NATIONAL CENTER FOR EDUCATION STATISTICS
Institute of Education Sciences

NAEP
NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS

U.S. Department of Education
Institute of Education Sciences
NCES 2007–466

The Nation's Report Card™

Problem Solving in Technology-Rich Environments

A Report From the NAEP Technology-Based Assessment Project, Research and Development Series

The National Assessment of Educational Progress

What Is The Nation's Report Card™?

The Nation's Report Card™ informs the public about the academic achievement of elementary and secondary students in the United States. Report cards communicate the findings of the National Assessment of Educational Progress (NAEP), the only continuing and nationally representative measure of achievement in various subjects over time. The Nation's Report Card™ compares performance among states, urban districts, public and private schools, and student demographic groups.

For over three decades, NAEP assessments have been conducted periodically in reading, mathematics, science, writing, history, geography, and other subjects. By making objective information available on student performance at the national, state, and local levels, NAEP is an integral part of our nation's evaluation of the condition and progress of education. Only information related to academic achievement and relevant variables is collected. The privacy of individual students is protected, and the identities of participating schools are not released.

NAEP is a congressionally mandated project of the National Center for Education Statistics within the Institute of Education Sciences of the U.S. Department of Education. By law, the Commissioner of Education Statistics is responsible for carrying out the NAEP project. The National Assessment Governing Board (The Governing Board) oversees and sets policy for NAEP. The Governing Board is an independent, bipartisan group composed of 26 representatives from throughout the United States, including state and local officials, educators, business leaders, and members of the general public.

The National Assessment Governing Board

NATIONAL CENTER FOR EDUCATION STATISTICS
Institute of Education Sciences

NAEP
NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS

Problem Solving in Technology-Rich Environments

A Report From the NAEP Technology-Based Assessment Project, Research and Development Series

U.S. Department of Education
Institute of Education Sciences
NCES 2007–466

August 2007

Randy Elliot Bennett
Hilary Persky
Andrew R. Weiss
Frank Jenkins
Educational Testing Service

In collaboration with
Douglas Forer
Bruce Kaplan
Michael Wagner
Lou Mang
Educational Testing Service

William Tirre
Project Officer
National Center for Education Statistics

U.S. Department of Education
Margaret Spellings
Secretary

Institute of Education Sciences
Grover J. Whitehurst
Director

National Center for Education Statistics
Mark Schneider
Commissioner

The National Center for Education Statistics (NCES) is the primary federal entity for collecting, analyzing, and reporting data related to education in the United States and other nations. It fulfills a congressional mandate to collect, collate, analyze, and report full and complete statistics on the condition of education in the United States; conduct and publish reports and specialized analyses of the meaning and significance of such statistics; assist state and local education agencies in improving their statistical systems; and review and report on education activities in foreign countries.

NCES activities are designed to address high-priority education data needs; provide consistent, reliable, complete, and accurate indicators of education status and trends; and report timely, useful, and high-quality data to the U.S. Department of Education, the Congress, the states, other education policymakers, practitioners, data users, and the general public. Unless specifically noted, all information contained herein is in the public domain.

We strive to make our products available in a variety of formats and in language that is appropriate to a variety of audiences. You, as our customer, are the best judge of our success in communicating information effectively. If you have any comments or suggestions about this or any other NCES product or report, we would like to hear from you. Please direct your comments to:

National Center for Education Statistics
Institute of Education Sciences
U.S. Department of Education
1990 K Street NW
Washington, DC 20006-5651

August 2007

The NCES World Wide Web Home Page address is http://nces.ed.gov.
The NCES World Wide Web Electronic Catalog address is http://nces.ed.gov/pubsearch.

Suggested Citation:

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 Project* (NCES 2007–466). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

For More Information

Content contact:
William Tirre
202-502-7361
William.Tirre@ed.gov

To obtain single copies of this report (limited number of copies available), or for ordering information on other U.S. Department of Education products, call toll free 1-877-4ED-PUBS (877-433-7827), or write:

Education Publications Center (ED Pubs)
U.S. Department of Education
P.O. Box 1398
Jessup, MD 20794-1398
TTY/TDD 1-877-576-7734
FAX 1-301-470-1244

The work upon which this publication is based was performed for the National Center for Education Statistics, Institute of Education Sciences, by Educational Testing Service and Westat.

Executive Summary

The Problem Solving in Technology-Rich Environments (TRE) study is the last of three field investigations in the National Assessment of Educational Progress (NAEP) Technology-Based Assessment Project, which explores 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. (See appendix B for detailed information about the TRE sample selection.) 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. 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 models 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 Scales and Results

Because the TRE study used measures that are experimental, data were analyzed to explore how well the TRE scenario scales captured the skills they were intended to summarize. For each scenario, the following measures were obtained: 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. Results were considered to be statistically significant if the probability of obtaining them by chance alone did not exceed the .05 level.

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. This report will discuss the ability of a nationally representative student sample to solve problems using technology in the TRE context. However, the results pertain to student performance in only two scenarios employing a limited set of technology tools and a range of science content sufficient only for demonstration purposes. Therefore, results cannot be generalized more broadly to problem-solving in technology-rich environments for the nation's eighth-graders.

The Search Scales and Results

TRE Search consisted of 11 items (or observables) and produced a total score and two subscores, scientific inquiry and computer skills.

- 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 provided overlapping but not redundant 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. No significant differences were found, however, for reporting groups categorized by gender.

The TRE Simulation Scenario Scales 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.

- The internal consistency of the four scales 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 provided overlapping but not redundant 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. No significant differences were found, however, for reporting groups categorized by gender or school location.

Foreword

The Research and Development series of reports has been initiated for the following goals:

1. To share studies and research that are developmental in nature. The results of such studies may be revised as the work continues and additional data become available.
2. To share results of studies that are, to some extent, on the cutting edge of methodological developments. Emerging analytical approaches and new computer software development often permit new, and sometimes controversial, analysis to be done. By participating in "frontier research," we hope to contribute to the resolution of issues and improved analysis.
3. To participate in discussions of emerging issues of interest to educational researchers, statisticians, and the federal statistical community in general. Such reports may document workshops and symposiums sponsored by the National Center for Education Statistics (NCES) that address methodological and analytical issues or may share and discuss issues regarding NCES practice, procedures, and standards.

The common theme in all three goals is that these reports present results or discussions that do not reach definitive conclusions at this point in time, either because the data are tentative, the methodology is new and developing, or the topic is one on which there are divergent views. Therefore, the techniques and inferences made from the data are tentative and are subject to revision. To facilitate the process of closure on the issues, we invite comment, criticism, and alternatives to what we have done. Such responses should be directed to:

Marilyn M. Seastrom
Chief Statistician
Statistical Standards Program
National Center for Education Statistics
1900 K Street NW, Suite 9000
Washington, DC 20006

Acknowledgments

The NAEP Problem Solving in Technology-Rich Environments (TRE) study was part of the Technology-Based Assessment (TBA) project, a collaborative effort led by the National Center for Education Statistics (NCES) and the National Assessment Governing Board, and carried out by Educational Testing Service (ETS) and Westat. The project was funded through NCES, in the Institute of Education Sciences of the U.S. Department of Education. We appreciate the support and guidance of Acting Commissioner of Education Statistics Grover J. Whitehurst, NAEP project director Suzanne Triplett, TBA project directors Holly Spurlock and William Tirre, and NCES consultants Vonda Kiplinger and Bob Evans.

NAEP is grateful to the students and school staff who participated in the study, and to the Westat staff who administered the study.

NAEP activities at ETS were directed by Stephen Lazer and John Mazzeo, with assistance from John Barone. The ETS management for the TBA project included Randy Bennett, James S. Braswell, Beth Durkin, Christine O'Sullivan, and Clyde Reese.

The TRE study was managed and coordinated by Hilary Persky of ETS. Development and scoring activities were conducted at ETS with contributions from Malcolm Bauer, Dwight Bawcom, Mary Lauko, Claudia Leacock, Christine O'Sullivan, and Andrew Weiss. Janet Stumper designed the Simulation scenario interface with the assistance of Debbie Pisacreta.

Data analysis activities were carried out by Kevin Bentley, Frank Jenkins, Gerry Kokolis, and John Willey. Katharine Pashley conducted the TRE database work.

The design and production of this report were overseen by Loretta Casalaina with assistance from Rick Hasney, Parul Kiran, and Susan Mills. Ming Kuang coordinated the documentation and data-checking procedures with assistance from Kit Chan and Janice Goodis. Carmen Payton and Joanne O'Brien reviewed tabular presentations for consistency with NCES standards. Arlene Weiner coordinated the editorial and proofreading procedures with assistance from Mary Daane, Alice Kass, Janice Lukas, and Linda Myers. The Web version of this report was coordinated by Rick Hasney.

This project could not have been completed without Westat, which conducted student sampling, administration, field support, and weighting. Westat's activities were managed by Dianne Walsh, Dward A. Moore, Jr., Brice Hart, and Brenda Ennis, with the assistance of Nia Davis. Sampling and weighting were conducted by Louis Rizzo and Tom Krenzke. Weighting systems work was completed by Bill Wall. The day-to-day management of Westat's technical and software systems was accomplished by Brice Hart with assistance from Fran Cohen and Karen Dennis.

Many thanks are due to the numerous reviewers, both internal and external to NCES and ETS. The comments and critical feedback of the following reviewers are reflected in the final version of this report: Tajuana Bates, Deven Carlson, Young Chun, Mary Crovo, Aaron Douglas, Ray Fields, Hilary Forster, Stephanie Germeraad, Arnold Goldstein, Steve Gorman, Fraser Ireland, Andrew Kolstad, Susan Loomis, Teresa Neidorf, Taslima Rahman, Linda Shafer, William Tirre, Alan Vanneman, Melissa Vrana, David Williamson, Matthias Von Davier, and Rima Zobayan.

Contents

List of Tables

Page

List of Figures

Page

Introduction

For more than 30 years, the National Assessment of Educational Progress (NAEP) has regularly collected, analyzed, and reported valid and reliable information about what American students know and can do in a range of subject areas. As authorized by the U.S. Congress, NAEP typically assesses nationally representative samples of students in grades 4, 8, and 12. Since 1990, NAEP has also assessed representative samples of students at grades 4 and 8 in states and other jurisdictions that participate in the NAEP state-by-state assessments. In 1988, Congress established the National Assessment Governing Board to oversee and set policy for NAEP.

In response to the ever-increasing importance of technology in educational and workplace settings, and to maintain its leadership role in the area of large-scale assessment, NAEP initiated the Technology-Based Assessment (TBA) Project in 1999. The TBA Project was intended to explore the many uses of new technology in NAEP, among them specific NAEP processes (e.g., item creation, test delivery), assessment of specific content domains, and assessment of technology skills.

The TBA Project focused on several key questions:

1. *What are the measurement implications of using technology-based assessment in NAEP?* Technology-based assessment may change the meaning of our measures in unknown ways. It may allow assessment of skills that could not be measured using paper and pencil or preclude measuring skills that could be tested by conventional means. It may allow the assessment of emerging skills, particularly those requiring students to employ new technology in learning and problem solving.
2. *What are the implications for equity?* If not carefully designed, technology-based assessment could inaccurately reflect the skills of some groups of students, especially those with differing degrees of access to computers. At the same time, it could increase participation of students with disabilities. It may also better reflect the skills of students who routinely use the computer to perform academic tasks like writing.
3. *What are the efficiency implications of using technology-based assessment compared with paper and pencil?* Along with other new technologies, the Internet may afford significant time and cost savings for large-scale assessments.
4. *What are the operational implications of technology-based assessment?* Moving from a paper-based program to an electronic one raises significant issues concerning school facilities, equipment functioning, administrator responsibilities, and school cooperation.

To answer these questions, the NAEP program undertook three empirical studies with students: Math Online (MOL; Sandene at al. 2005), Writing Online (WOL; Horkay et al. 2005), and Problem Solving in Technology-Rich Environments (TRE).

The MOL and WOL studies were designed to investigate the effects of delivering existing paper tests via computer. In contrast, the TRE study was designed to demonstrate and explore innovative uses of computers in NAEP by developing two sample extended problem-solving scenarios. This report describes the methodology, technology, and results of the TRE study.

The TRE Project was guided by several principles:

1. *TRE should use the computer to do what cannot easily be done on paper.* The TRE scenarios allow students to answer questions by searching electronic databases and by using a simulation tool to conduct experiments. All student actions are captured by computer for later scoring, allowing for evaluation of the processes used in problem solving. These capabilities could not be easily achieved with conventional paper-and-pencil testing. Chapter 1 of this report describes in detail the two grade 8 TRE problem-solving scenarios—the Search scenario and the Simulation scenario.
2. *TRE should represent the type of problem solving done with computers in educational and work environments.* TRE attempts to capture the multidimensionality characteristic of problem solving with technology by requiring students to demonstrate both science skills and basic facility with the computer. Further, technology in TRE is used as a means of solving substantive problems, rather than as an end in itself.
3. *To the degree possible, TRE should allow the disentangling of component skills.* The two TRE scenarios were intended to measure both basic computer skills and science skills in an integrated way; that is, students would need to use both skill sets simultaneously to solve the problems in the scenarios. For example, students were required to demonstrate mastery of searching for information in a World Wide Web environment, but this skill was to be used in a specific scientific domain that demanded the ability to select and synthesize relevant scientific material.

A consequence of this close integration of skills, however, is that a deficiency in one skill can prevent the expression of another. The TRE team sought to limit such occurrences in several ways. For example, to reduce the chances that limited computer skills would keep students from showing their science skills, tutorials were supplied to help students understand the scenario interfaces, common interface conventions were used (e.g., dialog boxes and wizards), and a computer-related help function was made available. To prevent lack of science skills from impeding the demonstration of computer skills, students were supplied with a science help tool to access basic information relevant to both scenarios; the Simulation interface tools were organized to facilitate a structured inquiry process built around designing experiments, running experiments, and interpreting results; certain choices in the Simulation scenario were constrained (e.g., the choice of variables to include on each graph axis); and the Simulation scenario began with a relatively simple problem. Finally, an interpretive framework was used that allowed for the simultaneous estimation of related proficiencies.

4. *TRE should be positioned so it can inform the development of a future assessment of emerging skills or of more traditional subject matter.* It should be possible to incorporate meaningful exercises using a simulation tool or electronic information search into existing NAEP subject-matter assessments; for example, a likeness of the TRE Simulation scenario could find a logical place in the NAEP science assessment to measure skills needed for scientific investigation. It should also be possible to use the TRE scenarios as models for measures of problem solving with technology generally.
5. *TRE should be an assessment, not instruction, but students should be able to learn from it incidentally.* Both scenarios involve discovery; hence, students may learn from working with the TRE scenarios in a way that participation in the typical large-scale assessment does not provide.

Overview of the Study

Educational Testing Service (ETS) assessment development and research staff created the two TRE scenarios with expert input and reviews from a TRE Development Committee. The committee was composed of science and technology educators and curriculum experts. (The membership of this committee can be found in appendix A.) NCES staff provided oversight and guidance as to the appropriate direction and nature of the scenarios. The development of the TRE scenarios was further informed by a variety of sources, among them the NAEP Science Framework (National Assessment Governing Board 2000) and current research in problem solving and scientific inquiry. Also important were various state and national science and technology standards, including the National Science Education Standards (National Academy of Sciences 1996) and the National Educational Technology Standards (International Society for Technology in Education 2002).

