J·T·L·A

The Journal of Technology, Learning, and Assessment

Volume 8, Number 8 · June 2010

Measuring Problem Solving with Technology: A Demonstration Study for NAEP

Randy Elliot Bennett, Hilary Persky,
Andy Weiss, & Frank Jenkins

www.jtla.org

A publication of the Technology and Assessment Study Collaborative
Caroline A. & Peter S. Lynch School of Education, Boston College

J·T·L·A

Volume 8, Number 8

Measuring Problem Solving with Technology: A Demonstration Study for NAEP

Randy Elliot Bennett, Hilary Persky, Andy Weiss, & Frank Jenkins

Editor: Michael Russell
russelmh@bc.edu
Technology and Assessment Study Collaborative
Lynch School of Education, Boston College
Chestnut Hill, MA 02467

Copy Editor: Jennifer Higgins
Design: Thomas Hoffmann
Layout: Aimee Levy

JTLA is a free online journal, published by the Technology and Assessment Study Collaborative, Caroline A. & Peter S. Lynch School of Education, Boston College.

Preferred citation:

Bennett, R.E., Persky, H., Weiss, A., & Jenkins, F. (2010). Measuring Problem Solving with Technology: A Demonstration Study for NAEP. *Journal of Technology, Learning, and Assessment, 8*(8). Retrieved [date] from http://www.jtla.org.

J·T·L·A

Abstract:

This paper describes a study intended to demonstrate how an emerging skill, problem solving with technology, might be measured in the National Assessment of Educational Progress (NAEP). Two computer-delivered assessment scenarios were designed, one on solving science-related problems through electronic information search and the other on solving science-related problems by conducting simulated experiments. The assessment scenarios were administered in 2003 to nationally representative samples of 8th-grade students in over 200 schools. Results are reported on the psychometric functioning of the scenarios and the performance of population groups. Implications are offered for using online performance assessment to measure emerging skills in NAEP and other large-scale testing programs.

J·T·L·A

Measuring Problem Solving with Technology: A Demonstration Study for NAEP[1]

Randy Elliot Bennett
Hilary Persky
Andy Weiss
ETS
Frank Jenkins[2]
Westat

Introduction

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 NAEP. The first two studies, Mathematics Online (MOL) and Writing Online (WOL), looked at the impact of delivering existing paper tests on computer, especially with respect to differences in psychometric functioning (Bennett, Braswell, Oranje, Sandene, Kaplan, & Yan, 2008; Horkay, Bennett, Allen, Kaplan, & Yan, 2006). The TRE study, in contrast, was intended as a demonstration of an instrument uniquely suited to the computer.

Four main intentions shaped the study. Those intentions were to demonstrate an instrument that:

1. *Measured important skills not easily tested on paper*. This intention was chosen to ensure that the instrument targeted substantively worthwhile proficiencies—i.e., ones NAEP constituencies would likely care about and that could not be adequately assessed in the traditional mode.
2. *Could be delivered successfully on computer by NAEP to 8th graders in a sample of schools throughout the nation*. Given that the instrument targeted worthwhile proficiencies, it had to be feasible to administer to at least a small national sample. Such an administration would help give an indication of logistical issues, as well as provide preliminary data for analysis.

3. *Held together reasonably well psychometrically*. Devising a substantively meaningful measure and successfully administering it do not guarantee that the instrument will function well technically. To demonstrate sound measurement, evidence should be provided to support at least a basic level of technical functioning.
4. *Produced credible results*. If the basic technical evidence turns out to be generally supportive, a final requirement is for results to be sensible. In particular, they should be consistent with what we know of student performance on other NAEP assessments or on additional measures of similar quality.

In this paper, we discuss these four intentions, their related outcomes, and some associated issues; review what the study did and did not do effectively; suggest how measures like those created for TRE might be used in NAEP; and offer some closing comments, including lessons learned.

Four Intentions

1. Demonstrate an Assessment that Measured Important Skills Not Easily Tested on Paper

For this demonstration project, we chose to focus on "problem solving with technology" because that skill seemed important by virtue of what workers in a knowledge economy, or students in higher education, must know and be able to do to succeed in a 21st Century world and because, by definition, that skill cannot be easily measured on paper. When planning for the study began in 2000, there was no NAEP content framework for problem solving with technology, so that skill needed to be conceptualized. In keeping with the nature of a demonstration project, the construct definition created is more an illustration of how the larger domain, and specific constructs within it, might be defined than a basis for an operational assessment. The process used obviously did not involve the extensive review of literature, analysis of curricula, or input from diverse constituencies that creation of a NAEP framework might entail.[3]

For this project's purposes, we conceptualized problem solving with technology as resulting from the intersection of *technology environment* and *content domain* (Figure 1, next page). Technology environment included the various software tools that might be brought to bear in problem solving, including such tools as databases, text editors, simulations, dynamic displays, and spreadsheets. Content domain included not only the problems that characterize a subject matter like biology, chemistry, physics, and his-

tory, but also the problem-solving processes commonly employed, which may differ somewhat from one domain to the next.

Figure 1: A Domain Conception for "Problem Solving with Technology"

	Technology Environment						
Content area	**Database**	**Text editor**	**Simulation**	**Dynamic visual display of information**	**Interactive feedback**	**Spread-sheet**	**Presentation and communi-cation tools**
Biology							
Ecology							
Physics							
Balloon science							
Economics							
History							

Note. The shaded area indicates the coverage of the scenarios.

One can assess proficiency in this universe in several ways. For example, one can emphasize technology environment. Following that approach, one might create a test by choosing a technology environment, like database, and asking students to use a search engine to locate information in response to a series of questions, each taken from a different content domain. Alternatively, one can sample from all sensible content-area by technology-environment pairs. In this approach, one might create a test that asks students to solve a biology problem using database search, a physics problem with a simulation tool, and an economics problem with a spreadsheet.

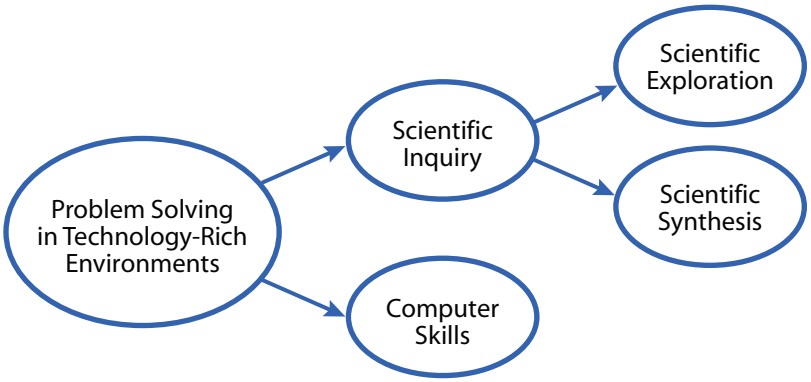
We chose a third approach which was to emphasize content domain in keeping with the view that, in real-world settings, problem solving with technology should be driven by the problem, not by the technology. We therefore selected a domain segment and then asked students to use more than one technology tool to respond to problems within that single domain segment. The domain segment we chose was the science surrounding helium gas balloons. We chose that segment because it represented a real application of fundamental physical principles, like mass and volume, and because we thought it would be interesting to 8th-grade students.

In our view, then, "problem solving with technology" results from the intersection of content-related and computer-related proficiencies. For this demonstration project, the key content-related proficiency we chose to emphasize was the problem-solving process of "scientific inquiry." We defined scientific inquiry narrowly to include being able to find information about a given topic, judge what information is relevant, plan and conduct experiments, monitor one's efforts, organize and interpret results, and communicate a coherent interpretation. Scientific inquiry was further separated into scientific exploration, intended to capture the cognitive activities involved in generating results, and scientific synthesis, primarily intended to reflect the organization of results in a meaningful response.

We defined computer-related proficiency (hereafter called "computer skills"), as being able to (fluently) carry out the largely mechanical operations of using a computer to do scientific inquiry; that is, find information, run simulated experiments, get information from dynamic visual displays, construct a table or graph, sort data, and enter text. From this perspective, the computer is a tool that helps individuals carry out cognitive activity in service of domain-related goals.

Figure 2 shows the relationships among these various proficiencies.

Figure 2: Problem Solving in Technology-Rich Environments and its Hypothesized Sub Proficiencies



It is important to note that although our conception of problem solving with technology included a component called, "scientific inquiry," the formulation of scientific inquiry embodied in TRE was a partial one (Olson & Loucks-Horsley, 2000, pp. 28-30). Full inquiry gives greater attention to question choice, explanations, and connections of those explanations with scientific knowledge than could be achieved in this project. The use of partial rather than full inquiry was consistent with the purpose of TRE, which was not as a science assessment on computer, but as a test of skill in using the computer for problem-solving (in a science-related context).

To measure problem solving with technology, we used two assessment "scenarios" called, Search and Simulation. Each scenario attempted to assess a different (but small) subset of the elements comprising our conception of problem solving with technology; that is, each scenario targeted only some of the components of scientific inquiry and computer skills delineated above. Each scenario contained extended tasks offering multiple opportunities to observe student behavior, and each scenario tried to more faithfully represent than do traditional tests the types of challenges individuals encounter in work and advanced academic settings.

The Search Scenario

The Search scenario presented the student with an environment for locating information electronically. The student was expected to employ the environment over a 40 minute period to answer one constructed-response question and four multiple-choice questions related to the uses and science of gas-balloon flight.

