

EIMAC Assessment Task Force Meeting Washington Marriott April 21, 2008

- Teri to probe states on whether future meetings should include more than the task force ongoing. Expanding the task force is useful for NCES. Is it useful for states?

National Assessment Governing Board Update Charles Smith, Executive Director

- Lou Fabrizio, NC: new member of the Board and recently elected as vice chair of the Committee on Standards, Design, and Methodology (COSDAM), member of Executive Committee
- National Assessment Governing Board Policy Task Force providing state input to NAGB re (1) NAEP reporting, (2) inclusions and accommodations, (3) assessment schedule, (4) grade 12 NAEP, and (5) the reading trend line
- Schedule of assessments uncertain but funding sustained plus funding for additional TUDA districts (7) and Grade 12 states (11) to begin in 2009. \$32 million requested by President
- Inclusions and accommodations - heavy media visibility
- Technical Panel on Preparedness chaired by Mike Hirst, identified 30 research projects. Interim report at August National Assessment Governing Board meeting, draft workplan.
- Two-day Bethesda conference with new TUDA districts
- More media focus on trend lines and disaggregated data. *Atlanta Journal Constitution* reported that students received timely report cards and progress on percent proficient. TUDA allowing more authentic comparisons between urban districts.
- Governor's conversation on NAEP frameworks and assessments, hearing states concerns re AYP.
- Outreach with Council of Great City Schools, governors, and editorial boards reaping meaningful output.

Q and A

- Range of topics within preparedness studies?
 - Content alignment, judgmental studies, statistical relationships, ACT/SAT prior work, college placement tests, patterns in equating college placement tests to success

National Center for Education Statistics (NCES) Update Peggy Carr, Associate Commissioner
--

Trial Urban District Assessment (TUDA)

- No date set for sending notifications regarding availability of data for TUDA.
- Those with new TUDAs in their states need to be prepared; TUDAs do not have a full-time NAEP state coordinator position to make sure they are up to par.
- Must keep TUDA people fully informed; don't want them to answer inquiries incorrectly.

Integrated Computer Tasks (ICTs)

- Only a small sample of students will receive ICTs (no state samples)
- 2009 administration in TUDA. Districts are having trouble getting equipment working.

Mapping Study 2007

- Mapping study will be released on an embargo basis for feedback.

GAO Study on NAEP Inclusion Rates

- GAO study is coming along; responded to feedback/suggestions of assessment task force.

Puerto Rico Results

- Puerto Rico required to participate in 4th and 8th grade assessments. NCES has worked with Title I to get reading assessment exempt. Puerto Rico has very low scores—as low as the lowest scoring district in any state in the rest of the country. Studies into Puerto Rico's assessment showed that there were many missing data points—students simply did not answer lots of questions.
 - Possible solution: include more low difficulty questions so that students have a chance to answer questions they know, rather than leave their tests blank. This allows for measurability of student performance at the bottom end of the distribution (also allows for more accurate measurements; no longer evaluating random guesses).
 - These kinds of modifications may be applicable to TUDA districts, which would allow for more measurement at the tail end of the distribution. It may also be helpful to modify tests in such a way that would allow for more measurement on the higher end.
- Looking for item characteristic curve (ICC) curves in Puerto Rico to make sure they behave the same way as in all other states. NCES will then identify a set of items that they can administer in Puerto Rico for a pilot phase to see if students can be asked more questions that they are able to answer. Identifying items with p values that are not as problematic, looking at item characteristic curve (ICC) to ensure normal behavior as other states. Taking items to Technical Panel in May for psychometric analysis and to identify set of items for Puerto Rico pilot to stabilize estimate with additional items
- Adjusting test as described for Puerto Rico would also bring students into testing process that would normally slide to 2% or other alternative assessments.

Future Assessments

- New reading assessment will be implemented in 2009. NCES is finishing up content alignment study. Items and essays judged in a blind review (presented items/essays and asked judges to assign them to either old or new framework). They will use results of this work to evaluate transition. Independent agency will produce published report on content alignment. Psychometric and statistical information to verify linking strategy. In progress.
- National Indian Education Study will be conducted in 2009. Findings from the 2007 study will be released soon. States involved: AK, AZ, MN, MT, NM, NC, ND, OK, OR, WA. Technical panel's work means significant revisions will be made in 2009.
- History, Civics 2009
- National High School Transcript Study 2009
- Special study to link SAT and NAEP in 2009 for 12th graders by student as requested by the National Assessment Governing Board (Preparedness Panel). Would like to work with ACT but no agreement in place so far