The scenarios were created for 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 between a sixth-grade and an eighth-grade level. NAEP project staff assumed that most grade 8 students have at least basic computer skills because the 2002 NAEP Writing Online data suggest that virtually all students use computers for schoolwork at least to some extent (Horkay et al. 2005). Further, because of the prevalence of experimental methodology and physics content in grade 8 science curricula, NAEP project staff assumed that members of the grade 8 population have had some basic exposure to scientific inquiry and to basic concepts of buoyancy, mass, and volume.[1]

The TRE study tested 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. (See appendix B for detailed information about the TRE sample selection.) Students were assigned randomly within each school to one of the scenarios—Search or Simulation. For both scenarios, data were collected about student demographics; students' access to, use of, and attitudes toward computers; and students' science coursetaking and activities in school. Additionally, before starting each scenario, students answered prior knowledge questions designed to determine the degree to which they had the computer and/or science knowledge and skills being assessed.

Staff members employed by Westat, the NAEP data collection contractor, administered the TRE scenarios and proctored all administrations using procedures generally similar to those employed for NAEP assessments. Testing was conducted either on school

[1] A range of state curricula surveyed by the authors included experimental activities and methods as well as mastery of the basic concepts of buoyancy, mass, and volume at the eighth-grade (middle school) level. Two typical examples are state middle school curricula for North Carolina and Massachusetts (North Carolina State Department of Education 2004; Massachusetts Department of Education 2001).

computers connected to the Internet or on laptop computers brought in by NAEP administrators. All computers, whether supplied by the school or by NAEP, had to meet minimum hardware and software specifications to ensure that the test would operate uniformly (see appendix C for these specifications). NAEP staff at ETS conducted the scoring and analysis of results.[2]

Analysis of student responses was conducted for two purposes. The first purpose was to evaluate the functioning of the TRE scenarios. The analyses included internal consistency, the relations of student scores to students' prior knowledge, the TRE scale intercorrelations, the correlations of each observable with the TRE subscales, the locations of the observables on the scales, the response probabilities for prototypic students (i.e., hypothetical students with different levels of proficiency), and the relations of relevant student background information to performance. The second purpose was to describe student performance on the scenarios in quantitative and qualitative terms. For differences in mean scores and for differences from zero of correlation coefficients, .05 was used as the level for deciding that a result was statistically significant, with score differences between group means evaluated for statistical significance using independent *t*-tests.

Chapter 1 of this report describes in detail two grade 8 TRE problem-solving scenarios—the Search scenario and the Simulation scenario. Chapter 2 describes how the TRE team used Evidence-Centered Design (ECD; Mislevy, Almond, and Lukas 2003; Mislevy et al. 2001) to help develop an interpretive framework for translating the multiplicity of actions captured from each student who took TRE into inferences about student proficiency. Chapter 3 describes TRE student responses to background questions concerning computer use, attitudes toward computers, and engagement in school science. Chapter 4 discusses how the evaluation rules, or scoring criteria, developed using ECD were applied to student performances by both machine and human scoring, and chapters 5 and 6 present the results of analyses of student performance. Finally, chapter 7 summarizes the TRE study results.

The appendixes that appear in this report are as follows: appendix A lists the members of the TRE Development Committee; appendix B discusses the TRE assessment sample selection process; appendix C identifies the computer specifications for schools that participated in the TRE assessment; appendix D presents the prior-knowledge computer and science questions students took before each scenario, and the background questions students responded to when they had completed the scenarios; appendix E shows the Simulation scenario tutorial and individual screens from the Computer and Science Help in the Simulation scenario; appendix F discusses the use of Bayesian estimation in the study; appendix G lists the rules used for the ETS automated scoring tool, c-rater, for scoring students' search queries; appendix H presents the Search and Simulation scenario scale scores and percentiles by student reporting groups; appendix I presents summary statistics for prior-knowledge measures and mean scale scores for background-question response options; appendix J shows student performance on observables for the Search and Simulation scenarios; and appendix K presents definitions for each of the TRE student reporting groups.

Limitations of the Study

Readers are reminded that the TRE project results pertain to student performance in only two scenarios. These scenarios employed a limited number of technology tools and a range of content sufficient for demonstration purposes only.

A second limitation is that the TRE study was not based on an existing NAEP content-area framework. As such, the conceptualization of the TRE construct domain used in this study did not involve the broad representation of diverse constituencies typical of NAEP assessment frameworks.

A third limitation is that the TRE assessment instruments and analysis methods were experimental ones drawing upon extended computer-delivered performance tasks and Bayesian modeling methods not previously used in NAEP assessments.

Because of these limitations, TRE study results should not be generalized to problem solving in technology-rich environments for the nation's eighth-graders, nor should they be used to draw general conclusions about the science knowledge or computer skills of those students.

[2] No analysis of performance on laptops vs. school computers was conducted because the meaning of any observed performance differences would be ambiguous. Since the assignment of students to computer type was not done at random but rather according to the fit of school technology infrastructure with the requirements of the test delivery system, performance differences could be caused by differences in other factors related to the quality of school technology (e.g., in socioeconomic status) and not by differences in the suitability of one or the other computer type for online assessment. Further, there were no measures of skill independent of computer type that could have been used to adjust statistically for pre-existing differences between groups. But see Horkay et al. 2005 for an analysis of performance differences on laptops vs. school computers for 8th-grade students.

Chapter 1: The TRE Construct Domain and Problem-Solving Scenarios

The TRE Construct Domain

There is no existing NAEP framework for the domain of “Problem Solving in Technology-Rich Environments.” As a result, that construct domain needed to be defined by the TRE team—i.e., the research scientists, test developers, and a Development Committee of technology and science education advisors who worked on the project (see appendix A for Development Committee membership). The domain definition process involved drawing upon a variety of sources, including national education standards in technology and science, relevant research literature, and the expertise and experience of the Development Committee. The resulting domain conceptualization, described below, served as the basis for creating the experimental measures used in this demonstration project. Readers should recognize that this conceptualization process did not involve the broad representation of diverse constituencies typical of NAEP assessment frameworks, and the conclusions drawn from TRE study results should, therefore, be limited accordingly.

The domain of “Problem Solving in Technology-Rich Environments” (TRE) was conceptualized as the intersection of content areas and technology environments. Problem solving with technology can occur in a range of content areas, such as biology, physics, economics, and history. Similarly, various technology environments such as databases, text editors, simulation tools, dynamic visual displays of information, spreadsheets, and presentation tools can be used to solve problems in these content areas.

The TRE team chose to sample from the universe of content areas and technology environments so that one content area—the physical science associated with helium gas balloons used for space exploration—carried through different technology environments. Using the same content across technology environments is consistent with the emphasis in the research literature on extended problem solving because the student remains situated in the same context throughout the assessment and, thus, has greater opportunity to apply response processes that might not be engaged by presenting a series of more elemental, unrelated tasks (Baxter and Glaser 1998; Nichols and Sugrue 1999). In addition, emphasizing content expresses the view that, in real-world settings, problem solving with technology is driven by the problem, and not by the technology.

Science was chosen as a content area because computers are used routinely as scientific problem-solving tools in advanced academic and work environments, and because these tools are increasingly being used in secondary school for instructional purposes. Further, a range of state middle school science standards, the National Education Technology Standards, and the National Science Education Standards typically cite scientific inquiry, problem solving with technology, and the use of simulation as key proficiencies (International Society for Technology in Education 1998; National Academy of Sciences 1996). The topic of helium gas balloons was selected because it is a working application of fundamental physical principles, like buoyancy and its relationship to mass and volume, in a context expected to be engaging to middle school students.

Figure 1-1 represents the TRE conception of problem solving with technology. In the figure, the TRE measure is indicated within the content area of phys-

Figure 1-1. Domain conception for problem solving in Technology-Rich Environments, grade 8: 2003

Content area	Database	Text editor	Simulation	Dynamic visual display of information[1]	Interactive feedback	Spreadsheet	Presentation and communication tools
Biology							
Ecology							
Physics							
Balloon science[2]	▨	▨	▨	▨	▨		
Economics							
History							

[1] A dynamic display changes in real time.

[2] The shaded area indicates the coverage of the scenarios.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

ics. The specific scenarios developed for the current study incorporate several technology uses within the same problem context, denoted by the shaded area. Note that very different measures would have resulted if the focus had been on a single technology use across different content areas. Also note that, because the measure is an example, it covers only a small portion of this hypothetical problem-solving domain, too small for any inferences to be made from study results to performance in problem solving in technology-rich environments generally.

In developing the Simulation scenario, the TRE team drew on the research of Glaser and associates, as well as that of others (Raghavan, Sartoris, and Glaser 1998; Schauble et al. 1991, 1992; Shute and Glaser 1990, 1991; White and Frederiksen 1998). The common theme running through this research is the "discovery environment." A discovery environment is a "microworld" where a student can experiment to construct an understanding of some underlying phenomenon, often physical in nature. Although these environments have primarily been used for instructional purposes, they also hold promise for assessment.

Among the more compelling models of such environments were the "Deformed Frog" scenario from the Knowledge Integration Environment (KIE) project at the Graduate School of Education at the University of California at Berkeley, which involves students in researching web-based information and testing hypotheses about what is causing deformation among frogs in North America (KIE 1997), and "Smithtown," developed at the University of Pittsburgh (Shute and Glaser 1990). In Smithtown, students learn basic macroeconomics concepts and scientific inquiry skills by conducting experiments in a simulation setting. Therefore, Smithtown was very helpful as a model for how to organize and present a computer-based tool for making and testing hypotheses. The "Jasper" series, developed by the Cognition and Technology Group at Vanderbilt University (although not a computer-based microworld), was an interesting model in which students must discover underlying mathematics and science concepts to solve hands-on design problems (Learning Technology Center 1992). While these projects are set in a variety of content areas, all of them offer students opportunities and sufficient context to form and test hypotheses and draw conclusions about underlying phenomena.

Research done by Schauble was particularly informative for the kinds of reasoning and strategies the TRE team wanted to measure, and what the team sought to avoid, namely, laboratory exercises in which students "follow prescribed procedures and hope to achieve the right answer" (Schauble et al. 1995, p. 133). This kind of activity is also criticized in the NAEP Science Framework:

> Many…so-called performance assessment scenarios…[are] reduced to "follow-the-instructions" problems. No inferences about a student's knowledge of science or its tools and procedures can be drawn from such exercises. (National Assessment Governing Board 2000, p. 33)

Instead, the TRE team sought to design scenarios that would feature (as far as possible in a large-scale assessment—versus a classroom—context) the kind of exploration characteristic of real-world problem solving.

Finally, the Search scenario was based on research about proficient and novice electronic information-finding behaviors of adolescents and adults (Fidel et al. 1999; Klein, Yarnall, and Glaubke 2001; Salterio 1996; Schacter, Chung, and Dorr 1998). Of particular use was a web-search study carried out by the National Center for Research on Evaluation, Standards, and Student Testing (CRESST), which suggested behaviors that might be used as markers of search proficiency (Klein, Yarnall, and Glaubke 2001, 2003). As with the Simulation scenario, the various documents describing standards for students' science and technology skills were also relevant because of their references to electronic information search as a desired proficiency (ISTE 1998; Riley, Holleman, and Roberts 2000).

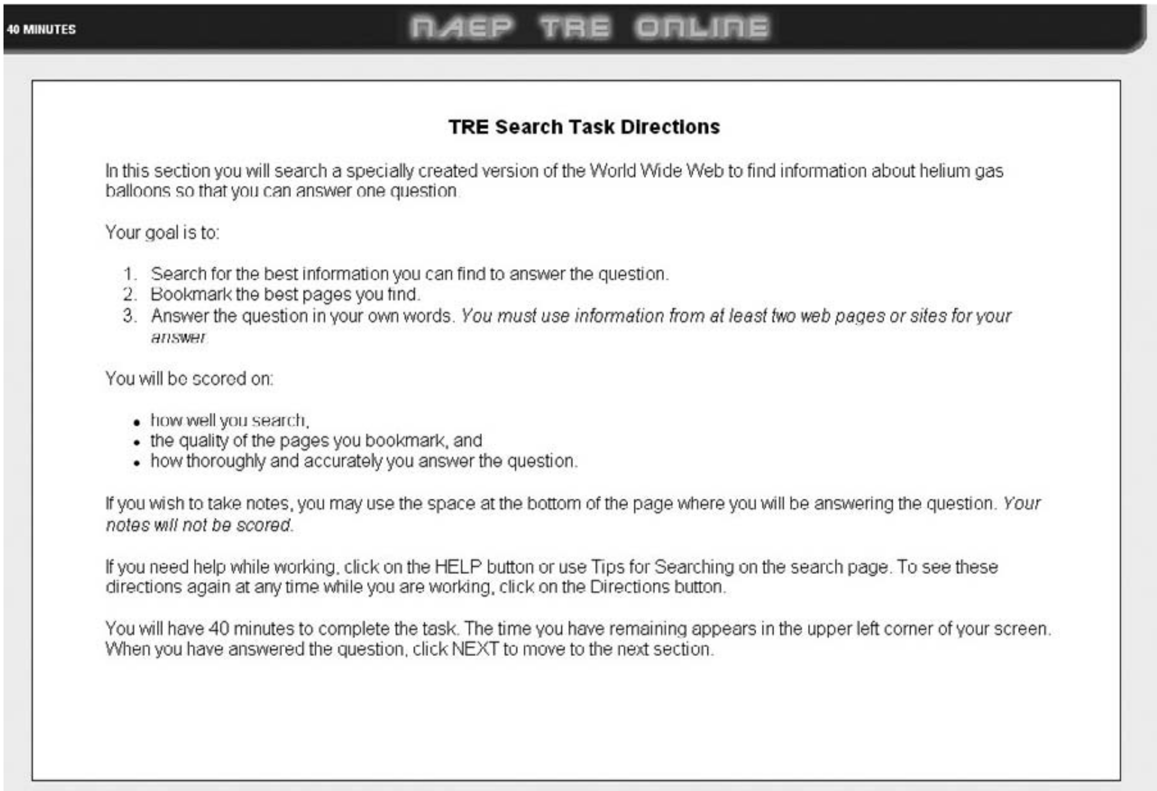
The TRE Problem-Solving Scenarios in Detail

The following section presents the two TRE scenarios in detail as a context for understanding the study. The discussion of the design and components of each scenario is accompanied by selected screen shots.

The TRE Search Problem-Solving Scenario

Figures 1-2 through 1-5 display the progression through the Search scenario. Students first received a set of prior science and computer knowledge questions (shown in appendix D) and worked through a brief (5 minute) tutorial (not shown) to introduce them to the Search interface. They were then shown the scenario directions presented in figure 1-2. The prior knowledge questions were intended to give a rough measure of students' degree of familiarity with the science and computer-related concepts being assessed. Although the Search interface was designed to be as close to a standard web search browser as possible, some features—such as buttons for reading directions and accessing the box to enter answers—are particular to the TRE software.

Figure 1-2. Computer screen with directions for TRE Search scenario, grade 8: 2003



TRE Search Task Directions

In this section you will search a specially created version of the World Wide Web to find information about helium gas balloons so that you can answer one question.

Your goal is to:

1. Search for the best information you can find to answer the question.
2. Bookmark the best pages you find.
3. Answer the question in your own words. *You must use information from at least two web pages or sites for your answer.*

You will be scored on:

- how well you search,
- the quality of the pages you bookmark, and
- how thoroughly and accurately you answer the question.

If you wish to take notes, you may use the space at the bottom of the page where you will be answering the question. *Your notes will not be scored.*

If you need help while working, click on the HELP button or use Tips for Searching on the search page. To see these directions again at any time while you are working, click on the Directions button.