The design of the Search scenario was based on prior work that looked at the information-search behavior of adults and young adults. This research literature offers some initial ideas as to what effective electronic information-search behavior might include (Fidel et al., 1999; Klein, Yarnall, & Glaubke, 2001, 2003; Salterio, 1996; Schacter, Chung, & Dorr, 1998). For example, individuals who are proficient at electronic information search appear to use, more often than less proficient individuals, such mechanisms as quotes or the "not" operator to reduce the number of irrelevant results. Proficient individuals also appear to frequently use the "Back" button to return to pages already visited, including to the listing of Search results. Finally, proficient searchers tend to use queries that are more precisely targeted to the topic of interest.

In addition to this research literature, the design of the Search scenario was based on standards for students' science and technology skills, which argue for the importance of electronic information search in academic and job environments. Among these standards were the National Academy of Sciences' *National Science Education Standards* (NAS, 1996), the

International Society for Technology in Education's *National Educational Technology Standards for Students* (ISTE, 2000), and the US Education Department's *National Educational Technology Plan* (Riley, Holleman, & Roberts, 2000).

Figure 3 shows the Search interface.[4] Students were introduced to this interface through a brief tutorial. Although the interface was designed to be as close to a standard Web browser as possible, some features—such as buttons for reading test directions and for entering answers—were particular to the TRE software.

On the left of the screen is a problem statement, which asks the student to find out and explain why scientists sometimes use helium gas balloons in place of other mechanisms like rockets and satellites for planetary space exploration. Below that problem statement is a summary of directions that students saw on earlier screens. To the right is a Web browser showing a search page into which the student may enter queries. Above the search page is a set of tools. The tools allow the student to go to pages already visited, return to the search page, bookmark, view the more detailed set of directions, get hints about how to solve the problem, and go to a form where notes or an extended response to the motivating question may be entered.

Students conducted their searches using a simulated World Wide Web. A simulated Web was chosen for two reasons. First, it was chosen to increase standardization because both the day the test was given and school-technology policy could affect what parts of the real Web were available to individual students. Second, a simulated Web was used to prevent visits to inappropriate sites from occurring under the auspices of NAEP.

The database that was used to populate this simulated Web consisted of some 5,000 pages pulled from the real World Wide Web. These pages included both relevant and irrelevant material. The information needed to answer the assessment questions was not available on any one page, so the student had to locate and visit multiple relevant pages and synthesize information across them to correctly respond.

To evaluate the relevance of the Web pages used to populate this database, which factored into scoring student performance, all pages were rated by a single judge for pertinence to the motivating problem on a 1–4 scale, where 4 indicated the most relevance. All pages designated as relevant or partly relevant (2, 3, or 4) were independently rated again by two other judges, with differences resolved by consensus. Pages assigned a "1" were not independently re-rated because these totally irrelevant pages were very easy to identify objectively (e.g., pages about party balloons).

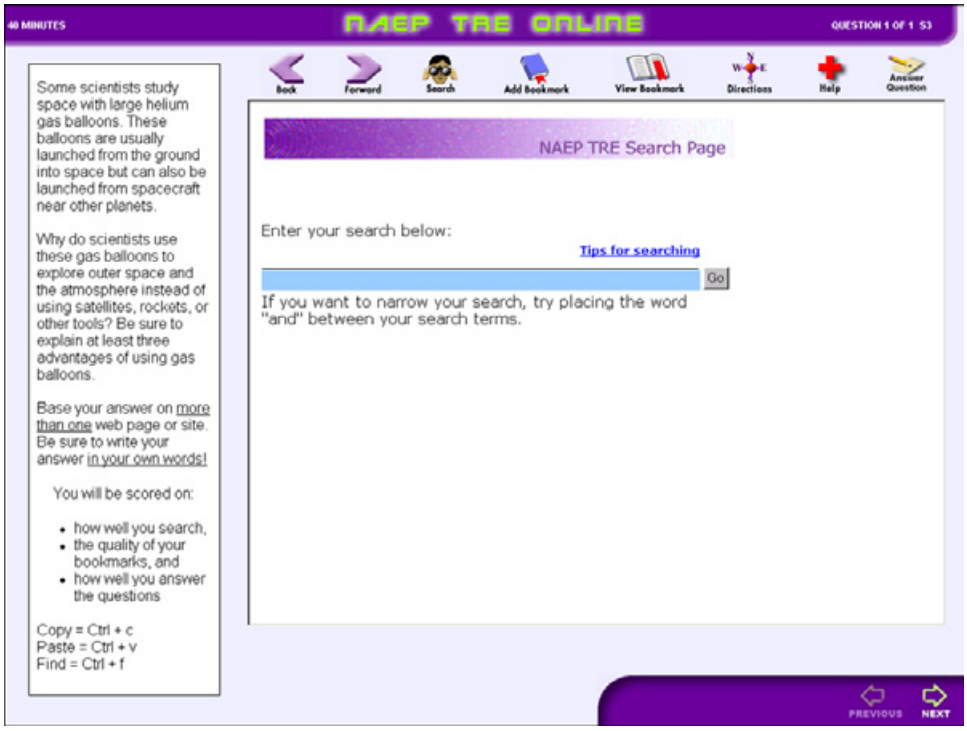
From the definition of problem solving with technology delineated earlier (Figure 2, page 7), five hierarchically organized scales were intended to be derived: a Total score scale segmented into Computer Skills and Scientific Inquiry, with the latter, in turn, segmented into Scientific Exploration and Scientific Synthesis. Preliminary analysis of the TRE Search data, however, suggested that there were too few items to sustain separate Scientific Exploration and Scientific Synthesis scales. As a consequence, only three scales were created: a Total score scale, a Scientific Inquiry subscale (combining the Scientific Exploration and Scientific Synthesis items), and a Computer Skills subscale.

Scores for each scale were generated not only from the constructed-response and multiple-choice answers students offered, but also from students' problem-solving behavior. Making sense of their problem-solving behavior, however, proved to be an enormous task because, in a computer-delivered test, every keystroke, every mouse click, and every resulting event can be recorded, only some of which turn out to be important. As a consequence, we used logical analysis, the results of previous research, and analysis of pilot-test data to determine which indicators should be employed as evidence of problem solving with technology. We judged each piece of evidence according to a rubric and combined the pieces—including students' constructed-response and multiple-choice answers—to form the three scores.

For Computer Skills, evidence of proficiency included the following six items or "observables:"[5]

- Use of advanced search techniques like quotes or the "not" operator to better focus search results;
- Use of the Back button to reorient or return to pages already visited;
- Number of searches for relevant hits (an indicator of search efficiency)
- Use of hyperlinks to dig down within a website to gather more information;
- Use of bookmarking to save pages so that information deemed relevant could be easily retrieved; and
- Use of deletion for unwanted filed pages to limit bookmarks to a manageable list.

Figure 3: The Search Scenario Interface



For the Scientific Inquiry score, the following five observables were taken as evidence of proficiency:

- Use of search terms relevant to the problem at hand;
- Average relevance of hits returned by the student's queries;
- Relevance of pages visited or bookmarked because, while generating relevant results is important, students must be able to distinguish those results from the irrelevant ones that will also inevitably be returned in any search list;
- Accuracy and completeness of the answer to the constructed-response question; and
- Number right on the four synthesizing multiple-choice questions that concluded the scenario and asked for other factual information about helium-gas balloons that also could be found on Web pages in the database.

Each of the pieces of evidence listed above was judged using a rubric that was pilot tested, empirically evaluated, and revised as appropriate. A simple example is shown in Table 1 (next page).

Table 1: A Rubric for Evaluating Bookmarking as Evidence of Computer Skill

If two or more pages were bookmarked, give full credit (2 points).
If only one page was bookmarked, give partial credit (1 point).
If no pages were bookmarked, give no credit (0 points).

This rubric essentially says that bookmarking constitutes one piece of evidence for Computer Skills, more frequent bookmarking suggests greater skill, and a couple of instances is sufficient as an indicator. It's worth noting that, in a computer-based performance assessment such as this one, what to consider as evidence of proficiency and how to evaluate that evidence is a judgment, in this case based on literature, logical analysis, and pilot results, but a judgment all the same.

The last step in scoring was to aggregate the pieces of evidence. A series of statistical models was employed to weight and combine the pieces of evidence to create the Total, Computer Skills, and Scientific Inquiry scores. An item response model was used to relate the latent proficiency measured to the probability of responding correctly to an observable; a structural equation model was employed to describe how the latent proficiencies measured by the three different scores influenced one another; and a conditioning model was used that, as in NAEP, employs background information to remove bias from the estimation of population-group means (Allen, Donoghue, & Schoeps 2001). Through these models, scores were derived and placed on an arbitrary scale with a mean of 150 and standard deviation of 35 that allowed rough comparison of the performance of various demographic groups.[6]

The Simulation Scenario

The Simulation scenario presented the student with an environment for asking "what-if" questions. The student was expected to use this environment over a sixty minute period for experimentally solving constructed-response and multiple-choice problems related to the science of gas-balloon flight.