States' Comments

- Some Task Force members raised the issue that loading up tests with easier items may lead to a perception that students are getting accommodations. Also, this may lead to a perception that students will look better than they should next to students that are taking a more difficult test. Concerns about affecting inclusion decisions
- In Puerto Rico, sounds like NCES is estimating a computer adaptive test by targeting items to students that they will do well on. Possibility of NAEP screening - diagnostic items to direct type of booklet to give students. [Response: NCES attempted this with foreign language test. No further plans]
- Why aren't items being developed from the beginning to improve accessibility, decrease construct-irrelevant-variance? [Response: Currently working on, focus on English language learners (ELLs), need to do more than ease language]
- NAEP work timely given peer review on states re updating standards. NCES to produce user's guide/handbook re how NAEP is bridging/building new assessments/frameworks - real value to states! [Response: There will be a field report.]

Reporting the 12th Grade State Assessment Results: Initial Planning **Jay Campbell, ETS**

- NAEP as common barometer between states, 1990
- Improvement in reporting at the state level over time including online tools, e.g., interactive State Reports
- Trial state assessment at 12th grade in 2009 -- 11 states: AR, CT, FL, ID, IL, IA, MA, NH, NJ, SD, and WV
- 12th Grade - more careful attention to what results mean
- New frameworks for reading and math at 12th grade based on both achievement and preparedness
- Anticipating issues

- No trend data - threatens relevance and interest. This may lead to inappropriate state to state comparisons
- Not all states, how to interpret data, some states will perform at the bottom
- Problems with student participation/inclusion, how to interpret findings
- Wendy Grigg, Director of Reporting at ETS
- Small group discussions re these challenges and others in 2009, how to address
- Is this an opportunity to re-tool current tools/reporting strategies?
- Hoping to use state feedback to improve reporting for 12th grade state results and all results

States' Comments: Grade 12 State NAEP

- Revisit achievement level descriptors which were never finalized. Different landscape than 1990 in terms of accountability and assessment systems.
- Link performance over time to provide context by state. Account for differences in population over time, in some states, there have been drastic population differences during this time frame (4th grade-2001, 8th grade-2005, 12th grade-2009)
- Concerns re another mapping study.

States' Comments: Grade 12 Sample

- Who is a 12th grader? Consider credits, time served, on-track for graduation; students with disabilities track is different from others; graduating 11th graders; students spending part of day in college
- Idea: disaggregate by age for 12th graders
- Participation: Who's not participating/contaminating data? Which students and why are some students not taking the NAEP at 12th grade? Some students will not take the test if already in college, may be the wealthier students but would like to be able to identify which students. [Response: College Board descriptive student questionnaire, possible to access database linked to student scores, where students will be going. Later discussion re 12th grade motivation, suspicion around results]
- Motivation: (1) students opting out completely and (2) students taking the test but mentally opting out in terms of motivation.
- Alternative settings more common at high school level (e.g., schools with students with disabilities)
- Students in poverty not accurately reflected by free lunch eligibility at high school level
- How will state NAEP sample feed into national sample? Separate or embedded? [Response: Expansion of national sampling, weighting to account for proportion to be represented in national sample]
- Meaningful/valid comparison between national sample and pilot. Don't want national sample to be 11-state sample

Reporting

- Clarify purpose and future plans. Is NAEP being used as a predictor for success? What's the road leading to? Be transparent.

- Leverage high school transcript study results re 12th grade NAEP reporting. Connect results to course taking patterns, courses offered by school, in order to describe types of schools and students included
- Helping non-pilot states be prepared for release. Reporters will ask why certain states didn't participate
- What do states have on hand re aggregate preparedness to confirm state and NAEP reporting? So much is unknown re 12th grade in terms of how states differ on data on preparedness. Few state-level agreements between K-12 and higher education. Many states don't test 12th graders. Nothing for states to confirm with NAEP data.
- Report in terms of pilot effort/experiment/trial. Emphasize that the test is voluntary. Treat as information gathering. Anticipate way to minimize state to state comparisons
- Tendency for reporters to line up 11 states and compare top and bottom.
- Analyze extent that results are distorted by motivation. Are there decision rules? Will be there caveats reported?
- Suggestion to report first trial in more descriptive way, not focus on achievement levels.

