose and value of NAEP and of full student participation in the assessment. These materials should be aimed at school personnel, state officials, and the general public, including the parents of students with disabilities and English language learners. The materials should emphasize that NAEP provides important information on academic progress and that all groups of students should be counted in the Nation's Report Card. The materials should state clearly that NAEP gives no results for individual students or schools, and can have no impact on student status, grades, or placement decisions.

8. Before each state and district-level assessment NAEP program representatives should meet with testing directors and officials concerned with SD and ELL students to explain NAEP inclusion rules. The concerns of state and local decision makers should be discussed.

IMPLEMENTATION GUIDELINES

For Students with Disabilities

1. Students with disabilities should participate in the National Assessment with or without allowable accommodations, as needed. Allowable accommodations are any changes from standard test administration procedures, needed to provide fair access by students with disabilities that do not alter the constructs being measured and produce valid results. In cases where non-standard procedures are permitted on state tests but not allowed on NAEP, students will be urged to take NAEP without them, but these students may use other allowable accommodations that they need.

2. The decision tree for participation of students with disabilities in NAEP shall be as follows:

NAEP Decision Tree for Students with Disabilities

BACKGROUND CONTEXT

1. NAEP is designed to measure constructs carefully defined in assessment frameworks adopted by the National Assessment Governing Board.

2. NAEP provides a list of appropriate accommodations and non-allowed modifications in each subject. An appropriate accommodation changes the way NAEP is normally administered to enable a student to take the test but does not alter the construct being measured. An inappropriate modification changes the way NAEP is normally administered but does alter the construct being measured.

STEPS OF THE DECISION TREE

3. In deciding how a student will participate in NAEP:

   a. If the student has an Individualized Education Program (IEP) or Section 504 plan and is tested without accommodation, then he or she takes NAEP without accommodation.

   b. If the student's IEP or 504 plan specifies an accommodation permitted by NAEP, then the student takes NAEP with that accommodation.

   c. If the student's IEP or 504 plan specifies an accommodation or modification that is not allowed on NAEP, then the student is encouraged to take NAEP without that accommodation or modification.

3. Students should be considered for exclusion from NAEP only if they have previously been identified in an Individualized Education Program (IEP) as having the most significant cognitive disabilities, and are assessed by the state on an alternate assessment based on alternate achievement standards (AA-AAS). All students tested by the state on an alternate assessment with modified achievement standards (AA-MAS) should be included in the National Assessment.

4. Students refusing to take the assessment because a particular accommodation is not allowed should not be classified as exclusions but placed in the category of refusals under NAEP data analysis procedures.

5. NAEP should report separately on students with Individualized Education Programs (IEPs) and those with Section 504 plans, but (except to maintain trend) should only count the students with IEPs as students with disabilities. All 504 students should participate in NAEP.

   At present the National Assessment reports on students with disabilities by combining results for those with an individualized education program (who receive special education services under the Individuals with Disabilities Education Act [IDEA]) and students with Section 504 plans under the Rehabilitation Act of 1973 (a much smaller group with disabilities who are not receiving services under IDEA but may be allowed test accommodations).* Under the Elementary and Secondary Education Act, only those with an IEP are counted as students with disabilities in reporting state test results. NAEP should be consistent with this practice. However, to preserve trend, results for both categories should be combined for several more assessment years, but over time NAEP should report as students with disabilities only those who have an IEP.

6. Only students with an IEP or Section 504 plan are eligible for accommodations on NAEP. States are urged to adopt policies providing that such documents should address participation in the National Assessment.

For English Language Learners

1. All English language learners selected for the NAEP sample who have been in United States schools for one year or more should be included in the National Assessment. Those in U.S. schools for less than one year should take the assessment if it is available in the student's primary language.

* NOTE: The regulation implementing Section 504 defines a person with a disability as one who has a physical or mental impairment which substantially limits one or more major life activities, has a record of such an impairment, or is regarded as having such an impairment. 34 C.F.R. § 104.3(j)(1).

One year or more shall be defined as one full academic year before the year of the assessment.

2. Accommodations should be offered that maximize meaningful participation, are responsive to the student's level of English proficiency, and maintain the constructs in the NAEP framework. A list of allowable accommodations should be prepared by NAEP and furnished to participating schools. Such accommodations may be provided only to students who are not native speakers of English and are currently classified by their schools as English language learners or limited English proficient (LEP).

3. Bilingual versions of NAEP in Spanish and English should be prepared in all subjects, other than reading and writing, to the extent deemed feasible by the National Center for Education Statistics. The assessments of reading and writing should continue to be in English only, as provided for in the NAEP frameworks for these subjects.

4. Staff at each school should select from among appropriate ELL-responsive accommodations allowed by NAEP, including bilingual booklets, those that best meet the linguistic needs of each student. Decisions should be made by a qualified professional familiar with the student, using objective indicators of English proficiency (such as the English language proficiency assessments [ELPA] required by federal law), in accordance with guidance provided by NAEP and subject to review by the NAEP assessment coordinator.

5. Schools may provide word-to-word bilingual dictionaries (without definitions) between English and the student's primary language, except for NAEP reading and writing, which are assessments in English only.

6. NAEP results for ELL students should be disaggregated and reported by detailed information on students' level of English language proficiency, using the best available standardized assessment data. As soon as possible, NAEP should develop its own brief test of English language proficiency to bring consistency to reporting nationwide.

7. Data should be collected, disaggregated, and reported for former English language learners who have been reclassified as English proficient and exited from the ELL category. This should include data on the number of years since students exited ELL services or were reclassified.

8. English language learners who are also classified as students with disabilities should first be given linguistically-appropriate accommodations before determining which additional accommodations may be needed to address any disabilities they may have.

RESEARCH AND DEVELOPMENT

The Governing Board supports an aggressive schedule of research and development in the following areas:

1. The use of plain language and the principles of universal design, including a plain language review of new test items consistent with adopted frameworks.

2. Adaptive testing, either computer-based or paper-and-pencil. Such testing should provide more precise and accurate information than is available at present on low-performing and high-performing groups of students, and may include items appropriate for ELLs at low or intermediate levels of English proficiency. Data produced by such targeted testing should be placed on the common NAEP scale. Students assessed under any new procedures should be able to demonstrate fully their knowledge and skills on a range of material specified in NAEP frameworks.

3. A brief, easily-administered test of English language proficiency to be used for determining whether students should receive a translation, adaptive testing, or other accommodations because of limited English proficiency.

4. The validity and impact of commonly used testing accommodations, such as extended time and small group administration.

5. The identification, measurement, and reporting on academic achievement of students with the most significant cognitive disabilities. This should be done in order to make recommendations on how such students could be included in NAEP in the future.

6. A study of outlier states and districts with notably high or low exclusion rates for either SD or ELL students to identify the characteristics of state policies, the approach of decision makers, and other criteria associated with different inclusion levels.

The Governing Board requests NCES to prepare a research agenda on the topics above. A status report on this research should be presented at the November 2010 meeting of the Board.

ies NATIONAL CENTER FOR EDUCATION STATISTICS
Institute of Education Sciences

State Participation in Voluntary NAEP Assessments

With the advent of the No Child Left Behind (NCLB) legislation, with its mandate for states and districts to participate in reading and mathematics assessments at grades 4 and 8, and the appointment of a NAEP State Coordinator in each state, it has become routine that recruitment of public schools and their districts for NAEP takes place at the state level. This is true of all NAEP assessments, not just those that are mandated by NCLB. Prior to 2003 NAEP staff generally negotiated directly with districts concerning participation in assessments that were conducted at the national level only.

Beginning with the 2005 assessments, a small number of states have declined to allow any of their schools to be recruited for participation in select national-only assessments. From 2005 to 2010 this was restricted to grade 12 (and age 17) assessments, but in the past three assessment years this has extended to non-mandated assessments at lower grades as well. In total, ten different states have refused to participate in one assessment or another. However, the greatest number of states to refuse any one assessment was five, which occurred for the 2009 grade 12 assessments in reading, mathematics, and science. These five states made up approximately 6.5 percent of the public school population, and were spread among the four Census regions. The only instances of large state refusals were in 2005, when New York declined to participate in the High School Transcript Study, and 2012, when Texas declined to participate in the grade 12 economics assessment. The states that have most consistently declined to participate are Maryland and Rhode Island, which have each declined to participate in every grade 12 and age 17 assessment since 2005.

Nonparticipation at the state level has implications for NAEP analysis procedures. In all past assessments NAEP results have been presented as representing the entire nation. To achieve this, nonresponse adjustments have been made at the school level, to compensate for the nonparticipating states, as well as any nonparticipating schools in other states. In states that do not participate, a school sample is selected in any case (generally the decision of the state not to participate is confirmed after the school sample has been selected). Thus nonresponse adjustments at the school level can be implemented, and can effectively compensate for the missing states. Nonresponse bias analyses are conducted for any assessment where the aggregate school response rate falls below 85 percent, and these have not revealed any major concerns arising from state level refusals.

Nevertheless, the threat to the validity of NAEP inferences resulting from state non-participation must be taken seriously. Trend analyses are also affected by the fact that, over time, different states have declined to participate in different assessments.

During the meeting information will be presented to COSDAM on how NAEP has addressed the issue of state non-participation in past analyses. The analysis implications and policy considerations involved in maintaining national trends will also be reviewed.

NAEP 12th Grade Preparedness Research

Based on the Program of Preparedness Research adopted by the Governing Board in March 2009, four categories of research studies were conducted to produce evidence to develop and support the validity of statements for NAEP reporting on the academic preparedness in reading and mathematics of 12th grade students for college and job training.

- content alignment studies;
- statistical relationship studies;
- judgmental standard setting studies; and
- surveys

Additionally, the Texas Commissioner of Higher Education offered the opportunity to conduct a benchmarking study with Texas higher education institutions, and a pilot study to examine the feasibility was conducted.

Based on discussions at its quarterly meetings in May 2012 and August 2012, the Governing Board has determined that the research studies completed to date should be made available through an online technical report. Dissemination through this format will be useful to the research community as well as policymakers and interested members of the general public. The online technical report was released on February 15, 2013. In addition, the NAEP 12th Grade Preparedness Commission will conduct a symposium in Washington, DC in Spring 2013 focused on the Board's preparedness research results and the plans for the next phase of the research.

The March 1, 2013 COSDAM briefing will primarily focus on:

- Progress updates related to the research being conducted in connection with the 2009 grade 12 NAEP results
    - □ College Course Content Analysis Progress Update (Attachment C-1)..........Page 14
    - □ Job Training Program Content Analysis Update (Attachment C-2) ....Page 21

Additionally, the following informational attachments are provided:

- The press announcement unveiling the newly released online technical report for the 2009 grade 12 NAEP preparedness research (Attachment C-3) ....Page 29

- Proposed research projects for phase 2 of the Board's preparedness research program (Attachment C-4) ....Page 33

- Background materials describing each study category (Attachment C-5) ....Page 35

Attachment C-1
College Course Content Analysis Progress Update

In September 2012, the Governing Board awarded a contract to the Education Policy Improvement Center (EPIC) to conduct research on entry level non-remedial college course content in order to (1) identify the prerequisite knowledge and skills in reading and mathematics for entry-level college courses and (2) determine the extent to which there is a match with the content of grade 12 NAEP. This project addresses academic preparedness for college only—a separate parallel research project addresses preparedness for job training (described below).

In this project, EPIC will determine the entry-level (introductory) credit-bearing courses most frequently taken by entering students that are reflective of college-level reading and mathematics demands and that satisfy general education requirements. These introductory courses should have no college-level prerequisite course requirements, and only non-remedial courses that satisfy general education requirements should be included in the analysis. Further, in cases where multiple versions of a course are offered for majors and non-majors, only the course for non-majors should be included.

Using course artifacts for a generally representative sample of institutions, EPIC will analyze the introductory course artifacts for commonalities and differences in the reading and mathematics prerequisites (i.e., the prerequisite KSAs) needed to qualify for placement into the course. From these analyses, EPIC will develop descriptions of the knowledge, skills, and abilities needed for students to qualify for placement into the introductory course, based on an analysis of the course artifacts. And as part of a set of comparative analyses, EPIC will then use these descriptions to review:

- the description of minimal requirements for placement into college-level coursework as developed in the NAEP preparedness judgmental standard setting (JSS) research
- KSAs represented by 2009 grade 12 items that map to the NAEP scale with a response probability of .67 and fall within the range of cut scores set by the two replicate panels in the JSS research
- 2009 and 2013 grade 12 NAEP items
- the KSAs represented by 2009 items that map in the range of the NAEP score scale from the mid-range of the Basic level to the mid-range of the Proficient level; and
- the NAEP achievement level descriptions.

A progress report is attached with more details on the project and a description of work completed to date.

College Course Content Analysis Study for NAEP Preparedness Research

Submitted by

Educational Policy Improvement Center (EPIC)

INTRODUCTION AND BACKGROUND

The National Assessment Governing Board (Governing Board) adopted a Program of Preparedness Research in March 2009. Part of this research includes an effort to examine the validity of findings obtained from the judgmental standards setting studies and to better understand the knowledge, skills, and abilities in reading and mathematics required for non-remedial, entry-level college courses that satisfy general education requirements leading to a Bachelor's degree. This College Course Content Analysis Study (CCCA) is intended to provide a clearer understanding of the prerequisite knowledge, skills, and abilities (KSAs) required for entry-level coursework across a nationally representative sample of colleges. This particular statement of work addresses academic preparedness for college only—a separate research project addresses preparedness for job training.

The CCCA study will address 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 (i.e., planning and analysis efforts relative to the 2013 grade 12 NAEP reading and mathematics assessments)?

METHODOLOGY

The Design Document is currently in development. It will guide the CCCA 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 reliability; and
- The process for ensuring reliability across raters providing artifact analysis.

This study comprises three primary phases:

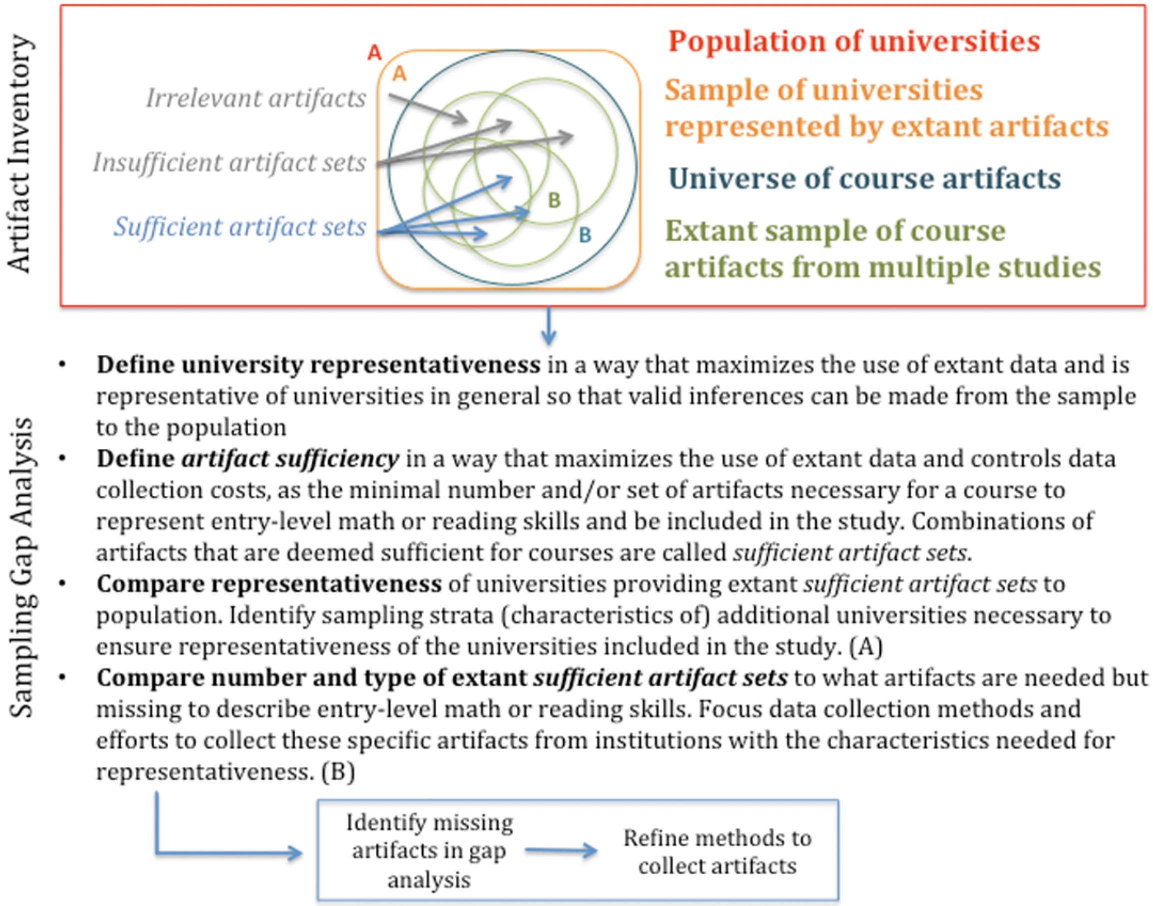
1. Identification and collection of course artifacts,
2. Review of course artifacts by Review Teams, and
3. Analysis and reporting.

Phase 1: Identification and collection of course artifacts

EPIC employs standard statistical methods and metrics necessary to monitor and demonstrate validity and reliability, as well as apply well-established processes for identification, collection and tracking course artifacts. EPIC will use both conceptual (information processing/document analysis) and technical (quantitative) analyses using established methods. EPIC relies on expert judgment models that are based on modified versions of the Delphi methodology to determine the mathematics and reading content of the NAEP that is prerequisite to entry-level college courses.

For the purposes of CCCA study, a *course artifact* is defined as a syllabus, textbook, assignment, or assessment. The CCCA sample of artifacts will be derived from extant artifacts and combined, if necessary, with newly gathered course artifacts. Extant artifacts contributing to the CCCA sample were extracted with permission from EPIC's repository of extant artifacts compiled during previous research on entry-level curricula at postsecondary educational institutions.

Illustration 1: Identification and collection of course artifacts


Artifact Inventory
Irrelevant artifacts
Insufficient artifact sets
Sufficient artifact sets
A
A
B
B
Population of universities
Sample of universities represented by extant artifacts
Universe of course artifacts
Extant sample of course artifacts from multiple studies
Sampling Gap Analysis

- **Define university representativeness** in a way that maximizes the use of extant data and is representative of universities in general so that valid inferences can be made from the sample to the population
- **Define *artifact sufficiency*** in a way that maximizes the use of extant data and controls data collection costs, as the minimal number and/or set of artifacts necessary for a course to represent entry-level math or reading skills and be included in the study. Combinations of artifacts that are deemed sufficient for courses are called *sufficient artifact sets.*
- **Compare representativeness** of universities providing extant *sufficient artifact sets* to population. Identify sampling strata (characteristics of) additional universities necessary to ensure representativeness of the universities included in the study. (A)
- **Compare number and type of extant *sufficient artifact sets*** to what artifacts are needed but missing to describe entry-level math or reading skills. Focus data collection methods and efforts to collect these specific artifacts from institutions with the characteristics needed for representativeness. (B)

Identify missing artifacts in gap analysis
Refine methods to collect artifacts

Phase 2: Review of course artifacts by Review Teams

EPIC uses convergent consensus for reviews by combining independent, individual judgments with panel consensus processes. This two-part approach allows for the capture and integration of responses from two types of experts: those knowledgeable in the content area, and those highly familiar with NAEP frameworks. Similar methods have been used in previous NAGB research describing the prerequisite KSAs for college and job training programs (Judgmental Standard Setting, Job Training Programs Curriculum Study).

Phase 3: Analysis and reporting

The analysis and reporting phase refers to 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.

Illustration 2 describes the project design and Illustration 3 demonstrates a timeline of the major project elements.