The design of the Simulation scenario drew heavily upon the research of Glaser and associates, as well as that of others (Raghavan, Sartoris, & Glaser, 1998; Schauble et al., 1991, 1992; Shute & Glaser, 1990, 1991; White & Frederiksen, 1998). The common theme running through this research is the "discovery environment," or "microworld," where a student can experiment to construct an understanding of some underlying phenomenon, often physical in nature. In "Smithtown" (Shute & Glaser,

1990), for example, students learn basic macroeconomics concepts and scientific-inquiry skills by conducting simulated experiments. Regardless of content area, a key characteristic of most such environments is that they offer students opportunities and sufficient context to form hypotheses, test them, and draw conclusions about governing principles. Although these environments have primarily been used for instructional purposes, they also hold considerable promise for assessment. For one, by emulating their character, we may be able to provide a more engaging assessment experience for students. In addition, we may be able to tap more of the desired reasoning and strategic knowledge, and less of the procedural direction-following characteristic of some types of stock laboratory exercises (National Assessment Governing Board, 2000, p. 33; Schauble et al. 1995, p. 133).

Figure 4 (next page) shows the interface with which the student worked.[7] A problem statement is given in the upper right-hand corner. There were three problems of increasing difficulty. Each problem asked the student to discover the relationship between or among physical quantities. The first problem asked the student to discover how the payload mass that a helium gas balloon can carry affects the altitude to which the balloon can rise in the atmosphere. This relationship, which is relatively easy to comprehend, is a negative linear one (i.e., the greater the mass, the lower the altitude). The second problem concerned the relationship between altitude and the amount of helium put into the balloon. Here the relationship is considerably more difficult because it takes the form of a step function. Up to a critical value, regardless of the amount of helium added, the balloon will not leave the ground. Once that value has been reached, the balloon will rise to a given altitude and no additional helium will cause it to rise any higher. The last problem centers upon the joint effect of mass and amount of helium on altitude, a more difficult set of relationships still, that takes the form of a series of step functions.

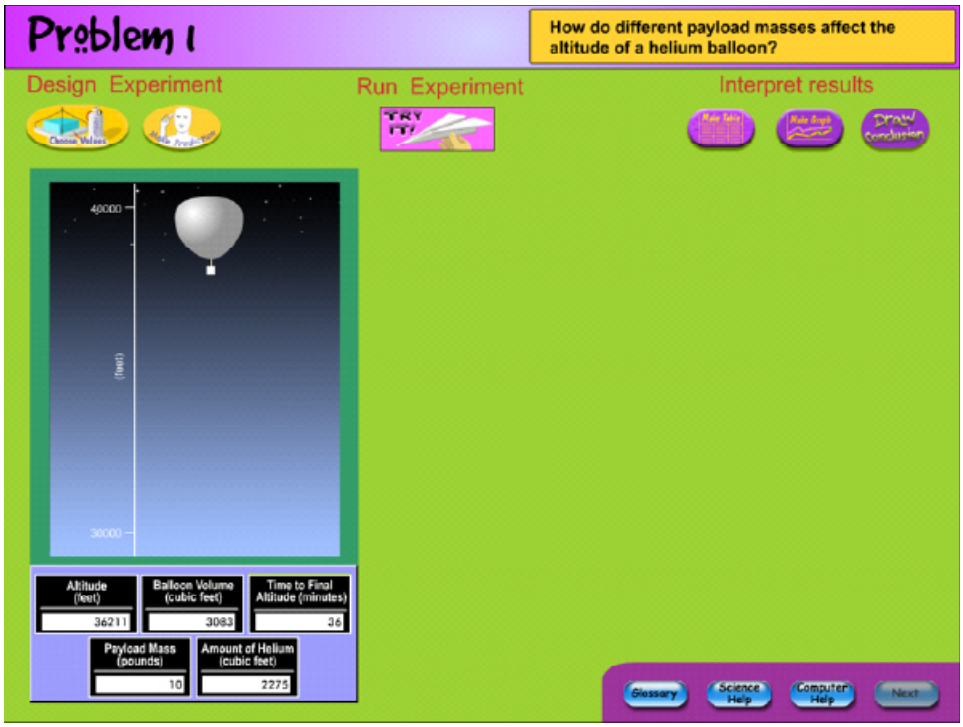
The student discovers these different relationships by conducting experiments using the tools arrayed across the upper portion of the screen in Figure 4. Going from left to right, the student may choose values for the independent variable (e.g., the mass to attach to the balloon or the amount of helium to put into it); may make a prediction about what will happen when the balloon is launched; may launch the balloon; may create a table to help make meaning of results; may create a graph; and may go to a form to enter an extended response to the motivating question.

The student may go through this process in any order, though obviously some orders will be more productive than other orders. In addition, the student may conduct as many or as few experiments as he or she sees fit.

The results of each experiment are shown in the flight window, which depicts the behavior of the balloon when it is launched. The balloon may rise, more or less rapidly, or may not rise at all, depending upon the mass attached to it or the amount of helium put into it. Results are also shown on the instrument panel below the flight window, which dynamically depicts the balloon's altitude, its volume, the time it takes to reach its final altitude, the payload mass it is carrying, and the amount of helium put into it.

Students were introduced to the Simulation interface through an interactive tutorial. The tutorial demonstrated the functioning of each interface component and walked the student through running an experiment. In addition to the tutorial, three forms of help were offered, accessed by buttons in the lower right-hand corner of the screen (Figure 4). These buttons, available throughout the test, brought up a glossary of science terms, science help, and computer help. Science help gave hints about the substance of the problem while computer help described the buttons and functions of the Simulation interface. In combination, the tutorial and three forms of help were intended to reduce the chances of a student not being able to demonstrate their scientific inquiry skills because of a low level of computer skill, or the converse.

Figure 4: The Simulation Scenario Interface



As noted, the Simulation scenario included three motivating problems. Each such problem presented one constructed-response question and one multiple-choice question. After the third motivating problem, several synthesizing multiple-choice questions were presented.

As was true for Search, preliminary analysis of the Simulation data suggested that the intended five scales could not be empirically supported. In the case of Simulation, the observables from the Scientific Synthesis scale and the Scientific Exploration scale could not be effectively combined to form a meaningful, higher-order, Scientific Inquiry scale. Consequently, a Total Simulation scale was created, along with three subscales: Computer Skills, Scientific Exploration, and Scientific Synthesis.

Table 2 (next page) shows the observables employed as evidence of proficiency. Observables differ somewhat from one problem to the next primarily because some observables were found to provide empirically redundant information with other observables and therefore were not included in the analysis for all motivating problems. Among the observables used as evidence of Computer Skills were keyboard fluency (the number of characters typed in each constructed response), fluency in using interface tools for experimenting and for drawing conclusions (the frequency with which interface tools were used in the wrong order), and how often computer help was used.[8] For Scientific Exploration, evidence included making a graph or table with variables suited to the problem, running experiments sufficient in number and range of the independent variable to support a meaningful conclusion, controlling for one variable in the last (multivariate) problem, and the frequency of using the glossary.[9] Finally, for Scientific Synthesis, evidence included the quality of the constructed-response answers and the correctness of responses to the multiple-choice questions.

Table 2: Observables Used as Evidence of Proficiency for TRE Simulation

Observable	Computer Skills	Scientific Exploration	Scientific Synthesis
Problem 1			
Keyboard fluency (number of characters in conclusion)	X		
Fluency in using interface tools for drawing conclusions	X		
Fluency in using interface tools for experimenting	X		
Degree of use of Computer Help	X		
Graph is useful to problem		X	
Choice of best experiments to solve problem		X	
Table is useful to problem		X	
Degree of use of Glossary		X	
Degree to which conclusions are correct and complete			X
Accuracy of response to final multiple-choice question			X
Problem 2			
Keyboard fluency (number of characters in conclusion)	X		
Fluency in using interface tools for drawing conclusions	X		
Choice of best experiments to solve problem		X	
Table is useful to problem		X	
Graph is useful to problem		X	
Degree to which conclusions are correct and complete			X
Accuracy of response to final multiple-choice question			X
Proportion of accurate predictions			X

(Table 2 continued, next page)

Table 2: Observables Used as Evidence of Proficiency for TRE Simulation (continued)

Observable	Computer Skills	Scientific Exploration	Scientific Synthesis
Problem 3			
Keyboard fluency (number of characters in conclusion)	X		
Use of computer interface (use of various interface functions, e.g., making tables and graphs)	X		
Fluency in using interface tools for drawing conclusions	X		
Proportion of experiments controlled for one variable		X	
Choice of best experiments to solve problem		X	
Graph is useful to problem		X	
Table is useful to problem		X	
Degree to which conclusions are correct and complete			X
Accuracy of response to final multiple-choice question			X
Conclusion			
Degree of correctness of responses to multiple-choice items			X

Note. The observables presented here represent the final set selected after preliminary data analysis. For a complete listing of all observables, see Bennett, Persky, Weiss, and Jenkins (2007, p. 71.)

As for the Search scenario, item-response, structural-equation, and conditioning models were used in the generation of scores. Also, as in the Search scenario, these scores were put on an arbitrary scale with a mean of 150 and standard deviation of 35.[10]

2. Demonstrate an Assessment that Could be Delivered on Computer Nationally

Data for the TRE study were collected in spring 2003.[11] The study samples comprised nationally representative groups of eighth-grade students selected through a multistage probability-based procedure. This procedure used counties and county equivalents, or groups of counties (primary sampling units, or PSUs), as the first-stage sampling units, and schools as the second-stage units. The third and final stage involved the selection of students within schools and their random assignment to either the Search scenario or the Simulation scenario. (Although it would have been preferable for each student to have taken both scenarios, the need to minimize burden on the respondents precluded this possibility.)