Preparedness

- Student questionnaire: what is currently asked regarding post-high school plans and should there be additional indicators?
- Clarify what students are prepared for (clear definition). Whose definition is it? NAEP's? K-12's? Higher education? May vary across these stakeholders.
- Is there a preparedness continuum? Are we assuming that it's low to high and that students are prepared for all previous points? From low end to high: jobs/apprenticeship, community college, 4-year college (state vs. Ivy), postgraduate, higher. Students prepared for college may not be prepared for an apprenticeship program. This cannot be assumed because different skills are needed for each post-secondary track.
- Alignment between preparedness and student plans. Follow students to determine what they did after high school. Preparedness should be data-driven, based on what students actually do after high school Can student enroll in credit-bearing coursework? Can't know until student in college, lag time.
- Care and thought on explaining preparedness and performance levels
- First time reporting outside of subject area being assessed. Preparedness more holistic than math/reading preparedness.
- Response from ETS and National Assessment Governing Board
 - Point taken about being clear about what prepared for. Preparedness will be restricted to academic (reading/math) skills, not soft skills. Defining as eligible for entry into credit-earning coursework or job training program associated with the workplace in the subject area.
 - Several approaches for indicators of preparedness, "reference points on the NAEP scale"; judgmental studies on standards from workplace/college; statistical studies (e.g., SAT/NAEP); college placement scores from various kinds of colleges/universities.
 - National Assessment Governing Board may choose several reference points or one indicator. Haven't collected any data yet. First set of studies: content alignment, pilots for judgmental studies and statistical relationship studies with SAT and hopefully military, workplace.

- In fall, perhaps focus major discussion on this issue with Susan and Charles.
- Assumption but hasn't been decided whether it will be achievement levels plus preparedness indicators as reference points on the scale.
- Timeline: not set, after August 2009 will have data.

12th Grade Motivation Studies

Maria Ivancin, American University

Current Studies: 4 current studies on 12th grade motivation due to effects of low participation/motivation on validity of 12th grade assessment

1. psychometric (existing IRT theory; Murray Aikens, University of Melbourne)
2. data analysis (existing data on omit rates and skip patterns re inferences on student motivation; Lynn Stokes, Southern Methodist)
3. experimental design (impact of varying levels of motivation on achievement, data collection complete in fall 2007, final report expected in May 2008; Henry Braun and Irwin Kirsch)
4. qualitative (led by Maria Ivancin) Student focus groups in 5 cities: Pittsburgh, Jackson, Albuquerque, Milwaukee, Sacramento. Sample diversity re ethnicity; urban/rural/suburban schools, post-high school plans; gender; class ranking. Explored motivation on no-stakes assessments and timing of NAEP in spring of senior year

Qualitative Study Findings

- Students excited about graduation, future. Anxiety re unknown. Some anxious to move on
- Proud of accomplishment. High status in school as senior
- Finished with requirements, did not have to work as hard. Expectations still high from teachers
- Didn't choose high school, lack of strong connection to high school. Strong ownership tended to be related to sports, other students, teachers, not school itself
- How motivated: internally, competition with others/themselves, recognition, fear of disappointment by family/friends. Not motivated by fear of punishment. Didn't want to be told what to do
- Most important influencer: parents. For some: boss, siblings, other relatives/friends/personal relationships
- Principals negative role for some, lack of personal relationship
- Students do not go to parents for help with school; students felt ownership of their work and didn't want to trouble their parents with issues going on with school.
- Standardized testing happens prior to senior year (or early in year, e.g., SAT), used to it, past state requirements
- Testing anxiety re timing constraints
- Like feedback from tests, like challenge and ability to demonstrate what you know
- Very little awareness of terms: NAEP, National Assessment of Educational Progress, or Nation's Report Card

- Aware that states compared and aware of international comparisons. Sensitive to state's fit within these comparisons. Inferiority complex for students in NM, MS. Felt affiliation with/representative of state
- Responded positively to concept of NAEP after reading description. Understood importance of NAEP, need for strong participation/engagement
- Better participation/engagement: early awareness of test. Felt that test sprung on them, didn't know what it was. Engagement less difficult to overcome than participation. Use respected teacher to promote. Self respect: will do good job if know that it's important. Wanted better sense of subject matter, kinds of questions
- Format fine: 90-minute, pencil/paper
- Biggest barrier: Spring timing is too late, major reason not to participate, thinking about other things and requirements are done. Suggested October or November of senior year as more appropriate
- Students did not want to be pulled out of classes because didn't want to have to make up class. Needs to be during regular school hours. Students wanted flexibility in schedule: which class to be pulled out of, time of day, and no make-up
- Did not see purpose without individual results. Wanted to use on college application if did well
- Different students wanted different kinds of recognition: at graduation, on college application, certificate signed by governor/President, part of community service requirement. No stickers/buttons/ribbons
- Little universal suggestions provided re incentives/motivation, students had diverse points of view. Incentives: food at test site, money or gift card if \$50 or more, coupon for prom ticket because more expensive. Not supported: restaurant gift cards, gift cards/coupons problematic based on which store, discounts/coupons for school events would not appeal to all