Illustration 2: Project design

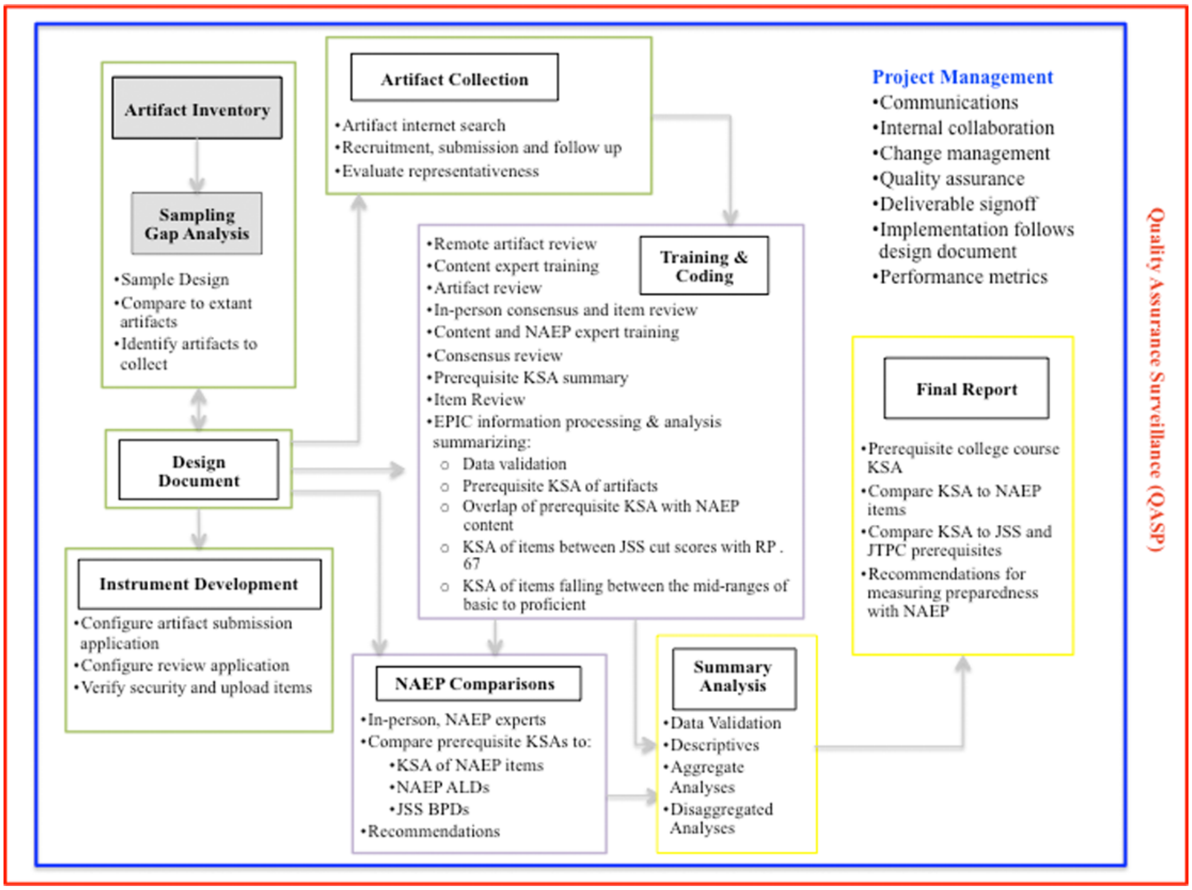

Artifact Inventory
Sampling Gap Analysis
•Sample Design
•Compare to extant artifacts
•Identify artifacts to collect
Artifact Collection
•Artifact internet search
•Recruitment, submission and follow up
•Evaluate representativeness
Project Management
•Communications
•Internal collaboration
•Change management
•Quality assurance
•Deliverable signoff
•Implementation follows design document
•Performance metrics
Training & Coding
•Remote artifact review
•Content expert training
•Artifact review
•In-person consensus and item review
•Content and NAEP expert training
•Consensus review
•Prerequisite KSA summary
•Item Review
•EPIC information processing & analysis summarizing:
○ Data validation
○ Prerequisite KSA of artifacts
○ Overlap of prerequisite KSA with NAEP content
○ KSA of items between JSS cut scores with RP . 67
○ KSA of items falling between the mid-ranges of basic to proficient
Design Document
Instrument Development
•Configure artifact submission application
•Configure review application
•Verify security and upload items
NAEP Comparisons
•In-person, NAEP experts
•Compare prerequisite KSAs to:
•KSA of NAEP items
•NAEP ALDs
•JSS BPDs
•Recommendations
Summary Analysis
•Data Validation
•Descriptives
•Aggregate Analyses
•Disaggregated Analyses
Final Report
•Prerequisite college course KSA
•Compare KSA to NAEP items
•Compare KSA to JSS and JTPC prerequisites
•Recommendations for measuring preparedness with NAEP
Quality Assurance Surveillance (QASP)

Illustration 3: Major project elements timeline

	FY 2012-13												FY 2013-14											
	O	N	D	J	F	M	A	M	J	J	A	S	O	N	D	J	F	M	A	M	J	J	A	S
Kickoff	■																							
Artifact Inventory & Sampling Gap Analysis	■	■	■	■	■																			
Design Document*	■	■	■	■	■	■																		
Instrument Development				■	■	■	■																	
Artifact Collection						■	■	■																
Coding & Review							■	■	■	■	■													
NAEP Comparisons											■	■	■											
Summary Analyses														■	■	■								
Draft & Final Report															■	■	■	■	■					
QASP	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■					

**Iterative process that can only be completed upon completion of Artifact Inventory and Sampling Gap Analysis.*

PROGRESS UPDATE

Identification and Collection of Course Artifacts

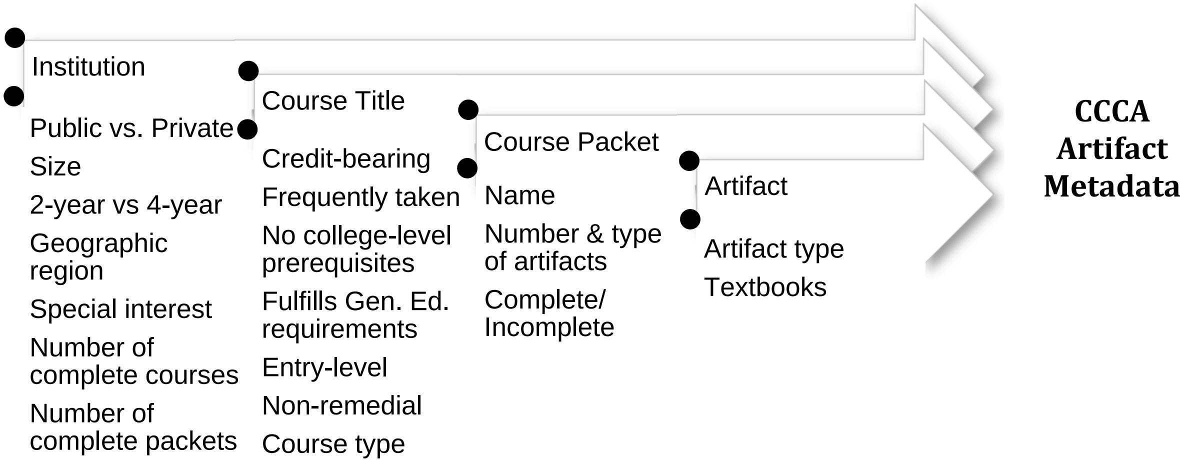
The remainder of the report summarizes progress made on the Artifact Inventory and Sampling Gap Analysis.

The Artifact Inventory, a descriptive report of extant artifacts contributing to the CCCA sample, has been completed. The Sampling Gap Analysis will be completed in the next few weeks. The Design Document will use the data and analysis from the Artifact Inventory and Sampling Gap Analysis to plan for the duration of the project. Extant artifacts that meet the CCCA sample criteria currently include 451 courses and 795 artifacts. The CCCA artifact sample requires that all courses included in the study meet the following criteria:

- Credit-bearing,
- Frequently taken,
- No college-level prerequisites,
- Fulfills general education requirements,
- Entry-level,
- Non-remedial,
- Not honors level,
- Not tailored for a specific major, and
- From academic year 2009-2010 or 2010-2011.

In addition to courses meeting the above criteria, the courses must be from a representative sample of institutions. Illustration 4 identifies the representativeness characteristics at the institution level, the inclusion criteria at the course level, and other relevant data to be collected and validated for each artifact in the CCCA sample.

Illustration 4: Representativeness, inclusion criteria, and other relevant data



Tables 1 and 2 contain descriptive information on the extant artifacts contributing to the CCCA sample. The data from the Artifact Inventory will be used in the Sampling Gap Analysis to determine whether additional data collection is necessary.

Table 1: Summary of Extant Artifacts Meeting CCCA Inclusion Criteria

Course Type (Math or Reading)	Course Title	# Courses	# Artifacts
Reading	Biology	50	111
Math	Calculus	8	15
Math	Chemistry	16	19
Math	College Algebra	13	16
Reading	Composition I	47	91
Reading	English Literature	20	64
Math	Statistics	**11**	16
Math	Physics	10	18
Reading	Introduction to Psychology	78	144
Reading	Introduction to Sociology	59	88
Reading	US Government	71	99
Reading	US History	68	114
	TOTAL	**451**	**795**

Table 2: Number of Course Artifacts by Course Type

Course Type		Syllabi	Assessments	Assignments	Other
Reading		392	77	168	74
Math		59	14	7	4
	TOTAL	**451**	**91**	**175**	**78**

Tables 3 and 4 show the number of courses in the current CCCA Sample by institution characteristics and the number of courses considered to have complete or incomplete packets of artifacts. Complete Courses/Course Packets are defined as those with a syllabus, at least one identified textbook, and at least one supplemental artifact (assignment or assessment). All supplemental artifacts in the extant artifact database were provided by the instructor and, therefore, considered non-textbook-based. That is, the artifacts were not extracted directly from the course textbook. A packet is deemed incomplete if it is missing any of the required artifacts.

Table 3: Math and Reading Courses by Institution Characteristics

	Math		Reading	
Institution Characteristic	Complete	Incomplete	Complete	Incomplete
Program Offered				
2-year	1	4	20	74
4-year	8	44	71	227
Institution Size				
Under 1,000	0	7	10	44
1,000 - 4,999	6	26	49	156
5,000 - 9,999	0	9	17	42
10,000 – 19,999	2	6	10	38
20,000 and above	1	0	5	21
Public/Private				
Public	3	24	49	142
Private	6	24	42	159
Geographic Region				
East	1	8	12	59
Midwest	4	19	28	78
Southeast	2	15	23	87
Southwest	1	3	11	30
West	1	3	17	47

Table 4: Number of Complete/Incomplete Course Packets by Course Type

Course Type	Complete Packets	Incomplete Packets	Total Packets
Reading	91	302	393
Math	10	48	58
TOTAL	**101**	**350**	**451**

The Artifact Inventory will be analyzed in the Sampling Gap Analysis to determine whether there is a sufficient and representative sample of institutions within the extant courses and artifacts. The Sampling Gap Analysis will determine if additional collection of data is necessary. Should the Sampling Gap Analysis determine that additional data collection is necessary, college instructors will be engaged in the process and other creative approaches will be explored to minimize their involvement.

A list of all textbooks is being compiled as part of the coding at the artifact level. The Design Document will address how textbook data will be used in the content reviews and analyses.

Attachment C-2
Job Training Program Content Analysis Progress Update

In October 2011, the Governing Board began work with WestEd and its subcontractor, the Education Policy Improvement Center (EPIC), to conduct follow-up research relative to the NAEP preparedness judgmental standard setting (JSS) research, wherein panelists reviewed NAEP questions and made judgments about the content knowledge needed by minimally prepared students. The research results from this project are intended to supplement the JSS research findings by providing a clearer understanding of the knowledge and skills required for entry- and exit-level coursework in designated occupational programs. By reviewing course artifacts such as syllabi, text books, and assignments, this study will help to determine if the knowledge, skills, and abilities (KSAs) required of students in the training programs are appropriately represented by the borderline preparedness descriptions (developed in the JSS research), by all the items on the 2009 NAEP, and by the 2009 NAEP items in the scale score ranges identified by panelists in the JSS research project.

Attached is a status report further detailing the methodology and providing a summary of the project's progress.

Job Training Program Course Content Study Update

BACKGROUND

The National Assessment Governing Board (Governing Board) adopted a Program of Preparedness Research in March 2009 that included judgmental standard-setting (JSS) studies to identify cutscores representing minimal academic preparedness on grade 12 NAEP with respect to entry into job-training programs and for placement in college credit-bearing courses. A total of 180 job training programs were represented in the judgmental standard setting studies focusing on five occupations:

Occupation	Number of Programs
Automotive master technician	41
Computer support specialist	31
Heating, ventilation, and air conditioning technician	31
Licensed practical nurse	40
Pharmacy technician	37

The Governing Board requested additional research to examine the validity of findings obtained from the JSS studies and to better understand the knowledge, skills, and abilities in reading and mathematics required for these occupational training programs. This additional research is intended to provide a clearer understanding of the knowledge, skills, and abilities (KSAs) required for entry- and exit-level coursework in designated job training programs within these occupations. This study will help to determine if the KSAs required of students in the training programs are appropriately represented by the borderline preparedness descriptions (BPDs) and by the NAEP items near the reference points developed in the JSS studies to represent the minimal level of academic knowledge and skills in the subject matter necessary for a student to be prepared to enter the job training course.

METHODOLOGY

This study addresses the following research questions:

1. What mathematics and reading KSAs are prerequisite to the introductory-level courses, and what mathematics and reading KSAs are taught in the introductory courses for the job-training programs for each occupation?
2. What mathematics and reading KSAs are students expected to have attained at the conclusion of the job-training programs for each occupation?

3. How do the prerequisites (KSA expectations for entry) for job training programs in each occupation relate to descriptions of minimal academic preparedness on NAEP (as described by the BPDs from the JSS studies)?
4. How do prerequisites (KSA expectations for entry) for job training programs in each occupation relate to the content assessed by NAEP (as determined by NAEP items representing minimal academic preparedness)?

This study comprises three primary phases:

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

PHASE 1
Identification and Collection of Course Artifacts

Programs from the five occupations used in the JSS studies have comprised the population of programs for this study; from this population, a minimum of 20 programs per occupation have been recruited from the 180 programs represented on the JSS panels.

Occupational job-training instructors who served on the JSS panels were recruited to participate in this study. These job training instructors were asked to identify courses that best address the objectives of this study and to submit artifacts for those courses. These instructors also had the option of nominating colleagues who teach one or more courses selected for the study to participate in this activity. Course artifacts were collected for all programs in each occupational area that agreed to participate, with course submission remaining open until either materials were obtained from a minimum of 20 programs or the population of programs had been exhausted.

Each participating program instructor was asked to (1) identify foundational textbooks for her/his program; (2) verify program and institution information (e.g., accreditation status, course sequencing, school and department admission requirements, degree accreditation, and credit requirements); and (3) submit course artifacts for one introductory course. Course artifacts may have been submitted via a web-based upload tool, email, facsimile, or physical mail.

Submitters from Pharmacy Technician programs were also asked to submit artifacts for one concluding course, and the concluding course response rates for pharmacy technician were sufficient to allow a review of these artifacts.

Introductory courses

Introductory courses differed across programs within an occupation, and across occupations, in terms of standardization and sequencing. As such, “entry-level” courses could embody one or more of numerous definitions, including (1) those that occurred lowest in the course sequence for a program, regardless of course title; (2) those that were core “Introduction to…” or “Foundations of…” courses that occurred across the majority of programs, and (3) those that were identified by instructors as being most representative of the mathematics and reading expectations for entry-level students in the program.

Because the study focuses on identifying the mathematics and reading skills expected upon entry into introductory-level courses in the job-training programs for each occupation, courses were selected for inclusion using the third definition.

Concluding courses

Concluding, or exit-level, courses also differed in level of standardization, and multiple options for identifying such courses also exist. For consistency, the same approach was used to identify the exit-level courses for inclusion in the study: instructors were asked to identify those courses that best represent the mathematics and reading knowledge and skills that students are expected to know upon program completion.

Course artifact sets

For each training program, a set of course materials was collected for introductory courses and a set for concluding courses. The following types of artifacts were submitted and assembled into a course packet (with only one of each type of artifact required):

1. Course syllabus
2. Textbook title(s) (with author and ISBN)
3. Textbook table of contents (instructor copied and uploaded or EPIC downloaded from publisher website)
4. Course exam (one or more), preferably the mid-term or earlier for introductory courses and the final exam for concluding courses
5. Text-based assignment (one or more), with corresponding passage, that best illustrates mathematics and reading KSAs needed by students—one or more for introductory courses and one or more for concluding courses
6. Stand-alone assignment (one or more) such as a lab, worksheet, problem sheet, essay, or group project that best represents mathematics and reading KSAs needed for students—one or more for introductory courses and one or more for concluding courses

Instructors representing institutions that offered more than one program within an occupational area were asked to complete a submission for one program and to complete submissions for additional degree programs if selected courses were different than those already submitted.

PHASE 2

Review of Course Artifacts

Phase 2 includes reviewing materials, referred to as "course artifacts," which were collected. Two Review Teams were recruited, one for mathematics and one for reading. Review teams were formed to include two faculty members from college mathematics departments who are experienced in the review of course artifacts and one job training faculty person who had served as a panelist in the NAEP judgmental standard setting studies. Each review team was to determine the knowledge, skills, and abilities (KSAs) that students need to have in entry-level courses in the job training program, i.e. prerequisites, and those that are taught in the program as new content.

The teams of content experts were trained to consistently and reliably apply a coding scheme to the course artifacts to identify prerequisite and taught content for each of the occupational training programs. These teams of three first reviewed the course artifacts independently, including a course syllabus, textbook table of contents (and often the actual textbook), course assignments, and an examination. The artifacts were coded for each NAEP grade 12 framework objective according to whether the artifacts provided evidence that the objective was a part of the training program curriculum. For this step, the knowledge, skills, and abilities described in the framework objective were coded as either "not applicable" to this course, "prerequisite" for this course, or "new content" taught in this course. Coding categories were constructed to record the evaluation of importance of the prerequisite objectives:

- Minimally important: Although a prerequisite, possessing this KSA will make minimal difference to student performance in this course.
- A little important: This KSA is a prerequisite, and if possessed, the student is likely to learn more and have higher performance in the course.
- Important: Without this KSA, students will struggle with the course material.
- Very important: Without this KSA students are not prepared for, and will be *unlikely* to complete, this course.

The review teams then met in person in early October 2012 to discuss each objective for each training program for which there was not agreement in the independent codings among the team of 3 members. The goal was to reach agreement on the coding of artifacts, but consensus was not a requirement. The results of this meeting were then compiled and presented as content maps to present to teams of NAEP content experts.

Once the Review Teams' review of course materials was complete, EPIC staff aggregated the individual ratings for each course within each program to summarize the mathematics and reading KSAs prerequisite to and taught in introductory-level courses and that students are expected to have attained at program completion. Aggregated responses were displayed in overall content maps (documents that summarize NAEP assessment framework content in a table format) describing the relationship between frameworks and prerequisite KSAs for each occupation. In addition to tabular data displays, the data was displayed using color shading, as well as summary statistics, to show the extent of overlap in content between standards and programs. Content maps, grouped by key characteristics, were also created for programs to show the impact of key program characteristics that impacted findings. EPIC staff reviewed the content maps to identify similarities and differences across program types within occupations and noted the differences in findings due to program characteristics. Final results were provided both overall and by key program characteristics. EPIC staff also computed descriptive statistics to summarize the Review Teams' demand ratings overall (by occupation) and by program type, in case program characteristics had an impact on the demand of occupational courses.

Review of Knowledge, Skills, and Abilities Required for Training Courses
Two NAEP Expert Teams, one team for mathematics and one for reading, each consisting of three experts, reviewed the prerequisite and taught KSAs (as identified by the Review Teams) in the context of NAEP. They were charged with describing the relationships between the prerequisite content and both the BPDs and the content on the 2009 NAEP, evaluating the results of the Review Team analyses to describe KSAs assessed by NAEP that are not included in the job-training programs and KSAs included in the job-training programs that are not part of the NAEP frameworks or assessments.

Comparison to BPDs
Using the Review Teams' determination of KSA requirements and course artifacts, the NAEP Expert Teams were tasked with synthesizing and describing the relationship between the content that is prerequisite to and taught in occupational programs and the content described in the BPDs for that program.

Comparison to NAEP items
Each NAEP Expert Team was also tasked with comparing KSAs identified for each program's introductory courses (drawing upon the content maps and BPD comparisons) to the NAEP item pools. Starting with a set of items near the cut scores identified in the JSS studies, they judged the correspondence between the course prerequisite KSAs and the KSAs needed to correctly respond to items with a .67 probability. They were asked to identify the items in the range of the cut score plus one standard deviation that are prerequisite to or required in the courses. They were also asked to examine items below the cut score and above the range in the first analysis to determine if the KSAs represented in the curricular requirements were largely above or below this range.

The total number of Introductory Course Packets reviewed was:

- Mathematics (Introductory):
  - Computer Support Specialist—10
  - HVAC—18
  - Pharmacy Technician—22

- Licensed Practical Nurse—14

- Reading (Introductory):
    - Computer Support Specialist—11
    - HVAC—14
    - Pharmacy Technician—22
    - Licensed Practical Nurse—15

The total number of Concluding Course Packets reviewed was:

- Mathematics (Concluding):
    - Pharmacy Technician— 17

- Reading (Concluding):
    - Pharmacy Technician— 19

Of the 180 programs represented in the JSS studies, 85 submitted entry level course materials related to mathematics content and 83 submitted materials related to reading. A total of 107 institutions participated in the Job Training Program Course (JTPC) Study. A total of 162 course packets were complete and reviewed for the study: 126 for introductory courses and 36 for concluding courses.

In addition to the course artifacts, per se, the Governing Board requested that information be collected to profile the institutions and programs in the study. From these institutions, the following information was collected.

Institutional Characteristics:

1. Institution name
2. Data year
3. Institution size
4. Geographic region, including the state names
5. Level of institution (e.g. four-year, two-year, less than two-year)
6. Control of institution (e.g. public, private not-for-profit, private for-profit)
7. Degree granting status
8. Degree of urbanization (e.g. City:Large, Suburb:Small, Rural:Fringe)
9. Campus setting (e.g. city, suburb, town, rural)
10. Admission policy (open/not)

Admission requirements:

1. Diploma or GED requirements
2. SAT or ACT requirements, including minimum scores accepted
3. Reading course placement exam requirements, and minimum score accepted
4. Mathematics course placement exam requirements, and minimum score requirements
5. Other required placement exam requirements, and minimum score requirements
6. Other school admission requirements (e.g. personal essay, interview).

Degree and Program Information:

1. Accreditation status
2. Accrediting organization
3. Minimum number of credits or hours required for program completion

PHASE 3
Analysis and Reporting

All project activities are complete, and analysis and reporting are currently underway. As part of a special analysis, the NAEP mathematics expert team met over the weekend of February 1-3, 2013, and the reading team will be convened February 22-24, 2013 in order to review content maps (documents that summarize NAEP assessment framework content in a table format). These content maps showed the objectives coded according to evidence of applicability, evaluation of evidence as prerequisite or new content, and ratings of importance of the knowledge, skills, and abilities included in the objectives to the success of the student in the course.