The selection procedure resulted in a sample of 270 schools, 222 of which participated, for a weighted cooperation rate of 85.1 percent. From the 222 participating schools, 2,409 students were selected to take part. After accounting for excluded students and non-respondents, the total number of students assessed was 2,134 (an average of about 10 students per school). Combining the effects of school nonparticipation and student nonparticipation resulted in an overall weighted participation rate of 79.6 percent, comparable to the weighted participation rate for the NAEP 2000 grade 8 science assessment of 78 percent.

All TRE administrations were done at school and were proctored by NAEP field staff. In addition to taking either the Search or Simulation scenario, students responded to three additional measures: (1) a background questionnaire, including questions about computer use, (2) a multiple-choice test of science knowledge, and (3) a multiple-choice test of computer-related knowledge. Relationships between TRE performance and student background are described briefly in a later section. Because of the limited reliability of the multiple-choice science and computer-knowledge tests, relationships with TRE performance are not reported here (see Bennett, Persky, Weiss, & Jenkins, 2007, pp. 55, 73, for those results).

When the TRE data files were returned from the field, it was found that 25 students did not have scenario data, which was presumed to be the result of technology failure. Additionally, one student, who was mistakenly coded as a non-respondent, actually did have scenario data but received no sampling weights. Missing and miscoded data resulted in a total number of 2,109 usable student records. Of this number, 1,077 student records were associated with the Search scenario and 1,032 with the Simulation scenario.[12]

How were these innovative assessments delivered to nationally representative samples with such an apparently small percentage of technology-

related problems? First, TRE was the third NAEP online study, following the 2001 Math Online study and the 2002 Writing Online study (Bennett, Braswell, Oranje, Sandene, Kaplan, & Yan, 2008; Horkay, Bennett, Allen, Kaplan, & Yan, 2006). Second, the NAEP field staff members responsible for data collection were trained to deal with basic technology-related issues, were backed by telephone technology support, and had often been participants in the two previous NAEP online studies. Thus, these individuals were experienced with resolving technology problems in an assessment-related environment. Third, standard Internet browser software was used, with common plug-ins and extensions, so most schools would be likely to already have the required software. Finally, because this was a demonstration project in which only 10 or so students per school had to be tested, field staff were able to use NAEP laptops when direct Internet delivery to school computers was not feasible. Laptops were used for about 40% of the participants.

3. Demonstrate an Assessment that Held Together Reasonably Well Psychometrically

The third intention for the study was an assessment that would hold together reasonably well psychometrically. In assessing psychometric functioning, the focus was more on detecting obvious problems rather than conducting a comprehensive validity analysis, which was not possible given the available financial resources and the need to minimize the time students spent participating in the study.

The analysis of basic psychometric functioning addressed the following questions:

- To what degree are scores internally consistent?
- Do the multiple scores computed for a scenario each provide distinct information?
- What student behaviors predict score on the constructed-response question(s) and are these predictions in the expected direction?
- How are scores related to reported computer use and are these relations in the expected direction?

A rationale for, and data relevant to, each of these questions are presented below, first for Search and then for Simulation. All results indicated as statistically significant are at the $p < .05$ level, unless otherwise stated.

Search Scenario Functioning

To what degree are scores internally consistent?

The Search scale should be comprised of items—or observables—that are positively related to one another. If not, each scale will be more a mélange than a coherent whole. To evaluate internal consistency, we used Coefficient alpha which, conceptually, is the mean correlation between all possible test halves and which ranges from 0 to 1.00. For the Search Total score, which consisted of 11 observables, the value of this statistic was .74. For the Scientific Inquiry score, which had 5 observables, the comparable value was .65. Finally, for the Computer Skills score, consisting of 6 observables, the value was .73.

When an assessment is composed of tasks that are related by virtue of a common stimulus, as was true for TRE Search, estimates of internal consistency may be artificially inflated (Sireci, Thissen, & Wainer, 1991). But if these values are approximately correct, they might be benchmarked against the internal consistency estimates for the NAEP science-assessment hands-on experiments. Although the hands-on experiments measure skills different from TRE Search and are somewhat shorter (30-minutes versus approximately 40 minutes for TRE Search), both measures take the form of extended performance tasks. For the 2000 NAEP science assessment, the mean weighted internal consistency taken across three hands-on blocks was .62 (B. Kaplan, personal communication, October 20, 2004), in the same general neighborhood as the values found for the TRE Search scores. In considering the magnitude of these estimates, it should be kept in mind that, in NAEP, these values are section estimates, not total test reliabilities, and that NAEP produces only group scores; no scores are computed for individuals.

Do the scores provide distinct information? This question is of interest because, if the three TRE Search scores are not reasonably distinct, there is little justification for computing three scores. The correlation of the Computer Skills and Scientific Inquiry scores (corrected for unreliability) was .57. The correlation of each of these scores with the Total score was .68. These values are in contrast with the correlations among the 1996 main NAEP eighth-grade science assessment scales, which ranged from .90 to .93 (Allen, Carlson, & Zelenak 1999).

Some degree of distinctiveness is also suggested by the correlations of the observables with each scale score (Table 3, next page). As should be apparent, in this student sample, the correlation for the scale to which an observable belongs (shown in bold) was, in most instances, noticeably higher than that observable's correlation with the other scale. For the Scientific Inquiry scale, performance was most highly related to the relevance of

the pages visited or bookmarked, the quality of the constructed response to the Search question, and the degree of use of relevant search terms (*r* range = .51 to .71). In contrast, scores on the Computer Skills scale were most highly associated with the use of hyperlinks, 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).

Table 3: Disattenuated Correlations of Search Observables with Each Scale Score

Observable	Computer Skills	Scientific Inquiry
Relevance of pages visited or bookmarked	.17	**.71**
Accuracy/completeness on CR question	.39	**.70**
Use of relevant search terms	.33	**.51**
Number right on final MC questions	.28	**.44**
Average relevance of hits to motivating problem	.20	**.34**
Use of hyperlinks to dig down	**.69**	.37
Use of Back button	**.65**	.36
Number of searches for relevant hits*	**.65**	.33
Use of bookmarking to save pages	**.60**	.45
Use of advanced search techniques	**.46**	.30
Use of deletion for unwanted filed pages	**.24**	.08

*The values for this observable were reversed (i.e., fewer searches received a higher score) to allow correlations with other variables to be positive.

Note. Values in bold indicate the scale to which an observable belonged. N = 672 to 1,077. All values are significantly different from zero at $p < .05$. All scale scores include the observable being correlated.

What student behaviors predict score on the constructed-response questions?

The motivating problem for the Search scenario was a constructed-response question that asked the student to find out and explain why scientists sometimes prefer to use helium gas balloons for planetary space exploration. Student responses to this question were rated by human judges once on a 3-point scale for accuracy and completeness; 25% of the responses were re-scored independently, with an exact agreement rate of 90% as a check on scoring reliability. An example of a response receiving a top score is shown in Table 4 (next page). The response accurately gives three advantages to using helium gas balloons for planetary atmospheric exploration.

Table 4: A Response Receiving a Top Score for the Search Scenario Constructed-response Question

> "One of the advantages of using a balloon is that is has a simple design and can hold a lot of weight. It also costs less to make a balloon rather than making a satelite. You can also launch them in the area you wish to conduct your experiment. It takes little time for it to be constructed as well. This is why it is better to have a balloon rather than a satelite or space shuttle."

Note. Response is the unedited, verbatim answer given by the student.

We should expect student Search behavior to be related to score on this question. That is, all other things equal, how a student searched should be associated with the quality of the student's answer. As Table 5 shows, 8 of the 9 search-related behaviors measured were significantly correlated with the constructed-response raw score in the expected direction.

Table 5: Observed Correlations of Search-related Behaviors with Constructed-response Raw Score

Observable	***r***
Relevance of pages visited or bookmarked	.55*
Use of bookmarking to save pages	.35*
Use of relevant search terms	.32*
Average relevance of hits to motivating problem	.21*
Use of hyperlinks to dig down	.21*
Use of advanced search techniques	.21*
Number of searches for relevant hits†	.20*
Use of back button	.19*
Use of deletion for unwanted filed pages	.03

*$p < .05$.

† The values for this observable were reversed (i.e., fewer searches received a higher score) to allow correlations with other variables to be positive.

The best predictors in this student sample were the relevance of pages visited or bookmarked, the use of bookmarking, and the use of relevant search terms. This set of behaviors makes sense: Given that a student doesn't know why scientists use gas balloons, the best way to find out is to pose relevant queries (i.e., search terms), visit relevant pages, and bookmark them so they can be easily located once the student is ready to compose an answer.

Are scores related to self-reported computer use?

Since the Search scenario was intended to measure problem solving with technology, students who said they used computers for related activities—like word processing or Internet exploration—should perform better than students who did not report such computer use. Because such data are observational, however, they should be viewed as suggestive only; positive associations may simply reflect more general relationships between student achievement, or motivation, and computer use (e.g., higher performers in general may use computers more frequently than lower performers do).

On all three Search scales, students who reported using a computer daily outside of school scored significantly higher statistically than students who reported using a computer less frequently; those who reported using a computer to find information on the Internet to a large extent scored significantly higher statistically than students who reported using it to find information on the Internet to a small extent; and those who said they used a word processor, regardless of extent, scored significantly higher statistically than students who reported not using a word processor at all.