Q and A

- Affect of moderator being enthusiastic/knowledgeable about NAEP. During actual test administration, principal/teacher/coordinator may not have that enthusiasm [Response: Written description very objective, handed out not read aloud. May have conferred enthusiasm just be focusing on the topic]
- Student selection [Response: Screening questionnaire to ensure quota on types of students, recruited by local focus group facilities, limited number of students per school]
- No dramatic differences based on demographics except NM/MS. Students in these states felt that the state always does badly, students wanted to help

NAEP SES Research Marilyn Binkley, NCES
--

- Current measure (National School Lunch Program, literary items, and parent education reported by student) neither reliable nor valid
- Free lunch eligibility not representative at high school level
- 2007 study components:

- original background variables (BV): race/ethnicity, home educational resources, parent education, National School Lunch Program (NSLP)
- extended/enhanced background questionnaire (EBQ)--would like feedback on items, revising items:
 - home composition
 - living in multiple homes
 - number of siblings
 - parent employment
 - household items
 - parent education
 - home ownership
 - parent education, parent occupation, household income, and household size from (1) Census 2000 long form--detailed, once every 10 years, labor force, income, wages, investments; and (2) American Community Survey--annual, sensitive to changes in community, aggregate of five years of data, will have summary of national data in 2009
- Early Childhood Longitudinal Study (has a parent survey, used to compare old and new background variable)
- geocoding (mapping student's address onto Census). To prevent NAEP's concerns about confidentiality of student addresses, state sends file including geocode composed of codes for state, district, race, and average data from 10 other students with same codes

Findings

- BV: some items function well but not all should be retained
- EBQ: new parent education measure is an improvement but not all indicators of parent wealth are stronger as validated by Census
- Geocoding: software is adequate but need better matching
- Haven't sampled grade 12 yet, some indicators only at grade 8
- Examining whether enough information from just zip code instead of extra step of having state develop block group
- Looking at overlap with ECLS-K to confirm results
- Is there a single SES measure on a continuum, constant across 4th, 8th, and 12th grades?

Proposed 2009 Study

- Improve geocoding
- How quickly and accurately can Census turn this around?
- Look at BV and EBQ, proposing to drop a number of items from BV and EBQ because outdated and not explaining variance
- Pilot in grades 4, 8, and 12

Q and A

- No decisions made. Considering variety of options: deciles, quartiles, combination of NSLP and other. Researchers don't want composite indicator to allow own manipulations. Related to what research says on nature of relationship between SES and achievement, e.g., conversations between parents and children at different SES levels

- Improvement over adjusted gross income by adding so many indicators? Confirmed by existing Census data. At block group, major variations not detected at community level (e.g., Bethesda-Chevy Chase neighborhood). In rural area (e.g. upstate New York), 10 Census tracts, picked up a lot of variability between block groups, more than from zip code. Not picking up Vermont area nuances between homes with a view (wealth) or on the river (poverty). Argument for using block group but must be balanced by operational hassle
- Implications for instituting a new SES measure? New trend line? [Response: Would like state input on these implications. If implement new measure, don't have to drop NSLP. Do others outside of this group need to be aware of this conversation, and if so, who?]
- Shift from income to more nuanced? If present to other groups, emphasize this shift, back to early notions about SES being more than just income, income as one component [Response: Literature refers to education, occupational prestige/status, income, and wealth. NSLP was best proxy but now other indicators under consideration, getting closer to research on SES, may need to weight certain factors]
- Reactions to EBQ pilot? [Response: Resistance anytime anything is added. Goal is to get rid of EBQ. After 2009, will be rolled together, question of what to keep/drop e.g., one set of parent education questions instead of two. When certain about what can get from Census, may be able to drop other indicators, need to be guided by actual data]
- Collect data on how many states tap into Census data? [Response: No states using microdata, able to match at fine level. Census interested re usefulness of American Community Survey]
- Talk to data folks re accurate student address: policy and burden. [Response: E-filing moving to state level, consideration about state burden, value added]
- Parent concerns re personal questions about parents. States using NSLP to avoid these concerns. [Response: If geocoding is best indicator (this has not been determined), will not need to have students report, less invasive.