In this special analysis, the NAEP experts are also tasked with developing descriptions of the knowledge, skills, and abilities that students need to demonstrate academic preparedness for entry level courses in the job training programs representing the five occupational areas in this research. The descriptions will be evidence-based statements of prerequisite knowledge and skills associated with the job training programs. The descriptions can then be compared to the descriptions that were developed by the JSS panelists to describe the minimal academic requirements associated with borderline preparedness levels. In addition, the descriptions can be compared across the occupational areas, compared to the NAEP achievement levels descriptions for reading and mathematics, and compared to descriptions based on other assessments of preparedness.

Additional analyses of the mathematics data are underway, and the data must undergo a thorough quality control check. The final results for the study of mathematics and reading entry-level course requirements will be available by the end of March 2013.

Attachment C-3
Newly Released Technical Report

The Board's First Phase of Preparedness Research is essentially complete. To begin the reporting process, the Board discussed specific staff-developed reporting options at the May 2012 Board meeting. At the August 2012 meeting, the Board decided to release an online technical report that would describe:

- the research conducted,
- the main research findings, and
- plans for future research based on the 2013 NAEP.

At the November 2012 Board meeting, staff presented draft research summaries for the Board's review in order to develop this online technical report. On February 11, 2013, the Board released the online technical report, which includes the following documents. A press announcement is attached with additional details.

Content Alignment

- Assessment Content Comparison: Methodology for Alignment Studies
- Preliminary NAEP and SAT Content Comparison: Mathematics
- Preliminary NAEP and SAT Content Comparison: Reading
- NAEP and WorkKeys Content Comparison: Mathematics
- NAEP and WorkKeys Content Comparison: Reading
- NAEP and ACT Content Comparison: Reading and Mathematics
- NAEP and SAT Framework Comparison: Mathematics
- NAEP and SAT Content Comparison: Mathematics
- NAEP and SAT Framework Comparison: Reading
- NAEP and SAT Content Comparison: Reading
- NAEP and ACCUPLACER Framework Comparison: Mathematics
- NAEP and ACCUPLACER Content Comparison: Mathematics
- NAEP and ACCUPLACER Framework Comparison: Reading
- NAEP and ACCUPLACER Content Comparison: Reading

Statistical Relationship

- Statistical Linking of National Results from NAEP and SAT
- Longitudinal Statistical Relationships for Florida NAEP Examinees: First-Year College Performance Outcomes

Judgmental Standard Setting

- Identification of Exemplar Occupations: Report
- Identification of Exemplar Occupations: Appendix A
- Identification of Exemplar Occupations: Appendix B
- NAEP 2009 Preparedness Standard Setting: Process Report
- NAEP 2009 Preparedness Standard Setting: Technical Report
- Paper: A Study of "Irrelevant" Items: Impact on Bookmark Placement and Implications
- for College and Career Readiness
- Paper: Preparing Job Trainers to Describe Knowledge, Skills, and Abilities Measured in an Academic Assessment

- Paper Appendix: Preparing Job Trainers to Describe Knowledge, Skills, and Abilities Measured in an Academic Assessment
- Paper: The Standard for Minimal Academic Preparedness in Mathematics to Enter a
- Job Training Program
- Paper: The Standard for Minimal Academic Preparedness in Reading to Enter a Job
- Training Program

Survey

- Survey on Postsecondary Course Placement Assessments: Report

Benchmarking

- Benchmarking Study with Texas College Freshmen: Methodology Report
- Benchmarking Study with Texas College Freshmen: Project Feasibility Report
- Benchmarking Study with Texas College Freshmen: Appendix A
- Benchmarking Study with Texas College Freshmen: Appendix B
- Benchmarking Study with Texas College Freshmen: Appendix C

NEWS RELEASE
CONTACT: Stephaan Harris (202) 357-7504, stephaan.harris@ed.gov

New Report on 12th Grade Preparedness Research:
Key Findings from the Nation's Report Card

WASHINGTON – (February 15, 2013) – The National Assessment Governing Board today released a web-based report on more than 30 research studies that examine whether the National Assessment of Educational Progress (NAEP) can be used to report on 12th graders' academic preparedness for college and job training. The website provides ready access to key findings from the research studies and links to the complete research reports.

"With the research reported today and other studies in the pipeline, we believe NAEP can make important contributions to the national conversation about the academic preparation of high school graduates for college and job training," said Dr. David Driscoll, Governing Board Chair.

NAEP is uniquely positioned to shed light on these issues as the nation's only representative-sample assessment of 12th grade students. The research reported today, and the Board's ongoing studies, can inform and help advance what we know about measuring the preparedness of high school seniors for postsecondary endeavors. At its core, the Board's research program addresses a series of questions related to the valid reporting of NAEP in terms of 12th grade academic preparedness.

This first wave of the Board's empirical studies provides important responses to these questions. Key findings are:

- Overall, the studies **found similar content in NAEP and the college admissions examinations SAT and ACT**, and somewhat less with the ACCUPLACER, a college placement exam.

- The **content comparison studies between NAEP and WorkKeys**, an exam used to assess job-related skills, found significant differences in focus and rigor. This difference suggests the need for further research to explore separately student preparedness for college and job training.

-more-

- In the **studies linking NAEP to the SAT**, the correlation was very strong between the two exams in mathematics and supports the use of a concordance in relating the two. The correlation between NAEP reading and the SAT was moderate and supports the use of other statistical approaches in relating the two exams.

- **Longitudinal analysis of Florida data** confirms findings of the national linking study of NAEP and the SAT, and generally indicates that the region around Proficient could be a reasonable benchmark for college academic preparedness. Florida's K-20 database was particularly informative in this initial phase of the Board's research.

"The achievement of this country's high school seniors is important for our economy and the future international competitiveness of the United States," said Cornelia Orr, the Board's Executive Director. "Taken as a whole, the Governing Board's ongoing research can help advance education reform efforts across the nation aimed at ensuring students leave high school well-prepared for postsecondary endeavors."

For more than a decade, the Board has been working to strengthen 12th grade NAEP. Efforts have included appointment of a blue-ribbon commission and changes in the NAEP 12th grade reading and math assessments. In addition the Board appointed an expert panel, which proposed a program of research studies. Subsequently the Board adopted a working definition of academic preparedness as the minimal reading and math achievement needed to qualify for credit-bearing college courses or job training programs without remediation. The research released today is intended to examine and provide support for the valid use of NAEP to report on the achievement of students in relation this definition.

The Board is using these newly-released research findings to inform ongoing preparedness studies based on results from the 2013 NAEP 12th grade assessments in reading and mathematics.

# # #

The National Assessment of Educational Progress, also referred to as The Nation's Report Card, is the only continuing, nationally representative measure of student achievement at grades 4, 8, and 12. *It has served as a national yardstick of student achievement since 1969, informing the public about what American students know and can do in various subject areas and comparing achievement between states, large urban districts, and various student demographic groups.*

The National Assessment Governing Board *is an independent, bipartisan board whose members include governors, state legislators, local and state school officials, educators, business representatives and members of the general public. Congress created the 26-member Governing Board in 1988 to oversee and set policy for NAEP.*

The National Assessment of Educational Progress (NAEP) is a congressionally authorized project sponsored by the U.S. Department of Education. The ***National Center for Education Statistics****, within the Institute of Education Sciences, administers NAEP. The Commissioner of Education Statistics is responsible by law for carrying out the NAEP project.*

Attachment C-4
Phase 2 Academic Preparedness Research Plans

Continued research plans call for NAEP-SAT, NAEP-ACT, and NAEP-EXPLORE statistical linking studies, more research partnerships with states, analysis of course content prerequisites for job training programs and freshman college courses, and efforts to partner with experts in military occupational training. A summary of each proposed research study follows. At the November 2012 Board meeting, COSDAM began discussion on these research plans.

National and State Statistical Linking Studies with the SAT and with the ACT

In 2013, the Governing Board will partner again with the College Board, as it did in 2009, to conduct a statistical linking study at the national level between NAEP and the SAT in reading and mathematics. Through a procedure that protects student confidentiality, the SAT 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 is planned in partnership with ACT, Inc.

In addition, the state-level studies, begun in 2009 with Florida, will be expanded in 2013. Again using a procedure that protects student confidentiality, the postsecondary activities of NAEP 12th grade test takers in the state samples in partner states will be followed for up to five years using the state longitudinal data bases. Five states will be partners in these studies: Florida, Illinois, Massachusetts, Michigan, and Tennessee. Others will be considered as time for completing the planning process and executing formal data sharing agreements permits. These studies will examine the relationship between 12th grade NAEP scores and GPA, placement into remedial versus credit-bearing courses, and scores on admissions and placement tests.

March 2013 Update: Illinois was added as an additional state partner.

Statistical Linking of Grade 8 NAEP and 8th Grade 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 two states, KY and TN. The objective is to determine the feasibility of identifying the point on the NAEP scales that indicate students are "on track" for being academically prepared for college and job training by 12th grade. As a foundation for the linking study, content alignment studies between 8th grade NAEP reading and mathematics and 8th grade EXPLORE would also be conducted.

March 2013 Update: No updates at this time.

Evaluation of NAEP Frameworks and Item Pools

The Governing Board is conducting a procurement (1) to design a comprehensive and multi-method evaluation of the grade 12 NAEP frameworks and item pools in both reading and mathematics as measures of academic preparedness for college and job training; and (2) based on the evaluation, to produce specific recommendations for changes that may be required to develop NAEP for 12th graders in reading and mathematics as valid measures of academic preparedness for placement in first year college courses without remediation in the subject areas and entry in job training programs that require at least three months of post-secondary training, but not a bachelor's degree in college.

Central to the validity of reporting preparedness of students on the NAEP grade 12 scale for reading and for mathematics is confirmation that the assessments actually measure the knowledge, skills, and abilities required for students to be academically prepared for college course work or for entry in job training programs. In this procurement, the Board seeks innovative, practicable design proposals for evaluations that will provide the foundation needed to make valid statements about academic preparedness.

March 2013 Update: The procurement process is ongoing.

Research Design Proposals for NAEP and Academic Preparedness for Job Training

Reporting on academic preparedness for college and job training is a challenging and important new direction for NAEP. Hence, the Governing Board is also conducting a procurement to seek proposals for research designs and studies that are feasible. The objective of the research is to advance the Governing Board’s efforts to identify locations on the 12th grade NAEP reading and mathematics scales that represent the knowledge and skills to qualify for training in various occupations.

March 2013 Update: The procurement process is on ongoing.

Attachment C-5
Overview of the Types of NAEP Preparedness Research

As part of the ongoing updates to COSDAM, this document includes an overview of each study type.

Content Alignment Studies

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.

A design document was developed by Dr. Norman Webb for the NAEP preparedness research alignment studies, and this design was implemented for the studies of the 2009 NAEP with the SAT and ACUPLACER in reading and mathematics. This design, with minor modifications, has also been used for the alignment of the 2009 NAEP with WorkKeys tests in these subject areas.

Content alignment studies for the first phase of the Board's Program of Preparedness Research have been completed for NAEP in reading and in mathematics with WorkKeys, the SAT, and ACCUPLACER. In addition, a content alignment study was designed and conducted by ACT for the ACT and NAEP in reading and mathematics before the content alignment design document was developed.

Studies to Establish Statistical Relationships

Highest priority is generally placed on these studies. Currently, two main sets of studies have been conducted under this heading. One set addresses ***statistical linking*** of NAEP with other assessments, and the other set examines ***longitudinal data*** for NAEP examinees.

For statistical linking, there has been a study to relate SAT scores in reading and in mathematics to the national sample of NAEP scores for grade 12. The objective was to provide a statistical linking of SAT and NAEP scores for all students in the 2009 grade 12 NAEP who had taken the SAT by June 2009. ETS staff reported that the match rate of approximately 33% of NAEP scores to SAT scores compares favorably to the national SAT participation rate of approximately 36% of public school students. The final sample used for linking the NAEP reading and SAT critical reading included approximately 16,200 students. For NAEP and SAT mathematics, the linking sample included approximately 15,300 students.

For longitudinal data, a series of analyses were conducted to examine statistical relationships for Florida's NAEP examinees. NAEP's 2009 state-representative sample of Florida 12th graders was used to match NAEP scores for reading and mathematics to student scores on several tests collected by the Florida Department of Education (FLDOE). The data sharing agreement with FLDOE provides access to scores for the SAT, ACCUPLACER, and WorkKeys. Additionally, ACT, Inc. has given permission to the Florida Department of Education to share ACT scores with the Governing Board for purposes of conducting the grade 12 preparedness research. We also plan to obtain employment data and salary data for Florida examinees, but access to those data was not included under the current data sharing agreement. A plan to allow for electronic transfer of data was developed to keep secure the identity of students, consistent with the NAEP legislation, FLDOE requirements, and requirements of each assessment program.

Records for roughly half of the Florida grade 12 NAEP examinees in 2009 could be matched to an ACT score and half to an SAT score. This match rate is consistent with other data for Florida students. The match of WorkKeys scores to the total 2009 state NAEP sample of 12th graders was only about 6%. FLDOE reported that around 89,300 Florida 12th graders were enrolled in vocational-technical programs in school year 2008-09. The match of WorkKeys examinees to NAEP examinees was not sufficient to warrant additional analyses for the 2009 cycle. The state of Florida has only recently implemented the testing of high school students in vocational programs with the WorkKeys exam, and we anticipate that the number of examinees will increase in subsequent years.

Judgmental Standard Setting Studies

A series of judgmental standard setting studies was planned to produce preparedness reference points on the NAEP scale for entry into job training programs and for placement in college credit-bearing courses. Within this category of studies, the Technical Panel for 12th Grade Preparedness Research placed highest priority on the judgmental studies related to preparedness for job training programs in 5-7 exemplar jobs. This priority is largely related to the paucity of national data available for statistical studies in these areas. The Governing Board has not assumed that academic preparedness for college and for job training are the same. Rather, our studies are aimed at determining the level of performance on NAEP that represents the reading and mathematics knowledge and skills needed to qualify for job training programs for each of the occupations included in our research studies and for placement in credit-bearing college courses that fulfill general education requirements for a bachelor's degree.

In order to maximize the standardization of judgmental standard setting (JSS) studies within and across post-secondary areas, a design document was developed to specify the number of panelists, the eligibility criteria for panelists, the procedures for drafting and finalizing borderline performance descriptions, the methodology to be implemented, feedback to be provided, key aspects to be evaluated, and reports to be produced. The methodology and basic procedures specified for the design of these studies were those implemented for the achievement levels-setting process for the 2006 grade 12 economics NAEP and for the 2009 science NAEP for grades 4, 8, and 12.

The five exemplar jobs approved by COSDAM for inclusion in these studies are as follows:

1. automotive master technicians
2. computer support specialists
3. heating, ventilation, and air conditioning technicians

4. licensed practical nurses
5. pharmacy technicians

A pair of replicate panels with 10 panelists each was convened for each subject and post-secondary area for a total of 24 operational panels.

Higher Education Survey

A survey of two-year and four-year post-secondary institutions was conducted in Fall 2011 to gather information regarding (1) the placement tests used and (2) the cut scores on those tests in reading and mathematics below which need was indicated for remedial/developmental courses in reading and mathematics, and at or above which placement in credit-bearing entry level courses was indicated. The sample of accredited postsecondary education institutions was nationally representative. A weighted response rate of 81% was achieved.

Benchmarking Studies

Benchmarking studies in the preparedness research context are studies in which NAEP is administered to groups of interest, e.g., college freshmen enrolled in credit-bearing college level courses that fulfill general education requirements for a four-year degree without the need for remediation. Determining the average NAEP performance of this group would then provide a “benchmark” score that can be considered as one of the reference points on the NAEP scale. A benchmarking study in combination with reference points from other studies in the Program of Preparedness Research can assist the Board in determining the areas of the NAEP scale that indicate preparedness. A benchmarking study of Texas college freshmen was planned, and it had the support of the Texas Commissioner of Higher Education and the cooperation of nine Texas higher education institutions. A small scale pilot study to evaluate the feasibility of the study design was implemented.

The Governing Board and the National Center for Education Statistics (NCES) collaborated on the implementation of this small scale pilot study, which was carried out by Westat, the NAEP sampling and administration contractor to NCES. The data collection phase for the pilot ended on October 15, 2010. Of the eligible sample of 1,234 students, 255 actually attended a NAEP session, for an overall response rate of 20.7 percent. As announced at the November 2010 meeting of COSDAM, NCES, Westat, and Governing Board staff met to discuss alternatives. Board staff decided that we will not proceed to the operational phase of this study due to low participation rates and the lack of feasible alternatives to increase participation.

No additional benchmarking studies are planned for the 2009 NAEP preparedness research.

OVERVIEW OF REFERENCED ASSESSMENTS

For additional background information, the following list presents a brief description of the assessments that the Technical Panel on 12th Grade Preparedness Research recommended for analysis in NAEP preparedness research. Many of these assessments are the primary focus of the proposed content alignment studies and statistical relationship 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.

- ACCUPLACER – ACCUPLACER is a computer adaptive test used for college course placement decisions in two-year and four-year institutions. It is produced by the College Board and includes assessments of sentence skills, reading comprehension, arithmetic, elementary algebra, college level math, and written essays.
- 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 WorkKeys –WorkKeys is a workplace focused set of tests that assess knowledge and skills in communication (business writing, listening, reading for information, writing) as well as problem solving (applied technology, applied mathematics, locating information, observation). There is also an interpersonal skills section of WorkKeys.
- 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 College Board has provided SAT score data to be used in research studies to establish a statistical relationship between the SAT and NAEP.

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

Status: Information and discussion

Objective: To provide background information and to discuss issues that should be included in the upcoming "issues paper," which will address achievement level setting for TEL in more detail.

Attachments: D-1 Abridged NAEP Framework for Technology and Engineering Literacy (National Assessment Governing Board, 2012)

D-2 Evidence-Centered Design for Certification and Licensure (Williamson, Mislevy, and Almond, 2004)

D-3 Standard Setting in Complex Performance Assessments: An Approach Aligned with Cognitive Diagnostic Models (Lissitz and Li, 2011)

Background

At the March 1, 2013 meeting, the Committee will have an opportunity to begin discussion on setting achievement levels for the 2014 NAEP TEL assessment. This discussion will support procurement and project planning for developing recommended achievement levels for TEL. Contract award is scheduled for late 2013.

As presented in the November 30, 2012 closed TEL briefing to the Board, this innovative assessment has many unique design features that will affect achievement level setting (ALS). Several of these features include computer-based delivery, cross-curricular content, combination of complex scenarios and discrete test items, and evidence-centered design, to name a few. Taken as a whole, these features will make TEL achievement level setting (ALS) more challenging than previous level-setting projects.

To inform COSDAM about the measurement and policy issues involved in the TEL ALS, Board staff plan to work with a consultant to develop an "issues paper" scheduled for discussion at the May 2013 COSDAM meeting. The outcome of both the March 2013 and May 2013 COSDAM discussions will inform the statement of work for the TEL level-setting procurement.

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
Late 2013	ALS procurement and contract award	Board staff, COSDAM
Early 2014	Operational administration – national sample, grade 8	NCES, NAEP contractors
2014 - 2015	Final phase of ALS process and Board action on TEL	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 Abridged TEL Framework in Attachment D-1; complete documents are at 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 (see Attachment D-2). 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" may also contribute to the scoring of student performance.

TEL Reporting

Based on the ECD approach, TEL reporting will be expanded beyond the traditional NAEP scores. It is expected that data from the complex performance tasks and discrete items will be reported in a number of ways:

- A composite 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.

Setting Standards on Complex Performance Assessments

Recent articles in the measurement literature state that traditional standard setting methods are not necessarily appropriate for assessments composed of complex performance tasks (see Attachment D-3). Assessments such as TEL, which consist of both performance tasks and discrete items, present unique challenges for setting standards. As noted earlier, TEL reporting will incorporate an extended array of response data (on constructed response tasks and discrete items) and observable data (from interactive computer tasks) as specified in the ECD model.

In its most recent ALS work on the computer-based 2011 NAEP Writing assessment, the Board contracted with Measured Progress to conduct standard setting based on a Body of Work methodology. The process included enhancements and research studies to enable the ALS process to be conducted entirely via computer. However, the NAEP Writing assessment consisted solely of student essays in response to writing prompts. In the case of Writing, all student responses were in the same format. The TEL assessment design is far more complex.

The article in Attachment D-3 is included to provide a general understanding of the challenges COSDAM and the Board will encounter in setting achievement levels for TEL. Committee members may wish to skim this article for purposes of the March 1st discussion. The authors present a brief overview of traditional standard setting methods and limitations in applying those methods to complex performance assessments. In addition, the article outlines measurement challenges in developing a valid and reliable standard setting procedure for use on ECD-based tests. Note that the specific standard setting procedure discussed (Cognitive Analytical Approach or CAA) is not directly applicable to the TEL context.

Attachment D

Potential Discussion Questions for COSDAM

- Given the emerging field of setting achievement levels on ECD-based complex performance assessments, what additional background materials are needed to inform the COSDAM/Board decision on the most appropriate method for ALS on the TEL assessment?

- What are some specific challenges the TEL ALS issues paper should address related to setting standards on complex performance tasks?

- Is there any particular body of research in this area that the issues paper should draw upon?

- To what extent should research studies be built into the TEL ALS project?

- Are there examples of ALS exploratory work the Board should undertake in collaboration with the two assessment consortia? (The two assessment consortia are also planning to set performance standards on complex ECD-based tests.)