Statistically significant positive relations were also found between Search performance and the following aspects of students' reported background: using e-mail, talking in chat groups, and having a computer in the home that the student uses. For some uses of the computer, however, more use was not associated with higher performance on the Search scales. For example, students who reported using the computer to make drawings or create artwork to a large extent scored statistically significantly *lower* on average on all three TRE Search scales than students who reported engaging in these activities to a small extent or not at all. This finding may perhaps be indicative of a distinction between the typical uses that students interested in art and those who are more scientifically oriented make of computers.

Finally, worth mentioning is that the relationships with computer use appeared to hold over all Search scenario scores. The associations did not, however, serve to differentiate the three scales in any notable way as relationships with external variables sometimes do.

Simulation Scenario Functioning

To what degree are scores internally consistent?

For the Simulation Total score, which consisted of 28 observables, coefficient alpha was .89. For the Scientific Exploration score, which had 11 observables, alpha was .78. For Scientific Synthesis, with 8 observables, internal consistency was .73. Finally, the Computer Skills score had 9 observables and an internal consistency of .74.[13] By way of comparison, these values are higher than the average reliability for the shorter hands-on blocks used in the 2000 NAEP science assessment. As noted earlier, for the NAEP 2000 science assessment, the mean weighted internal consistency taken across three such blocks was .62 (B. Kaplan, personal communication, October 20, 2004).

Do the scores provide distinct information?

Table 6 gives the disattenuated correlations among the Simulation scores. As the table shows, the Computer Skills, Scientific Exploration, and Scientific Synthesis scores correlate about equally with the Simulation Total score (of which all three subscales are a part). In addition, the correlations of the subscales with each other are in the middle .70s. These values contrast with correlations among the 1996 main NAEP eighth-grade science assessment scales ranging from .90 to .93 (Allen, Carlson, & Zelenak 1999).

Table 6: Disattenuated Correlations among the Simulation Scales

TRE Scale	Computer Skills	Scientific Exploration	Scientific Synthesis
Total	.75	.74	.76
Computer Skills	—	.73	.73
Scientific Exploration		—	.74

Note. N = 1,032. All correlations are significantly different from zero at $p < .05$.

Table 7 (pages 26 & 27) gives the disattenuated correlations of each observable with the three TRE subscales. Each observable was intended to measure proficiency on one scale (i.e., Computer Skills, Scientific Exploration, or Scientific Synthesis), with the assigned scale indicated by the bold values. Although the distinctions between the scales are not as sharp as they were for TRE Search, visual inspection suggests that, in general, the Simulation observables correlate in this student sample more with the scale they were intended to measure than with the other scales. In this student sample, the Scientific Exploration skill scale score was most

highly associated with what experiments students chose to run in order to solve each of the Simulation problems, whether students constructed tables and graphs that included the relevant variables for Simulation problems 1 and 2, and the degree to which experiments controlled for one variable for Simulation problem 3 (r range =.49 to .74). For the Scientific Synthesis scale, the observable most highly associated with performance was the degree of correctness and completeness of conclusions drawn for each Simulation problem (r range = .67 to .72). Lastly, performance on the Computer Skills scale was most highly associated with the number of characters in the conclusions drawn by students for each Simulation problem (r range = .72 to .78). Students who evidenced greater keyboard fluency (through entering longer answers to the constructed-response question that concluded each Simulation problem) tended to receive higher Computer Skills scores than students who entered shorter ones.

What student behaviors predict score on the constructed-response questions?

The Simulation scenario included three constructed-response (CR) questions, each to be solved by conducting experiments intended to help students discover the relationship between or among a set of physical quantities. Responses to the first Simulation problem were scored on a 3-point scale, whereas 4-point scales were used for the other two problems. Student responses were rated by human judges once for accuracy and completeness; 25% of the responses were re-scored independently, with an exact agreement rate of 89%-95%, depending upon the question, as a check on scoring reliability.

Each question and an example of a student response receiving a top score are shown in Table 8 (page 28). To receive a top score, the response had to accurately describe the functional relationship between or among the variables.

Table 7: Disattenuated Correlations of Simulation Observables with Each Scale Score

Observable	Computer Skills	Scientific Exploration	Scientific Synthesis
Problem 1			
Degree to which conclusions are correct and complete	.57	.56	**.69**
Accuracy of response to final multiple-choice question	.22	.26	**.31**
Graph is useful to problem	.45	**.60**	.52
Choice of best experiments to solve problem	.35	**.53**	.40
Table is useful to problem	.41	**.50**	.44
Degree of use of Glossary	–.17	**–.17**	–.19
Keyboard fluency (number of characters in conclusion)	**.72**	.49	.54
Fluency in using interface tools for drawing conclusions*	**–.32**	–.25	–.28
Fluency in using interface tools for experimenting*	**–.28**	–.24	–.27
Degree of use of Computer Help	**–.26**	–.22	–.24
Problem 2			
Degree to which conclusions are correct and complete	.59	.61	**.72**
Accuracy of response to final multiple-choice question	.31	.31	**.37**
Proportion of accurate predictions	.22	.22	**.25**
Choice of best experiments to solve problem	.45	**.64**	.52
Table is useful to problem	.41	**.52**	.44
Graph is useful to problem	.40	**.49**	.44
Keyboard fluency (number of characters in conclusion)	**.78**	.52	.55
Fluency in using interface tools for drawing conclusions*	**–.27**	–.21	–.23

(Table 7 continued, next page)

Table 7: Disattenuated Correlations of Simulation Observables with Each Scale Score (continued)

Observable	Computer Skills	Scientific Exploration	Scientific Synthesis
Problem 3			
Degree to which conclusions are correct and complete	.52	.52	**.67**
Accuracy of response to final multiple-choice question	.36	.36	**.43**
Proportion of experiments controlled for one variable	.51	**.74**	.56
Choice of best experiments to solve problem	.44	**.56**	.46
Graph is useful to problem	.32	**.42**	.35
Table is useful to problem	.14	**.21**	.20
Keyboard fluency (number of characters in conclusion)	**.76**	.53	.59
Use of computer interface (use of various interface functions, e.g., making tables and graphs)	**.42**	.54	.42
Fluency in using interface tools for drawing conclusions*	**–.21**	–.19	–.20
Conclusion			
Degree of correctness of responses to multiple-choice items	.47	.48	**.58**

*Fluency was measured in terms of the frequency with which interface tools were used incorrectly, such that fewer errors indicated greater fluency..

Note. Values in bold indicate the scale to which an observable was assigned. All correlations are significantly different from zero at $p < .05$. N range = 221 to 1,032. All scale scores include the observable being correlated.

Table 8: Answers Receiving Top Scores for Each of the Three Simulation Scenario Problems

Problem 1	
Question	How do different payload masses affect the altitude of a helium balloon? Support your answer with what you saw when you experimented.
Student response	The lower the payload mass, the higher the altitude the balloon reaches. For example, when you had 10 pounds of payload mass, the balloon rose to 36211. When you had 30 lbs. of payload mass the balloon rose 28640 ft. When you had 50 lbs. of payload mass the balloon rose 22326 ft.
Problem 2	
Question	How do different amounts of helium affect the altitude of a helium balloon? Support your answer with what you saw when you experimented.
Student response	The amount of helium affects the balloon altitude. There must be at least 2500 cubic feet of helium for the balloon to even rise. After 2500 cubic feet the baloon altitude stays constant even if you add more helium. When i used less helium than 2500 cubic feet the balloon did not gain any altitude. But after the 2500 cubic feet mark the balloons altitude stayed at approximately 10000 feet even after i tried almost 3000 cubic feet of helium
Problem 3	
Question	How do amount of helium and payload mass together affect the altitude of a balloon? Support your answer with what you saw when you experimented. Refer to at least two masses.
Student response	The greater the payload mass is the lower the maximum altitude for that balloon will be, and the more helium it will require to lift it off the ground. For a 10 pund payload mass it took 910 cubic feet of helium to get it a little bit off the ground. 975 cubic feet lifted the 10 pound payload mass to its maximum hieght of 36211 feet above ground. With 50 pounds of payload mass 1700 cubic feet was needed to lift the payload 2 feet off the ground. At least 2400 cubic feet of helium was needed for the 50 pound payload mass to reach its maximum hieght of22326 feet above ground. During experimenting with the 110 pound payload mass 2400 cubic feet of helium was required for a tiny lift off the ground, and at least 2616 cubic feet of helium was needed to reach its maximum height of 7918 feet above ground.

Note. Responses are the unedited, verbatim answers given by students.

Table 9 (page 29) shows how different process-related observables were associated with scores on these questions. In general, students who wrote longer answers tended to receive higher scores, a result related at least in part to the fact that longer responses tended to be more detailed. Apart from response length, the results show statistically significant positive relationships between CR scores and process-related behaviors that can help students develop better answers. For example, students who chose a better set of experiments for any given Simulation problem generally tended to receive higher scores for responses to the constructed-response question than did students who chose a less adequate set of experiments (e.g., too few experiments, too narrow a range on the independent variable). Further, students who made graphs and tables appropriate to

Simulation problems 1 and 2 were more likely to receive higher scores for their conclusions to those problems than students who did not make such graphs and tables. Finally, students who controlled for one variable in their experiments for Simulation problem 3 were inclined to attain higher scores on the constructed-response question.