NAEP Interactive Computer Tasks

Jay Campbell, ETS and Dianne Walsh, Westat

Jay Campbell

- Science interactive computer tasks (ICTs) at grades 4, 8, and 12 implemented in 2009
 - 1 - extended = 30 minutes. Students demonstrate knowledge and engage in inquiry. Answer multiple choice and constructed response. Conduct science experiments. Manipulate, observe, explain. Will capture every keystroke/way in which students interact with task. Report processes to public.
 - 2 - short = 15 minutes. Abbreviated experiments. Simulation of lab equipment. Video clip on events occurring over weeks. Multiple choice and constructed response. Not capturing online behaviors.
- Measurement challenges
 - Developing tasks that match the framework: building tasks that measure both science content and process; developing substantive tasks that can be completed within 30 minutes

- Developing interfaces that are appropriate and engaging and that do not introduce construct irrelevance or take too long to learn
- Anticipate changes prior to administration in 2009
- Challenges re collecting data on content and process
- Already know not enough time to complete tasks
- Need collaborative input (ETS, NCES, NESSI, Westat, Pearson, Fulcrum IT, and external advisors) as part of development discussion. Different from process of developing paper and pencil items
- Small 2008 pilot (273 schools): complicated challenges at school level re hardware, firewalls, connectivity; security protocol too sensitive
 - Two goals
 - Pilot test tasks: evaluate student engagement in and understanding of tasks, try out scoring processes, evaluate item performance.
 - Learn about issues related to delivery of “simulation-based assessments” on nationally representative samples.
- Adjustments made to increase standard implementation. Removed reliance on school hardware and connectivity
 - Desktop settings less sensitive
 - Switched from internet to “NAEP on a stick” -- both test and responses on flash drives
 - Using NAEP laptops
 - IT support for field staff
- 2011 state-by-state, large scale implementation in reading
- More input from external panels, states, ETS Design and Analysis Committee (DAC), Massachusetts Institute of Technology (MIT) to revamp tasks as needed over next 6 months; lock-down content no later than September 1 so that enough time is given for stress testing (want 100% success rate for 2009).
- Expect special report in 2009 re learnings after main release of reports

Dianne Walsh

- Staffing
 - Experience with schools and students
 - Basic network infrastructure
- Training
 - PowerPoint, live demo online, lab simulation, addressing common failures in the lab
 - Help Desk, field staff for troubleshooting, Pearson technical professionals
- Resources required at school
 - School/district computer specialist during planning visit and on standby on test day
 - Access to computer lab
 - Debriefing with school staff, students, and school coordinator. Positive about student engagement but negative about time occupying computer lab and lack of notice

Q and A

- NAEP laptop and flash drive for each student: one extended and two of three possible short tasks
- 2009 pilot: 4,000 students per grade in next pilot
- Maine implementing online assessments. Similar challenges despite same hardware for each student
- Suggest practice test available prior to test administration in order to familiarize students with navigation
- Encrypted flash drives for security, similar security to test booklets; using NAEP hardware is additional security
- Tasks demonstrated were the only ones developed to date. Need to develop number of items in order to have longitudinal data over time.
- Cost estimate is high due to data collection and technicians. State concern about moving forward with more schools without getting bugs out (98% success rate) at least at one grade level. Assessing one grade makes sense especially at high school level. At lower grades, makes sense to do hands-on task. Hoping for interim report because state being pressured to move to ICT [Response: Would like to share interim report. In addition to ICTs for science, there is a paper and pencil component and hands-on tasks (HOTs). Governing Board wanted NCES to try with all three grade levels (4, 8, and 12). Governing Board will determine amount of money that can be spent]
- Problem on task on soil porosity: samples were switched around
- Will you move forward even if not more effective than paper and pencil? What is the rationale for ICT? [Response: not worthy unless something learned through process. Not entirely empirical. Should it be well-correlated than paper and pencil, should it be un-correlated? Need to analyze data to know]
- Would like update at fall meeting re how students performed
- Helpful to states that NCES is leading this pilot, potential for more authentic assessment. Capitalize on what you can measure through manipulation and understanding students' thought processes instead of what you can measure on paper
- State in pilot: extremely high student engagement, high student preference
- Addressing requirement expected that science NAEP will measure inquiry
- Follow up in the future