Attachment D-1

2014 Abridged
Technology and Engineering Literacy Framework

FOR THE 2014 NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS

NATIONAL ASSESSMENT GOVERNING BOARD

Introduction

We live in a world that is, to a large extent, shaped by technology: The computers and smart phones we use, the cars and planes we travel in, the homes and offices we inhabit; our food, clothes, entertainment, and medical care—all are created and driven by technology. Technology is also at the root of critical challenges we face as a society, such as the quest to link experts throughout the world, the search for sustainable energy, the ability to deal with global pandemics, and the development of environmentally friendly agriculture to feed a growing world population.

Until now, however, technology has not been a focus of instruction and assessment in our educational system, particularly at the elementary and secondary levels. Because of the growing importance of technology and engineering in the educational landscape, and to support America's ability to contribute to and compete in a global economy, the National Assessment Governing Board initiated development of the first national assessment in Technology and Engineering Literacy. Relating to national efforts in science, technology, engineering, and mathematics (STEM) fields, the NAEP Technology and Engineering Literacy assessment measures the "T" and "E" in STEM, augmenting longstanding NAEP assessments in science and mathematics.

NAEP Technology and Engineering Literacy (TEL) Assessment

The National Assessment of Educational Progress (NAEP), otherwise known as The Nation's Report Card, informs the public about the academic achievement of elementary and secondary students in the United States. Report cards communicate the findings of NAEP, a continuing and nationally representative measure of achievement in various subjects over time. For more than 35 years, NAEP has assessed achievement by testing samples of students most often in the fourth, eighth, and 12th grades. The results have become an important source of information on what U.S. students know and are able to do in a range of subject areas.

To create the new assessment, the National Assessment Governing Board sought a framework of technological literacy knowledge and skills that identifies the understandings and applications of technology principles that are important for all students. The framework defines "literacy" as the level of knowledge and competencies needed by all students and citizens. More than testing students for their ability to "do" engineering or produce technology, then, the assessment is designed to gauge how well students can apply their understanding of technology principles to real-life situations. At grade 4, for example, all students are expected to identify types of technologies in their world, design and test a simple model, explain how technologies can result in positive and negative effects, and use common technologies to achieve goals in school and in everyday life. By grade 12, students are expected to select and use a variety of tools and media to conduct research, evaluate how well a solution meets specified criteria, and develop a plan to address a complex global issue. To learn more, see a video clip ("Ecosystems") in the interactive framework of a sample scenario for grade 8 showing a student investigation of how organisms in an ecosystem are affected by a pollutant.

Technological literacy at grades 4, 8, and 12 is a pathway promoting further study and occupational pursuits. The Governing Board assembled a broad array of individuals and organizations to create a test of students' abilities to grasp and apply technology principles. The resulting framework is the culmination of a long, complex process that drew on the contributions of thousands of individuals and organizations including technology experts, engineers, teachers, researchers, business leaders, testing experts, and policymakers.

The 2014 NAEP Technology and Engineering Literacy Assessment will provide important results and information that can be used to determine whether our nation's students have the essential knowledge and skills needed in the technology and engineering areas. Policymakers, educators, and the public can use data from the initial assessments as tools for monitoring certain aspects of student achievement in technology and engineering literacy over time.

Definitions of Technology, Engineering, and Technology and Engineering Literacy

Any assessment of students' technology and engineering literacy must start with a clear idea of exactly what technology and engineering literacy means. That in turn requires clear definitions of technology and engineering.

Technology is any modification of the natural world done to fulfill human needs or desires.

This definition sees technology as encompassing the entire human-made world, from paper to the Internet. Technology also includes the entire infrastructure needed to design, manufacture, operate, and repair technological artifacts, from corporate headquarters and engineering schools to manufacturing plants and media outlets.

Engineering is a systematic and often iterative approach to designing objects, processes, and systems to meet human needs and wants.

This framework defines technology and engineering literacy in a broad fashion:

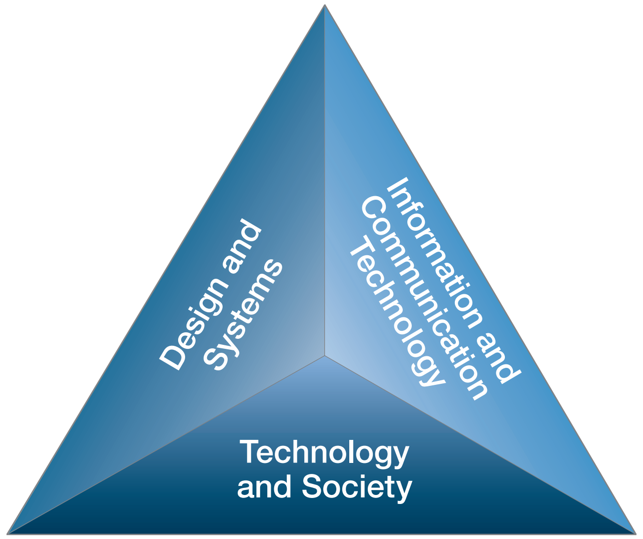
Technology and engineering literacy is the capacity to use, understand, and evaluate technology as well as to understand technological principles and strategies needed to develop solutions and achieve goals.

Thus—as with scientific, mathematical, and language literacy—technology and engineering literacy involves the mastery of a set of tools needed to participate intelligently and thoughtfully in society.

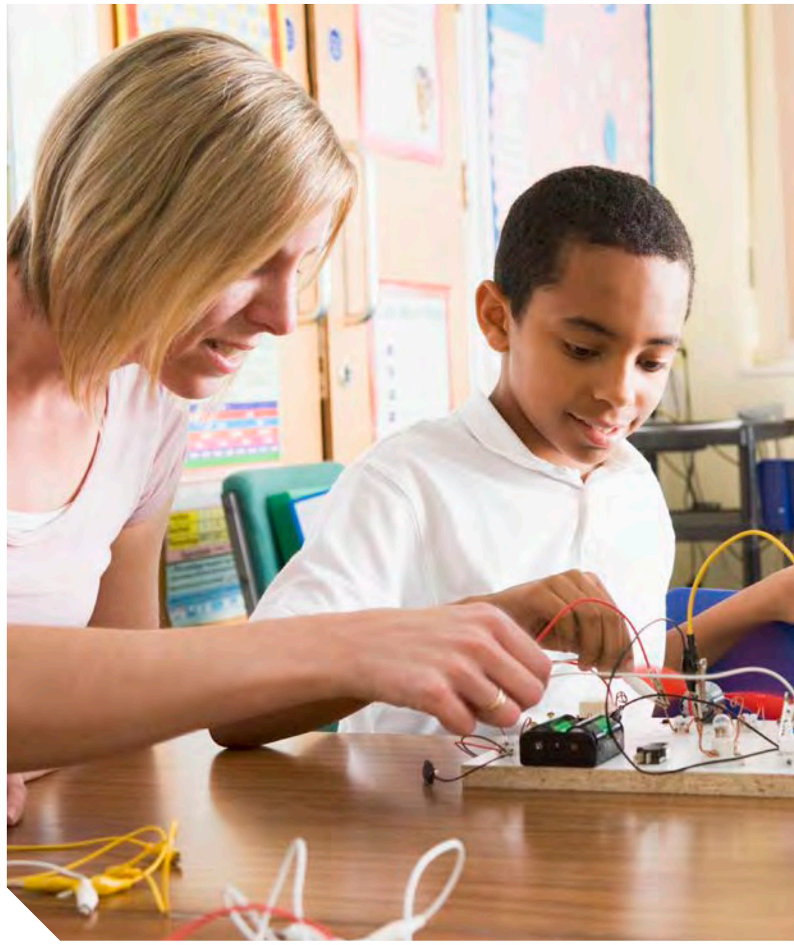
Three Areas of Technology and Engineering Literacy

Recognizing that it is not possible to assess every aspect of technology and engineering literacy, the TEL assessment framework targets the nature, processes, and uses of technology and engineering that are essential for 21st century citizens.

The assessment objectives are organized into three major areas: *Technology and Society; Design and Systems; and Information and Communication Technology (ICT).* Each broad category is further broken down into discrete areas to be assessed.



The interconnected relationship among these three major assessment areas can be illustrated as a three-sided pyramid in which each side supports the other two. For example, in order to address an issue related to technology and society, such as clean water, energy needs, or information research, a person who is literate in technology and engineering must understand technological systems and the engineering design process and be able to use various information and communication technologies to research the problem and develop possible solutions.



Area 1. Technology and Society

deals with the effects that technology has on society and on the natural world and with the sorts of ethical questions that arise from those effects.

The four sub-areas in which students are assessed include:

A. *Interaction of Technology and Humans* concerns the ways in which society drives the improvement and creation of new technologies and how technologies serve society as well as change it. **Fourth-graders** are expected to know that people's needs and desires determine which technologies are developed or improved. For example, cell phones were invented, produced, and sold because people found it useful to be able to communicate with others wherever they were. **Eighth-graders** are expected to understand how technologies and societies co-evolve over significant periods of time. For example, the need to move goods and people across distances prompted the development of a long series of transportation systems from horses and wagons to cars and airplanes. **By 12th grade**, students are expected to realize that the interplay between culture and technology is dynamic, with some changes happening slowly and others very rapidly. They should be able to use various principles of technology design—such as the concepts of trade-offs and unintended consequences—to analyze complex issues at the interface of technology and society and to consider the implications of alternative solutions.

B. *Effects of Technology on the Natural World* is about the positive and negative ways that technologies affect the natural world. **Fourth-graders** are expected to know that sometimes technology can cause environmental harm. For example, litter from food packages and plastic forks and spoons discarded on city streets can travel through storm drains to rivers and oceans where they can harm or kill wildlife. **Eighth-graders** are expected to recognize that technology and engineering decisions often involve weighing competing priorities, so that there are no perfect solutions. For example, dams built to control floods and produce electricity have left wilderness areas under water and affected the ability of certain fish to spawn. **By 12th grade**, students should have had a variety of experiences in which technologies were used to reduce the environmental impacts of other technologies, such as the use of environmental monitoring equipment.

C. *Effects of Technology on the World of Information and Knowledge* focuses on the rapidly expanding and changing ways that information and communication technologies enable data to be stored, organized, and accessed and on how those changes bring about benefits and challenges for society. **Fourth-graders** should know that information technology provides access to vast amounts of information, that it can also be used to modify and display data, and that communication technologies make it possible to communicate across great distances using writing, voice, and images. **Eighth-graders** should be aware of the rapid progress in development of ICT, should know how information technologies can be used to analyze, display, and communicate data, and should be able to collaborate with other students to develop and modify a knowledge product. **By 12th grade**, students should have a full grasp of the types of data, expertise, and knowledge available online and should be aware of intelligent information technologies and the uses of simulation and modeling.

D. *Ethics, Equity, and Responsibility* concerns the profound effects that technologies have on people, how those effects can widen or narrow disparities, and the responsibility that people have for the societal consequences of their technological decisions. **Fourth-graders** should recognize that tools and machines can be helpful or harmful. For example, cars are very helpful for going from one place to another quickly, but their use can lead to accidents in which people are seriously injured. **Eighth-graders** should be able to recognize that the potential for misusing technologies always exists and that the possible consequences of such misuse must be taken into account when making decisions. **By 12th grade**, students should be able to take into account both intended and unintended consequences in making technological decisions.

Area 2. Design and Systems covers the nature of technology, the engineering design process by which technologies are developed, and basic principles of dealing with everyday technologies, including maintenance and troubleshooting.

The four sub-areas in which students are assessed include:

A. *Nature of Technology* offers a broad definition of technology as consisting of all the products, processes, and systems created by people to meet human needs and desires. **Fourth-graders** are expected to distinguish natural and human-made materials, to be familiar with simple tools, and to recognize the vast array of technologies around them. **Eighth-graders** should know how technologies are created through invention and innovation, should recognize that sometimes a technology developed for one purpose is later adapted to other purposes, and should understand that technologies are constrained by natural laws. **By 12th grade,** students should have an in-depth understanding of the ways in which technology coevolves with science, mathematics, and other fields; should be able to apply the concept of trade-offs to resolve competing values; and should be able to identify the most important resources needed to carry out a task.

B. *Engineering Design* is a systematic approach to creating solutions to technological problems and finding ways to meet people's needs and desires. **Fourth-graders** should know that engineering design is a purposeful method of solving problems and achieving results. **Eighth-graders** should be able to carry out a full engineering design process to solve a problem of moderate difficulty. **By 12th grade,** students should be able to meet a complex challenge, weigh alternative solutions, and use the concept of trade-offs to balance competing values.

C. *Systems Thinking* is a way of thinking about devices and situations so as to better understand interactions among components, root causes of problems, and the consequences of various solutions. **Fourth-graders** should know that a system is a collection of interacting parts that make up a whole, that systems require energy, and that systems can be either living or non-living. **Eighth-graders** should be able to analyze a technological system in terms of goals, inputs, processes, outputs, feedback, and control, and they should be able to trace the life cycle of a product from raw materials to eventual disposal. **By 12th grade,** students should be aware that technological systems are the product of goal-directed designs and that the building blocks of any technology consist of systems that are embedded within larger technological, social, and environmental systems. They should also be aware that the stability of a system is influenced by all of its components, especially those in a feedback loop.

D. *Maintenance and Troubleshooting* is the set of methods used to prevent technological devices and systems from breaking down and to diagnose and fix them when they fail. **Fourth-graders** should know that it is important to care for tools and machines so they can be used when they are needed. Students should also know that if something does not work as expected, it is possible to find out what the problem is in order to decide if the item should be replaced or how to fix it. **Eighth-graders** should be familiar with the concept of maintenance and should understand that failure to maintain a device can lead to a malfunction. They should also be able to carry out troubleshooting, at least in simple situations. **By 12th grade,** students should know that many devices are designed to operate with high efficiency only if they are checked periodically and properly maintained. They should also have developed the capability to troubleshoot devices and systems, including those that they may have little experience with.

Area 3. Information and Communication Technology

includes computers and software learning tools, networking systems and protocols, hand-held digital devices, and other technologies for accessing, creating, and communicating information and for facilitating creative expression.

The five sub-areas in which students are assessed include:

A. *Construction and Exchange of Ideas and Solutions* concerns an essential set of skills needed for using ICT and media to communicate ideas and collaborate with others. **Fourth-graders** should understand what is expected from members working as part of a team and should realize that teams are better than individuals at solving many kinds of problems. **Eighth-graders** should know that communicating always involves understanding the audience—the people for whom the message is intended. They should also be able to use feedback from others, and provide constructive criticism. **By 12th grade**, students are expected to have developed a number of effective strategies for collaborating with others and improving their teamwork. They should be able to synthesize information from different sources and communicate with multiple audiences.

B. *Information Research* includes the capability to employ technologies and media to find, evaluate, analyze, organize, and synthesize information from different sources. **Fourth-graders** should be aware of a number of digital and network tools that can be used for finding information, and they should be able to use these tools to collect, organize, and display data in response to specific questions and to help solve problems. **Eighth-graders** should be aware of digital and network tools and be able to use them efficiently. They should be aware that some of the information they retrieve may be distorted, exaggerated, or otherwise misrepresented, and they should be able to identify cases where the information is suspect. **By 12th grade**, students should be able to use advanced search methods and select the best digital tools and resources for various purposes. They should also be able to evaluate information for timeliness and accuracy.

C. *Investigation of Problems* concerns the use of information and communication technology to define and solve problems in core school subjects and in practical situations. **Fourth-graders** should be able to use a variety of information and communication technologies to investigate a local or otherwise familiar issue and to generate, present, and advocate for possible solutions. **Eighth-graders** should be able to use digital tools to identify and research a global issue and to identify and compare different possible solutions. **By 12th grade**, students should be able to use digital tools to research global issues and to fully investigate the pros and cons of different approaches. They should be able to design and conduct complex investigations in various subject areas using a variety of digital tools to collect, analyze, and display information and be able to explain the rationale for the approaches they used in designing the investigation as well as the implications of the results.

D. *Acknowledgement of Ideas and Information* involves respect for the intellectual property of others and knowledge of how to credit others' contributions appropriately, paying special attention to the misuse of information enabled by rapid technological advances. **Fourth-graders** should understand that it is permissible to use others' ideas as long as appropriate credit is given. They should also know that copyrighted materials cannot be shared freely. **Eighth-graders** should be aware of general principles concerning the use of other people's ideas and know that these principles are the basis for such things as school rules and federal laws governing such use. They should know about the limits of fair use of verbatim quotes and how to cite sources. **By 12th grade**, students should understand the fundamental reasons for intellectual property laws and should know acceptable practices for citing sources when incorporating ideas, quotes, and images into their own work.

E. *Selection and Use of Digital Tools* includes both knowledge and skills for choosing appropriate tools and using a wide variety of electronic devices, including networked computing and communication technology and media. **Fourth-graders** should know that different digital tools have different purposes and they should also be able to use a variety of digital tools that are appropriate for their age level. **Eighth-graders** should be familiar with different types of digital tools and be able to move easily from one type of tool to another—for example, creating a document or image with one tool and then using a second tool to communicate the result to someone at a distant location. **By 12th grade**, students should be competent in the use of a broad variety of digital tools and be able to explain why some tools are more effective than others that were designed to serve the same purpose, based on the features of the individual tools.

Although these elements are central to the design of the NAEP Technology and Engineering Literacy Assessment, they are not sufficient to describe the kinds of reasoning to be expected from students, the context or subject matter that will be used to construct test items, or the overall shape of the entire assessment. The assessment targets and the sub-areas within each describing what students should be able to do foreshadow the cross-cutting practices—ways of thinking and reasoning—for which the TEL is designed.

Practices and Contexts for Technology and Engineering Literacy

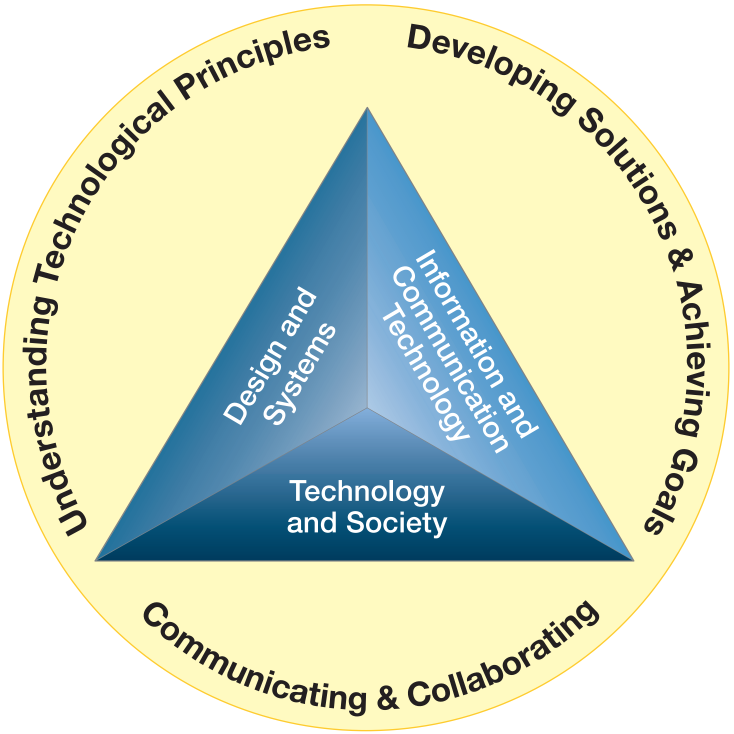
In all three areas of technology and engineering literacy, students are expected to be able to apply particular ways of thinking and reasoning when approaching a problem, and they are expected to do so in various contexts.

The practices can be grouped into three broad categories: *Understanding Technological Principles; Developing Solutions and Achieving Goals; and Communicating and Collaborating.*

Understanding Technological Principles focuses on students' knowledge and understanding of technology and their capability to think and reason with that knowledge.

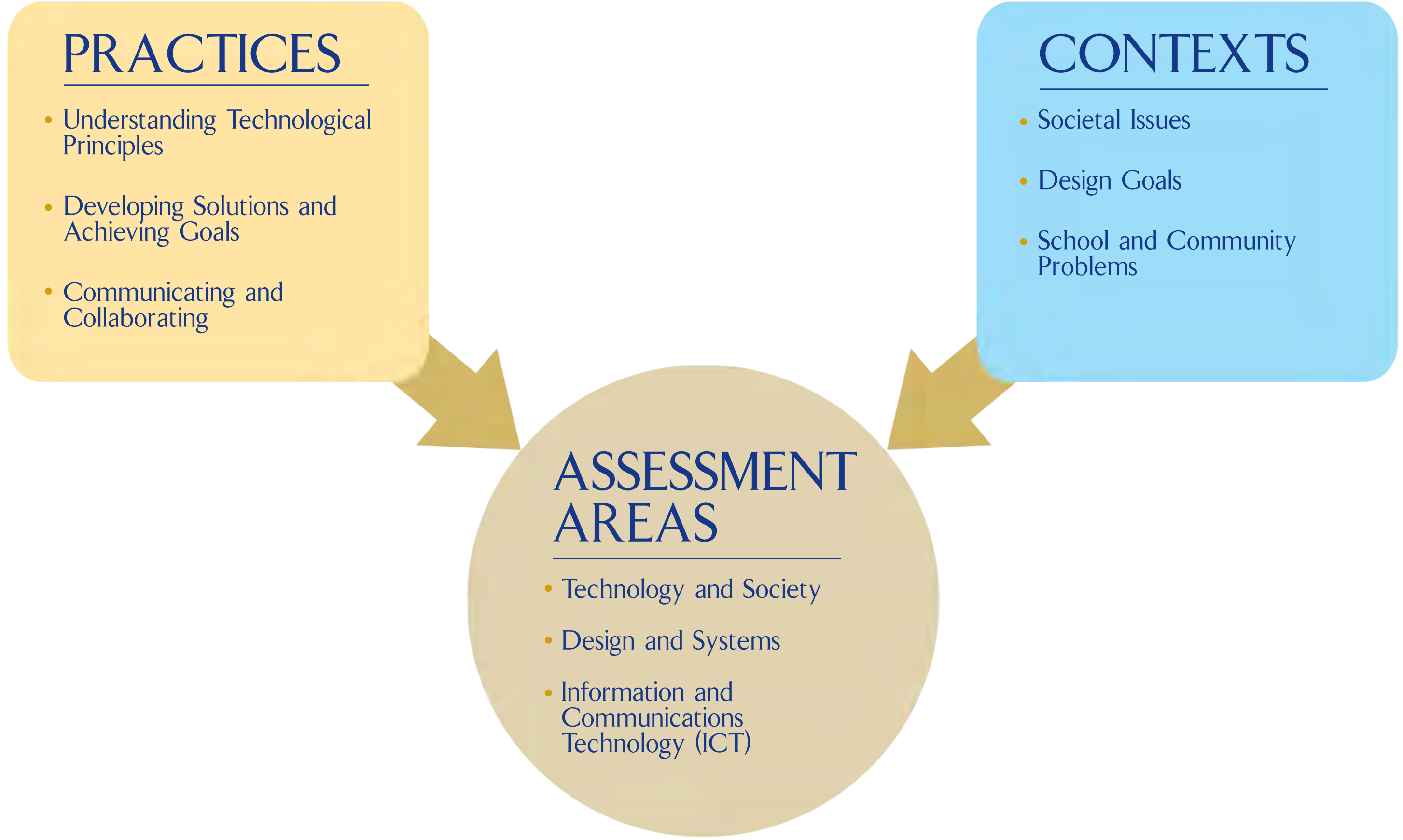
Developing Solutions and Achieving Goals refers to students' systematic application of technological knowledge, tools, and skills to address problems and achieve goals presented in societal, design, curriculum, and realistic contexts.