Table 9: Observed Correlations of Simulation Process-related Observables with the Constructed-response Raw Score

Observable	*r*
Problem 1	
Keyboard fluency (number of characters in conclusion)	.48
Graph is useful to problem	.45
Table is useful to problem	.37
Choice of best experiments to solve problem	.32
Fluency in using interface tools for drawing conclusions*	–.23
Fluency in using interface tools for experimenting*	–.18
Degree of use of Computer Help	–.15
Degree of use of Glossary	–.14
Problem 2	
Keyboard fluency (number of characters in conclusion)	.50
Choice of best experiments to solve problem	.47
Graph is useful to problem	.39
Table is useful to problem	.35
Fluency in using interface tools for drawing conclusions*	–.16
Proportion of accurate predictions	.15

(Table 9 continued, next page)

Table 9: Observed Correlations of Simulation Process-related Observables with the Constructed-response Raw Score

Observable	*r*
Problem 3	
Proportion of experiments controlled for one variable	.45
Keyboard fluency (number of characters in conclusion)	.44
Choice of best experiments to solve problem	.43
Use of computer interface (use of various interface functions, e.g., making tables and graphs)	.31
Graph is useful to problem	.24
Table is useful to problem	.12
Fluency in using interface tools for drawing conclusions*	–.11

*Fluency was measured in terms of the frequency with which interface tools were used incorrectly, such that fewer errors indicated greater fluency..

Note. All correlations are significantly different from zero at $p < .05$. The constructed-response question for Simulation problem 1 was scored on a 1–3 scale. The constructed-response questions for problems 2 and 3 were each scored on a 1–4 scale.

Are scores related to self-reported computer use?[14]

As might be expected from the nature of the Simulation scenario, students who reported using computers more frequently for a variety of activities outperformed their peers who reported using computers less frequently for these activities. As for the Search scenario, the associations with these computer activities, in most cases, carried across all four Simulation scales and did not bring out differences in the meaning of scores from one scale to the next.

Of particular note is that students who reported using a word processor, regardless of extent, performed significantly better statistically on all Simulation scales than students reporting not using a word processor at all. In addition, students who reported using a computer to make charts, tables, and graphs to a small or moderate extent performed better on all scales than students who reported that they did not do so at all. These two types of computer experience could have helped students in the Simulation scenario as some degree of text processing skill was required to answer the constructed-response questions and because the quality of those answers was associated with making an appropriate table or graph. Finally, students who reported playing computer games to a moderate or large extent had higher Scientific Exploration scores than students who reported that

they did not play such games at all. This result may reflect the fact that the observables assigned to the Scientific Exploration scale resemble the activities involved in some complex computer games. Manipulating conditions, keeping track of choices made and their outcomes, observing and interpreting dynamic displays, and creating and manipulating tables and graphs are effective strategies for solving problems in many such computer games.

Aside from computer experience that might be considered directly relevant, engagement in other computer activities also was associated with performance in the Simulation scenario. For instance, students who reported finding information on the Internet to a large extent had higher scores on all Simulation scales than their peers who reported doing so to a small extent. In addition, on all Simulation scales, students who reported using a computer outside of school daily outperformed students who reported doing so less frequently. Finally, the presence of a computer at home was positively and statistically significantly associated with student performance on all scales.

4. Demonstrate an Assessment that Produced Credible Results

The last intention of the project was to demonstrate an assessment that produced credible results, primarily in the context of what we know of the previous NAEP performance of 8th grade students. The focus is, therefore, on the overall pattern of results rather than on the specific results themselves which, because of the narrow scope of the TRE scenarios, have limited meaning.

The analyses focus on the performance of NAEP demographic groups, including ones categorized by gender, race/ethnicity, parents' education level, eligibility for free or reduced-price school lunch, and school location. Comparisons were made within each demographic variable using an independent-samples *t*-test, correcting for the number of tests run for the category via the false discovery rate procedure (Benjamini & Hochberg, 1995). Because of this correction, and because sample sizes within groups were often small, some seemingly large differences may not be statistically significant.

Search Scenario Results

Table 10 (next page) gives the mean scores and standard errors for gender groups. The scores are on a scale with a mean of 150 and standard deviation of 35. In this case, there was no statistically significant difference between males and females. Although this result runs counter to the stereotype of males being more computer-familiar, and more computer-profi-

cient, than females, it is consistent with data from other sources. In 2003, the same year these TRE results were collected, there was no measurable difference between males and females in self-reported, overall computer- or Internet-use rates, according to the National Center for Education Statistics (DeBell & Chapman, 2006, p. v). Also, there were no statistically significant differences between gender groups for year 10 in the Australian National Assessment Program 2005 Information and Communication Technology (ICT) Literacy study and, although there were statistically significant differences at year 6, the differences were both small and in favor of females (MCEETYA, 2007, p. 60; MCEECDYA, 2010, p. 39).

Table 10: Search Mean Scores and Standard Errors for Gender

Group	N	Total	Scientific Inquiry	Computer Skills
Male	517	148 (2.4)	149 (2.7)	147 (2.5)
Female	560	151 (2.3)	150 (2.3)	152 (1.9)

Table 11 shows the results for groups categorized by race/ethnicity. With respect to such groups, as well as to ones categorized by socio-economic status, there are well-documented differences in school performance, commonly referred to as "the achievement gap" (Barton, 2003). For TRE Search performance, this unwelcome gap also appears to exist. On all three scales, White students performed significantly higher statistically than either Black or Hispanic students. The Black-White difference, in particular, was dramatic, one standard deviation or more depending on the scale, and of just about the same size as found, for example, on the 2005 NAEP science assessment and the 2007 reading assessment (NCES, 2006, p. 20; NCES, 2007, p. 29). In addition, on the Computer Skills scale, the mean score for Hispanic students was significantly higher statistically than the mean for Black students.

Table 11: Search Mean Scores and Standard Errors for Race/Ethnicity

Group	N	Total	Scientific Inquiry	Computer Skills
White	643	161 (1.9)	160 (1.6)	158 (1.7)
Black	185	121 (3.8)	125 (2.8)	128 (3.3)
Hispanic	188	139 (3.4)	137 (4.8)	142 (3.4)

Table 12 shows results for groups categorized by parents' highest education level. NAEP asks how far the student's mother and father progressed in school, and uses the higher parental level for this categorization. As is typical for NAEP results, students who reported higher levels of parental education outperformed their peers who reported lower parental education levels on TRE Search. For example, on all three scales, those reporting that a parent had graduated college scored significantly higher statistically than those reporting that a parent had graduated high school (or that a parent had not finished high school).

Table 12: Search Mean Scores and Standard Errors for Parents' Highest Education Level

Group	N	Total	Scientific Inquiry	Computer Skills
Not finish HS	72	133 (3.7)	135 (4.3)	139 (4.5)
Grad HS	214	142 (4.4)	143 (2.9)	145 (3.1)
Post HS	202	155 (3.0)	154 (2.7)	154 (2.6)
Grad College	497	157 (2.4)	156 (2.4)	155 (2.4)

Table 13 gives results for students categorized by eligibility for school lunch, a proxy for poverty level. "Not eligible" indicates a student-group from relatively good economic circumstances, while the other designations denote progressively more impoverished groups. Again, the ordering is the unwanted but expected one based on previous NAEP results in such related areas as science and reading (NCES, 2006, p. 21; NCES, 2007, p. 31); that is, the score means decrease as poverty level increases. A similar result for socioeconomic status was reported from the Australian National Assessment Program 2005 ICT Literacy study for years 6 and 10 (MCEETYA, 2007, p. 62).

Table 13: Search Mean Scores and Standard Errors for Eligibility for School Lunch

Group	N	Total	Scientific Inquiry	Computer Skills
Not eligible	656	160 (1.6)	158 (2.0)	158 (1.8)
Reduced-price	70	145 (4.3)	148 (3.7)	147 (4.4)
Free lunch	300	129 (2.5)	131 (2.6)	133 (2.5)

Finally, the results for school location are shown in Table 14. Students differed in their performance only for the Search total score. On this scale, students attending central city schools scored significantly lower statistically than students attending urban fringe/large town schools and students attending rural schools.

Table 14: Search Mean Scores and Standard Errors for School Location

Group	N	Total	Scientific Inquiry	Computer Skills
Central city	288	142 (3.1)	142 (3.4)	144 (2.7)
Urban fringe/ large town	436	152 (2.4)	151 (2.8)	152 (2.2)
Rural	353	153 (3.1)	154 (3.4)	152 (3.4)

Simulation Scenario Results

The results for TRE Simulation generally mirror the findings for TRE Search reported above. Tables 15 through 19 (pages 34-35) give those results. Consistent with self-reported computer use data (DeBell & Chapman, 2006, p. v), there was no measureable difference between the gender groups on any scale. However, for the other NAEP reporting groups, there were differences in the expected directions based on previous NAEP science and reading performance (e.g., NCES, 2006, pp. 20, 21; NCES, 2007, pp. 29, 31). For each Simulation score scale, there were statistically significant differences among the racial/ethnic groups: White students received higher scores on all four scales than did their Black and Hispanic peers. For all scales, students reporting that a parent graduated from college outperformed students reporting that their parents did not finish high school and also outperformed students reporting that a parent graduated from high school. Similarly, for all scales, those students not eligible for free or reduced-price lunch received higher mean scores than students eligible for reduced-price lunch. Last, there were no measureable differences for any scale by school location.