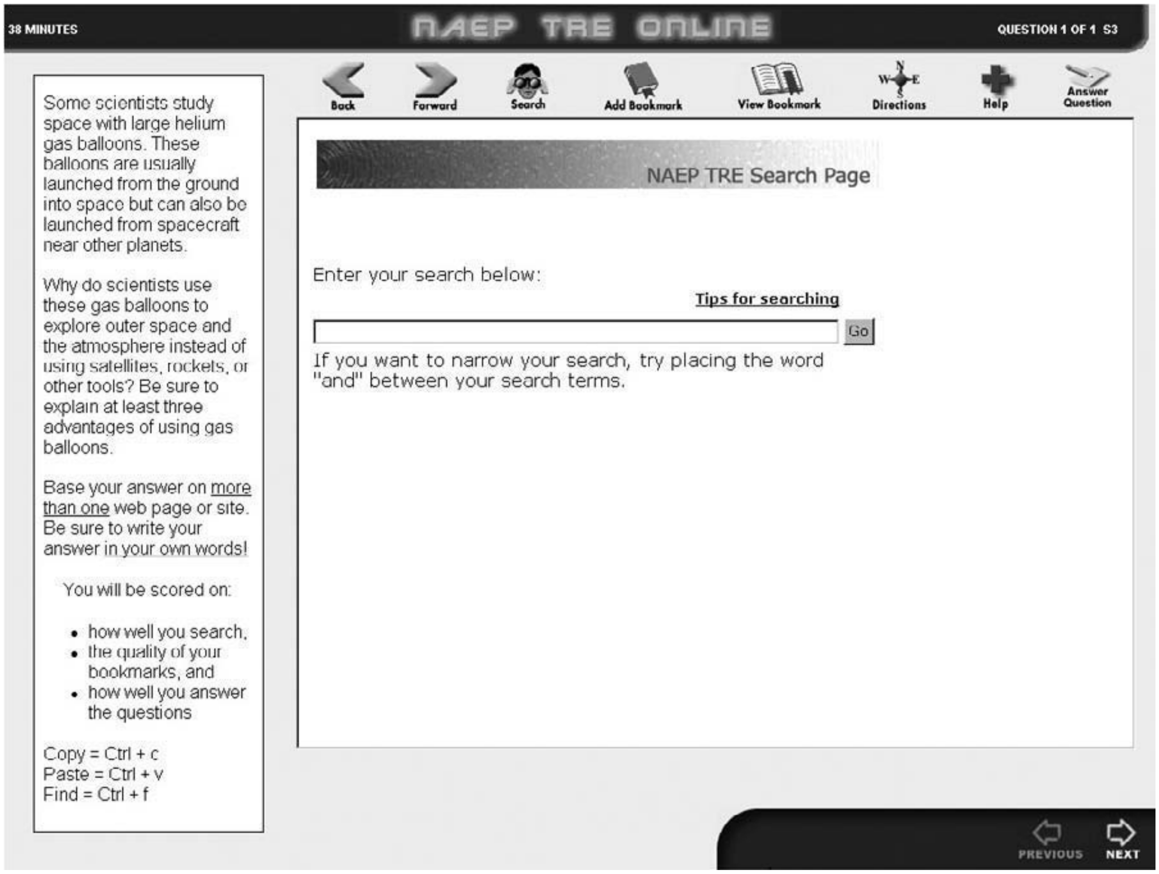
You will have 40 minutes to complete the task. The time you have remaining appears in the upper left corner of your screen. When you have answered the question, click NEXT to move to the next section.

NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

The directions, shown in figure 1-2, were designed to introduce students to the tasks they would be performing and to let them know the basis on which their responses would be evaluated: their searching adequacy, the quality of the information they located, and the quality of their answers to the question (referred to in this report as the "problem") posed to motivate their searching.

After the directions screen, students moved to the Search interface (see figure 1-3) to which they had been introduced in the tutorial. The problem intended to motivate students' searching was located, and always visible, on the left-hand side of the screen. Also always visible was a summary of scoring criteria for students' work. On the right side was a web browser created for the purposes of this TRE scenario. At the top of the browser was a toolbar that included buttons for moving back and forth among pages,

Figure 1-3. Computer screen with TRE Search motivating problem in left pane and web browser in right pane, grade 8: 2003



Some scientists study space with large helium gas balloons. These balloons are usually launched from the ground into space but can also be launched from spacecraft near other planets.

Why do scientists use these gas balloons to explore outer space and the atmosphere instead of using satellites, rockets, or other tools? Be sure to explain at least three advantages of using gas balloons.

Base your answer on more than one web page or site. Be sure to write your answer in your own words!

You will be scored on:

- how well you search,
- the quality of your bookmarks, and
- how well you answer the questions

Copy = Ctrl + c
Paste = Ctrl + v
Find = Ctrl + f

NAEP TRE Search Page

Enter your search below:

Tips for searching

If you want to narrow your search, try placing the word "and" between your search terms.

NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

returning to the search page, bookmarking, viewing bookmarks, getting more extensive directions, receiving science help, and going to a page to take notes or answer the motivating problem. In the center of the browser page were a space for entering queries and a link to tips for searching.

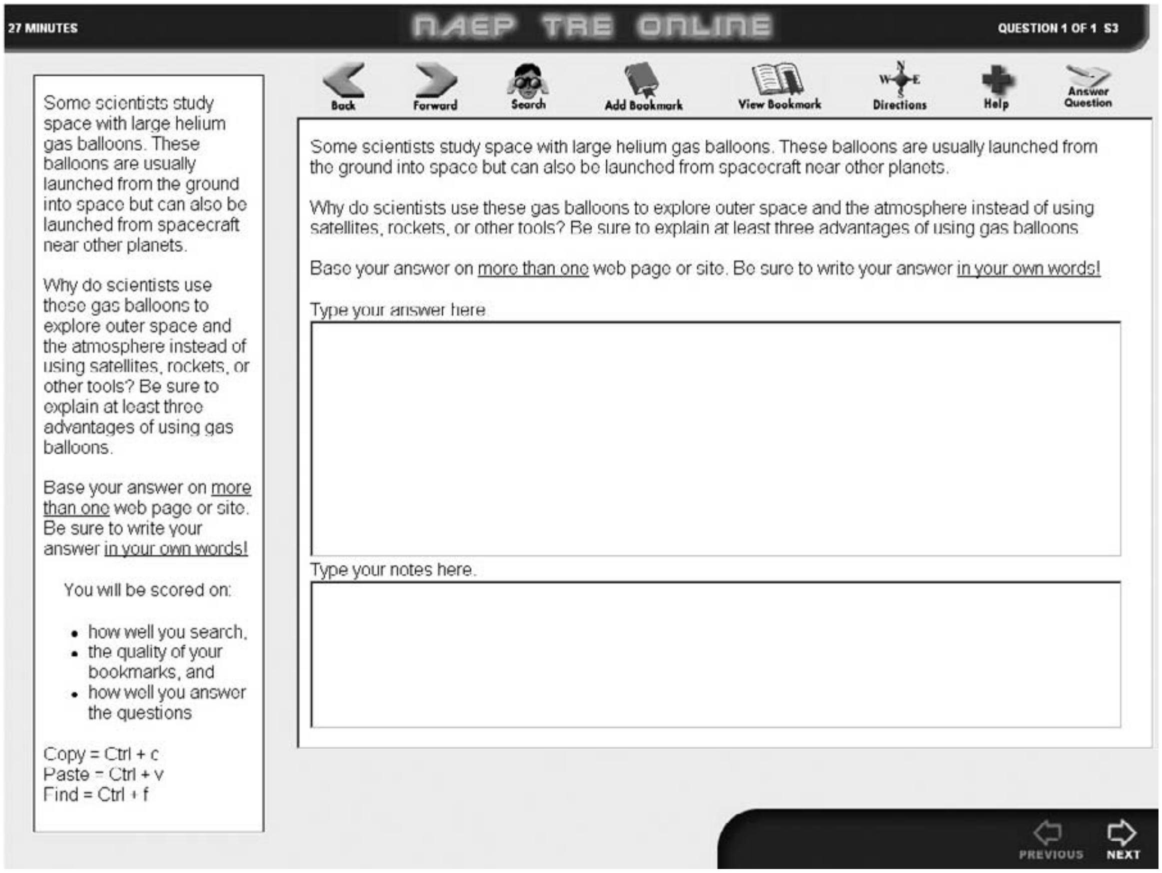
The motivating problem in the left-hand column of the screen, shown again in figure 1-4, was developed over many iterations and pilots of the scenario with students. The problem was designed to be open enough to encourage searching, and yet specific enough so that reasonably skilled searching would supply substantive information to answer it within the 40-minute time allotted for the Search scenario.[3]

Skilled searching using relevant terms from the motivating problem and methods for focusing searches (e.g., quotations, use of "near" and "or") yielded a list of pages, including some suitable for answering the question. Unskilled searching that employed only generic terms from the motivating problem (e.g., "balloon"), on the other hand, yielded less relevant or irrelevant hits.

To ensure that the TRE universe was as authentic as possible and would yield results ranging from the very irrelevant to the highly relevant, with many gradations in between, skilled and unskilled searches were run to identify the kinds of pages students would find by searching the real World Wide Web (WWW). Web pages ranged from those pertaining to party bal-

[3] The motivating problem refers to balloons being launched "into space" because that is how scientists often speak of the upper parts of the atmosphere where the balloons operate. To date, only one balloon has been launched in the atmosphere of another planet (Venus), but several countries have considered using balloons to explore the atmosphere of other planets.

Figure 1-4. Computer screen with spaces for note-taking and for answering TRE Search motivating problem, grade 8: 2003



Some scientists study space with large helium gas balloons. These balloons are usually launched from the ground into space but can also be launched from spacecraft near other planets.

Why do scientists use these gas balloons to explore outer space and the atmosphere instead of using satellites, rockets, or other tools? Be sure to explain at least three advantages of using gas balloons.

Base your answer on more than one web page or site. Be sure to write your answer in your own words!

You will be scored on:

- how well you search,
- the quality of your bookmarks, and
- how well you answer the questions

Copy = Ctrl + c
Paste = Ctrl + v
Find = Ctrl + f

Some scientists study space with large helium gas balloons. These balloons are usually launched from the ground into space but can also be launched from spacecraft near other planets.

Why do scientists use these gas balloons to explore outer space and the atmosphere instead of using satellites, rockets, or other tools? Be sure to explain at least three advantages of using gas balloons.

Base your answer on more than one web page or site. Be sure to write your answer in your own words!

Type your answer here.

Type your notes here.

NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

loons, which would be returned to students who used undirected search queries such as "balloons," to pages from NASA describing uses of gas balloons in space research. Once many thousands of pages had been collected, a NAEP staff member assigned scores to each page based on the relevance of the page to the Search scenario motivating problem. The scores ranged from a low of 1 to a high of 4. Two additional NAEP staff members also rated all the pages considered by the first rater to have at least some relevance, i.e., all pages scored at least a "2." Any differences in scores assigned were resolved among the raters to ensure that pages were properly scored for relevance.[4] Ultimately, a sample of some 5,000 pages from the World Wide Web was selected and used as the TRE web universe.

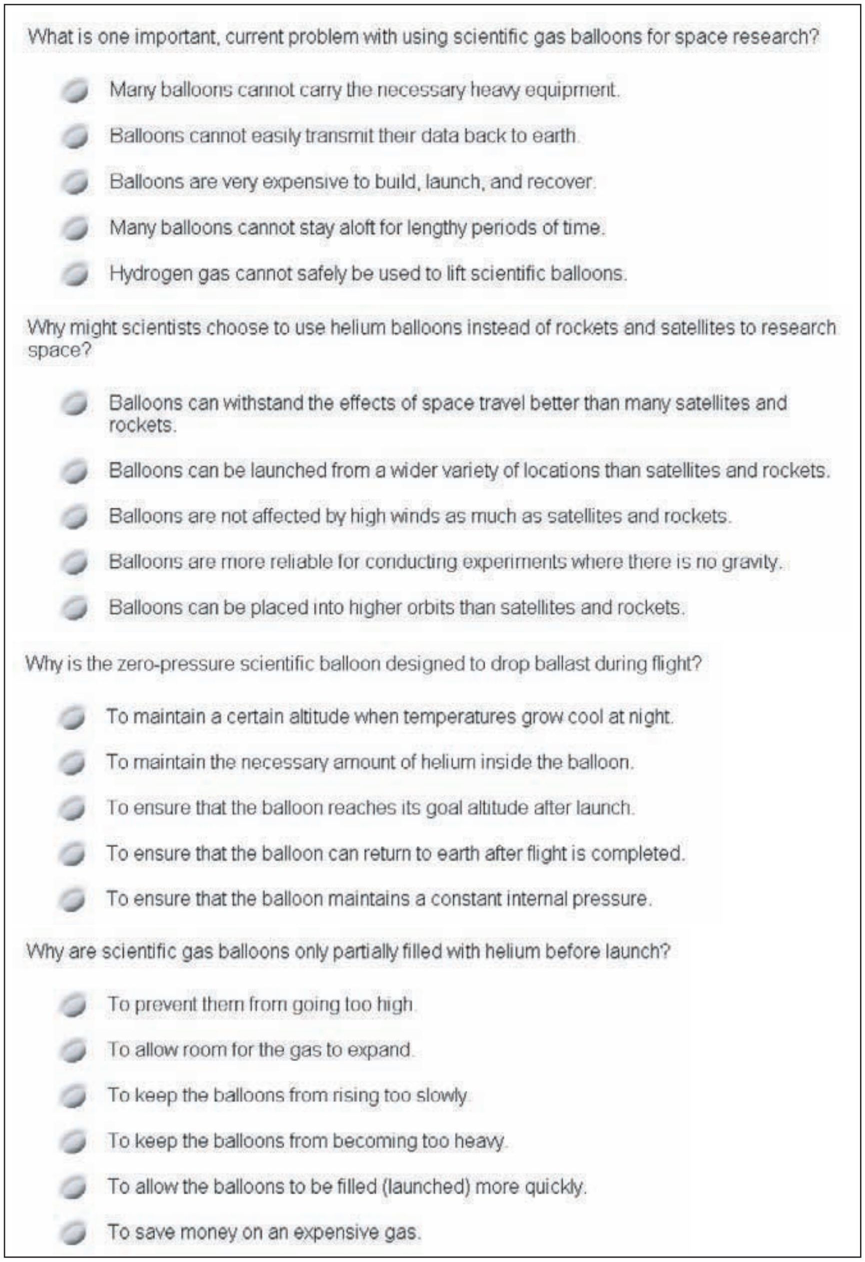
To maximize authenticity, students could use the tool bar to cycle among searching, bookmarking, and other activities, including responding to the motivating problem. Figure 1-4 shows the box for entering both the response and any notes made while searching. Students were permitted to take notes but were told in the initial directions that their notes would not be scored.[5]

[4] Ratings were done by NAEP assessment development staff members and associates with graduate degrees according to criteria defined by the rating group. Because group discussions of exemplars indicated that irrelevant pages were easily agreed upon, only pages receiving a score of at least "2" were independently rescored.

[5] As in any real-world, information-search task, students in the TRE study could have used non-technological alternatives like paper-and-pencil or memory in place of electronic note-taking. The extent to which such alternatives were used could not be determined.

To ensure that a response was collected from each student, students could not leave the Search task without entering some text into the answer space. Once students had made some attempt to answer the question, they were given (assuming time had not run out) the option of reviewing their work. They were then moved to a set of four multiple-choice questions designed to test how well they had synthesized the information they had found about the use of helium gas balloons in space exploration. As with the motivating question, students could search while answering. Figure 1-5 displays the synthesizing questions.

Figure 1-5. TRE Search synthesizing questions and answer options, grade 8: 2003



What is one important, current problem with using scientific gas balloons for space research?

○ Many balloons cannot carry the necessary heavy equipment.
○ Balloons cannot easily transmit their data back to earth.
○ Balloons are very expensive to build, launch, and recover.
○ Many balloons cannot stay aloft for lengthy periods of time.
○ Hydrogen gas cannot safely be used to lift scientific balloons.