Communicating and Collaborating centers on students' capabilities to use contemporary technologies to communicate for a variety of purposes and in a variety of ways, working individually or in teams.


Understanding Technological Principles
Developing Solutions & Achieving Goals
Communicating & Collaborating
Design and Systems
Information and Communication Technology
Technology and Society

These practices are applied across all three major assessment areas. For example, communicating effectively and collaborating with others are necessary skills for understanding the effects of technology on the natural world, designing an engineering solution to a technological problem, and achieving a goal using information and communication technologies.

As crucial to the assessment as the practices are the **contexts**—the situations and types of problems in which assessment tasks and items will be set.



PRACTICES

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

CONTEXTS

- Societal Issues
- Design Goals
- School and Community Problems

ASSESSMENT AREAS

- Technology and Society
- Design and Systems
- Information and Communications Technology (ICT)

Elements of the NAEP Technology and Engineering Literacy Assessment

The practices expected of students are general, cross-cutting reasoning processes that students must use in order to show that they understand and can use their technological knowledge and skills. The contexts in which technology and engineering literacy tasks and items appear will include typical issues, problems, and goals that students might encounter in school or practical situations. Together, the assessment targets, practices, and contexts provide a structure for the generation of tasks and items.

Below are examples of the types of tasks and items that result when these three elements are combined. The table shows how the three practices—Understanding Technological Principles, Developing Solutions and Achieving Goals, and Communicating and Collaborating—can be used to classify the general types of thinking and reasoning intended by the assessment targets in the three major assessment areas of Technology and Society, Design and Systems, and Information and Communication Technology.

Classification of types of assessment targets in the three major assessment areas according to the practices for technology and engineering literacy

	Technology and Society	Design and Systems	Information and Communication Technology
Understanding Technological Principles	**Analyze** advantages and disadvantages of an existing technology **Explain** costs and benefits **Compare** effects of two technologies on individuals **Propose** solutions and alternatives **Predict** consequences of a technology **Select** among alternatives	**Describe** features of a system or process **Identify** examples of a system or process **Explain** the properties of different materials that determine which is suitable to use for a given application or product **Analyze** a need **Classify** the elements of a system	**Describe** features and functions of ICT tools **Explain** how parts of a whole interact **Analyze** and compare relevant features **Critique** a process or outcome **Evaluate** examples of effective resolution of opposing points of view **Justify** tool choice for a given purpose
Developing Solutions and Achieving Goals	**Select** appropriate technology to solve a societal problem **Develop** a plan to investigate an issue **Gather and Organize** data and information **Analyze and Compare** advantages and disadvantages of a proposed solution **Investigate** environmental and economic impacts of a proposed solution **Evaluate** trade-offs and impacts of a proposed solution	**Design and Build** a product using appropriate processes and materials **Develop** forecasting techniques **Construct and Test** a model or prototype **Produce** an alternative design or product **Evaluate** trade-offs **Determine** how to meet a need by choosing resources required to meet or satisfy that need **Plan** for durability **Troubleshoot** malfunctions	**Select and Use** appropriate tools to achieve a goal **Search** media and digital resources **Evaluate** credibility and solutions **Propose and Implement** strategies **Predict** outcomes of a proposed approach **Plan** research and presentations **Organize** data and information **Transform** from one representational form to another **Conduct** experiments using digital tools and simulations
Communicating and Collaborating	**Present** innovative, sustainable solutions **Represent** alternative analyses and solutions **Display** positive and negative consequences using data and media **Compose** a multimedia presentation **Produce** an accurate timeline of a technological development **Delegate** team assignments **Exchange** data and information with virtual peers and experts	**Display** design ideas using models and blueprints **Use** a variety of media and formats to communicate data, information, and ideas **Exhibit** design of a prototype **Represent** data in graphs, tables, and models **Organize, Monitor, and Evaluate** the effectiveness of design teams **Request** input from virtual experts and peers **Provide and Integrate** feedback	**Plan** delegation of tasks among team members **Provide and Integrate** feedback from virtual peers and experts to make changes in a presentation **Critique** presentations **Express** historical issues in a multimedia presentation **Argue** from an opposing point of view **Explain** to a specified audience how something works **Address** multiple audiences **Synthesize** data and points of view

Content and Design

To identify what students know and can do with regard to technology and engineering, the NAEP TEL framework calls for the assessment to be totally computer-based. In 2014 the NAEP TEL assessment will be conducted at grade 8 with a national sample of students in public and private schools. The assessment will include tasks and items sampled from the domain of technology and engineering literacy achievement identified by the intersection of the three major areas of technology and engineering literacy and the cross-cutting practices at grades 4, 8, and 12—grades that will participate in the TEL assessment in future years.

Allowing students to demonstrate the wide range of knowledge and skills detailed in the NAEP Technology and Engineering Literacy Assessment targets will require a departure from the typical assessment designs used in other NAEP content areas. Thus students will be asked to perform a variety of actions using a diverse set of tools in the process of solving problems and meeting goals within rich, complex scenarios that reflect realistic situations. Consequently, this assessment will rely primarily on scenario-based assessment sets that test students through their interaction with multimedia tasks that include conventional item types, such as selected response items, and also monitor student actions as they manipulate components of the systems and models that are presented as part of the task.

Because of their capability to replicate authentic situations examinees may encounter in their lives, scenarios have the potential to provide a level of authenticity other types of assessment tasks cannot provide. At the same time, the choice to use these complex tasks reduces the number of measures that can be included in any one test and causes many of the measures to be interdependent because they are related to the same scenario. To counteract this interdependency and ensure reliability, the NAEP assessment of technology and engineering literacy will also include sets of discrete items that produce independent measures.

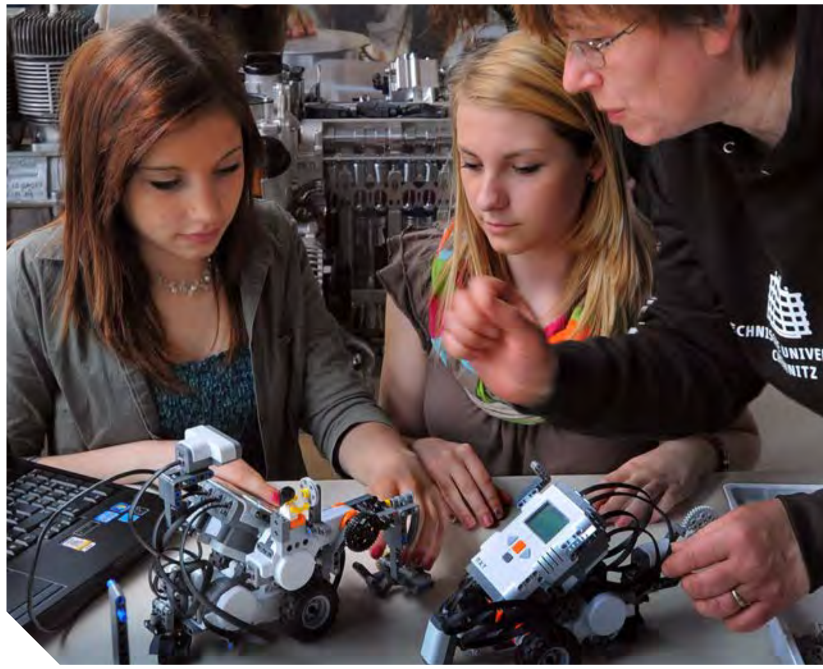
Scenario-Based Assessment Sets

There will be two types of scenario-based assessment sets, one long and one short. The long scenarios will take students approximately 25 minutes. The short scenarios will take students about 12 to 15 minutes to respond. The two types of scenarios have common characteristics, but they differ in the complexity of the scenario and the number of embedded assessment tasks and items to which a student is asked to respond.

A set of sample video clips demonstrates the types of interactivity and functionality of tools that students might be expected to use as they respond to short and long scenarios that will be developed for the Technology and Engineering Literacy Assessment.

Discrete Item Sets

Discrete item sets will include conventional selected response items and short constructed response items. The discrete item sets will comprise approximately 10-15 stand-alone items in either selected or constructed response format to be completed within a 25-minute block. Each discrete item would provide a stimulus that presents enough information to answer the particular question posed in the stem of the item. Items in discrete sets will be selected response items (e.g., multiple choice) or short constructed response items in which a student writes a text-based response.



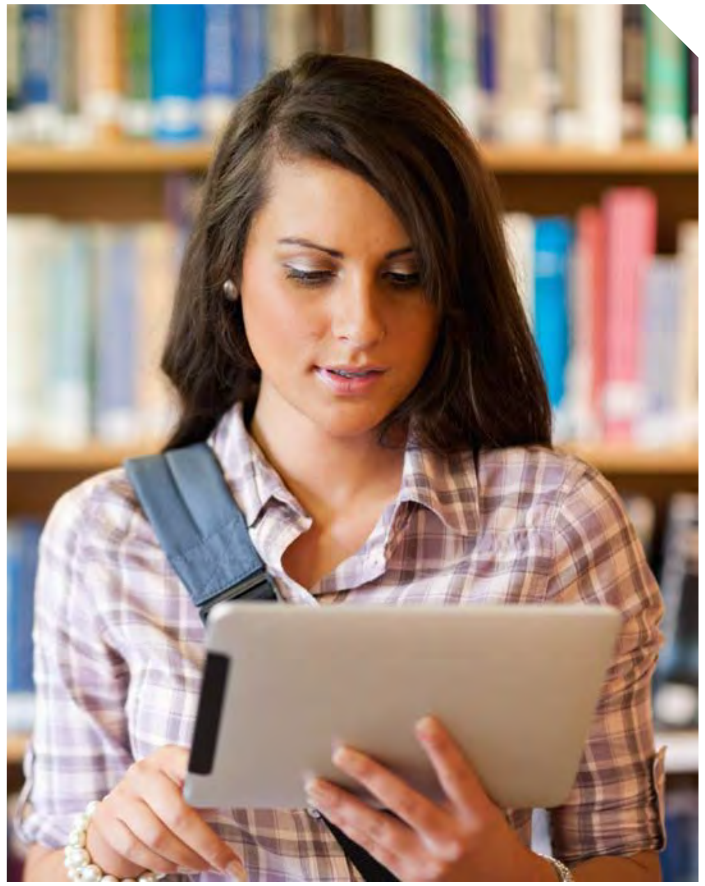
Background Variables

Background data on students, teachers, and schools are needed to fulfill the statutory requirement that NAEP include information, whenever feasible, for various subgroups of students at the national level including gender, race/ethnicity, eligibility for free or reduced-price lunch, English language learners, and students with disabilities. Therefore, students, teachers, and school administrators participating in NAEP are asked to respond to questionnaires designed to gather demographic information. Information is also gathered from non-NAEP sources, such as state, district, or school records. For the 2014 NAEP Technology and Engineering Literacy Assessment, only student and school information will be collected as many students will not have taken a separate course in technology and engineering literacy taught by a specific teacher.

In addition to demographic information, background questionnaires include questions about variables related to opportunities to learn and achievement in technology and engineering literacy. The variables are selected to be of topical interest, to be timely, and to be directly related to academic achievement and current trends and issues in technology and engineering literacy. Questions do not solicit information about personal topics or information irrelevant to the collection of data on technology and engineering literacy achievement.

Achievement Levels

The Governing Board uses student achievement levels of *Basic, Proficient,* and *Advanced* to report results of NAEP assessments. The achievement levels represent an informed judgment of "how good is good enough" in the various subjects that are assessed. Technology and Engineering Literacy achievement levels specific to the 2014 NAEP Technology and Engineering Literacy Framework will be developed to elaborate the generic policy definitions of *Basic, Proficient,* and *Advanced* achievement. Preliminary achievement level definitions have been developed for each of the three areas to be reported separately in the assessment and they will be used to guide item development and initial stages of standard setting for the 2014 NAEP Technology and Engineering Literacy Assessment.



The preliminary achievement level definitions will be revised when actual student responses have been collected and analyzed. The Governing Board will convene panels of experts to examine the preliminary achievement level definitions and to recommend final achievement level definitions for each grade level.

Conclusion

For generations students have been taught about technology and have been instructed in the use of various technological devices, but there has been no way to know exactly what students understand about technologies and their effective uses. The exploding growth in the world of technology led the Governing Board to sponsor the development of a framework for a National Assessment of Technology and Engineering Literacy. The Governing Board hopes that this TEL Framework will serve as a significant national measure of what students know and can do in technology and engineering, and support improvements in student achievement.

To view the complete Technology and Engineering Literacy Framework for the 2014 NAEP, or to view an interactive version of the framework, please visit http://nagb.org/publications/frameworks.htm or call us at 202.357.6938.

NATIONAL ASSESSMENT GOVERNING BOARD

The National Assessment Governing Board is an independent, bipartisan board whose members include governors, state legislators, local and state school officials, educators, business representatives, and members of the general public. Congress created the 26-member Governing Board in 1988 to set policy for the National Assessment of Educational Progress (NAEP).

NATIONAL ASSESSMENT GOVERNING BOARD

For more information on the National Assessment Governing Board, please visit www.nagb.org or call us at 202-357-6938.

EVIDENCE-CENTERED DESIGN FOR CERTIFICATION AND LICENSURE

David M. Williamson, Robert J. Mislevey, Russell G. Almond
Educational Testing Service

What is Evidence-Centered Design?

Evidence-Centered Design (ECD) (Almond, Steinberg, & Mislevy, 2002; Mislevy, Steinberg, & Almond, 2003) is a methodology applied at Educational Testing Service that emphasizes an evidentiary chain of reasoning for assessment design. This approach results in a more complete representation of the design rationale for an assessment, better targeting of the assessment for its intended purpose, and a more substantial basis for a construct-representation validity argument supporting use of the assessment. The approach encourages test developers to design with intent and provides several advantages:

Clarity of purpose – representation of assessment goals and the relevance of design decisions to those goals.

Interrelated design – modeling the interactions of design decisions and how changes in one aspect of design affect other design elements.

Evidentiary requirements – explication of what constitutes relevant evidence of ability and how such evidence bears on assessment-based decision-making.

Validity – a documented chain of reasoning and rationale underlying design decisions and their relevance to the criterion of interest.

Innovation – a guide for developing assessments targeting elusive domain constructs or using emerging technologies and new item types.

The foundations of ECD stem from validity theory (Messick, 1989), psychometrics (Mislevy, 1994), philosophy (Toulmin, 1958), and jurisprudence (Wigmore, 1937). They adapt the evidence-oriented approach to evaluating the degree to which conclusions about people can be made on the basis of collected evidence. The ECD process centers around four key questions:

1. ***Claims***: Who is being assessed and what will be declared about them as a result?

2. ***Proficiencies***: What proficiencies must be measured to make appropriate decisions?

3. ***Evidence***: How will we target, recognize, and interpret evidence of these proficiencies?

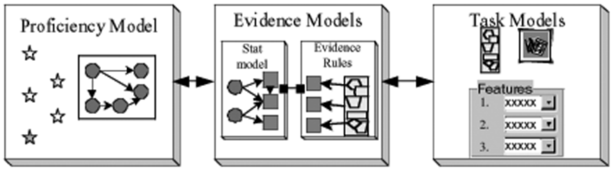
4. ***Tasks***: Given practical constraints, what situations will elicit the kind of evidence needed?

Addressing these questions results in three fundamental assessment design models, represented here as Figure 1. These ECD models include:

- ***Proficiency Model*** – defines the claims and constructs of interest for the assessment and their interrelationships.
- ***Evidence Models*** – define how observations of behavior are considered as evidence of proficiency.
- ***Task Models*** – describe how assessment tasks must be structured to ensure opportunities to observe behaviors constituting evidence.

These interrelated models comprise a chain of reasoning for an assessment design that connects the design of assessment tasks to evidence of proficiencies targeted by the assessment, which in turn are formally associated with claims made on the basis of assessment results.

Figure 1: Fundamental Models of Evidence-Centered Design



David M. Williamson, Robert J. Mislevey and Russell G. Almond are _______ with Educational Testing Service.

The following presents each of these models in turn with some discussion of their implications in the context of certification and licensure testing.

Proficiency Model

The proficiency model is really a combination of the formal assessment *claims* to be made on the basis of assessment and the *proficiencies* measured by the test. *Claims* are the specific arguments being made about people on the basis of assessment results. *Proficiencies* are measured knowledge, skills and abilities of people that provide the basis for making claims.

In order to make such claims or to identify important proficiencies, one must first have a good understanding of the population being served. Therefore, a precursor to claim specification is a definition of the examinee population, the users of the test results, and the intended use of test results in decision-making by these users. In certification and licensure testing the decision being made on the basis of the assessment is typically straightforward: either to issue or withhold the credential in question. Based on this definition, the sole users of test results are the issuing body of the credential.[1] However, since the credential itself represents a claim about the examinee made by the credentialing organization, it is typical to consider the interests of the users of these credentials (e.g., potential employers, the general public selecting their services, state licensure boards, etc.) when establishing claims. The examinee population is typically defined as individuals who have met some educational and/ or practice prerequisites and are seeking the credential in question. Implicit in this definition is the perceived value of the credential and how it benefits the personal and professional interests of the examinee.

The understanding of assessment use and population being served drives the specification of claims being made on the basis of assessment results. These claims are represented as stars in the Proficiency Model portion of Figure 1. For example, in licensure testing a common global claim made on the basis of assessment might be something like, "Can engage in professional practice without representing a risk to the health, safety or well-being of the public." Often, such a global claim about ability is supported by a number of subclaims that make explicit statements intended to directly support the overall claim of the assessment. Often these are based on elements of the domain of practice that are ultimately reflected in test content. In this way, it is typical for the claims associated with an assessment design to be organized as a claim hierarchy that elaborates the various arguments that a test score represents about individual ability. As such, the specific claims chosen for an assessment design are often directly related to needs of score reporting or delivery of instruction.

The proficiencies of individuals being measured by the assessment follow from the claims. The claims express the goals of assessment design as states of knowledge about aspects of proficiency and represent the declarations that must be supported by test results. In order to support these arguments, certain levels of ability must be demonstrated during the assessment. It is these proficiencies and the levels required to make certain claims that are specified in the proficiency structure. Assume, for example, that for a certification of computer network engineers there is a claim that such persons are adept at troubleshooting technical problems in network connectivity. It might be reasonable to expect that supporting this claim would require declarative knowledge (recall) of computer network hardware and their technical capabilities and interconnectivity protocols. It might also be reasonable to expect that supporting this claim requires an ability to employ a logical and efficient cognitive strategy to determine the cause of common network problems. Therefore, two proficiencies that might be implied by such a claim could include "hardware connectivity knowledge" and "strategic troubleshooting." These proficiency variables are inherently latent (not directly observable) and are therefore the target of the inference process of the assessment. These various proficiencies of interest are represented symbolically in Figure 1 by the set of circles and arrows in the Proficiency Model section. The circles represent various proficiency variables of interest and the arrows reflect known relationships between proficiencies (e.g., correlations or prerequisite relationships) and conditional independence relationships between variables.

The specification of claims and the description of proficiencies that one must possess to support these claims are related to traditional approaches to professional domain analysis. Often this is conducted through traditional job analyses, or in the case of assessments emphasizing strategic problem-solving, cognitive task analysis (Mislevy, Steinberg, Breyer, Almond, & Johnson, 1999).

Evidence Models (Conceptual)

Operationally, the evidence model specifies the manner in which observations during assessments are used to update estimates of ability. However, during the initial phases of assessment design, the evidence models are specified from a purely optimal domain perspective in order to drive task model development. This conceptual specification begins by imagining that there are no constraints or limitations to the ability to observe and track behaviors in naturalistic settings for a domain of interest. The task is to specify the situations and observable behaviors that are most revealing in terms of distinguishing among levels of ability in the proficiency model. The specification of what these crucial

[1] Some organizations also define the unsuccessful examinee as a user of results when diagnostic information is provided in order to guide further study.

situations are and what important behaviors can be observed will drive both the evidence model development for scoring and the specification of task models (discussed below). We will revisit the Evidence Model from the scoring perspective after the section on Task Models below.