Table 15: Simulation Mean Scores and Standard Errors for Gender

Group	N	Total	Scientific Exploration	Scientific Synthesis	Computer Skills
Male	545	149 (2.7)	152 (2.7)	151 (2.5)	147 (3.7)
Female	487	150 (3.1)	147 (2.4)	149 (2.8)	153 (3.7)

Table 16: Simulation Mean Scores and Standard Errors for Race/Ethnicity

Group	N	Total	Scientific Exploration	Scientific Synthesis	Computer Skills
White	644	161 (1.9)	160 (1.6)	161 (1.9)	159 (3.3)
Black	171	127 (3.8)	131 (3.9)	128 (4.5)	132 (4.1)
Hispanic	168	128 (4.7)	130 (4.1)	130 (3.8)	132 (4.2)

Table 17: Simulation Mean Scores and Standard Errors for Parents' Highest Education Level

Group	N	Total	Scientific Exploration	Scientific Synthesis	Computer Skills
Not finish HS	66	121 (5.1)	127 (3.8)	125 (4.1)	125 (3.7)
Grad HS	199	141 (3.3)	142 (3.1)	142 (3.1)	143 (3.5)
Post HS	180	150 (2.8)	151 (3.3)	150 (3.9)	149 (4.4)
Grad College	493	161 (2.4)	159 (2.6)	160 (2.2)	160 (3.7)

Table 18: Simulation Mean Scores and Standard Errors for Eligibility for School Lunch

Group	N	Total	Scientific Exploration	Scientific Synthesis	Computer Skills
Not eligible	625	160 (2.1)	158 (1.4)	159 (1.7)	158 (3.2)
Reduced-price	70	143 (4.7)	146 (5.9)	146 (5.5)	146 (6.4)
Free lunch	289	127 (3.2)	131 (3.9)	130 (3.6)	131 (4.0)

Table 19: Simulation Mean Scores and Standard Errors for School Location

Group	N	Total	Scientific Exploration	Scientific Synthesis	Computer Skills
Central city	254	145 (3.7)	147 (3.1)	146 (3.4)	146 (4.1)
Urban fringe/ large town	443	151 (3.5)	150 (3.4)	151 (3.7)	151 (4.0)
Rural	335	151 (3.3)	151 (3.3)	152 (3.5)	151 (3.9)

What Did the TRE Study Appear to Do and Not Do Effectively?

In our view, the TRE study produced a serviceable definition of problem solving with technology, broke it down into measurable process and product components, and created two demonstration performance-assessment scenarios to measure selected aspects of that construct. The study did not ground those assessment scenarios in a NAEP content framework, since such a framework was not then available, nor did it do the extensive literature review, curriculum analysis, or gathering of public input more typical of NAEP framework efforts. The study also did not cover problem solving with technology, scientific inquiry, or computer skills as broadly as would a NAEP assessment. For the Search scenario, one motivating problem dealing with only one topic was used. For Simulation, three problems were used but all shared the same context and required largely overlapping inquiry processes. Finally, the Search scenario did not use the real Internet, which has better search tools and far more relevant (and, also, irrelevant) information than the simulated World Wide Web that was employed. Using the real Internet might have produced different results (though not necessarily more meaningful ones due to the standardization concerns noted earlier).

The study was able to deliver the TRE scenarios on computer to a national sample of students with participation rates comparable to paper NAEP and without any significant technology problems, arguably a non-trivial accomplishment, especially for 2003. However, the project did not deliver those assessment scenarios to either a large sample of schools or to a large sample of students. Such delivery undoubtedly would have involved far greater technological, and logistical, challenges than the ones faced.

The study did produce scores that appeared to function in a reasonable way psychometrically. Internal consistency, scale intercorrelations, relations of process observables with performance on the constructed-response questions, and associations with background information generally supported the meaning of TRE scores. Further, population-group results were basically consistent with findings from NAEP assessments in associated content areas like science and reading which, logically, should be related to TRE performance.

All the same, the study did not provide convincing evidence of the validity of scores for making strong claims about students' skills, as the psychometric analyses conducted were only the most basic ones. As a consequence, the results should not be taken as estimates of problem solving with technology, scientific inquiry, computer skills, or electronic-information-search skill for the nation's 8th-grade students.

How Might Such a Measure Be Used in NAEP?

Measures like those demonstrated in the TRE study could potentially be used in NAEP or other large-scale assessments in several ways. One possibility is as part of a content assessment. The interactive computer tasks administered as a special study alongside the 2009 NAEP science assessment provide a working example. Growing out of the TRE study, this small set of extended online tasks was intended to complement the traditional paper NAEP survey test. In future years, both the extended tasks and the more traditional NAEP survey test could be given online. The rationale for using both measures is that the extended tasks allow students to engage with one or more extended problems in depth, providing construct representation—particularly in terms of certain cognitive processes—that cannot be attained with traditional multiple-choice and short-answer questions. The shorter, more traditional questions, on the other hand, afford the broad domain coverage needed for dependably generalizing from the total sample of tasks administered to performance on other item samples and to other non-test situations.

A second possibility for using measures like those demonstrated in TRE is as part of an information and communications technology assessment. Such an assessment might be built of multiple scenarios covering a range of substantive contexts and using a variety of technology tools. That collection of scenarios--perhaps 20 in total--would be arranged in much the same matrix-sample fashion as employed with the typical NAEP assessment, each student taking, for example, a pair of scenarios. An assessment of this type would be far more challenging than the model above in which a small number of extended online tasks are used to complement a traditional paper assessment. Instead of being the online complement to a paper test, an assessment composed entirely of online scenarios would be considerably more expensive to create and to deliver.

Conclusion

The TRE study suggests that we can measure—at least on a small scale—some important skills that cannot be assessed through traditional paper-and-pencil, multiple-choice tests (e.g., electronic information search). However, the study also made clear that going beyond traditional testing is extremely challenging. It is challenging because, for emerging domains like problem solving with technology, by definition there isn't a well-developed research base, or a widely accepted content framework in which to ground test development. Second, going beyond traditional testing is challenging because designing performance tasks for computer is a relatively new activity and there is little knowledge among test devel-

opers about how to do it effectively and efficiently. Third, many schools do not yet have the technology infrastructure to deliver such highly interactive tests to large numbers of students securely and efficiently. Finally, students produce extensive information when taking such tests and we are only just beginning to learn which of the literally hundreds of pieces of information produced are worth attending to. Further, we are only just beginning to learn how to defensibly score the pieces of information that do prove meaningful.

Although measuring emerging domains like problem solving with technology is a considerable challenge, it's also true that the need for such measures is not going to diminish. Quite the contrary, the need for these measures will only grow as problem solving with technology and related emerging skills increasingly become central to survival in a global economy. The 2012 NAEP Technology Literacy assessment (WestEd, undated), the Australian National Assessment Program 2011 Information and Communication Technology Literacy assessment (MCEECDYA, 2009), and the Cisco/Intel/Microsoft (undated) Assessment and Teaching of 21st Century Skills collaboration appear indicative of this trend. Consequently, if policy makers are to render sensible decisions about how to improve education, assessment agencies will have to learn how to assess these skills in technically defensible, fair, affordable, and logistically feasible ways. Last, we are unlikely to learn how to measure these emerging skills effectively if we are not willing to invest in the long-term programs of research and development needed to build, pilot, and iteratively refine such innovative measures.

Endnotes

1. Funded by the National Center for Education Statistics, Institute of Education Sciences, US Department of Education under contract number ED-02-CO-0023. Sampling and data collection were conducted by Westat. We are indebted to Malcolm Bauer, Dan Eignor, Irv Katz, and two anonymous reviewers for their helpful comments on an earlier draft of this paper.
2. Frank Jenkins was on the ETS staff at the time this study was conducted.
3. A NAEP Technology and Engineering Literacy framework (WestEd, undated) was recently released for guiding development of a 2014 national assessment that will include (but go well beyond) constructs similar to the ones described here.
4. Additional screens from the Search scenario can be accessed at: http://nces.ed.gov/nationsreportcard/studies/tba/tre/scenarios.asp.
5. The observables presented here represent the final set selected after preliminary data analysis. That analysis resulted in dropping two observables and assigning one observable originally connected with both scales to only the Scientific Inquiry scale. For further information, see Bennett, Persky, Weiss, and Jenkins (2007, p. 54.).
6. See Bennett, Persky, Weiss, and Jenkins (2007, p. 144-154) for a description of the statistical models used in scoring.
7. The Simulation scenario can be tried at: http://nces.ed.gov/nationsreportcard/studies/tba/tre/scenarios.asp.
8. Using computer help and using interface tools in the wrong order were hypothesized as suggesting lower proficiency.
9. Using the glossary was hypothesized as suggesting lower proficiency.
10. See Bennett, Persky, Weiss, and Jenkins (2007, p. 144-154) for a description of the statistical models used in scoring.
11. Prior to this data collection, several pilot administrations were conducted, including ones focused on usability and examinee response processes. Among other things, these pilots helped in refining the user interface to reduce sources of irrelevant variance, in revising scoring rubrics, and in selecting process observables to serve as evidence of proficiency.
12. The difference in the numbers of students taking Search and Simulation was due to 26 students taking the wrong scenario. Sampling weights were subsequently adjusted so that their data could be used.
13. As was true for TRE Search, the Simulation observables may not be completely independent, so the internal consistency estimates for the scales may be inflated.
14. Because these data are observational, they should be viewed as suggestive only.

References

Allen, N.L., Carlson, J.E., & Zelenak, C.A. (1999). *The 1996 NAEP technical report* (NCES 1999–452). Washington, DC: National Center for Education Statistics, U.S. Department of Education.

Allen, N.L., Donoghue, J.R., & Schoeps, T.L. (2001). *The NAEP 1998 technical report* (NCES 2001–509). Washington, DC: National Center for Education Statistics, U.S. Department of Education.

