Results and Lessons Learned from the NAEP Problem Solving
in Technology-Rich Environments Study[1,2,3]

Randy Bennett
Hilary Persky
Andrew Weiss
and
Frank Jenkins[4]

ETS
Princeton, NJ 08541
USA

September 2008

Abstract

The Problem Solving in Technology-Rich Environments Study was the last in a series of three field investigations funded by the National Center for Education Statistics to explore the use of new technology in the US National Assessment of Educational Progress (NAEP). This demonstration study evaluated the performance of nationally representative samples of 8th graders on two computer-delivered, extended problem-solving scenarios. One scenario measured skill in electronic information search and the other assessed skill in using "what-if" simulation to discover physical relationships. Each scenario was taken by a different sample of approximately 1,000 students. Performance was judged by evaluating the quality of answers to open-ended and multiple-choice questions, and by assessing aspects of the process used to reach those answers. Study results were related to instrument functioning and to the relative performance of population groups.

Results and Lessons Learned from the NAEP Problem Solving
in Technology-Rich Environments Study

The Problem Solving in Technology-Rich Environments (TRE) study was the last of
three field investigations in the National Assessment of Educational Progress (NAEP)
Technology-Based Assessment Project, which explored the use of new technology in
administering NAEP. The TRE study was designed to demonstrate and explore an
innovative use of computers for developing, administering, scoring, and analyzing the
results of NAEP assessments. The prior two studies, Mathematics Online (MOL) and
Writing Online (WOL), compared online and paper testing in terms of issues related to
measurement, equity, efficiency, and operations.

In the TRE study, two extended scenarios were created for measuring problem solving
with technology. These scenarios were then administered to nationally representative
samples of students. The resulting data were used to describe the measurement
characteristics of the scenarios and the performance of students.

The context for the problem-solving scenarios was the domain of physical science. The
TRE *Search* scenario required students to locate and synthesize information about
scientific helium balloons from a simulated World Wide Web environment. The TRE
Simulation scenario required students to experiment to solve problems of increasing
complexity about relationships among buoyancy, mass, and volume; students viewed
animated displays after manipulating the mass carried by a scientific helium balloon and
the amount of helium contained in the balloon. Both scenarios targeted grade 8 students
who were assumed to have basic computer skills; basic exposure to scientific inquiry and
to concepts of buoyancy, mass, and volume; and the ability to read scientifically oriented
material at a sixth-grade level or higher.

In the TRE study, data were collected from a nationally representative sample of grade 8
students in the spring of 2003. Over 2,000 public school students participated, with
approximately 1,000 students taking each assessment scenario. Students were assigned
randomly within each school to one of the scenarios—Search or Simulation. Students
took the scenarios on school computers via the World Wide Web or on laptop computers
taken into the schools. For both scenarios, data were collected about student
demographics; students' access to computers, use of computers, and attitudes toward
them; and students' science coursetaking and activities in school.

Methodology

The TRE study used Evidence-Centered Design (ECD) (Mislevy, Almond, and Lukas
2003) to develop the interpretive framework for translating the multiplicity of actions
captured from each student into inferences about what populations of students know and
can do. (Populations were the targets of inference because NAEP does not award scores
for individual students.) In ECD, the key components of the interpretive framework are
student and evidence models. The student model represents a set of hypotheses about the
components of proficiency in a domain and their organization. The evidence model

shows how relevant student actions are connected to those components of proficiency, including how each relevant action affects belief in student standing on each proficiency component. The structure provided by ECD is particularly important for complex assessments like TRE, for which meaningful inferences must be drawn based on hundreds of actions captured for each student.

For the purposes of TRE, the student model represented the components of student proficiency in the domain of problem solving in technology-rich environments. Two primary components were postulated: scientific inquiry and computer skills. Scientific inquiry was defined as the ability to find information about a given topic, judge what information is relevant, plan and conduct experiments, monitor efforts, organize and interpret results, and communicate a coherent interpretation. Computer skills were defined as the ability to carry out the largely mechanical operations of using a computer to find information, run simulated experiments, get information from dynamic visual displays, construct a table or graph, sort data, and enter text.

Evidence of these skills consisted of student actions called "observables." Observables were captured by computer and judged for their correctness using scoring criteria called "evaluation rules," and summary scores were created using a modeling procedure that incorporated Bayesian networks (Mislevy et al. 2000). Bayesian networks belong to a class of methods particularly suited to the TRE scenarios because these methods account for multidimensionality and local dependency, neither of which is explicitly handled by the measurement models typically used in NAEP assessments.

The TRE Scenario Scores and Results

Because the TRE study used measures that are experimental, data were analyzed to explore how well the TRE scenario scores captured the skills they were intended to summarize. For each scenario, the following measures were obtained from scores on the scenario: internal consistency; the relations of student scores to students' prior knowledge; the TRE scale intercorrelations; the correlations of each observable with each subscale; the locations of the observables on the scales; the response probabilities for prototypic students (i.e., hypothetical students with low, medium, and high levels of proficiency); and the relations of relevant student background information to performance.

Readers are reminded that the TRE project was intended as an exploratory study of how NAEP can use technology to measure skills that cannot be easily measured by conventional paper-and-pencil means. Because the results pertain to student performance in only two scenarios employing a limited set of technology tools and range of science content, results cannot be generalized more broadly to problem-solving in technology-rich environments for the nation's eighth-graders.