Why might scientists choose to use helium balloons instead of rockets and satellites to research space?

○ Balloons can withstand the effects of space travel better than many satellites and rockets.
○ Balloons can be launched from a wider variety of locations than satellites and rockets.
○ Balloons are not affected by high winds as much as satellites and rockets.
○ Balloons are more reliable for conducting experiments where there is no gravity.
○ Balloons can be placed into higher orbits than satellites and rockets.

Why is the zero-pressure scientific balloon designed to drop ballast during flight?

○ To maintain a certain altitude when temperatures grow cool at night.
○ To maintain the necessary amount of helium inside the balloon.
○ To ensure that the balloon reaches its goal altitude after launch.
○ To ensure that the balloon can return to earth after flight is completed.
○ To ensure that the balloon maintains a constant internal pressure.

Why are scientific gas balloons only partially filled with helium before launch?

○ To prevent them from going too high.
○ To allow room for the gas to expand.
○ To keep the balloons from rising too slowly.
○ To keep the balloons from becoming too heavy.
○ To allow the balloons to be filled (launched) more quickly.
○ To save money on an expensive gas.

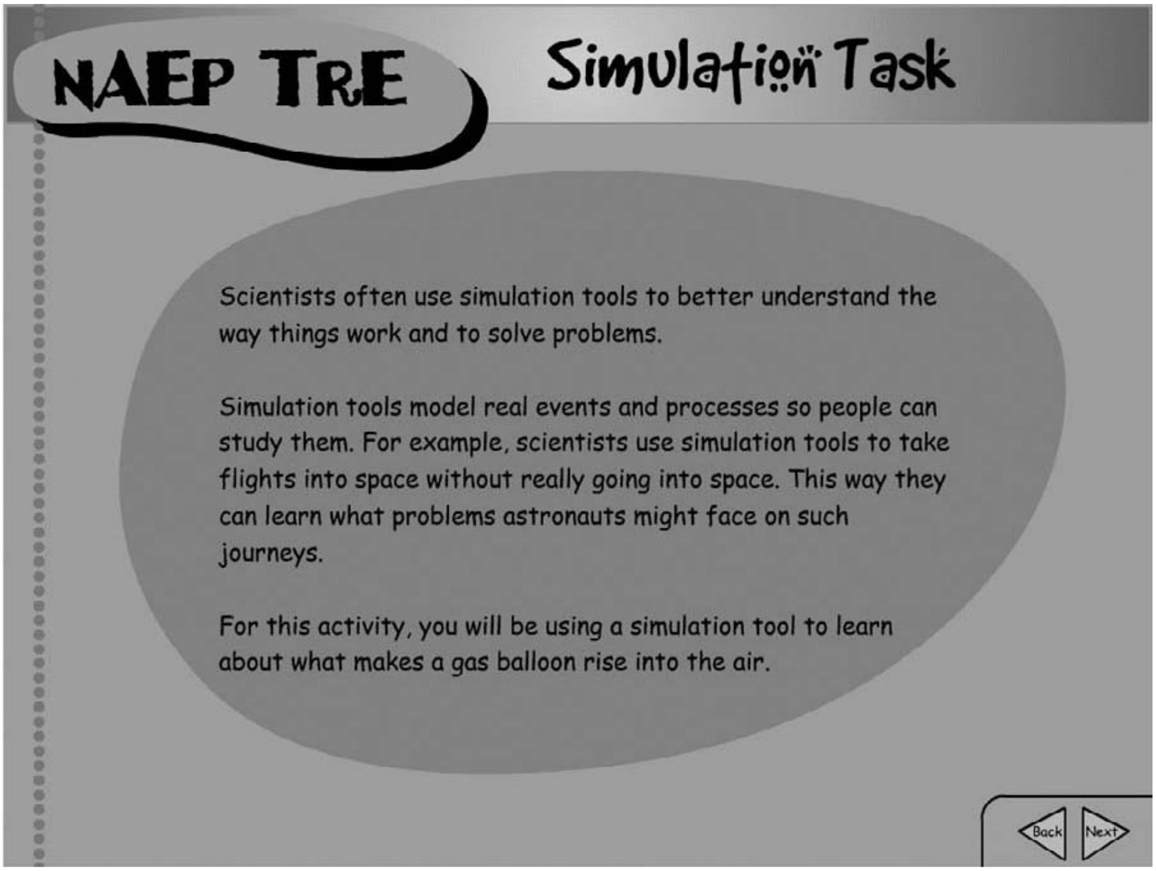
NOTE: TRE = Technology-Rich Environments. Questions were presented individually, one per screen, and not as shown here.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Students had to answer all four multiple-choice synthesizing questions before they could leave the Search scenario. After completing the scenario, students responded to background questions intended to gather information about their demographic characteristics, school science classes and activities, and computer familiarity. (The full text of the background questions is available in appendix D.) A detailed discussion of the percentages of students in various background-question response categories appears in chapter 3 of this report.

The TRE Simulation Problem-Solving Scenario

Figures 1-6 through 1-22 illustrate the progression students followed through the Simulation scenario. Figures 1-6 and 1-7 display the introduction that students received after they responded to a set of prior science and computer knowledge questions, as they did for the Search scenario. The introductory pages told students the purpose of simulation tools generally and what kind of simulation tool they would be working with during the course of the scenario, and then explained how they would be applying the simulation tool. “Back” and “Next” buttons on the lower right-hand side of the screen allowed students to navigate among the Simulation scenario pages, so they could review the introductory pages.

Figure 1-6. Computer screen introducing use of simulation tools in science for the TRE Simulation scenario, grade 8: 2003



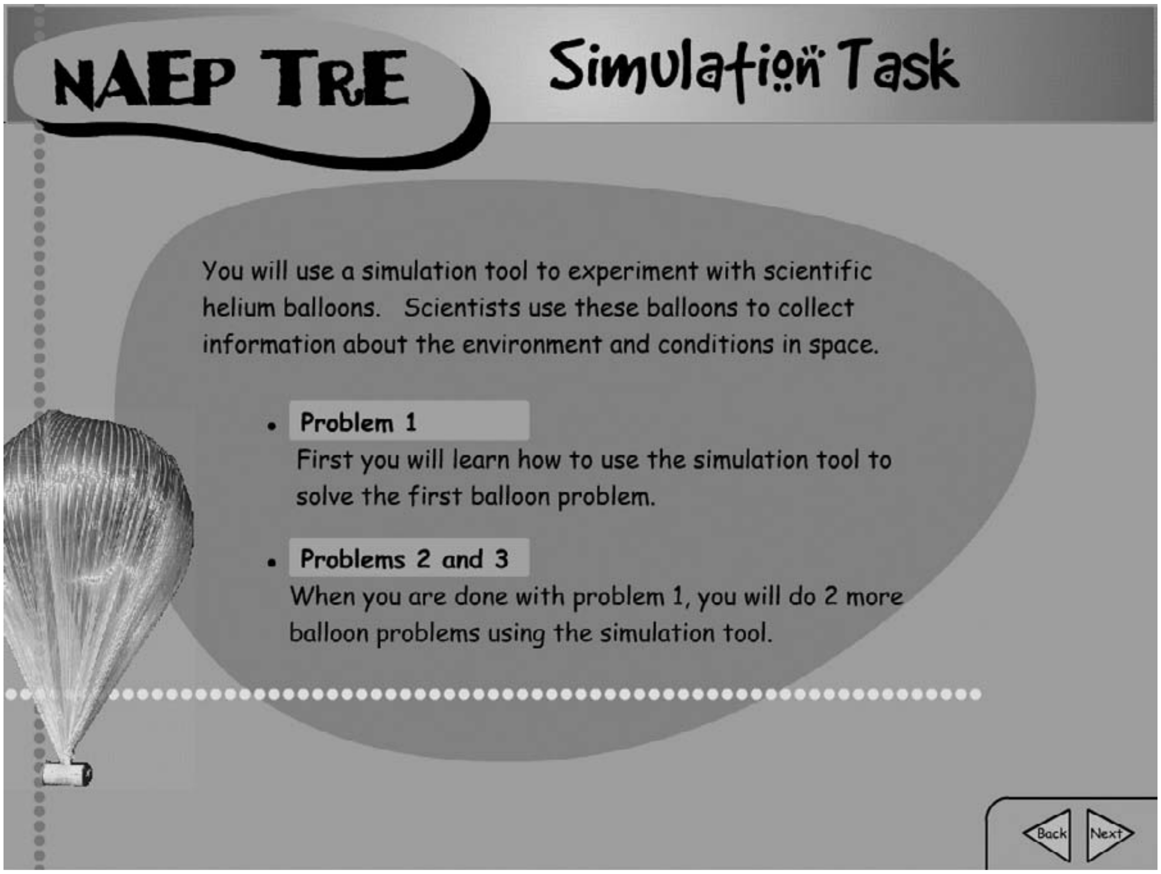
Scientists often use simulation tools to better understand the way things work and to solve problems.

Simulation tools model real events and processes so people can study them. For example, scientists use simulation tools to take flights into space without really going into space. This way they can learn what problems astronauts might face on such journeys.

For this activity, you will be using a simulation tool to learn about what makes a gas balloon rise into the air.

NOTE: TRE = Technology-Rich Environments..
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure 1-7. Computer screen introducing content of the TRE Simulation scenario, grade 8: 2003



NAEP TRE

Simulation Task

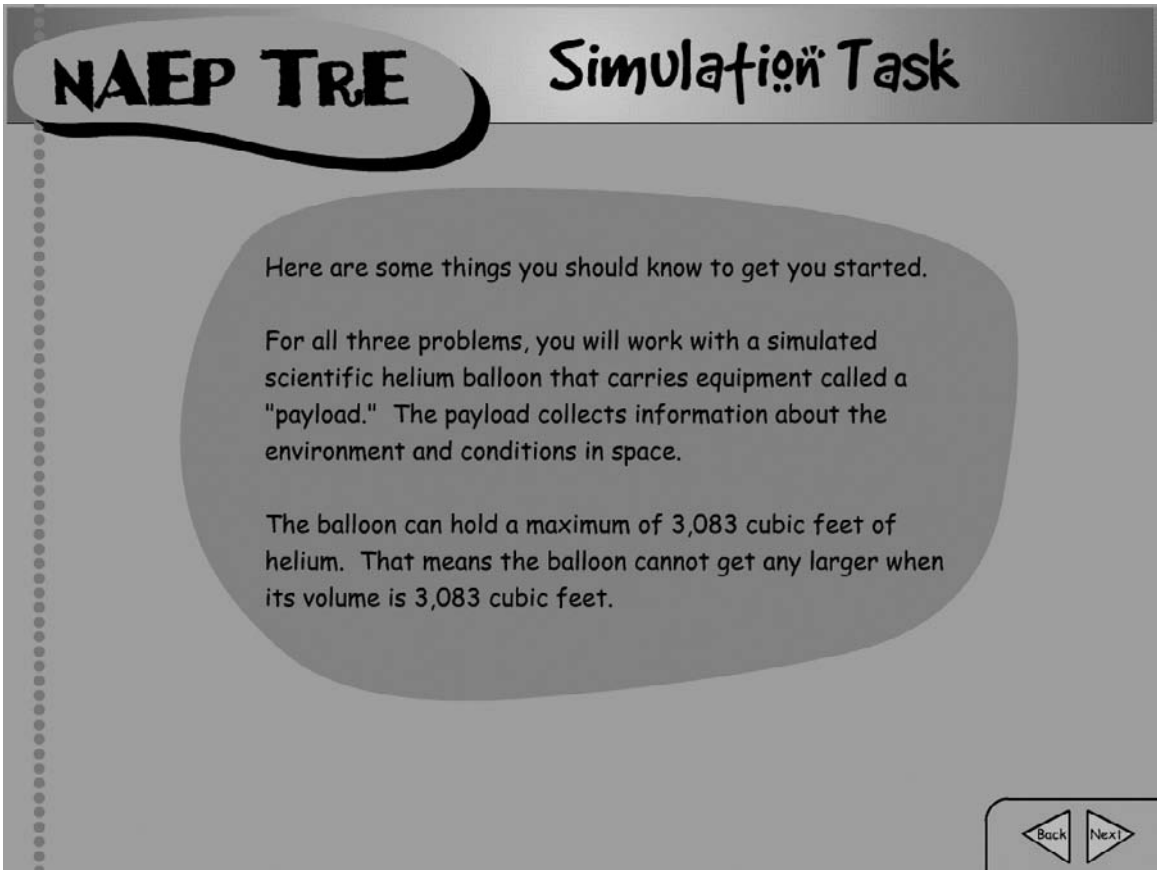
You will use a simulation tool to experiment with scientific helium balloons. Scientists use these balloons to collect information about the environment and conditions in space.

- **Problem 1**
  First you will learn how to use the simulation tool to solve the first balloon problem.
- **Problems 2 and 3**
  When you are done with problem 1, you will do 2 more balloon problems using the simulation tool.

NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Moving at their own pace (with the understanding that they had 60 minutes to complete the scenario), students were given some conditions and definitions, as shown in figure 1-8, to keep in mind as they proceeded. These included definitions of “scientific balloon” and “payload,” and the maximum volume of the scientific balloon with which students experimented.

Figure 1-8. Computer screen with conditions and definitions for the TRE Simulation scenario, grade 8: 2003



Here are some things you should know to get you started.

For all three problems, you will work with a simulated scientific helium balloon that carries equipment called a "payload." The payload collects information about the environment and conditions in space.

The balloon can hold a maximum of 3,083 cubic feet of helium. That means the balloon cannot get any larger when its volume is 3,083 cubic feet.

NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

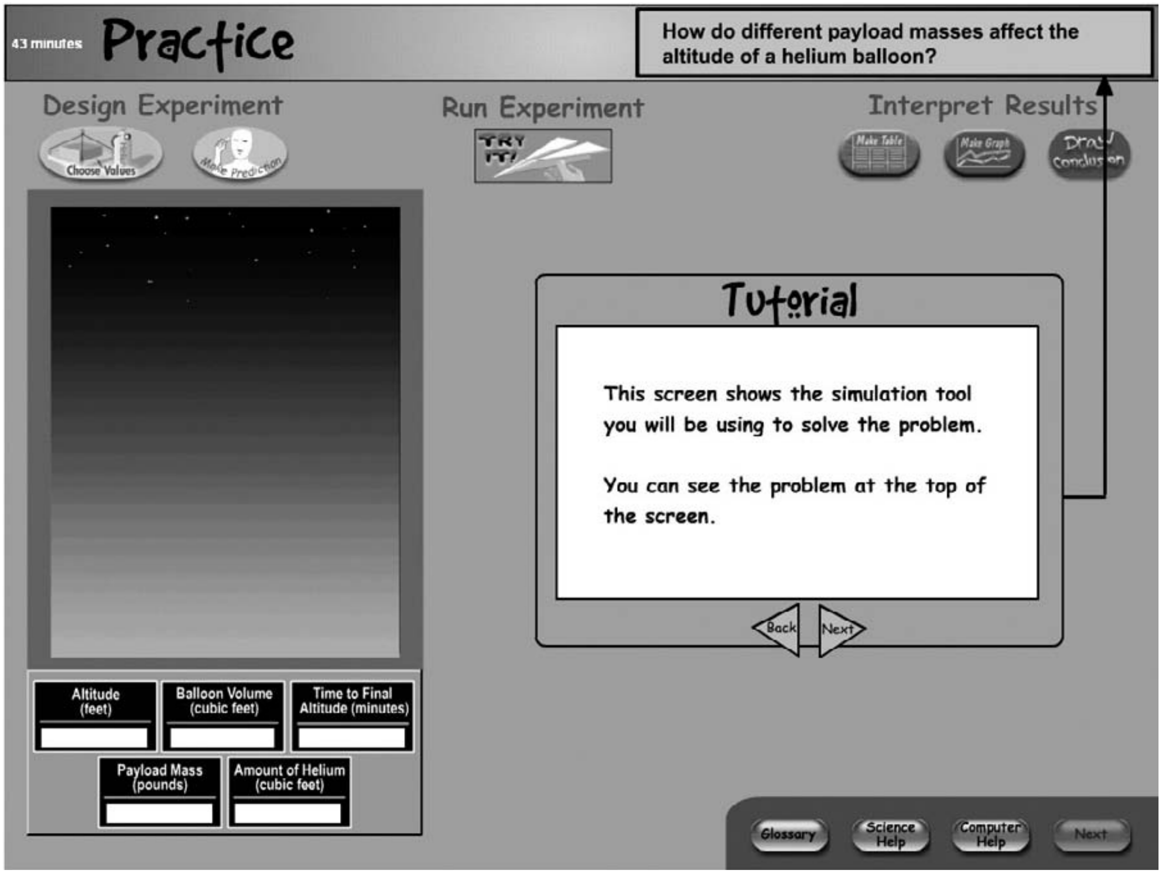
Figure 1-9 displays the first page of the tutorial for the Simulation tool interface. (See appendix E for screens in the Simulation tutorial.) During the tutorial, students were introduced to each component of the Simulation tool and were directed to run an experiment and make a prediction about the results, with the option of repeating the various steps of the tutorial. (Note that the screen clearly indicated "Practice" in the upper left-hand side, so students knew they were not yet being scored for their performances.)