Task Models

Task models are detailed descriptions of families of tasks with similar characteristics. These task models establish the framework, or the blueprint, for producing tasks (or items) that address particular targeted areas of the overall test blueprint. The conceptual evidence model helps to specify the characteristics of these task models that best distinguish among levels of ability. The task models, as pictured in Figure 1, consist of several variable elements: task design *features* (symbolized by the set of drop-down menu variables in the lower portion of the task model figure); *presentation material* (symbolized by the video screen icon in the upper right portion of the task model figure); and *work products* (symbolized by the jumble of shapes in the upper left of the task model figure). Task features describe the intent, construction, and associated design elements and options for a task. Presentation material defines what is presented to an examinee as part of a particular task (e.g., any graphics, any text, a question prompt, options to select from among, etc.). The work products are the resultant examinee data captured as a result of the examinee's interaction with the task, regardless of whether that data is directly used in scoring or not.

As an example of a portion of a task model, assume that for a test of basic math there was a claim of "Can add two integers" and an associated proficiency called "basic addition" (both very fine-grained examples). An item model (one of many) targeting such an ability might have elements such as those that appear as Table 1.

Table 1: Example Portion of a Task Model

Task Model Variable	***Possible Values (implication)***
Ability target	basic addition, sum of two integers
Difficulty factors	• single-digit integers with single-digit outcome (easier) • single digit integers with two-digit outcome (moderate difficulty) • two-digit integers with two-digit outcome (harder)
Reading load	• none (easier) • few simple words (moderate) • word problem (harder)
Presentation material	• Equation form of a problem (easier) • Word problem embedding problem (harder)
Work product	• multiple-choice (easier) • free response (harder) – show work (complex scoring) – response only (simple scoring)

Note that along with specification of various aspects of the task, it also indicates how different potential specifications can be expected to impact the difficulty of the item. Figure 2 and Figure 3 present the presentation material (in this case assuming paper presentation) for two items, both of which could be produced from the task model excerpted above. Note that while each is consistent with the task model above, each has characteristics that would tend to make it more or less difficult for the examinee, as well as to deliver (particularly assuming computerized presentation material rather than paper) and to score. These design decisions have implications for how a set of tasks discriminates among different levels of ability targeted by the assessment.

Figure 2

4 + 5 = ___

a) 1
b) 5
c) 9
d) 45

Figure 3

If you place a box that is 12 inches high on top of another box that is 23 inches high, how high are the two boxes together? Show your work and write your answer below.

Another characteristic to note is that for both multiple-choice and the free-response task there is an implication for a need to extend the level of detail of the task model. For the multiple-choice item, this extension would address how the distracters are developed (note, for example, that option (a) in Figure 2 is the answer someone would obtain if they subtracted 4 from 5 instead of adding) and their order in presentation. For a free-response story problem task, any use of word problems operationally would require further specification of the permissible vocabulary, sentence structures, topics, and representation of actors in the text (as well as considering the impact of the potential confound of reading ability with pure math ability on the measurement of proficiency model variables).

Obviously, the work products for these two examples in Figures 2 and 3 differ in that the former consists of an indication of which option is selected, while the latter consists of an indicated response and the calculations the examinee executed to determine the answer. Part of the consideration of such work products includes the medium used to collect the response. The item in Figure 2 is almost equally viable in paper and computerized format, while for the item in Figure 3, it is more difficult to capture the work products in computerized administration than in paper-and-pencil administration.

The set of task models for an assessment design can be organized hierarchically to facilitate the degree to which the test developer must exercise control over the types of tasks used. For example, the math test item of the general form:

{single-digit integer} + {single-digit integer} = [single-digit integer]

is a sub-category of the more general form:

{integer} {operation} {integer} = [integer]

Depending on the degree of control which must be exercised in test authoring (based on the test blueprint) and the intent of the item usage, a hierarchy of task models can be developed with varying degrees of specificity of the model design. For example, in cases where the prediction of the specific difficulty of an item is important, the test designer may wish to exercise a relatively high degree of control. This is often the case in efforts that use task modeling as the basis for automatic item generation (Embretson, 1998; Embretson, 1999; Williamson, Johnson, Sinharay, & Bejar, 2002; Newstead, Bradon, Handley, Evans, & Dennis, 2002), in which a computer generates items according to a given task model with no human intervention.

Evidence Models (Scoring)

Evidence Models specify how the evidence contained in task data informs belief about Proficiency Model variables. Evidence Models for scoring rely on the Proficiency Model as the fixed target for inferences and the Task Models and tasks authored from them, as the mechanism for producing data to be used in scoring. The Evidence Model for scoring, as presented in Figure 1, consists of two subcomponents:

- *evidence rules* – determine what elements of the task performance constitute evidence and summarizes their values
- *statistical model* – aggregates evidence to update estimates of ability in the proficiency model

Evidence rules transform elements of the work product (the record of examinee task performance) into observables; summary representations of work used by the statistical model to update estimates of proficiency. This process is called *evidence identification* in ECD terminology. In Figure 1, this process is represented in the Evidence Rules portion of the Evidence Models diagram. This figure illustrates how work products from the Task Model are parsed to produce observables, symbolized by the three squares (for three observables). With most multiple-choice questions this process seems almost trivial. For each question there is only one observable, the value of which is determined by comparing the response indicated by the examinee with a predetermined key and representing the observable as a simple 1, for correct, when the response is the same as the key, or 0, for incorrect, when the response does not match the key. In other situations the determination of observables requires more effort, such as for the task presented in Figure 3. Some evidence rules would be required to establish both the correctness of the final answer and for representing the degree of adequacy of shown work in computing the answer. Note that the determination of the value of observable variables also implies using some elements of the work product and ignoring other elements. For example, most scoring of multiple-choice items ignores the particular choice the examinee made if the choice was incorrect, while others might infer the nature of misunderstandings examinees may have when they select particular incorrect answers. Also, in computerized testing environments it is common to collect information on the amount of elapsed time an examinee took to respond to a question despite the fact that this is seldom used in the evidence rules for scoring. The representation in Figure 1 implies three observables obtained from this particular work product.

The statistical model portion of the Evidence Model uses the values of observables to update estimates of ability. In number-right scoring this statistical model is a simple summation function in which the prior value plus the value of the observable (1 or 0) equals the new value. In models using item response theory (IRT) this updating is controlled by the parameters associated with the item for which the observable is being used as evidence and by the fundamental statistical relationships for updating ability estimates from observations under the IRT model being applied. In most common applications (e.g., number right, IRT, etc.) there is a single proficiency variable for ability and a single observable variable from each item. In Figure 1, however, we illustrate the case where three observables are produced from an item and these are used to update two proficiency variables. These two proficiency variables, in turn, represent two of the five proficiency variables that make up the Proficiency Model.

Such a representation illustrates the value of such models for more complex assessment designs, such as for computerized simulations that use automated scoring, while still representing the fundamental structure and critical models for design of traditional assessments.

Summary of ECD Model Interactions

In review, the ECD process provides a framework for assessment design that emphasizes a systematic consideration of multiple models for design and their interaction. These begin with the fundamentals of assessment purpose (specification of populations being served, decisions being made, known assessment constraints, etc.) from which formal claims are developed. These claims drive the specification of a Proficiency Model. The implications of the Proficiency Model and claims in combination drive the evidential needs of the assessment, formally represented as the Evidence Model. These needs are actualized in the design of assessment tasks, the blueprints for which are expressed as Task Models.

Once tasks from these models are developed and fielded, the scoring process is essentially a reversal of the development process. The administered tasks result in work products with pre-established properties. These work products are parsed according to the evidence rules of the Evidence Model to produce observables. The statistical model of the Evidence Models is applied to draw inferences about proficiencies on the basis of these observables. Finally, the ultimate values of proficiency variables establish what assessment claims can be supported on the basis of the assessment. These reported claims, in turn, are used by the consumers of score reports to make informed decisions.

Conclusion

This work has presented the basic concepts of ECD and made an argument for the relevance and value of such an approach for any assessment design process, whether for a paper-and-pencil assessment using multiple-choice tasks or a computerized assessment using complex simulations and automated scoring. It is hoped that through wide adoption of such a process, the process of assessment design can be improved, both by formalizing processes that good assessment designers perform implicitly, and by encouraging consideration of issues not previously addressed in formal assessment design. It is also hoped that such resultant design rationales strengthen the quality and the validity arguments for use of such measures for their intended purpose.

References

Almond, R. G., Steinberg, L. S., & Mislevy, R. J. (2002). Enhancing the design and delivery of assessment systems: A four-process architecture. *Journal of Technology, Learning, and Assessment, 1*(5). Available from http://www.jtla.org.

Embretson, S. (1998). A cognitive design system approach to generating valid tests: application to abstract reasoning. *Psychological Methods*, 3 (3), 380 – 396.

Embretson, S. E. (1999). Generating items during testing: Psychometric issues and models. *Psychometrika*, 64, 407–433.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd Ed.) (pp. 13-103). New York: American Council on Education/Macmillan.

Mislevy, R.J. (1994). Evidence and inference in educational assessment. *Psychometrika*, 59, 439-483.

Mislevy, R.J., Steinberg, L.S., & Almond, R.G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1, 3-62.

Mislevy, R.J., Steinberg, L.S., Breyer, F. J., Almond, R.G. & Johnson, L. (1999). A cognitive task analysis, with implications for designing a simulation-based assessment system. *Computers and human behavior*, 15, 335-374.

Newstead, S.E., Bradon, P., Handley, S., Evans, J., and Dennis, I. (2002). Using the psychology of reasoning to predict the difficulty of analytical reasoning problems. In Kyllonen, P. and Irvine, S.H. (Eds.) *Item Generation for Test Development.* Lawrence Erlbaum Associates: Mahwah, NJ.

Toulmin, S.E. (1958). *The uses of argument.* Cambridge: Cambridge University Press.

Wigmore, J.H. (1937). *The science of judicial proof* (3rd Ed.). Boston: Little, Brown, & Co.

Williamson, D. M., Johnson, M. S., Sinharay, S., & Bejar, I. (2002, April). *Applying Hierarchical model calibration to automatically generated items.* Paper presented at the American Educational Research Association, New Orleans, LA.

Psychological Test and Assessment Modeling, Volume 53, 2011 (4), 461-485

Standard setting in complex performance assessments: An approach aligned with cognitive diagnostic models[1]

Robert W. Lissitz[2] *& Feifei Li*[3]

Abstract

With the increased interest in student-level diagnostic information from multiple performance assessments, it becomes possible to create multivariate classifications of knowledge, skills and abilities (KSAs). In this paper, a systematic, multivariate and non-compensating standard setting approach, called the cognitive analytical approach (CAA), is proposed for performance assessment with complex tasks.

CAA is based on the framework of evidence-centered design (Mislevy, Steinberg, & Almond, 2003) that supports a chain of reasoning from design and development to delivery of an assessment. In CAA, the performance standards are established simultaneously with domain-modeling, test specifications, and item writing rather than after the assessment has been completed; the cut scores are evaluated iteratively along with the test design and development phases. CAA has the benefits of ensuring the validity of the performance standards, reducing the cognitive load of standard setting, including the complexity of the tasks, and facilitating the vertical articulation of KSAs. In this paper, we elucidate the theoretical and practical rationale of CAA and demonstrate its procedures and results with an illustrative example.

Key words: Standard setting, cognitive diagnostic models, analytical approach, evidence centered design, performance centered, multidimensional standards

[1] An earlier version of this paper was presented at the Annual Meeting of NCME in Denver, Colorado (Lissitz and Li, 2010). The two authors should share equally in any compliments and any criticism that this paper might receive.

[2] *Correspondence concerning this article should be addressed to:* Robert W. Lissitz, PhD, 1229 Benjamin Building, University of Maryland, College Park, MD 20742, USA; email: RLissitz@umd.edu

[3] University of Pennsylvania

Introduction

Setting performance standards is critically important because they are used to determine which examinees will be certified, licensed, or graduated. In the context of No Child Left Behind that is mandated by the Federal Government (NCLB, 2001), individual students' academic achievements are evaluated through state testing. As a result of the evaluation, each student is assigned a Performance Level Label (PLL) based on these performance standards. One example set of PLLs could be "basic", "proficient", and "advanced". Cut scores are intended to divide students into each performance category. These standard-based labels have become an effective means of communicating the results to a variety of audiences, including parents, teachers, administrators and policymakers, and the proportion of proficient or above proficient students in a school/district may be used to determine whether the school is performing satisfactorily over time.

Despite its significance in testing and the educational system, the procedure of standard setting is often seen as arbitrary (Glass, 1978), because little consensus is often reached on the best choice of procedures, and the results of standard setting cannot be easily validated post hoc (Kane, 1994). In addition to producing defensible and valid performance standards by selecting an appropriate method and following the rigorous procedural guidelines, some scholars argue that the results of the standard setting should be evaluated in a validity framework (Hambleton & Pitoniak, 2006; Cizek, 1996). Some of them also suggested that performance standards be set in line with the design model of the assessment so that the tests could be developed on the targeted constructs and created to fit the standard (Bejar, Braun, & Tannenbaum, 2007; Bejar, 2008; Kane, 1994).

In addition to the need for a cognitive framework, there has recently been an increasing interest in the finer-grained student-level diagnostic information from performance assessment (DiBello, Roussos, & Stout, 2007). For example, NCLB requires the parents, teachers and principals receive a diagnostic report to ensure the student obtains the necessary level of knowledge, skills and abilities (KSAs) (Goodman & Hambleton, 2004). The fine-grained diagnostic feedback makes it possible for the individuals, instructors or the program managers to identify the deficiencies in abilities that are revealed by the content standards and implement interventions to remedy those skills that have not yet been mastered.

For the traditional standard setting methods that fall in a test-centered vs. examinee-centered classification (e.g., bookmark, Angoff), a single unidimensional continuum is assumed along which either the difficulty of items or the ability of the examinees can be rank ordered. In contrast, current performance assessments with complex tasks require the tasks be developed based on a well-established cognitive model so as to ensure the link with the KSAs of interest and draw sensible inferences from the scores. For items that involve multiple KSAs, a single continuum or even a composite scale may not capture multiple KSAs that underlie a complex task.

In response, new standard-setting methods for multidimensional tests have been created for educational assessments that include constructed-response items such as writing samples and short answer questions. These new methods either involve the review of candidate work or the review of the score profiles (Hambleton & Pitoniak, 2006). When the panelists are required to select the borderline work or rank order the work based on their quality, standards are set with respect to the overall quality of the examinees' performance across all questions. In contrast, it might be more informative to create classifications for each of the multiple KSAs and profile the examinees. In the standard setting methods involving the review of the score profiles, the standards are presented as score vectors, the purpose of which is to capture multiple KSAs of tests containing complex multidimensional tasks (Jaeger, 1995a, 1995b; Plake, Hambleton, & Jaeger, 1997). Although there is evidence indicating the feasibility and reliability of this type of method, the implementation procedure is challenging as it is not easy to explain the statistical models and the overall process to the panelists.

Some researchers (Roussos, DiBello, Stout, Hartz, Henson, & Templin, 2007) proposed using probabilistic diagnostic models to estimate the cut scores and classify the students. This is regarded as an objective standard setting approach in which the classifications are subject to the properties of the items and the performance of the population. However, if the number of examinees is not large enough, the model will be unidentified. In addition, the probabilistic diagnostic models are quite complicated statistical approaches that may not be appropriate for most of the audiences that use the score reports.

The shortcomings of each of the approaches above have limited their contribution to the standard-setting for complex performance assessment. According to Hambleton and his colleagues (Hambleton, Jaeger, Plake, & Mills, 2000), standard-setting for performance assessment is not nearly as well developed, and none of the methods have been fully researched and validated. Standard 4.21 in the Standards for Educational and Psychological Testing (American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME), 1999) states that "When cut scores defining pass-fail or proficiency categories are based on direct judgments about the adequacy of item or test performances or performance levels, the judgmental process should be designed so that judges can bring their knowledge and experience to bear in a reasonable way." (4.21, p 60) It stresses the importance of designing a process where panelists can optimally use the knowledge that they have to influence the process.

The purpose of this article is to propose a systematic, multivariate and non-compensating (i.e., one higher skill does not compensate for another lower skill) standard setting approach for performance assessment with complex tasks, termed the "Cognitive Analytical Approach" (CAA). CAA is created based on the framework of evidence-centered design (Mislevy, et al., 2003; Kane, 2004). In CAA, the performance standards are established simultaneously with domain modeling, test specifications and item writing; the cut scores are evaluated iteratively before and after the test development phases. By using this procedure, we expect to ensure the validity of the tests and performance standards, reduce the cognitive complexity of standard setting, and facilitate the vertical articulation of KSAs. In this paper, we intend to answer the following questions 1) What is the theoretical rationale for the CAA approach? 2) Why might the CAA be appropriate for standard setting in cognitive diagnostic assessments, compared with other approaches? 3) How should the CAA result be presented and 4) How should CAA results be used?

To address these questions, we first briefly illustrate the theoretical components for this standard setting approach, including the theories of cognitive diagnostic assessment design. Next, we make an argument for the hypothesis that CAA will outperform the traditional or the existing complex performance assessment standard setting methods by comparing and analyzing the properties and assumptions of these methods. Then, we present the framework of the CAA as well as its properties. We finally exemplify CAA with a proposed standard-setting procedure and discuss its utility in real applications.

Rationale of Cognitive Analytical Approach

Validity of standard and cut score

Performance standards and cut score are defined as distinct but related concepts (Kane, 1994; Waltman, 1997). Performance standards refer to the minimally adequate level of KSAs that students must demonstrate for some purpose, while *cut score* is a point on a score scale that forms the boundaries between contiguous levels of student performances. The cut scores that differentiate examinees on performance levels define an ordinal scale that adds more interpretation to the existing information compared to raw scores or scale scores alone. The evaluative labels (i.e., PLL) defined by the cut scores suggest substantial differences between the performance levels. Examinees assigned with a PLL are assumed to have met the required KSAs described in the Performance Level Descriptor (PLD) corresponding to that PLL and should have demonstrated the evidence of that level of proficiency in the assessment. The appropriateness of the standards, cut scores, and the claims based on them need to be validated by the evidence shown in details in The Standards for Educational and Psychological Tests ([AERA, APA, & NCME], 1999). However, as noted by Kane (2001), like policy decisions, the standards are hard to validate, especially by comparing with external criteria, so are the consequences from the decisions of the standard setting.

We may never be able to set a “correct” cut score. Nevertheless, a clear set of performance standards makes it easier to state the PLDs and set the cut score. Kane (2001) has pointed out that procedural evidence was especially important in evaluating the appropriateness of performance standards and that the standards tend to be more convincing if they have been set in a reasonable way by knowledgeable people who know the process of standard setting and the purpose for which the standards are being set. To ensure the validity and defensibility of the standards, guidelines of standard setting were recommended to be used, which include the steps to select an appropriate standard-setting method, choose a panel, arrange the activities in the panel meeting, collect evidence of validity, and conduct technical analysis (Hambelton & Pitoniak, 2006; Cizek, 1996). The importance of building the link between the assessment and standard setting is stressed, for example, by choosing the standard-setting method based on the type of items or the computation of test scores, and connecting the standard-setting methods with KSAs being assessed (Hambleton & Pitoniak, 2006; Cizek, 1996).

Some researchers further took the stance of setting the standards before the tests were designed and administered (Bejar, et al., 2007; Bejar, 2008; Kane, 1994). Kane (1994) advocated specifying the performance standard and then developing the test according to the standards. Bejar et al. (2007) proposed creating the performance standard on an assessment framework that was consistent with the theories of diagnostic assessment design (Mislevy, et al., 2003). By this approach, it is more likely that the standards will cover the constructs of interest. They argued that this approach tended to lead to more valid and reliable standard setting results.

Cognitive diagnostic assessment design

The CAA standard setting approach requires a thoughtful integration of educational policy, learning theory and curricular considerations in the process of constructing a framework to guide the development of performance standards. Each of the steps requires judgment. By following this framework, the judgments and decisions can be based on logical, articulated models and credible evidence. The evidence-centered design (ECD) framework described by Mislevy and his colleagues (Mislevy & Haertel, 2006) is an overarching and systematic framework for diagnostic assessment design. ECD incorporates models of learning throughout the assessment process and simultaneously provides support for a systematic approach to standard setting and therefore we believe it to be more likely to lead to improved learning (Mislevy and Haertel, 2006).

ECD is aimed at providing an evidentiary argument for inference about what the examinees know, can do or have acquired from what we observe them say, do or make in a few assessment circumstances (Mislevy, et al., 2003; Mislevy & Haertel, 2006). The construct-centered approach advocated by Messick (1994) supports a chain of reasoning in ECD to construct a valid assessment and develop rational scoring rubrics. This approach consists of finding a representation of constructs related to instructions or societal values, behaviors or performances revealing those constructs, and the tasks or situations that elicit those behaviors. ECD applies to the processes of designing, implementing, and delivering an educational assessment. Its key concepts and entities, and knowledge representations and tools thread through the layers of domain analysis, domain modeling, conceptual assessment framework, assessment implementation, and assessment delivery. The layered framework of ECD affords intra-field investigations while simultaneously providing structures that facilitate communication across various kinds of expertise.