The Search Scores and Results

TRE Search consisted of 11 items (or observables) and produced a total score and two subscores, scientific inquiry and computer skills. Scores were reported on a scale with a mean of 150, standard deviation of 35, and range from 0 to 400.

- The internal consistency of the three TRE Search scores (total, scientific inquiry, and computer skills) ranged from .65 to .74, as compared to .62 for the typical main NAEP science assessment hands-on task block, which, although measuring skills different from TRE, also includes extended, problem-solving tasks.
- The Search scores appeared to provide separable information; the (disattenuated) intercorrelation of the subscores was .57. This value contrasts with intercorrelations of .90 to .93 for the main NAEP science assessment scales.
- The scientific inquiry skill scale score was most related in the student sample to the following scale observables: the relevance of the World Wide Web pages visited or bookmarked, the quality of the constructed response to a question designed to motivate students to search for and synthesize information from the Web, and the degree of use of relevant search terms (*r* range between performance on the observable and scale score = .51 to .71).
- The computer skills scale score was related in the student sample primarily to the following scale observables: the use of hyperlinks, the use of the Back button, the number of searches needed to get relevant hits (an efficiency measure), and the use of bookmarking (*r* range = .60 to .69).
- Statistically significant differences in performance were found on one or more TRE Search scales for NAEP reporting groups categorized by race/ethnicity, parents' highest education level, students' eligibility for free or reduced-price school lunch, and school location. These differences were in the same direction as those typically found on NAEP assessments in such logically related areas as reading and science. No significant differences were found, however, for reporting groups categorized by gender.

The TRE Simulation Scenario Scores and Results

The TRE Simulation scenario consisted of 28 observables and produced a total score and three subscores: scientific exploration, scientific synthesis, and computer skills. Scores were reported on a scale with a mean of 150, standard deviation of 35, and range from 0 to 400.

- The internal consistency of the four scores ranged from .73 to .89, as compared to .62 for the typical main NAEP science assessment hands-on task block, which, although measuring skills different from TRE, also includes extended, problem-solving tasks.
- The Simulation scores appeared to provide separable information; the (disattenuated) intercorrelations of the subscores ranged from .73 to .74. These values contrast with intercorrelations of .90 to .93 for the main NAEP science assessment scales.

- The scientific exploration skill scale score was most related in the student sample to three scale observables: which experiments students chose to run to solve the Simulation problems, whether students constructed tables and graphs that included relevant variables for solving the problems, and the degree to which experiments controlled for one variable in the one problem demanding controlled experimentation.
- The scientific synthesis scale score was primarily related in the student sample to the degree of correctness and completeness of conclusions drawn for each Simulation problem.
- Performance on the computer skills scale was related in the student sample mainly to the number of characters in the written responses students gave for each of the three Simulation problems.
- Statistically significant differences in performance were found on one or more TRE Simulation scales for NAEP reporting groups categorized by race/ethnicity, parents' highest education level, and students' eligibility for free or reduced-price school lunch. These differences were in the same direction as those typically found on NAEP assessments in such related areas as reading and science. No significant differences were found, however, for reporting groups categorized by gender or school location.

Conclusion

The TRE study was able to define and create measures of problem solving with technology; successfully deliver those measures to nationally representative samples of 8th grade students on computer; produce scores that behaved in reasonable ways psychometrically; and provide results for population groups that were basically consistent with the performance of those groups on related NAEP assessments. These study outcomes suggest that we can successfully measure aspects of 21st century skills that cannot be measured on paper.

Although the TRE study achieved its goals as a demonstration effort, moving to an operational assessment of problem solving with technology will be extremely challenging. It will be extremely challenging because there is no NAEP framework (or other widely accepted syllabus) in which to ground test development; designing performance tasks for computer is a relatively new activity and there is little knowledge among exam developers about how to do it effectively and efficiently; many schools still do not have the technology infrastructure needed to deliver these measures to large numbers of students efficiently and securely; and students produce extensive information when taking such tests. We are just beginning to learn which of the literally hundreds of pieces of information produced is worth attending to and how to defensibly score the information that does prove to be worthwhile.

While measuring problem solving with technology is a considerable challenge, it is also true that the need for such measures is not going to diminish. On the contrary, the need for such measures will only grow as technology becomes increasingly central to survival in a global economy. Consequently, if policy makers are to make sensible decisions

about how to improve education, examination boards will need to learn how to assess this problem-solving-with-technology skill, and related 21st century proficiencies, that can't be effectively measured on paper. Finally, we will never learn to assess these emerging skills if we are not willing to invest the time, effort, and money in trying and, sometimes, in failing.

References

Bennett, R.E., Persky, H., Weiss, A.R., and Jenkins, F. (2007). *Problem Solving in Technology-Rich Environments: A Report From the NAEP Technology-Based Assessment Projec*t (NCES 2007–466). U.S. Department of Education. Washington, DC: National Center for Education Statistics. Retrieved June 30, 2008 from http://nces.ed.gov/nationsreportcard/pdf/studies/2007466.pdf

Mislevy, R.J., Almond, R.G., and Lukas, J.F. (2003). *A Brief Introduction to Evidence-Centered Design* (RR-03-16). Princeton, NJ: Educational Testing Service.

Mislevy, R.J., Almond, R.G., Yan, D., and Steinberg, L.S. (2000, March). *Bayes Nets in Educational Assessment: Where Do the Numbers Come From?* (CSE Technical Report 518). Retrieved January 25, 2005, from http://www.cse.ucla.edu/CRESST/Reports/TECH518.pdf.