The Simulation tool interface in many aspects resembled instructional software and simulation games students might already have encountered. For example, the top of the interface featured a task bar for designing, running, and interpreting experiments, and the "Back" and "Next" buttons enabled students to navigate among screens.

The problem to solve was displayed in the upper right-hand corner. It asked students to determine the relationship between payload mass and balloon altitude. To design an experiment to explore this relationship, students clicked on the Choose Values button in the Design Experiment area. A prediction could then be made about the results of the experiment. Although making predictions was optional, the interface alerted students that they could not make predictions without having first chosen values for experiments. When students were ready to run an experiment, clicking Try It caused the instrument display to activate and caused the balloon in the flight box to rise or remain stationary, depending on the value of the payload mass chosen.

Students could construct tables or graphs if they wished to keep track of experimental results by

Figure 1-9. Computer screen with the TRE Simulation scenario tutorial, grade 8: 2003



How do different payload masses affect the altitude of a helium balloon?

This screen shows the simulation tool you will be using to solve the problem.

You can see the problem at the top of the screen.

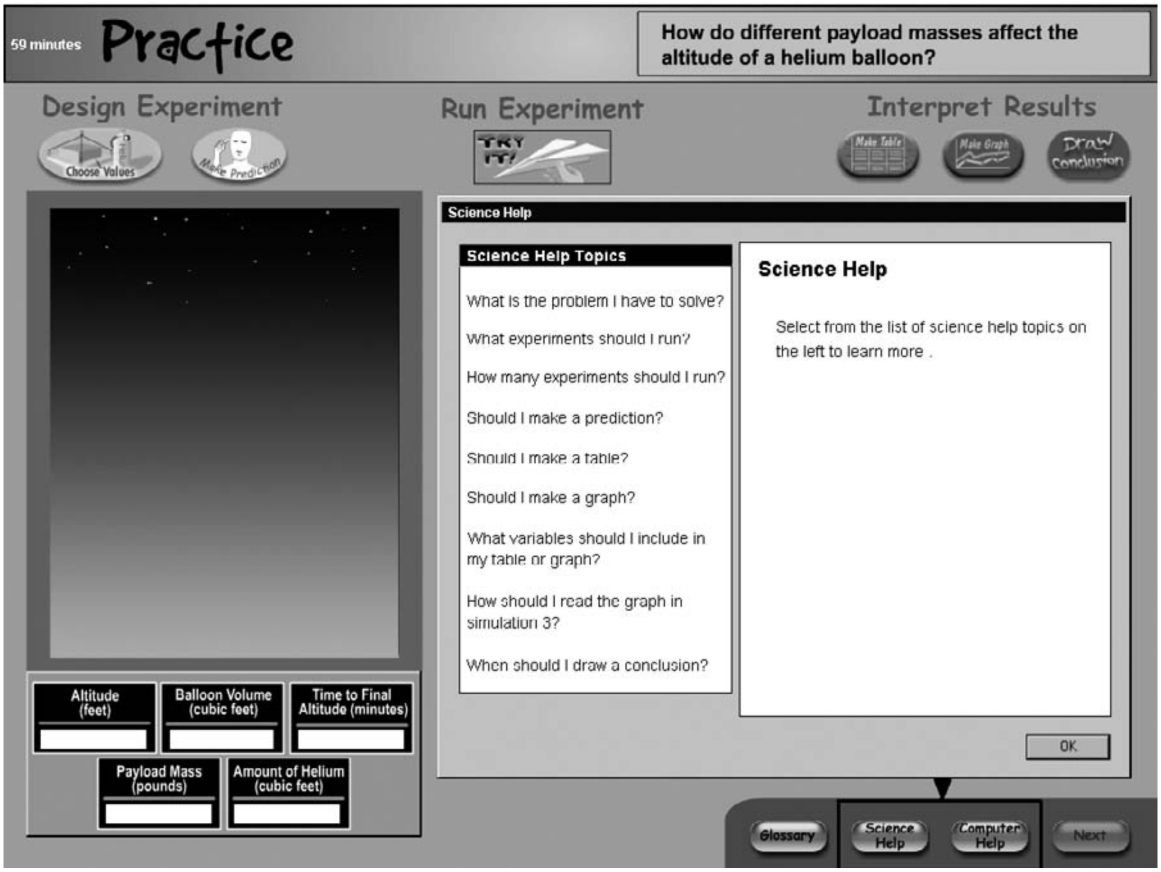
NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

clicking on the appropriate buttons under Interpret Results. The interface then presented results for all experiments run to that point. (Although it can be argued that students should not have been able to access data they did not explicitly record, the automatic recording of data is typical in scientific simulation environments.)

Students were able to watch the balloon rise in the flight box, and could observe changes in the values of dependent variables (altitude, balloon volume, and time to final altitude) in the instrument panel below that box. Values for the independent variables (payload mass and amount of helium) were also displayed in the instrument panel. When students were ready to draw conclusions, they clicked on the Draw Conclusions button under Interpret Results to bring up a box where they could enter a response to the problem featured on the upper right-hand part of the screen. Students could continue to experiment and use tables and graphs while they responded to the question.

Three forms of help were offered, as indicated by the buttons in the lower right-hand corner. These buttons brought up a glossary of science terms, science help, and computer help. Science Help gave hints about the substance of the problem. The menus for Science Help are shown in figure 1-10. Computer Help described the buttons and functions of the Simulation tool interface. (See appendix E for Science and Computer Help screens.)

Figure 1-10. Computer screen with Science Help for the TRE Simulation scenario, grade 8: 2003



How do different payload masses affect the altitude of a helium balloon?

What is the problem I have to solve?

What experiments should I run?

How many experiments should I run?

Should I make a prediction?

Should I make a table?

Should I make a graph?

What variables should I include in my table or graph?

How should I read the graph in simulation 3?

When should I draw a conclusion?

Science Help

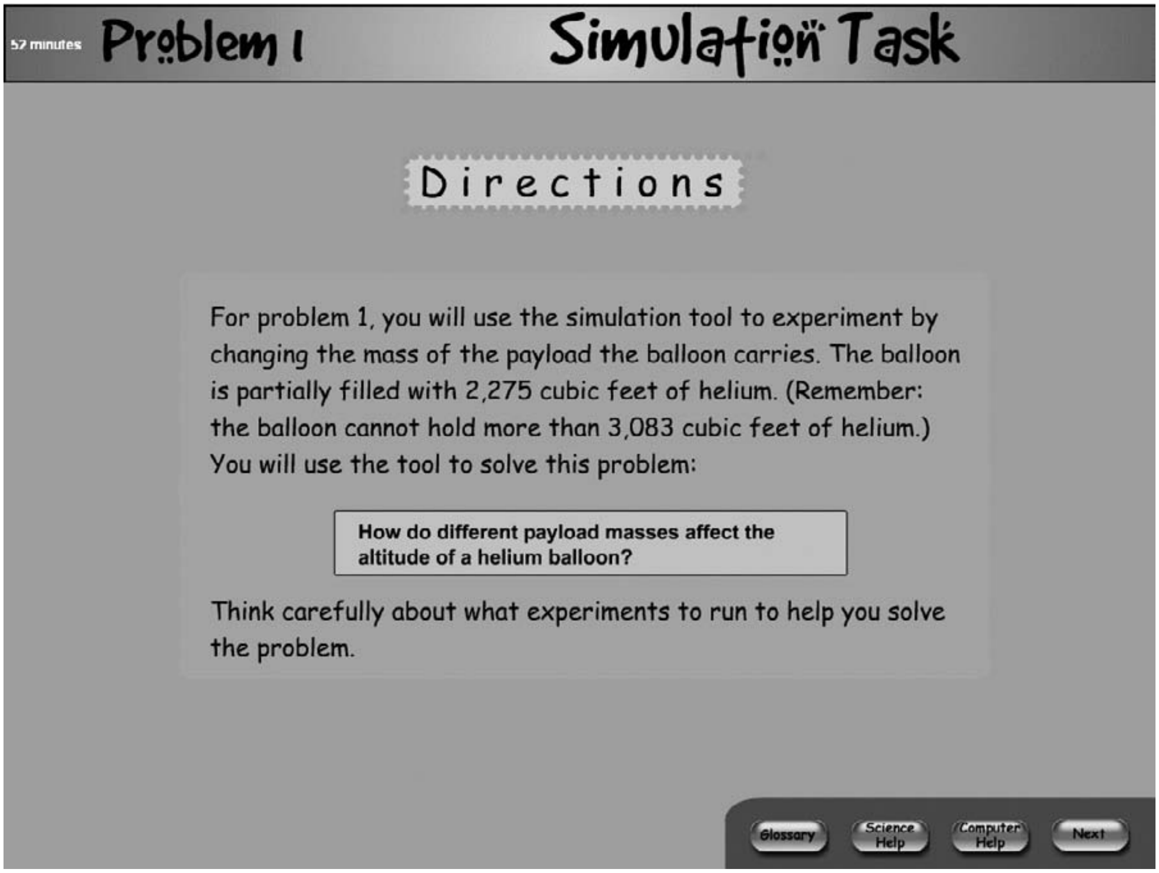
Select from the list of science help topics on the left to learn more .

NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

After students completed the tutorial, they were presented with directions for the first problem in the simulation, shown in figure 1-11. The problem asked the student to determine the relationship between the amount of mass that a balloon can carry and the height to which the balloon can rise in the atmosphere. The only available independent variable was mass, and the values of mass that the student could select were restricted. (The balloon held a constant amount of 2,275 cubic feet of helium.) These constraints were imposed because of assessment time limitations and concern that the problem might otherwise be too difficult for significant numbers of eighth-graders. Note that the directions reminded the students that the balloon could hold only 3,083 cubic feet of helium.

Figure 1-12 displays the menu of possible masses from which students could choose for experimentation in problem 1.

Figure 1-11. Computer screen with directions for TRE Simulation scenario problem 1, grade 8: 2003



Directions

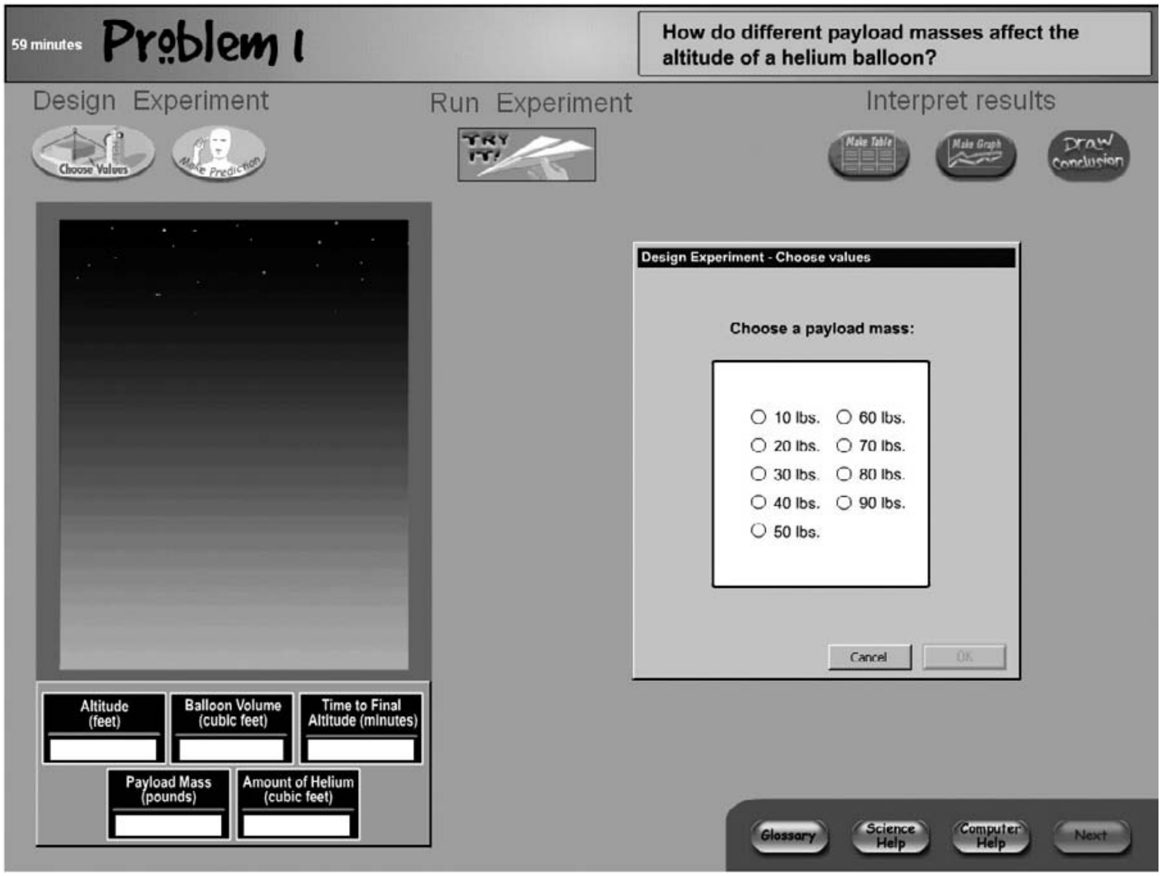
For problem 1, you will use the simulation tool to experiment by changing the mass of the payload the balloon carries. The balloon is partially filled with 2,275 cubic feet of helium. (Remember: the balloon cannot hold more than 3,083 cubic feet of helium.) You will use the tool to solve this problem:

How do different payload masses affect the altitude of a helium balloon?

Think carefully about what experiments to run to help you solve the problem.

NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure 1-12. Computer screen with the menu of values for the independent variable, payload mass, in TRE Simulation scenario problem 1, grade 8: 2003



How do different payload masses affect the altitude of a helium balloon?