Barton, P.E. (2003). *Parsing the achievement gap: Baselines for tracking progress*. Princeton, NJ: Educational Testing Service. Retrieved February 11, 2010 from http://www.ets.org/Media/Education_Topics/pdf/parsing.pdf.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B, 57*, 289–300.

Bennett, R.E., Braswell, J., Oranje, A., Sandene, B, Kaplan, B., & Yan, F. (2008). Does it matter if I take my mathematics test on computer? A second empirical study of mode effects in NAEP. *Journal of Technology, Learning and Assessment, 6*(9). Retrieved October 7, 2008 from http://escholarship.bc.edu/jtla/vol6/9/.

Bennett, R.E., Persky, H., Weiss, A.R., & Jenkins, F. (2007). *Problem solving in technology-rich environments: A report from the NAEP Technology-Based Assessment Project* (NCES 2007-466). Washington, DC: National Center for Education Statistics, US Department of Education. Retrieved October 7, 2008 from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2007466.

Cisco/Intel/Microsoft. (Undated). *Assessment call to action: Transforming education: Assessing and teaching 21st century skills*. Retrieved December 30, 2009 from http://download.microsoft.com/download/6/E/9/6E9A7CA7-0DC4-4823-993E-A54D18C19F2E/Transformative%20Assessment.pdf.

DeBell, M., & Chapman, C. (2006). *Computer and Internet use by students in 2003* (NCES 2006–065). Washington, D.C.: US Department of Education, National Center for Education Statistics. Retrieved December 25, 2009 from http://nces.ed.gov/pubs2006/2006065.pdf.

Fidel, R., Davies, R.K., Douglass, M.H., Holder, J.K., Hopkins, C.J., Kushner, E.J., Miyagishima, B.K., & Toney, C.D. (1999). A visit to the information mall: Web searching behavior of high school students. *Journal of the American Society for Information Science, 50*(1), 24–37.

Horkay, N., Bennett, R.E., Allen, N., Kaplan, B, & Yan, F. (2006). Does it matter if I take my writing test on computer? An empirical study of mode effects in NAEP. *Journal of Technology, Learning and Assessment, 5*(2). Retrieved October 7, 2008 from http://escholarship.bc.edu/jtla/vol5/2/.

International Society for Technology in Education (ISTE). (1998). *National Educational Technology Standards for Students*. Eugene, OR: Author.

Klein, D.C.D., Yarnall, L., & Glaubke, C. (2001). *Using technology to assess students' web expertise* (CSE Technical Report 544). Los Angeles: UCLA-CRESST. Retrieved April 29, 2005, from http://www.cse.ucla.edu/CRESST/Reports/TECH544.pdf.

Klein, D.C.D., Yarnall, L., & Glaubke C. (2003). Using technology to assess students' web expertise. In H.F. O'Neil, Jr. and R.S. Perez (Eds.), *Technology Applications in Education* (pp. 305–320). Mahwah, NJ: Erlbaum.

Ministerial Council on Education, Employment, Training and Youth Affairs (MCEETYA). (2007). *National Assessment Program – ICT literacy years 6 & 10 report 2005*. Victoria, Australia: Author. Retrieved December 30, 2009 from http://www.mceecdya.edu.au/verve/_resources/NAP_ICTL_2005_Years_6_and_10_Report.pdf.

Ministerial Council for Education, Early Childhood Development and Youth Affairs (MCEECDYA). (2009). *National Assessment Program: Information and Communication Technology Literacy*. Retrieved December 30, 2009 from http://www.mceetya.edu.au/mceecdya/nap_ict_literacy,12183.html.

Ministerial Council for Education, Early Childhood Development and Youth Affairs (MCEECDYA). (2010). *National Assessment Program – ICT Literacy Years 6 & 10 report 2008*. Retrieved June 3, 2010 from http://www.mceecdya.edu.au/verve/_resources/NAP-ICTL_report_2008.pdf.

National Academy of Sciences (NAS). (1996). *National Science Education Standards*. Washington, DC: National Academies Press. Retrieved February 16, 2005, from http://www.nap.edu/readingroom/books/nses/html/1.html.

National Assessment Governing Board. (2000). *Science assessment framework for the 1996 and 2000 National Assessment of Educational Progress*. Washington, DC: Author. Retrieved January 25, 2005, from http://www.nagb.org.

National Center for Education Statistics (NCES). (2006). *The nation's report card: Science 2005 (NCES 2006-466)*. Washington, DC: US Department of Education, National Center for Education Statistics. Retrieved December 30, 2009 from http://nces.ed.gov/nationsreportcard/pdf/main2005/2006466.pdf

National Center for Education Statistics (NCES). (2007). *The nation's report card: Reading 2007 (NCES 2006-496)*. Washington, DC: US Department of Education, National Center for Education Statistics. Retrieved December 30, 2009 from http://nces.ed.gov/nationsreportcard/pdf/main2007/2007496.pdf.

Olson, S., & Loucks-Horsley, S. (2000). *Inquiry and the National Science Education Standards: A guide for teaching and learning* (pp. 28–30). Retrieved on January 25, 2005, from http://www.nap.edu/books/0309064767/html/.

Raghavan, K., Sartoris, M.L., & Glaser, R. (1998). Why does it go up? The impact of the MARS Curriculum as revealed through changes in student explanations of a helium balloon. *Journal of Research in Science Teaching, 35*, 547–567.

Riley, R.W., Holleman, F.S., & Roberts, L.G. (2000). *E-Learning: Putting a world-class education at the fingertips of all children (The National Educational Technology Plan)*. Washington, DC: U.S. Department of Education. Retrieved February 18, 2005, from http://www.ed.gov/about/offi ces/list/os/technology/reports/e-learning.pdf.

Salterio, S. (1996). Decision support and information search in a complex environment: Evidence from archival data in auditing. *Human Factors, 38*(3): 495–505.

Schacter, J., Chung, G.K.W.K., & Dorr, A. (1998). Children's Internet searching on complex problems: Performance and process analysis. *Journal of the American Society for Information Science, 49*(9): 840–849.

Schauble, L., Glaser, R., Raghavan, K., & Reiner, M. (1991). Causal models and experimentation strategies in scientific reasoning. The *Journal of the Learning Sciences, 1*(2), 201–238.

Schauble, L., Glaser, R., Raghavan, K., & Reiner, M. (1992). The integration of knowledge and experimentation strategies in understanding a physical system. *Applied Cognitive Psychology, 6*, 321–343.

Shute, V.J., & Glaser, R. (1990). A large scale evaluation of an intelligent discovery world: Smithtown. *Interactive Learning Environments, 1*, 51–77.

Shute, V.J., & Glaser, R. (1991). An intelligent tutoring system for exploring principles of economics. In R.E. Snow and D. Wiley (Eds.), *Improving inquiry in social science: A volume in honor of Lee J. Cronbach* (pp. 333–336). Hillsdale, NJ: Erlbaum.

Sireci, S.G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28*, 237–247.

WestEd. (undated). *Technology and Engineering Literacy Framework for the 2014 NAEP (Pre-Publication Edition)*. Retrieved May 19, 2010 from http://www.nagb.org/publications/frameworks/prepub_naep_tel_framework_2014.pdf.

White, B.Y., & Frederiksen, J.R. (1998). Inquiry, modeling, and metacognition: Making science accessible to all students. *Cognition and Instruction,16*(1), 3–118.

Author Biographies

Randy Elliot Bennett is Norman O. Frederiksen Chair in Assessment Innovation in the Research & Development Division at Educational Testing Service in Princeton, New Jersey. His research focuses on integrating advances in cognitive science, measurement, and technology to create new approaches to assessment. He is co-author of "Technology and Testing" (with Fritz Drasgow and Ric Luecht) in Educational Measurement (4th Edition) and "What Does it Mean to Be a Nonprofit Educational Measurement Organization in the 21st Century" (http://www.ets.org/Media/Research/pdf/Nonprofit.pdf). He is currently leading a research program, "Cognitively Based Assessment of, for, and as Learning (CBAL)" (http://www.edutopia.org/assessment-reinventing-standardized-tests), that centers on creating an innovative model for K–12 assessment. He can be contacted at rbennett@ets.org.

Hilary Persky is currently Assessment Lead in the NAEP Test Development Group at Educational Testing Service (ETS). She manages the NAEP writing and arts assessments, and oversees staff coordinating the science and social science assessments. In recent years she has focused on developing innovative computer-based performance assessment, both inside of NAEP and elsewhere at ETS.

Andy Weiss is an Assessment Specialist at Educational Testing Service in Princeton, NJ, where he has worked since 1996. He is the coordinator of test development for the NAEP social science assessments and also works on item development for the AP, CLEP, and GED programs. He holds an M.A. in American History from Cornell University and a B.A. in History from SUNY Binghamton.

Frank Jenkins is a senior statistician at Westat, Inc in Rockville, MD. He received his Ph.D. in educational research from the department of Education at Michigan State University. His interests include hierarchical linear modeling, Bayesian modeling and, measurement. He has worked on a number of large-scale assessments of educational programs in the United States including NAEP, Head Start and the Pre-Elementary Education Longitudinal Study. He can be contacted at FrankJenkins@Westat.com.

J·T·L·A

The Journal of Technology, Learning, and Assessment

www.jtla.org

Technology and Assessment Study Collaborative
Caroline A. & Peter S. Lynch School of Education, Boston College