Domain analysis is intended to abstract substantive information of the concepts, terminology, and knowledge representation of the domain to be assessed. Many cognitive models provide a good starting point at this stage, for example, Bloom's taxonomy (Bloom, Engelhart, Fust, Hill, & Krathwohl, 1956) that differentiates learning into the hierarchical levels of knowledge, recall, application, analysis, synthesis, and evaluation, for another example, Anderson's ACT (Adaptive Components of Thought) theory that describes the phases of acquiring declarative knowledge and procedural knowledge respectively (Anderson, 1976;1983). *Domain modeling* adopts the terminologies and reasoning from Toulmin's diagram for the assessment argument, that is, providing an explanation of the claims about a student or his/her proficiency demonstrated by the tasks created from the design pattern. The *conceptual assessment framework* (CAF) consists of student models, evidence models and task models, where technical specifications are designated. A *Student model* expresses the KSAs that the assessment designer is intending to measure in a domain of tasks, a multidimensional student model for instance. A *task model* describes the environment that elicits the student behaviors to provide evidence. The *evidence model* connects the student model and the task model, namely, evaluating the information extracted from the work products through scoring and synthesizing the evaluation data to obtain the values on measurement variables through particular measurement models such as IRT.

Standard setting in the framework of ECD

The standard setting design proposed by Bejar et al. (2007) is characterized by taking account of performance standards in the early stage of ECD and developing the performance standards for several grades simultaneously. The articulation of performance standards at an early stage is important to inform the rest of the assessment development, in addition to serving as the basis for the cut scores that become the realization of those performance standards. A conceptual model is depicted in Figure 1. The essence of this approach is that the content standards (i.e., description of what students are expected to learn) and a competency model (i.e., mechanism of how students learn) inform the formulation of performance standards, which in turn can inform the test development process. Standard setting interacts with domain analysis and modeling. The content standards, the educational policy and the learning constructs are transformed into more concrete assessment elements, influencing evidence models and task models in CAF and consequently the test specifications and PLDs. In this way, standard setting is aligned with the framework set up by ECD. Cut-score setting is an iterative process that is subject to pragmatic and psychometric constraints, informed by the plausible theory-driven maximal discrimination region on the scale, tested by the field trials, and continuously adjusted by these earlier obtained data, as appropriate.

We agree that involving standard setting at an early stage of assessment design is an efficient approach to keep performance standards in line with the content standards as well as the cognitive framework. By this means, it is more likely to reduce the cognitive load for standard setting in complex performance tasks and ensure the validity of the test development and performance standards. Therefore, we use this approach to guide our CAA standard setting.

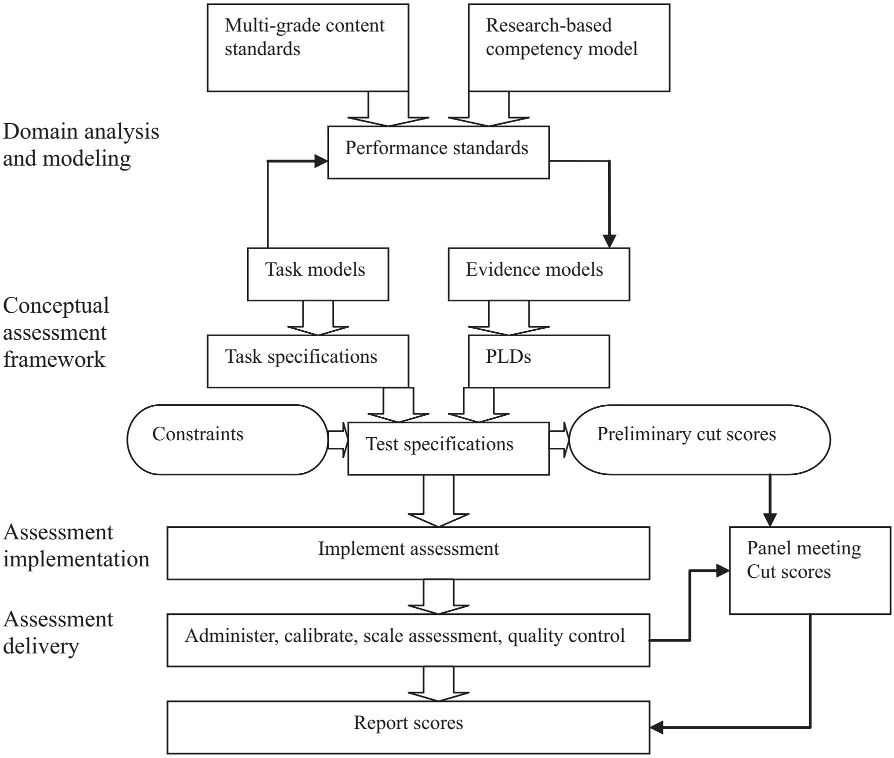

Multi-grade content standards
Research-based competency model
Domain analysis and modeling
Performance standards
Task models
Evidence models
Conceptual assessment framework
Task specifications
PLDs
Constraints
Test specifications
Preliminary cut scores
Assessment implementation
Implement assessment
Panel meeting
Cut scores
Assessment delivery
Administer, calibrate, scale assessment, quality control
Report scores

Figure 1:
Flowchart of standard-setting in the framework of evidence-centered design
(adapted from Bejar, et al., 2007)

CAA compared with other standard setting methods

CAA is meant to be implemented simultaneously with content specification and test development. It is designed to capture the non-compensatory, multiple dimensions for performance assessment. There might be no resolution to the argument about whether CAA is adequate or appropriate for a test, because it is hard to quantify the personal and societal costs and benefits associated with any particular performance standard (Kane, 1994). It might be virtually impossible to validate a claim that any performance standard is correct, but by validation we can justify that one standard-setting method is better than any of the others (Cizek, 2001). Five types of evidence are usually considered to evaluate the validity: explicitness, practicability, implementation of procedures, panelist evaluations and documentation (Hambleton & Pitoniak, 2006).

Traditional standard setting methods

Traditional standard-setting methods are usually classified into a dichotomy: test-centered methods versus examinee-centered methods (Cizek, 2001). Test-centered standard-setting methods require panelists to make judgments on the expected levels of performance by borderline examinees on each assessment task (e.g., Angoff, modified Angoff, bookmark); while examinee-centered standard-setting methods require panelists, who know the students, to place the students into performance categories, without any knowledge about their actual performance on the test (e.g. contrasting groups). This dichotomy can be applied as follows to three very popular approaches to standard setting:

1. Modified Angoff: the judges estimate the probability that a minimally proficient or minimally advanced student will get the item correct. One alternative approach is to rate the item as more likely to be answered correctly or incorrectly by a student who is minimally proficient or is minimally advanced. This procedure assumes that ordering items by the probability of getting an item correct (difficulty level) is also ordering the level of KSA. The sum of the numbers or probabilities across all the items in the test is the cut-point as determined by that standard setter for that test. Averaging across standard setters provides the recommended cut-score for each of the three levels of student performance. Cut scores are set iteratively. In each round, the standard setters are usually informed about the impact data, that is, how the cut scores they have recommended are going to affect the classification of the population of students who have taken or will take the test.
2. Bookmark: a number of items are examined and are organized from easiest to most difficult by the p value in classical testing theory or item difficulty parameter b value in IRT. The task for the standard setting panelists is to place a "bookmark" between the hardest items that a basic student would get right and the easiest item that the basic student would not get right. Again, this approach asks the standard setters to use the PLDs to determine the placement of the cutoff and their work is informed by discussion with other members of the standard setting team. The difficulty of each item is central to this procedure and organizes the items by difficulty in what is called an "ordered item booklet."
3. Traditional contrasting groups method: the judges, who are familiar with a group of examinees, are asked to use the PLDs to identify examinees who are clearly above a particular performance standard and those who are apparently below that performance standard based on their knowledge about the examinees' overall performances or proficiencies. Then the test score distributions of these two groups are plotted and the cut score is placed at the point where the two distributions intersect (Cizek, 2001). Ordering by test scores implies, again, a reliance on the difficulty of the test items that aggregate to that total score to define the KSA of the construct(s) for which the standard is being determined.

It is noteworthy that the traditional standard setting methods have in common a single scale along which either the item difficulties or the levels of ability are rank ordered. The item difficulty can be the p-value or the IRT difficulty parameter, response probability or an average item score for constructed response items. This scale is analogous to the item difficulty/ability scale in IRT. For an examinee-centered standard setting such as the contrasting group method, the students are ordered by and within PLDs along a single continuum of the skill. The assumption is that the total test score is a monotonic function of the latent ability. In other words, students with higher value on the latent ability will score higher on their performance based on the total test. While for test-centered approaches, such as the bookmark method, items are placed along a single continuum of difficulty and a marker is placed to differentiate students who are able to answer items difficult enough to be considered proficient, yet not able to answer items so difficult as to be considered expert. If the items are assumed to be monotonic, they make sense lined up against the continuum implied by and within the PLDs.

This single scale embedded in the traditional standard setting provides a means of communication to panelists. The panelists must consider the ability of the students along a continuum that is adequately captured by difficulty or some essential variant of difficulty. It is implied that the placement of a student in a PLL should depend upon the difficulty of the items that he or she can answer correctly or with higher probability. It suggests that difficulty is a proxy for or monotonic to the ordering of the PLLs from the lowest level (e.g., being able to answer easiest items) to the highest level (e.g., being able to answer hardest items). This ignores the fact that there could be and almost certainly are multiple scales underlying a complex performance task. Even a composite scale cannot precisely capture several attributes at one time, unless they at least mirror each other monotonically. Cognitively simple knowledge level items can be very difficult for a variety of reasons and in fact might be much harder than more complex reasoning items.

Finally, there is considerable evidence that difficulty is not the same as cognitive complexity, and it is cognitive complexity that is at least the conceptual focus of standard setting. In other words, schools are not usually interested in whether students can answer "hard" knowledge items rather than analysis and synthesis items. It is the difference between students who operate cognitively at the knowledge level versus those that operate at more advanced cognitive levels that is of interest. Several papers have shown that assessment items that may be ordered in difficulty, do not necessarily order the same way by their cognitive complexity. Papers by Arend, Colom, Botella, Contreraa, Rubio, Snatacreu (2003), Spilsbury, Stankov and Roberts (1990), Stankov (2000) and Stankov and Raykov (1995) are examples of work in that area.

Standard setting methods for complex performance assessment

Over the last 10 years, many assessment programs have added constructed-response (CR) items, with a hope to deliver a test that is closer to real learning situations. CR questions require the examinees to produce the response in their own words. CR questions vary in cognitive and format complexity to a larger extent than multiple-choice questions. For instance, complex CR questions could require examinees to integrate knowledge and apply to a real-life situation, or provide a rationale to justify their responses.

One example in which a CR item is scored multidimensionally is the trait scoring for some writing assessments (e.g., writing test in Arizona Instruments to Measure Standards (AIMS)). A set of rubrics is created for latent traits (i.e., idea/content, organization, voice, word choice, sentence fluency, conventions). The answer is scored by rater's judgments regarding the performance on each trait. However, when it comes to the standard setting, a composite scale is created by averaging the trait scores to allocate an overall cut score. This provides no classification information on each of the traits for diagnostic purposes.

The new test format presents the need for appropriate standard setting methods to accommodate such complexity. Presented in this section are methods that could deal with tests containing constructed-response items. These methods involve either review of candidate work or review of score profiles. For the former type of standard setting, the product work could be viewed either item by item (Loomis & Bourque, 2001), or section by section (Plake & Hambleton, 2001; Plake, Hambleton, & Jaeger, 1997), or holistically (Jaeger & Mills, 2001), depending on the properties of the test, such as the type of items, the total number of CR questions, the complexity of the questions whatever type they are, or the actions required by the examinees (Cizek, 2001).

1. Item by item approaches: for each question, panelists are asked to select from a set of examinee performances the work that best represents the performance of minimally competent candidates. In some cases, the actual scores assigned to the papers are revealed to the panelists. Then the panelists make an estimate of the performance of the minimally competent candidate on each question. These standards are then aggregated to obtain the overall performance standard for the full test. However, since this approach takes place after the test is administered, it may be difficult for the panelists to adjust their performance standards from round to round. On the one hand, it may not be easily interpretable to the panelists how one score may represent borderline performance at a given performance standard, and another score represents borderline performance at another level. On the other hand, there may be a lack of papers at a given score point. Therefore, it may take a long time to prepare work representative at different levels.

2. Holistic approaches: Like the *Body of Work* (Kingston, et. al., 2001) they require panelists to view the samples of examinee performance holistically. Panelists are provided with more examinee papers representing a more focused score range around a cut point. The values in this range suggest where the minimum performance standard would be likely to fall. The score point where panelists seem indifferent to pass-fail decisions is chosen as the passing score. This process is repeated for each performance standard of interest. The limitation of this type of methods is that there is a maximal booklet length beyond which panelists cannot make valid and reliable judgments about the materials (Hambleton, Jaeger, et al., 2000). The researchers had observed that when panelists were presented with the complete work of examinees, they tended to skip over some of that work and key in on selected questions or the first part of the students' work.

3. Hybrid approaches: *Analytic judgment method* for instance, the panelists' ratings are based on components of the test, rather than on the entire test. Breaking up the test book-

let into smaller collections of test items was done to reduce the cognitive complexity of the rating task by reducing it to judging more modest sets of items. Panelists sort candidate papers into ordered performance categories. The ratings can be transformed into performance standards by using a boundary method (i.e., averaging the scores of papers assigned to the high end of one performance category and the low end of the next higher performance); the performance standards established for each set of test items are then summed in order to obtain performance standards for the total test. However, this set of approaches does not set standards for multiple dimensions in particular. Again, the procedure depends upon the reasonableness of adding scores and that depends upon their being at least monotonically related.

We may notice that no matter how the work is reviewed item-by-item or holistically, scores assigned to the performance of borderline candidates are ultimately aggregated across the test and result in a set of standards to evaluate the overall performance. Proficiency is measured on a composite scale that is directly related to number correct or some weighted average or sum of sub-scores. In contrast, methods that involve review of the score profile address the standard setting for the complex exercises that are scored multidimensionally and focus on the cognitive structure that underlies the test. That focus is retained as long as the process does not involve adding the multidimensional scores together to form a single composite.

In the *Judgmental policy-capturing (JPC) method,* the panelists' task is to review hypothetical score profiles and rate a large number of vectors of scores, and the standards are inferred from a statistical analysis of their ratings. In one of the variations implemented for the National Board for Professional Teaching Standards Certifications (Jaeger, 1995), each exercise and the entire assessment were scored multidimensionally. The panelists were provided with information about their own ratings of profiles and the ratings of the entire panel. This approach was claimed to be feasible and reliable (Jaeger, 1995b), but Hambleton (1998) also noted that it is challenging to find statistical models that fit the panelists' ratings and explain the overall process to panelists for deriving a performance standard. *Dominant profile method (DPM)* is another approach where the panelists, who are fully aware of the questions and the meaning of the scores, derive decision rules that capture the score levels across the profile components. With a large number of possible score profiles, it is hard to reconcile panelists' views into a consensus.

Methods and procedures

The following is an illustrative example that we have created around the standard setting flowchart proposed by Bejar et al. (2007). Using CAA, we demonstrate how one might establish the performance standards, task models, and evidence models before defining the task specifications and PLDs, which in turn precede formulating the test specifications and item development. Setting cut scores is now an iterative process along with test construction and does not depend upon test performance or upon difficulty level or its aggregation. When agreement is attained on task models and constraints, the blueprint of the test specification can be finalized, and the preliminary cut scores corresponding to the performance standards are determined. Preliminary cut scores can be evaluated after the assessment implementation and adjusted in light of the data available, if it is desired to do so, but that is not necessary. Impact data are often of great interest to policy considerations, but are not of much interest at a conceptual level. The standards could be specified across the grades by systematically basing new learning on the preceding acquired skills, but here we focus on the CAA procedure for one grade to illustrate its process and application.

Purpose of the assessment

The current standard setting is assumed to take place in a large-scale performance assessment that is intended to diagnose the Knowledge, Skills and Abilities (KSA) in mathematics for regular students in 6th grade. Students are required to draw on a broad body of mathematical knowledge and apply a variety of mathematical skills and strategies. In order to function as a citizen and a worker in the contemporary society, a person should have the ability to explore, to conjecture, to reason logically, to communicate in mathematics effectively, and to apply a wide repertoire of methods to solve problems.

Domain analysis

Through the analysis, we have the following list of content-related standards for 6th grade: (1) numbers and operations, (2) data analysis, probability and discrete analysis, (3) patterns, algebra and functions, (4) geometry and measurement, (5) structure and logic. Other abilities we call structural KSAs are (1) communication, (2) problem-solving, (3) reasoning proof, (4) connections, and (5) representation that are embedded throughout the teaching and learning of all mathematical content.

Content standards

Learning objectives represent the expectations in regard to each content area. The skills necessary to meet those expectations are identified. We take the Measurement component in Geometry and Measurement for example (Table 2). Key skills required for each objective are listed, some of which come from the previous learning objectives. For example, to estimate the measure of objects using a scale drawing or map, the required KSAs are, in brief, the content-related KSA of fractions, and the structural KSAs of problem-solving, reasoning, representation and communication.

Proficiency level descriptors (PLDs)

PLDs identify the evidence that is determinant to the proficiency levels. The evidence can evolve from an abstract expectation to a more concrete form of descriptions for “advanced”, “proficient” and “basic” levels. As is shown in Table 3, the PLDs are labeled first in an abstract form, and then in a concrete form as in Table 4. The set of all PLDs corresponding to the learning objectives are derived, but not all the learning objectives are covered in Table 4. Three sets of PLDs are shown for three different proficiency levels, but notice that these are for a specific skill/concept and there may be many such in a test to which CAA is applied.

Test specifications

As the PLDs are elaborated it is necessary to create tasks that can be expected to elicit evidence linked to the PLDs. Tasks could take forms ranging from multiple choice items to open-ended questions. The task model can be built upon the structural variables and content-related variables defined in the content standards. Each of the task models can be represented in a variety of ways. Given a specific instance of a task model we can describe the structural attributes, including the PLDs, that the task was designed to elicit. One of the conveniences brought by CAA is to designate individual items to discriminate between the adjacent proficiency levels before data collection and analysis. Another advantage is to specify the KSAs involved in the design of a particular item. At the early stage of developing CAA, we assume that each item is rated with a score vector on the designated KSAs assessed. At this stage, we may focus on the test tasks where KSAs are non-compensatory (the score on one KSA is independent of the score on another one) in accomplishing a correct answer to the item. An example of test specification is presented in Table 1, a possible structure of test specification. The analysis inherent in CAA might even suggest that additional items need to be written to permit more accurate measurement associated with specific score vectors.

Table 1:
Table of test specifications

	KSA1	KSA2	KSA3	KSA4	……	KSAm
Item 1	A/P		P/B			
Item 2		A/P		P/B		A/P
Item 3		A/P	A/P			
Item 4	P/B			A/P		
Item 5	A/P					
Item 6						P/B
⋮						
Item n		P/B	A/P			

A=Advanced, P=Proficiency, B=Basic

Table 2:
Learning objectives and the key KSAs. ([4]Arizona mathematics standard articulated by grade level, grade 6, 2008)

Strand 4: Geometry and Measurement Geometry involves the development of students' reasoning, higher-order thinking, and justification skills culminating in work with proofs. Geometric modeling and spatial reasoning offer ways to interpret and describe physical environments and can be important tools in problem solving. Students use geometric methods, properties and relationships, transformations, and coordinate geometry as a means to recognize, draw, describe, connect, analyze, and measure shapes and representations in the physical world. Measurement is the assignment of a numerical value to an attribute of an object, such as the length of a pencil. At more sophisticated levels, measurement involves assigning a number to a characteristic of a situation, as is done by the consumer price index. A major emphasis in this strand is becoming familiar with the units and processes that are used in measuring attributes.	
Measurement Understand and apply appropriate units of measure, measurement techniques, and formulas to determine measurements. In Grade 6, students build upon their prior knowledge of measurement to determine the appropriate unit of measure, tool, and necessary precision to solve problems. They convert within systems of measurement to solve problems. They use scale drawings to estimate the measure of an object. Students also apply formulas for area and perimeter to solve problems and explore the relationship between volume and area.	
Performance Objectives	***Key KSAs***
Students are expected to:	
PO 1. Determine the appropriate unit of measure for a given context and the appropriate tool to measure to the needed precision (including length, capacity, angles, time, and mass). Connections: M06-S1C3-02, SC06-S1C2-04	*(M06-S5C2-01) Analyze a problem situation to determine the question(s) to be answered. (M06-S1C3-02) Multiply and divide fractions. (SC06-S1C2-04) Perform measurements using appropriate scientific tools (e.g., balances, microscopes, probes, micrometers).
PO 2. Solve problems involving conversion within the U.S. Customary and within the metric system. Connections: M06-S1C1-03, M06-S1C3-02	*(M06-S5C2-03) Apply a previously used problem-solving strategy in a new context. (M06-S1C1-03) Demonstrate an understanding of fractions as rates, division of whole numbers, parts of a whole, parts of a set, and locations on a real number line. (M06-S1C3-02) Make estimates appropriate to a given situation and verify the reasonableness of the results.