Design Experiment - Choose values

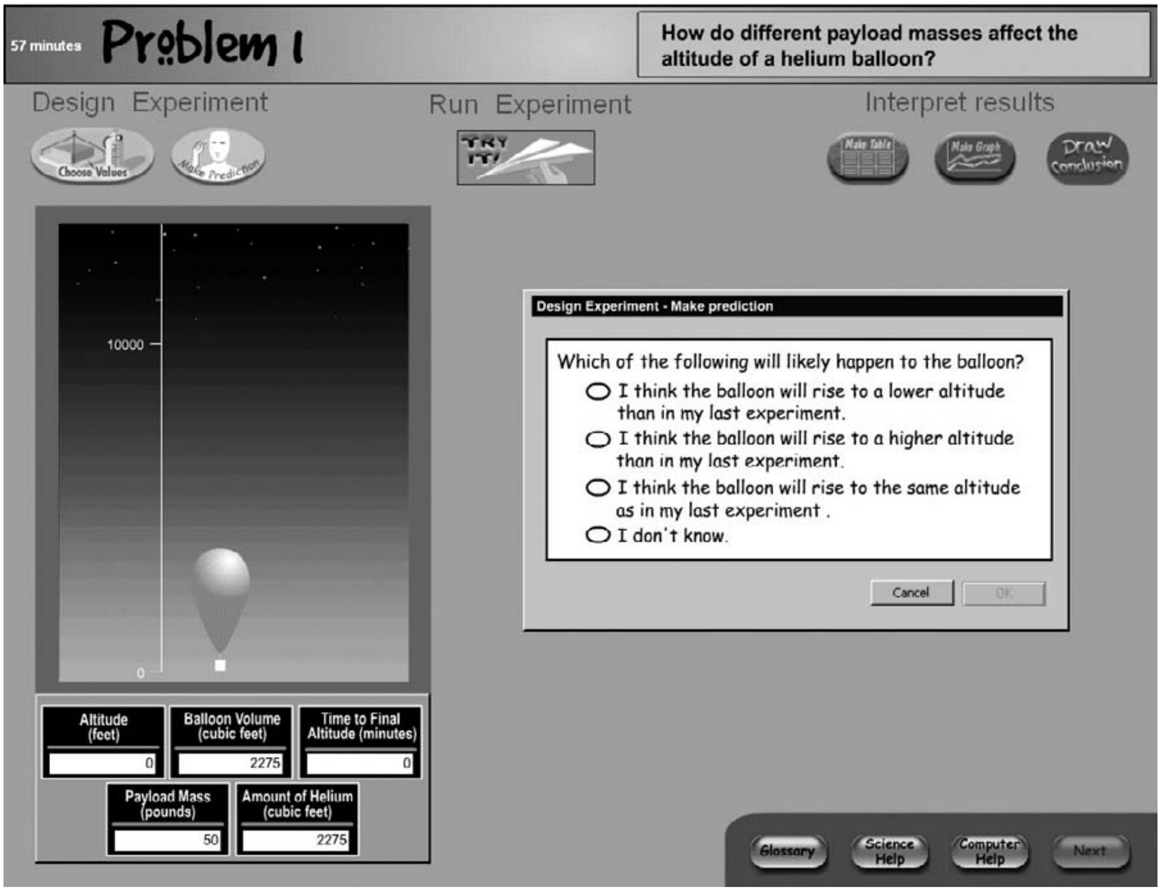
Choose a payload mass:

○ 10 lbs. ○ 60 lbs.
○ 20 lbs. ○ 70 lbs.
○ 30 lbs. ○ 80 lbs.
○ 40 lbs. ○ 90 lbs.
○ 50 lbs.

NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

After choosing a value for mass, students could choose to make a prediction. Figure 1-13 displays the four possible options. By comparing the current experiment to the previous one, the options were intended to encourage students to think in terms of patterns of results: in this case, the impact on balloon altitude of varying the payload masses. (Although more might have been learned by requiring students to key-enter predictions and interim hypotheses about the relationship between mass and balloon altitude, limited assessment time discouraged this more in-depth approach.)

Figure 1-13. Computer screen with the prediction options in TRE Simulation scenario problem 1, grade 8: 2003



57 minutes **Problem 1**

How do different payload masses affect the altitude of a helium balloon?

Design Experiment - Make prediction

Which of the following will likely happen to the balloon?

○ I think the balloon will rise to a lower altitude than in my last experiment.

○ I think the balloon will rise to a higher altitude than in my last experiment.

○ I think the balloon will rise to the same altitude as in my last experiment .

○ I don't know.

Altitude (feet)	Balloon Volume (cubic feet)	Time to Final Altitude (minutes)
0	2275	0

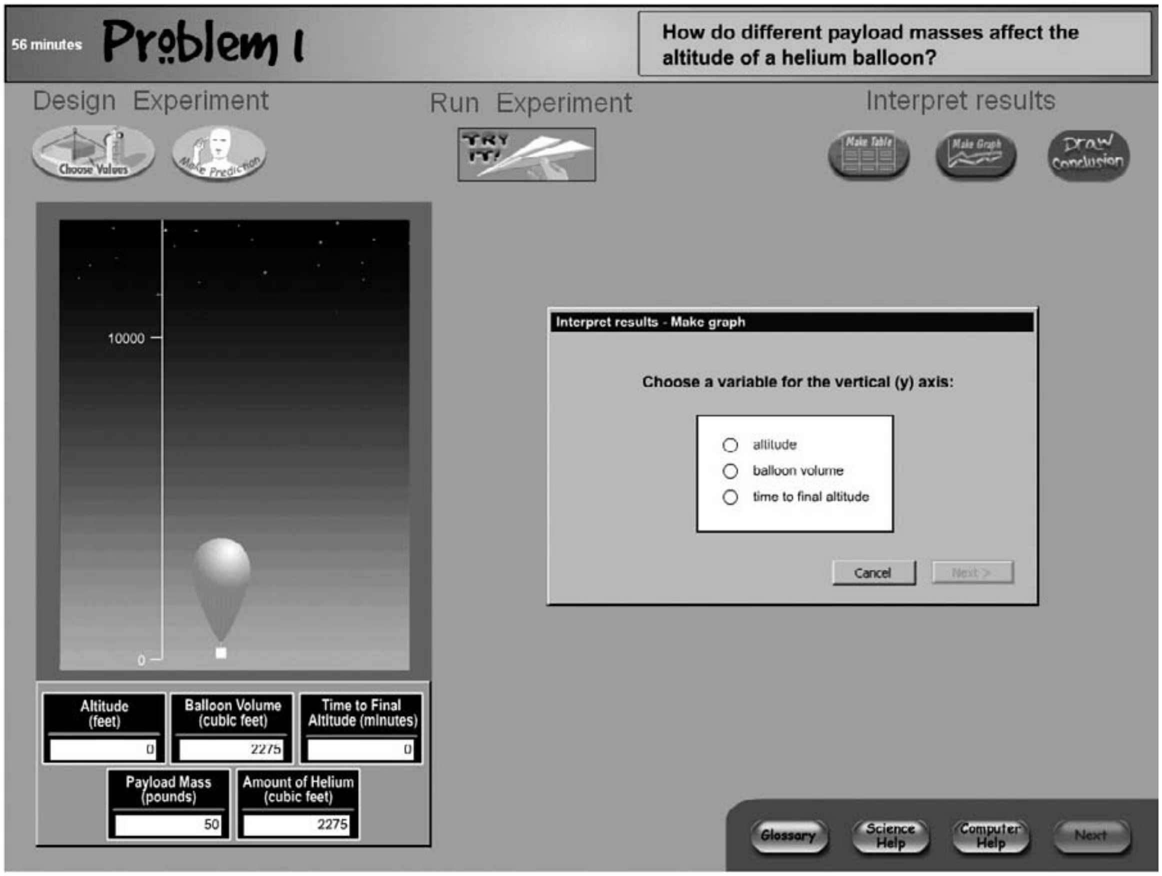
Payload Mass (pounds)	Amount of Helium (cubic feet)
50	2275

NOTE: TRE = Technology-Rich Environments.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

To help interpret data, students could make a graph, a table, or both. Clicking on the Make Graph button opened a dialog box that asked students to select a variable for the vertical axis (see figure 1-14) and then, in a subsequent box, for the horizontal axis. Note that students had leeway to get into trouble by choosing less relevant or incorrect variables for either graph axis; this design allowed an opportunity to determine whether students created interpretive tools related to the problem they were supposed to be solving.

Figure 1-14. Computer screen with dialog box for creating a graph in TRE Simulation scenario problem 1, grade 8: 2003



How do different payload masses affect the altitude of a helium balloon?

Interpret results - Make graph

Choose a variable for the vertical (y) axis:

○ altitude
○ balloon volume
○ time to final altitude

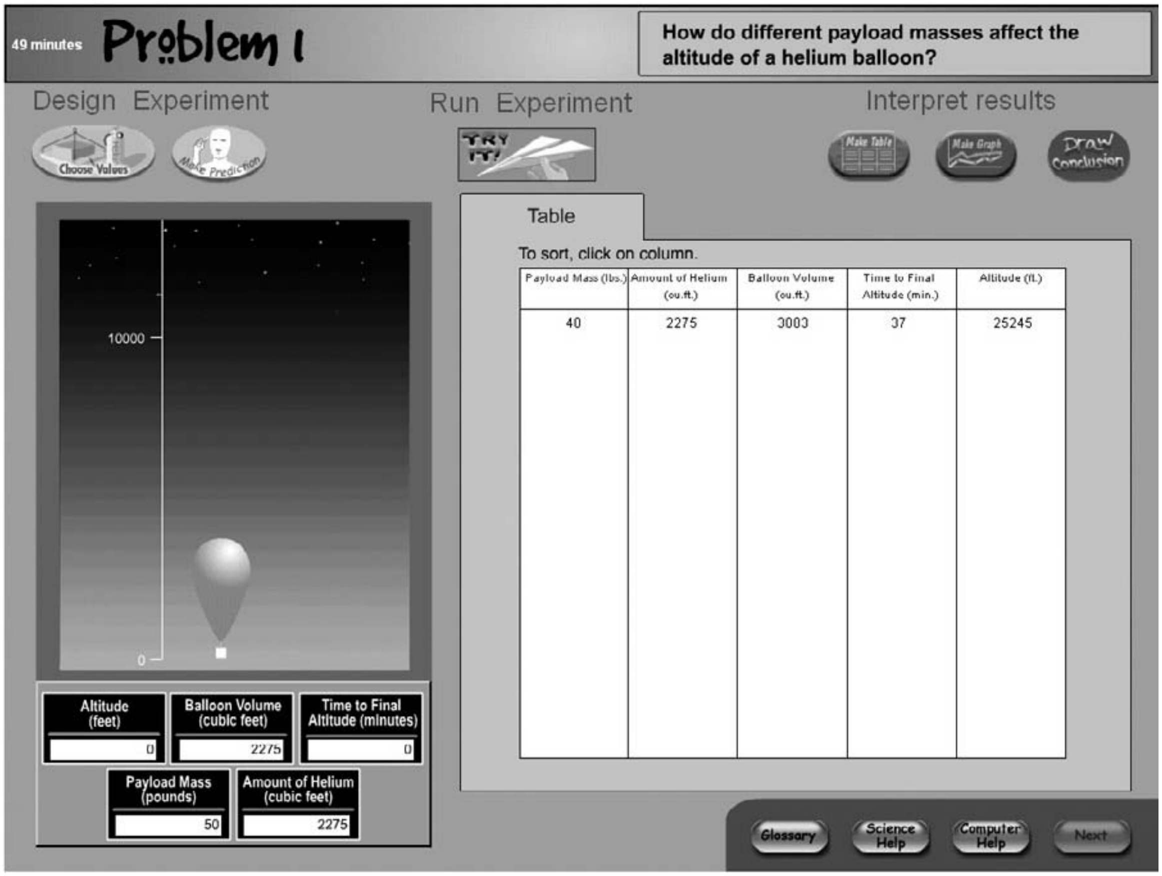
Altitude (feet)	Balloon Volume (cubic feet)	Time to Final Altitude (minutes)
0	2275	0

Payload Mass (pounds)	Amount of Helium (cubic feet)
50	2275

NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Similarly, students could construct a table by choosing from the variables tracked in the instrument panel. The resulting displays may, therefore, have contained relevant information, some relevant and some irrelevant information, or only irrelevant information. If, for example, a student chose to include all five variables, the table would appear as in figure 1-15. A more helpful table for problem 1 would be limited to the dependent and independent variables necessary to solve the problem—altitude and mass. For each subsequent experiment that students chose to conduct, a line of data was added to the table automatically. Students could sort the table on any variable by clicking on the appropriate column heading.

Figure 1-15. Computer screen with a table of results for one experiment conducted in TRE Simulation scenario problem 1, grade 8: 2003



How do different payload masses affect the altitude of a helium balloon?

Table

To sort, click on column.

Payload Mass (lbs.)	Amount of Helium (cu.ft.)	Balloon Volume (cu.ft.)	Time to Final Altitude (min.)	Altitude (ft.)
40	2275	3003	37	25245

Altitude (feet): 0

Balloon Volume (cubic feet): 2275

Time to Final Altitude (minutes): 0

Payload Mass (pounds): 50

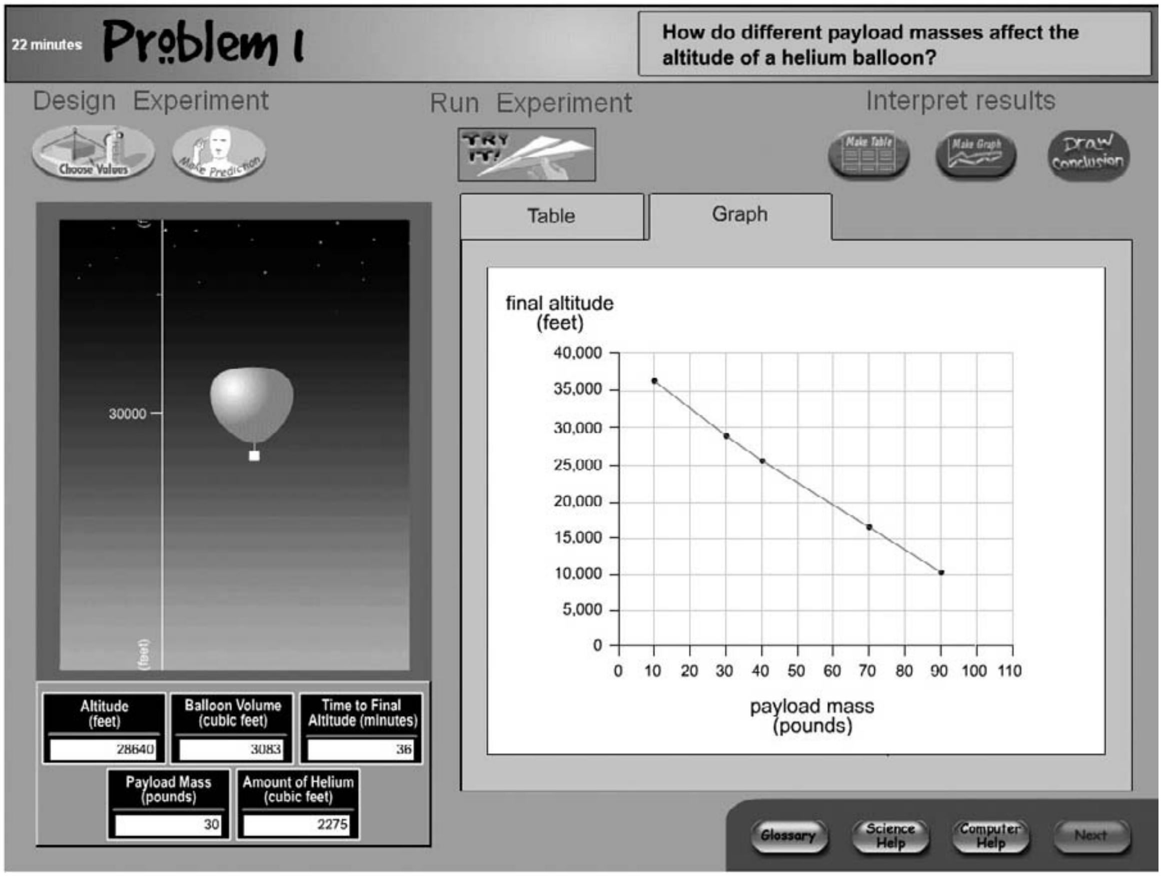
Amount of Helium (cubic feet): 2275

NOTE: TRE = Technology-Rich Environments.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Note that the relationship to be discovered in problem 1 was a virtually linear, negative one: as mass increases, the altitude the balloon can achieve decreases. Figure 1-16 shows the display that would result from creating a graph with the relevant variables and experiments with a sufficient range of masses.[6]

Figure 1-16. Computer screen with a graph of the relationship between altitude and mass in TRE Simulation scenario motivating problem 1, grade 8: 2003



22 minutes Problem 1

How do different payload masses affect the altitude of a helium balloon?

Altitude (feet): 28640

Balloon Volume (cubic feet): 3083

Time to Final Altitude (minutes): 36

Payload Mass (pounds): 30

Amount of Helium (cubic feet): 2275

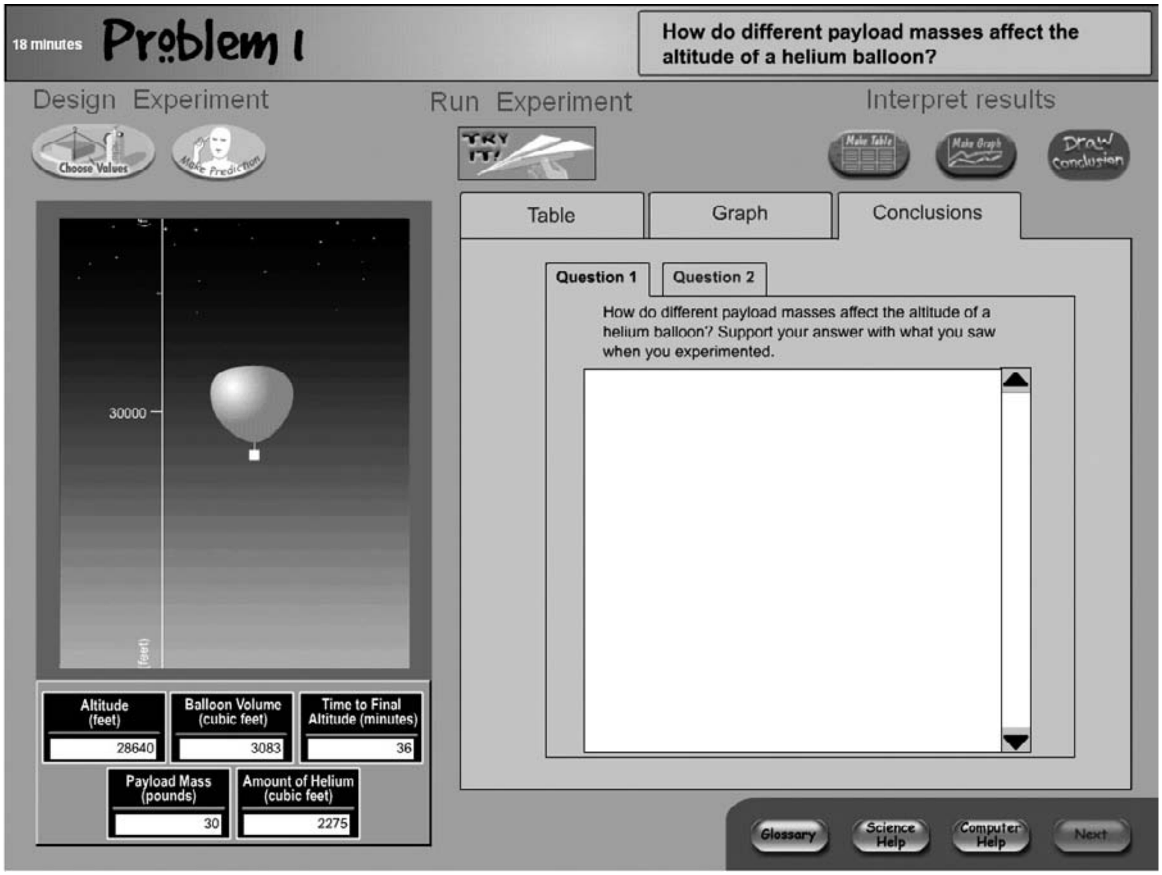
NOTE: TRE = Technology-Rich Environments.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

[6] Students in the TRE study could have used non-technological alternatives like paper-and-pencil in place of creating an electronic table or graph. The extent to which such alternatives were used could not be determined.

When ready, students could click on the Draw Conclusions button to bring up a text-entry box, as shown in figure 1-17. This box called for students to construct a response to the problem about the relationship between payload mass and altitude and to support the answer with experimental observations. Before completing the response, students could choose to revisit an existing table or graph, construct new tables or graphs, or conduct more experiments.

Figure 1-17. Computer screen with the box for answering the TRE Simulation scenario motivating problem 1, grade 8: 2003



How do different payload masses affect the altitude of a helium balloon?

How do different payload masses affect the altitude of a helium balloon? Support your answer with what you saw when you experimented.

Altitude (feet): 28640

Balloon Volume (cubic feet): 3083

Time to Final Altitude (minutes): 36

Payload Mass (pounds): 30

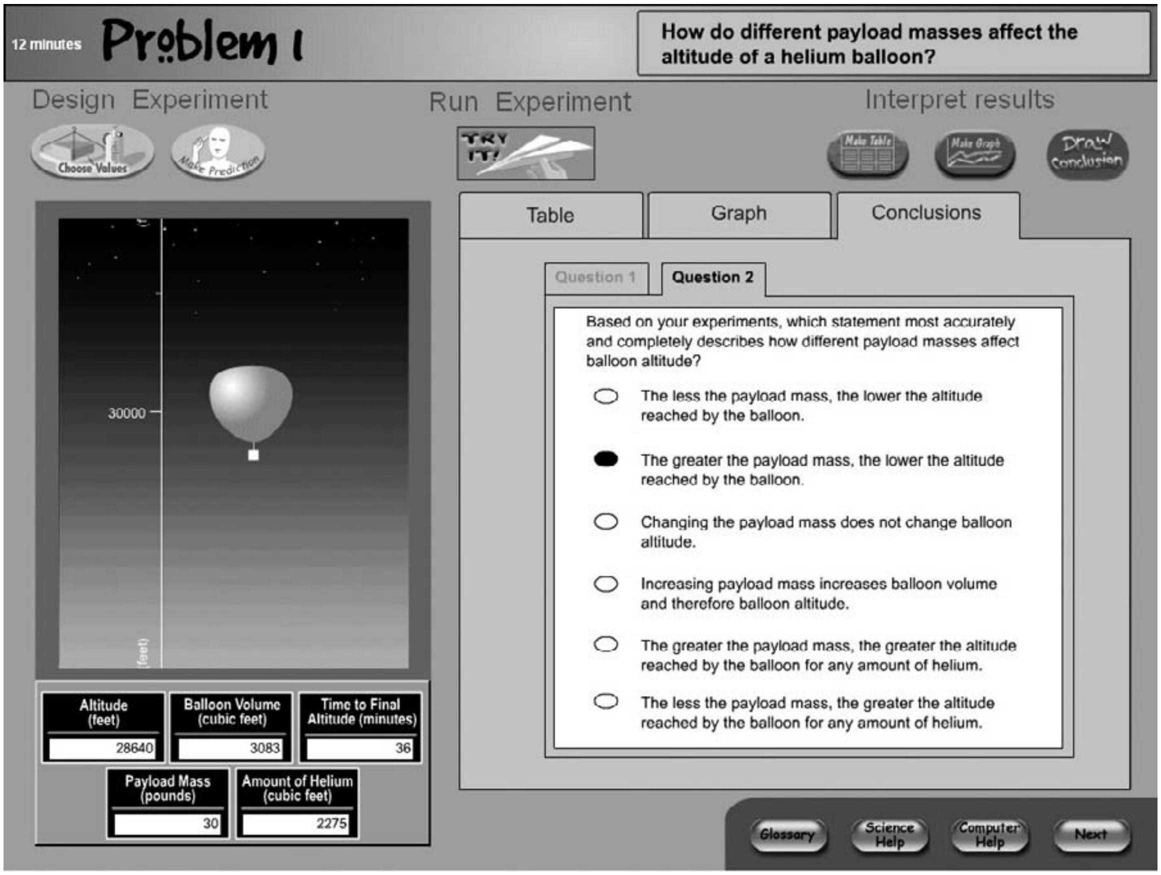
Amount of Helium (cubic feet): 2275

NOTE: TRE = Technology-Rich Environments.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Having completed their written responses, students were required to respond to a multiple-choice question (see figure 1-18), which provided an alternative measure for those individuals unable to express adequately their understanding of the mass-altitude relationship in writing.

Figure 1-18. Computer screen with the multiple-choice question concluding TRE Simulation scenario problem 1, grade 8: 2003



How do different payload masses affect the altitude of a helium balloon?

Based on your experiments, which statement most accurately and completely describes how different payload masses affect balloon altitude?

○ The less the payload mass, the lower the altitude reached by the balloon.

● The greater the payload mass, the lower the altitude reached by the balloon.

○ Changing the payload mass does not change balloon altitude.

○ Increasing payload mass increases balloon volume and therefore balloon altitude.

○ The greater the payload mass, the greater the altitude reached by the balloon for any amount of helium.

○ The less the payload mass, the greater the altitude reached by the balloon for any amount of helium.

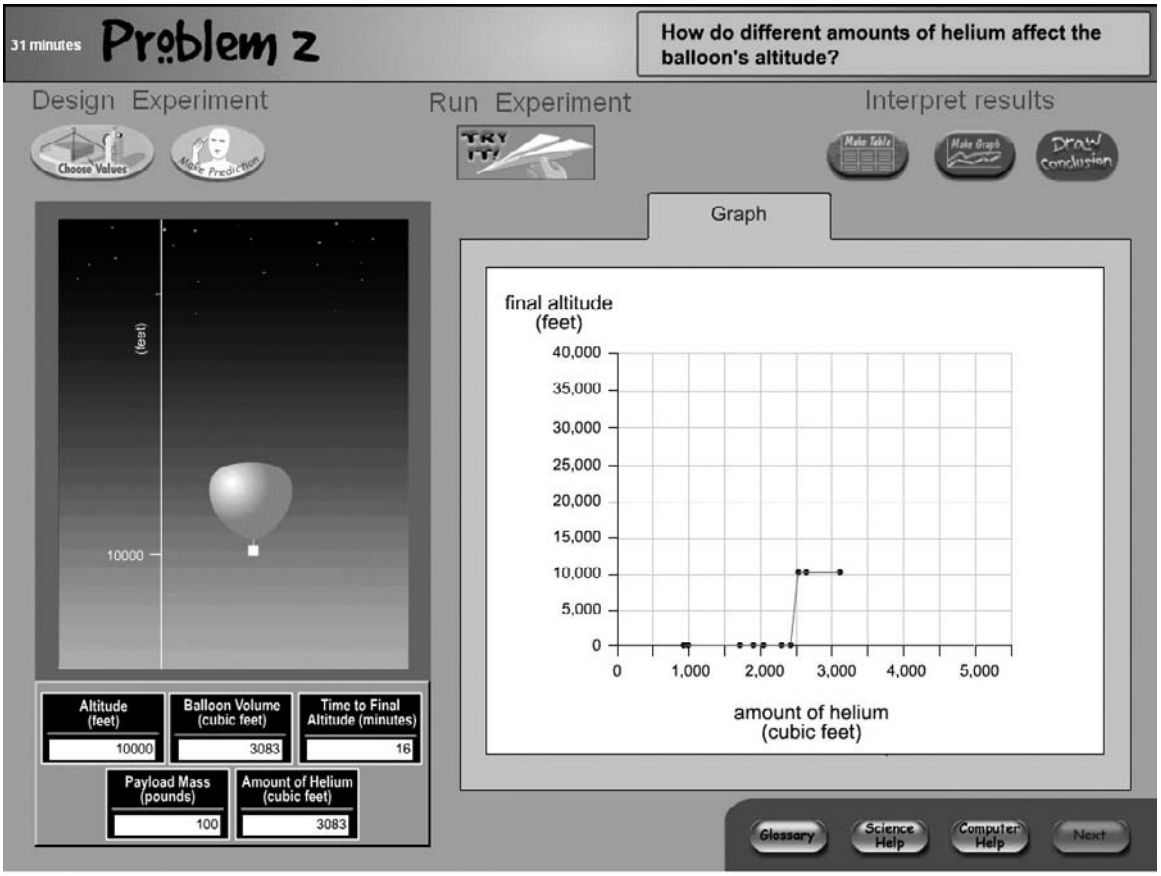
Altitude (feet)	Balloon Volume (cubic feet)	Time to Final Altitude (minutes)
28640	3083	36

Payload Mass (pounds)	Amount of Helium (cubic feet)
30	2275

NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

The second Simulation problem asked students to determine the relationship between the amount of helium put in the balloon and the altitude that the balloon could reach. This time, the payload mass the balloon carried was fixed. Problem 2 was conceptually more difficult because the relationship students had to discover was not linear. Rather, the relationship took the form of a step function. That is, until a critical amount of helium was put in the balloon, the balloon did not leave the ground; once that critical amount of helium was achieved, the balloon would rise to a maximum altitude, then go no higher regardless of how much more helium was put into it. To recognize the relationship, students had to choose a sufficient number and range of values and not draw conclusions prematurely; a premature conclusion would lead them to assume falsely either that the amount of helium did not matter, or that the balloon would continue to rise higher as it was filled with more helium. Figure 1-19 displays what the graph

Figure 1-19. Computer screen with a graph of the relationship between altitude and amount of helium in TRE Simulation scenario problem 2, grade 8: 2003



How do different amounts of helium affect the balloon's altitude?

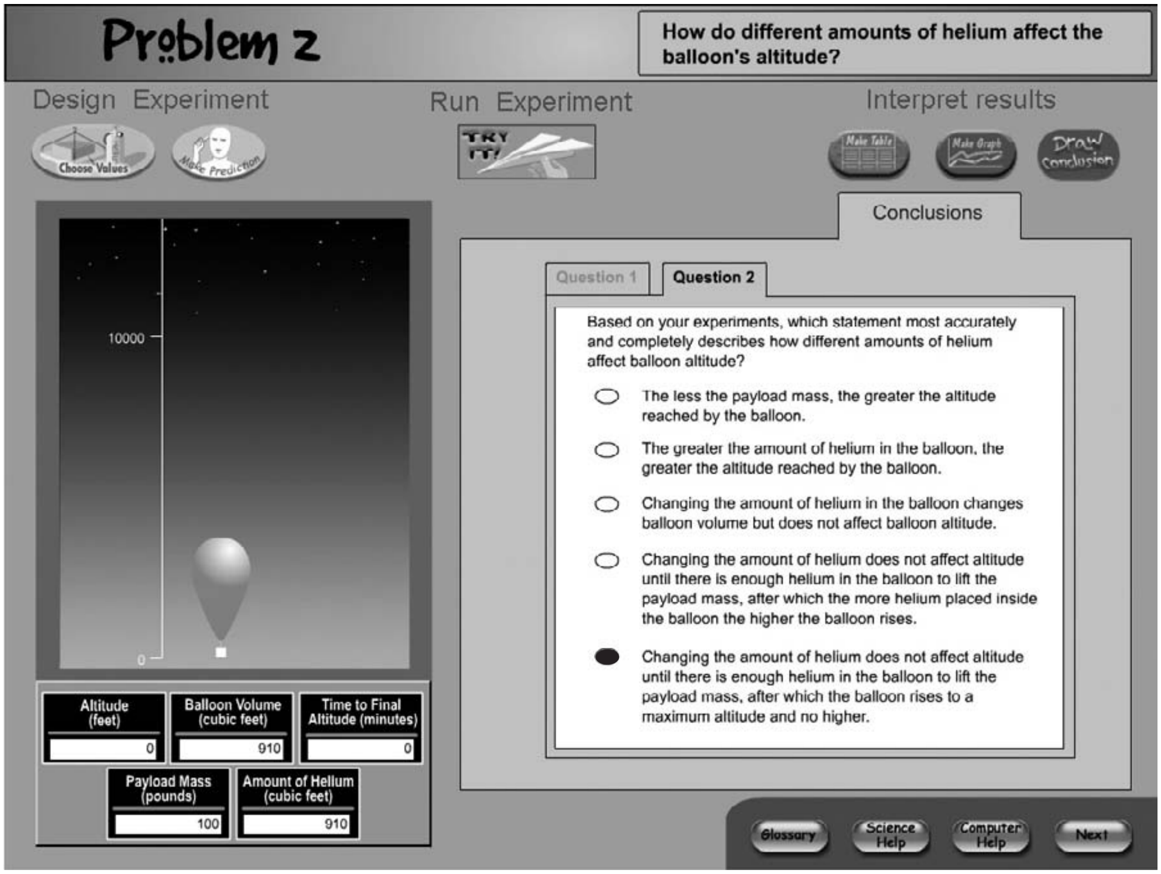
Altitude (feet)	Balloon Volume (cubic feet)	Time to Final Altitude (minutes)
10000	3083	16

Payload Mass (pounds)	Amount of Helium (cubic feet)
100	3083

NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

looked like with the relevant variables and sufficient experiments to reveal the step function. Figure 1-20 shows the multiple-choice question that students were asked to answer after they entered the constructed response to problem 2.