[4] http://www.ade.state.az.us/standards/math/Articulated08/Gradeleveldocs/MathGrade6.pdf

PO 3. Estimate the measure of objects using a scale drawing or map. Connections: M06-S1C1-03, M06-S1C3-02, SS06-S4C1-03	*(M06-S5C2-03) Analyze and compare mathematical strategies for efficient problem solving; select and use one or more strategies to solve a problem. (M06-S1C1-03) Demonstrate an understanding of fractions as rates, division of whole numbers, parts of a whole, parts of a set, and locations on a real number line. (M06-S1C3-02) Make estimates appropriate to a given situation and verify the reasonableness of the results. (SS06-S4C1-03) Interpret maps, charts, and geographic databases using geographic information
PO 4. Solve problems involving the area of simple polygons using formulas for rectangles and triangles. Connections: M06-S1C3-02, M06-S3C3-04, M06-S5C1-02	*(M06-S5C2-02) Identify relevant, missing, and extraneous information related to the solution to a problem. *(M06-S5C2-04) Apply a previously used problem-solving strategy in a new context. (M06-S1C3-02) Make estimates appropriate to a given situation and verify the reasonableness of the results. (M06-S3C3-04) Evaluate an expression involving the four basic operations by substituting given fractions and decimals for the variable. (M06-S5C1-02) Create and justify an algorithm to determine the area of a given compound figure using parallelograms and triangles.
PO 5. Solve problems involving area and perimeter of regular and irregular polygons. Connections: M06-S1C3-02, M06-S3C3-04, M06-S5C1-02	*(M06-S5C2-04) Apply a previously used problem-solving strategy in a new context. (M06-S1C3-02) Make estimates appropriate to a given situation and verify the reasonableness of the results. (M06-S3C3-04) Evaluate an expression involving the four basic operations by substituting given fractions and decimals for the variable. (M06-S5C1-02) Create and justify an algorithm to determine the area of a given compound figure using parallelograms and triangles.
PO 6. Describe the relationship between the volume of a figure and the area of its base.	*(M06-S5C2-04) Apply a previously used problem-solving strategy in a new context.

Table 3:
Performance level descriptors on the general KSAs

Performance Level	Descriptor
Advanced	The student exceeds the expectations for demonstrating an independent and accurate understanding of the specified math skills/concepts. The student demonstrates the ability to apply the skills/concepts to an authentic task and/or environment with analysis and reflection by: – solving a real world problem (e.g., determining what fraction of a dozen eggs are needed to bake a cake if 3 are needed) – applying math skill/concept in the natural environment (e.g., store, home, technical education class, science class, home economics, etc.) to solve a problem – communicating an in-depth explanation that analyzes or reflects on the problem (e.g., demonstrate how left over pieces of one pizza can be combined with pieces of another pizza to create a whole pizza and explain how that works)
Proficient	The student demonstrates an independent and accurate understanding of the specified math skills/concepts. Occasional inaccuracies, which do not interfere with conceptual understanding, may be present. The student demonstrates the ability to apply the skills/concepts to an authentic task and/or environment by: – solving a real world problem (e.g., determining what fraction of a dozen eggs are needed to bake a cake if 3 are needed; determine the perimeter of a table to determine the amount of ribbon needed to decorate the sides; reproduce two dimensional shapes to complete an art project; construct a bar graph showing class election results; etc.) – applying math skill/concept in the natural environment (e.g., store, home, technical education class, science class, home economics, etc.) to solve a problem. – using relevant details (e.g., uses measurements, elements of 2 D shapes, data, numbers, etc.) – using math vocabulary (e.g., fractions, whole, area, perimeter, rectangle, square, data, graph, pattern, etc.) – using a model or explanation to demonstrate a concept or solve a problem (e.g., create a chart showing fractional parts; draw a floor plan of a clubhouse and provide area; categorize shapes according to elements; create a bar graph and answer questions; etc.)
basic	The student demonstrates basic understanding of the specified math skills/concepts. Inaccuracies may interfere with or limit the conceptual understanding. The student demonstrates some understanding without applying the skills/concepts to an authentic task and/or environment by:

	– solving a problem (e.g., identify fractions on worksheet, figure area problems; match element to 2 D shape; complete numerical pattern; etc.) – using relevant details (e.g., uses measurements, elements of 2 D shapes, data, numbers, etc.) – using math vocabulary (e.g., fractions, whole, area, perimeter, rectangle, square, data, graph, pattern, etc.) **or by:** – using a model or explanation to demonstrate a concept or solve a problem (e.g., create a chart showing fractional parts; draw a floor plan of a clubhouse and provide area; categorize shapes according to elements; create a bar graph and answer questions; etc.)
Below basic	The student demonstrates little or no understanding of the math skills/concepts. Inaccuracies interfere with the conceptual understanding. The student demonstrates this by: – inaccurate use of details (e.g., uses measurements, elements of 2 D shapes, data, numbers, etc.) – inaccurate or no use of math vocabulary (e.g., fractions, whole, area, perimeter, rectangle, square, data, graph, pattern, etc.)

Table 4:
[5]Arizona mathematics standard performance level descriptors on specific learning objectives (grade 6)

Students at the "Advanced" level generally know the skills required at the "Proficient" and "Basic" levels and are able to:	Students at the "Proficient" level generally know the skills required at the "Basic" level and are able to:	Students at the "Basic" level generally know and are able to:
• Use prime factorization to determine greatest common factor and least common multiple. • Express the inverse relationships between exponents and roots for perfect squares and cubes. • Apply and interpret the concepts of addition and	• Convert between fractions, decimals, percents, and ratios. • Express a whole number as the product of its prime factors. • Demonstrate an understanding of fractions as rates or as division of whole numbers. • Compare and order integers, positive fractions, positive decimals, and positive percents.	• Express that a number's distance from zero on the number line is its absolute value. • Apply properties to solve numerical problems. • Make estimates appropriate to a given situation and verify

[5] http://www.ade.state.az.us/standards/AIMS/PerformanceStandards/6thMathPLD.pdf

subtraction with integers using models. • Provide a mathematical argument to explain operations with two or more fractions or decimals. • Build and explore tree diagrams where items repeat. • Investigate and solve problems using Hamilton paths and circuits. • Create and solve two-step equations with fractions and decimals. • Solve problems involving supplementary, complementary, and vertical angles. • Solve problems involving area and perimeter of regular and irregular polygons. • Describe the relationship between the volume of a figure and the area of its base. • Create, analyze, and justify algorithms for multiplication and division of fractions and decimals and area of compound figures. • Make and test conjectures based on information collected from explorations and experiments. • Solve simple logic problems and justify solution methods and reasoning.	• Multiply and divide decimals or fractions. • Simplify numerical expressions using the order of operations. • Use benchmarks as meaningful points of comparison for rational numbers. • Interpret, describe, and analyze displays of data. • Determine theoretical probability and apply it to predicting experimental outcomes. • Analyze numerical patterns using all four operations. • Describe the relationship between two quantities in a function. • Use an algebraic expression to represent a quantity. • Evaluate an expression by substituting given fractions and decimals for the variable. • Solve problems involving the relationship among the circumference, diameter, and radius of a circle. • Identify the missing coordinate of a polygon on the coordinate plane. • Solve problems involving conversion within the U.S. Customary and within the metric system. • Solve problems involving the area of simple polygons using formulas for rectangles and triangles. • Evaluate situations and select strategies to find and apply solutions to problems. • Compare sets of data by analyzing trends. • Explore counting problems using Venn diagrams with three attributes.	the reasonableness of the results. • Identify a simple translation or reflection of a 2-dimensional figure on a coordinate plane. • Graph ordered pairs in any quadrant of the coordinate plane. • Determine the appropriate unit of measure for a given context. • Estimate the measure of objects using a scale drawing or map.

Panelists' selection

We assume that the panelists consist of teachers, non-teacher educators, test developers, and the general public. Panelists are usually recruited statewide through a stratified sampling, and we will assume that has occurred in this hypothetical example. In many standard setting applications, sampling would try to have no less than 30% of the panelists from ethnic minority groups and no less than 25% of them being males. In this case, we will retain the usual three rounds of panelists' meetings to set cut scores, once before the tests are developed and once after the tests are written and administered. Notice, that unlike the usual standard setting, these rounds are separated by long time periods of intense test related work. The cut scores are finalized as the third and last round of discussions. For illustrative purposes, let's assume that 18 panelists are distributed to three tables with 6 at each table and stratification is utilized to maintain balance along various dimensions of interest such as race, gender, geographic region and SES level of the typical student at the participant's school.

Orientation and discussions

Round 1. Panelists receive an overview of CAA method in a 60-minute presentation. The presentation describes the purpose of the diagnostic assessment, the basic concepts and framework of ECD, and interpretations of PLDs. The role of standard setting prior to the test development is explained and its value and interpretation is made clear.

KSA Review. Panelists are presented with a learning objective table such as Table 2, showing learning objectives in each content domain area and the KSAs necessary to meet a learning objective. When the KSA review is complete, panelists should have a detailed, structured understanding of the assessment and expected student achievement.

PLD Review. Panelists also review the PLD tables in both abstract (Table 3) and concrete (Table 4) forms. The abstract PLDs describe the expectations on general KSAs (e.g., analyzing, application, problem-solving and communication), which are less related to the specific content. In contrast, the concrete form provides PLDs on a sample of learning objectives.

Test Specifications Review. Panelists are also instructed to study the test-specification table as shown in Table 1, where the items are represented as capable of discriminating between adjacent performance levels on KSAs based on the attributes of the learning objectives. Panelists are asked to think of a task, preferably in the form of an item that exemplifies the content knowledge, skill or ability given in the specifications. Panelists are also asked to share items with the whole group for discussions.

Preliminary cut-score setting. With clear test blueprints such as summarized in Table 1, the next step is to obtain preliminary cut scores. At this stage in the development process, prior information on the PLDs has been accumulated and, moreover, the PLDs can be associated with the learning expectations linked to the performance levels. Based on the characteristics on the KSA continuum, there are items that are more likely to discriminate between “advanced” and “proficient”, and items that are likely to discriminate between “proficient” and “basic”. It is feasible, and desirable, to associate performance levels with possible performance on a test, even though the test has not been fully implemented or administered.

Panelists spend the next five hours of meeting time identifying the knowledge, skills, and abilities, and the learning objectives students must have to qualify for “advanced” or “proficient”. They also read the test specifications as presented in Table 1. For each KSA, panelists can make their decisions on the cut scores by aggregating the ratings on the same KSAs across items (i.e., calculating the number of “A” or “P” or “B) and fill in a cut-score table (Table 5). The specification table may be revised if more or less information on a particular KSA is needed. For example, the cut-score of “proficient” and “basic” is all “Ps” on problem-solving, and the cut-score of “advanced” is at least four out of five “As” and one “P” on the same KSA.

Test Development. Test developers generate tasks that best discriminate the levels designated in the table of specifications and written items. Tests are administered and scored on KSAs (the score points differentiate “advanced” and “proficient”, “proficient” and “basic”). For example, for the learning objective “to determine the appropriate unit of measure for a given context and the appropriate tool to measure to the needed precision (including length, capacity, angles, time, and mass)”, panelists, in their first round of discussion, decide that this would involve the content-related KSAs of fractions and using measurement tools, and the structural KSAs of problem-solving and reasoning. The KSAs on fractions and measurement tools are likely to discriminate the proficient students from the advanced ones, while on problem-solving and reasoning, these items are expected to differentiate “proficient” and “basic” students. Based on these task features, an item could be constructed as follows: In your science class, you want to measure leaf width and plant heights to determine the effects of different kinds of fertilizers. What tools and units of measure would you use to make the measurements? To what degree of precision should you measure? Explain and justify your choices.

Round 2. Panelists are convened again after the test design implications from round one are implemented and they have a brief review on the KSAs that each item measures. They are presented with sample papers with a wide range of proficiency levels. The panels, again, keeping in mind the performance level descriptors on each KSA and using a table like Table 1 to rate the performance as “A”, or “P” or “B” for each KSA, decide the minimum number of “As”, “Ps” and “Bs” for each proficiency level of a KSA (Table 5).

Table 5:
Cut-score table

	KSA1	KSA2	KSA3	KSA4	……	KSAm
Basic/proficient	4Ps&1A	6Ps	5Ps & 1B	7Ps		5Ps&2As
Proficient/Advanced	5As	5As&1P	6As	5As&2Ps		6As&1P

Round 3. Cut-score results from the Round 1 and Round 2 are provided for comparison purposes. Panelists are shown the numerical values of the Round 1 and Round 2 medians. Panelists could see the change in the median from Round 1 to Round 2, and give cut score recommendations. This is an iterative process. Discussions take place to explore and try to resolve salient differences of opinion within each group. Panelists will be provided with results from other groups and discussions will continue until a consensus is (hopefully) achieved for the whole group.

The procedure illustrated above is one of the possible procedures of CAA. Other variations could be a bookmark like procedure that orders sets of items along the specific scales of KSA and places a bookmark at the borderline that divides the proficiency levels for each KSA, or an Angoff procedure that requires judgments on the probabilities of correct answer for the minimally proficient candidates, again on the relevant items measuring a specific scale. Notice that the essential multidimensionality of the test is maintained in the standard setting. The procedure illustrated above in detail represents a hybrid approach that integrates both a test-centered component in Round 1 and an examinee-centered component in Round 2. This hybrid approach enables the performance standards to be determined in what we argue as a more sensible way. Other variations on the essential ideas of CAA can be implemented, as the client (state) might choose.

Discussions and conclusions

In CAA, the performance standards are established simultaneously with domain modeling and test specifications; the standards and cut scores are evaluated iteratively along with the test design and development phases. CAA has the benefits of ensuring the validity of the performance standards, reducing the cognitive load of standard setting, including the complexity of the tasks, and facilitating the vertical articulation of KSAs. In this paper, we elucidate the theoretical and practical rationale of CAA and demonstrate its procedures and results with an illustrative example that we have created to show how this process might unfold.

CAA that is specifically tailored for cognitive diagnostic assessment is a thoughtful integration of educational policy, learning theories and curricular considerations in the process of constructing a framework to guide the development of performance standards. At the first stage, the learning objectives are translated into proficiency models and then linked to PLDs. The standards are set in regard to each learning objective while the test specifications are also determined. Once the tests are created and implemented, judgment is required again to reevaluate the performance standards and transform them into a set of cut scores. One of the major advantages of this approach is that with the guidance of ECD, the cognitive structure is maintained to be consistent and coherent across the stages from the domain modeling to score reporting. By this means, we would have more convincing evidence for the construct relevant validity since the test is designed to adhere to this structure.

CAA is innovative and appropriate for the cognitive diagnostic assessment compared with the existing standard setting methods. The traditional standard setting methods assume a unidimensional scale along which the abilities or the item difficulty values are rank ordered. This simplified cognitive pattern facilitates the communication with the standard setters, but it is an incorrect and misleading assumption with respect to the latent structure for complex performance tasks tapping into multiple skills. The contemporary standard setting methods for complex performance assessments which fall in the categories of analytical or holistic methods treat each item as a distinctive instrument measuring a sub-domain of skills, but items are still assumed to be unidimensional and the standard setting procedures result in an overall cut score on the composite scale that is a fiction. The current standard setting methods that involve the creation and review of score profiles tend to result in a large number of score patterns, which make it cognitively challenging to reduce to a smaller number of performance standards that are conceptually sensible.

In contrast, CAA considers a pattern of constructs to be assessed at the very beginning, and designates the constructs to determine the test specifications. CAA becomes an integral part of the planning and design process. In other words, the dimensions and their standards are designed into the test at the beginning. The performance levels that each item is intended to discriminate are specified as part of the development process. This approach recognizes the multidimensional latent structure of CR items and MC items and facilitates setting cut scores on several constructs at a time. The participation of test developers helps to ensure the consistency of assessment design and standard setting. The standards are set in a way consistent with how the learning objectives are labeled and items are scored. That is, the panelists are able to express their standards in terms of the number of “As”, “Ps” or “Bs”, which is explicit and determined prior to any test administration and data collection. In addition, CAA provides a systematic approach to develop the standards for different grades, and thus has the potential for setting standards across grades. We have not explicitly addressed this application, but creating panels from different subject matter areas and especially different grades can be used to create vertically moderated standards (Lissitz and Huynh, 2003).

We take account of the assessments with CR items in this study. Further research could investigate complex performance assessments that involve both multiple choice and CR items, make a distinction between the different test formats and update the standard setting methods accordingly. We could also examine the utility of other variations of CAA that use bookmark or modified Angoff procedures adapted for this purpose.

Some researchers (Roussos, et al., 2007) proposed model-driven classifications using probabilistic diagnostic models to estimate the cut scores to classify the students at different levels. On the one hand, this is an objective approach to obtain the classifications from the data and model. On the other hand, some of the parameters in the diagnostic models are specified based on the cognitive theory, such as those in the Q matrix that connect the latent attributes and the items, and many other assumptions are imposed to make the estimation possible. In addition, model identification will be an issue especially for a small-scale performance assessment where the examinee pool is not big enough to ensure all parameters can be accurately estimated. Importantly, the probabilistic diagnostic models are grounded in probability theory and applications of Bayesian statistics and might not be accessible or interpretable for most of the audiences that receive the score reports or the classification results. Finally, such models are usually implemented after the test data are obtained and our approach is designed to be a part of the test construction process. CAA can be regarded as complementary to the probabilistic diagnostic approach. They are both based on a certain kind of cognitive diagnostic framework, but through different classification procedures. However, it would be interesting to compare the results of the standard setting by human judgment with the model-driven classifications.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Anderson, J. R. (1976). *Language, Memory, and Thought*. Hillsdale, NJ: Erlbaum.

Anderson, J. R. (1983). *The Architecture of Cognition*. Cambridge, MA: Harvard University Press.

Arend, I., Colom, R., Botella, J., Contreras, M.J., Rubio, V., & Santacreu, J. (2003). Quantifying cognitive complexity: evidence from a reasoning task. *Personality and Individual Differences. 35*(3), 659-669

Bejar, I. I., Braun, H. I., & Tannenbaum, R. J. (2007). A prospective, progressive, and predictive approach to standard setting. R. Lissitz (Ed.), *Assessing and modeling cognitive development in school*. Maple Grove, MN: JAM Press.

Bejar, I. I. (2008). Standard setting: What is it? Why is it important? R&D Connection. 7. www.ets.org

Bloom, B. S. (ed.), Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of Educational Objectives: Handbook I: Cognitive Domain*. New York: David McKay

Cizek, G. J. (1996). Standard-setting guidelines. *Educational Measurement: Issues and Practice, 15*(1), 12-21.

Cizek, G.J. (2001).Conjectures on the rise and call of standard setting: An introduction to context and practice. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 3-17). Mahwah, NJ: Erlbaum.

DiBello, L. V., Roussos, L. A., & Stout, W. (2007). Review of Cognitive Diagnostic Assessment and a Summary of Psychometric Models. *Handbook of Statistics*, 26, 1-52.

Glass, G. V. (1978). Standards and criteria. *Journal of Educational Measurement*, 15, 237-261.

Goodman, D. P. & Hambleton, R. K. (2004). Student Test Score Reports and Interpretive Guides: Review of Current Practices and Suggestions for Future Research. *Applied Measurement in Education, 17*(2), 145-220.

Hambleton, R. K., Jaeger, R. M., Plake, B. S., & Mills, C. N. (2000). Setting performance standards on complex educational assessments. *Applied Psychological Measurement, 24*, 355-366.

Hambleton, R. K., & Pitoniak, M. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp.433-470). Westport, CT: Praeger.

Jaeger, R. M. (1995a). Setting performance standards through two-stage judgmental policy capturing. *Applied Measurement in Education, 8*, 15-40.

Jaeger, R. M. (1995b). Setting standards for complex performances: An iterative, judgmental policy-capturing strategy. *Educational Measurement: Issues and Practice, 14*(4), 16-20.

Kane, M. T. (1994). Validating the performance standards associated with passing scores. *Review of Educational* Research, 64(3), 425-461.

Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 53-88). Mahwah, NJ: Lawrence Erlbaum.

Kane, M. T. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research and Perspectives, 2*(3), 135-170.

Kingston, N. M., Kahl, S. R., Sweeney, & Bay, L. (2001). Setting Performance Standards Using the Body of Work Method. In G. J. Cizek (Ed.), *Setting Performance Standards: Concepts, Methods and Perspectives* (pp. 219-248). Mahwah, NJ: Lawrence Earlbaum Associates.

Lissitz, R. W., & Huynh, H. (2003). *Vertical Equating for State Assessments: Issues and Solutions in Determination of Adequate Yearly Progress and School Accountability*. Practical Assessment Research and Evaluation.

Lissitz, R. W., & Li, F. (2010). *Standard setting in complex performance assessments: An approach aligned with cognitive diagnostic models.* National Council on Measurement in Education, Denver.

Loomis, S. C., & Bourque, M. L. (2001). From tradition to innovation: Standard setting on the National Assessment of Educational Progress. In G. J. Cizek (Ed*.), Standard setting: Concepts, methods, and perspectives* (pp. 175-217). Mahwah, NJ: Erlbaum.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments*. Educational Researcher,* 1994(23), 13-23

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*(1), 3-62.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice, 25*(4), 6-20.

Plake, B. S., & Hambleton, R. K. (2001). The analytic judgment method for setting standards on complex performance assessments. In G. J. Cizek. (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 283-312). Mahwah, NJ: Lawrence Erlbaum.

Plake, B. S, Hambleton, R. K., & Jaeger, R. M. (1997). A new standard setting method for performance assessments: The dominant profile judgment method and some field-test results. *Educational and Psychological Measurement, 57*, 400-411.

Roussos, L. A., DiBello, L. V., Stout, W., Hartz, S. M., Henson, R. A., & Templin, J. L. (2007). The fusion model skills diagnosis system. In J. P. Leighton, & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications*. (pp. 275-318). New York, NY: Cambridge University Press.

Spilsbury, G., Stankov, L., & Roberts, R. D. (1990). The effect of a test's difficulty on its correlation with intelligence. *Personality and Individual Differences, 11*(10), 1069-1077.

Stankov, L. (2000). Complexity, Metacognition, and Fluid Intelligence. *Intelligence, 28*(2), 121-143.

Stankov, L., & Raykov, T. (1995). Modeling complexity and difficulty in measures of fluid intelligence. *Structural Equation Modeling, 2*(4), 335-366.

Waltman, K. K. (1997). Using performance standards to link statewide achievement results to NAEP. *Journal of Educational Measurement, 34,* 101-121.