Figure 1-20. Computer screen with multiple-choice question on the relationship between altitude and amount of helium in TRE Simulation scenario problem 2, grade 8: 2003



How do different amounts of helium affect the balloon's altitude?

Based on your experiments, which statement most accurately and completely describes how different amounts of helium affect balloon altitude?

○ The less the payload mass, the greater the altitude reached by the balloon.

○ The greater the amount of helium in the balloon, the greater the altitude reached by the balloon.

○ Changing the amount of helium in the balloon changes balloon volume but does not affect balloon altitude.

○ Changing the amount of helium does not affect altitude until there is enough helium in the balloon to lift the payload mass, after which the more helium placed inside the balloon the higher the balloon rises.

● Changing the amount of helium does not affect altitude until there is enough helium in the balloon to lift the payload mass, after which the balloon rises to a maximum altitude and no higher.

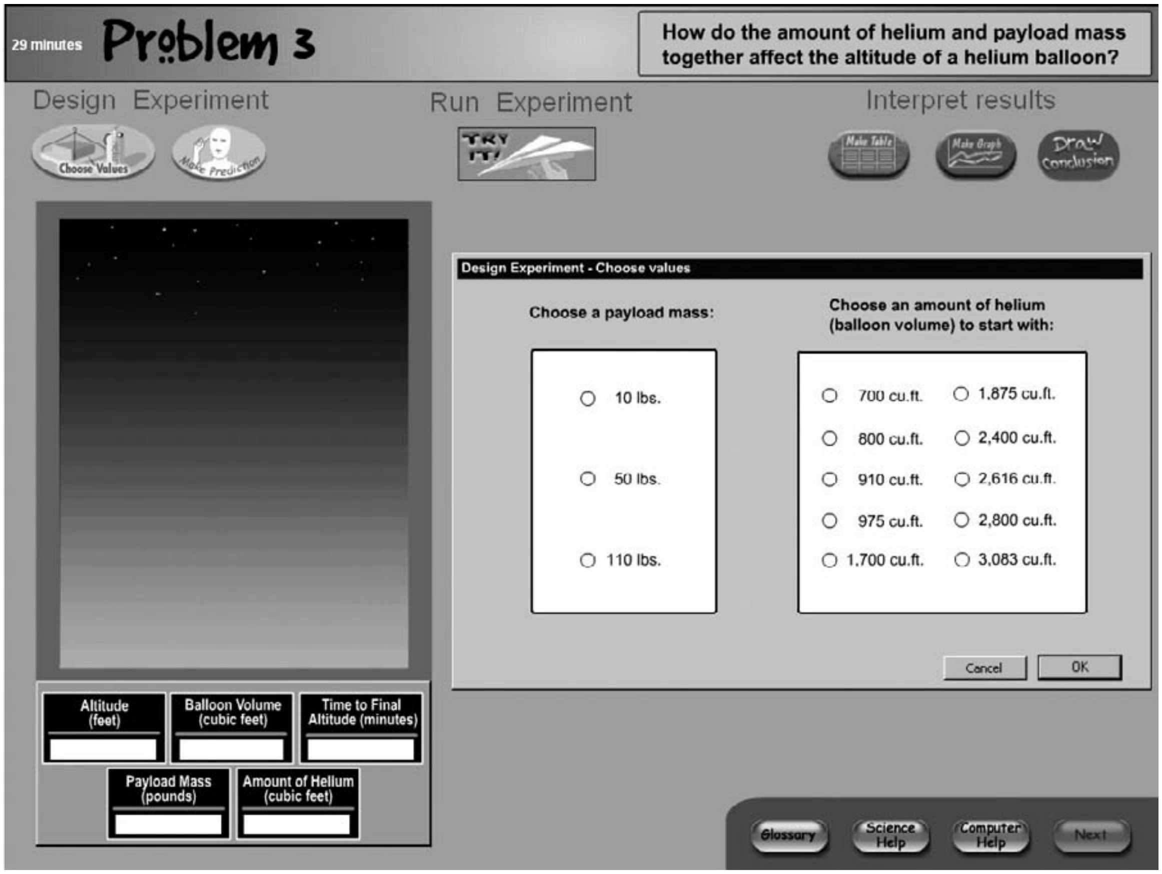
Altitude (feet)	Balloon Volume (cubic feet)	Time to Final Altitude (minutes)
0	910	0

Payload Mass (pounds)	Amount of Helium (cubic feet)
100	910

NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Problem 3, the final Simulation problem, was the most conceptually complex, as it required students to discover how payload mass and amount of helium worked together to determine the altitude that the balloon could reach. Thus, students not only had to think about which experiments to run and how many, but they also had to control for one independent variable while manipulating the other. To limit the complexity of the problem, the number of masses students could vary was reduced to three, as shown in figure 1-21.

Figure 1-21. Computer screen with the dialog box menu of choices for the independent variables in TRE Simulation scenario problem 3, grade 8: 2003



How do the amount of helium and payload mass together affect the altitude of a helium balloon?

Design Experiment - Choose values

Choose a payload mass:

○ 10 lbs.
○ 50 lbs.
○ 110 lbs.

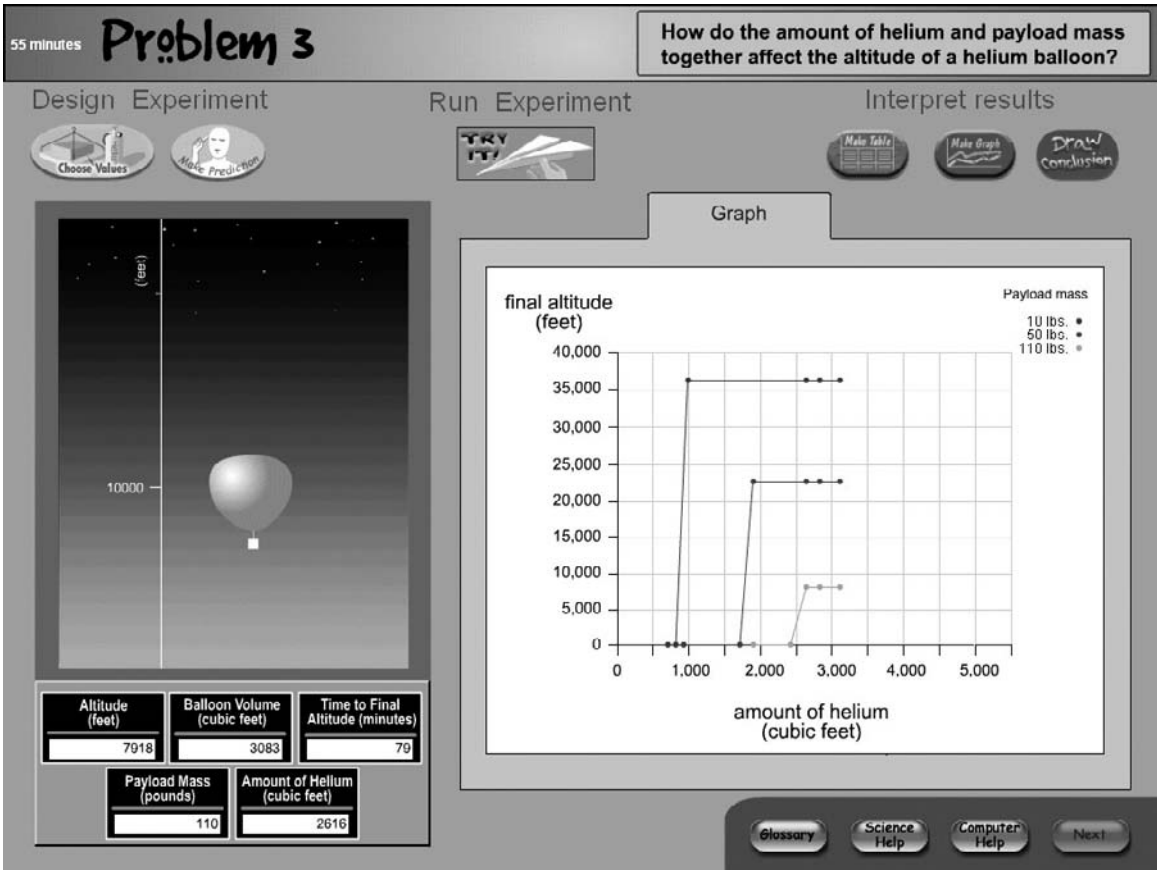
Choose an amount of helium (balloon volume) to start with:

○ 700 cu.ft.	○ 1,875 cu.ft.
○ 800 cu.ft.	○ 2,400 cu.ft.
○ 910 cu.ft.	○ 2,616 cu.ft.
○ 975 cu.ft.	○ 2,800 cu.ft.
○ 1,700 cu.ft.	○ 3,083 cu.ft.

NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

In problem 3, students had to discover a nonlinear relationship that took the form of a series of step functions, one for each mass. Figure 1-22 displays what the graph looked like if a student had constructed the correct data display and had run a sufficient number of experiments to reveal all three functions. Note that the maximum altitude for each step function decreased as payload mass increased.

Figure 1-22. Computer screen with graph of the relationship of altitude with mass and amount of helium in TRE Simulation scenario problem 3, grade 8: 2003



Altitude (feet)	Balloon Volume (cubic feet)	Time to Final Altitude (minutes)
7918	3083	79

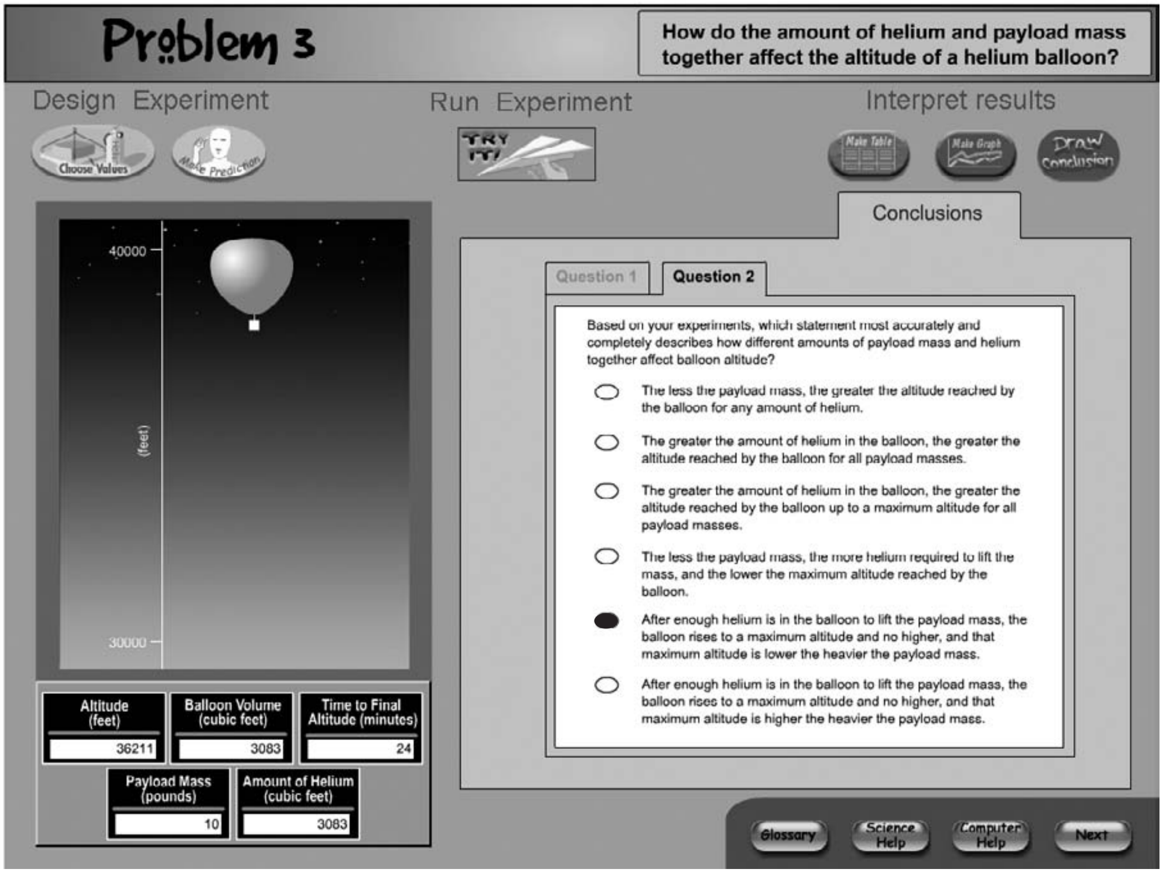
Payload Mass (pounds)	Amount of Helium (cubic feet)
110	2616

NOTE: TRE = Technology-Rich Environments.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

After entering a constructed response describing the relationship they discovered, the students were asked to respond to a multiple-choice question intended to probe the same relationship. The question is shown in figure 1-23.

Figure 1-23. Computer screen with multiple-choice question on the relationship of altitude with mass and amount of helium in TRE Simulation scenario problem 3, grade 8: 2003



How do the amount of helium and payload mass together affect the altitude of a helium balloon?

Based on your experiments, which statement most accurately and completely describes how different amounts of payload mass and helium together affect balloon altitude?

○ The less the payload mass, the greater the altitude reached by the balloon for any amount of helium.

○ The greater the amount of helium in the balloon, the greater the altitude reached by the balloon for all payload masses.

○ The greater the amount of helium in the balloon, the greater the altitude reached by the balloon up to a maximum altitude for all payload masses.

○ The less the payload mass, the more helium required to lift the mass, and the lower the maximum altitude reached by the balloon.

● After enough helium is in the balloon to lift the payload mass, the balloon rises to a maximum altitude and no higher, and that maximum altitude is lower the heavier the payload mass.

○ After enough helium is in the balloon to lift the payload mass, the balloon rises to a maximum altitude and no higher, and that maximum altitude is higher the heavier the payload mass.

Altitude (feet)	Balloon Volume (cubic feet)	Time to Final Altitude (minutes)
36211	3083	24

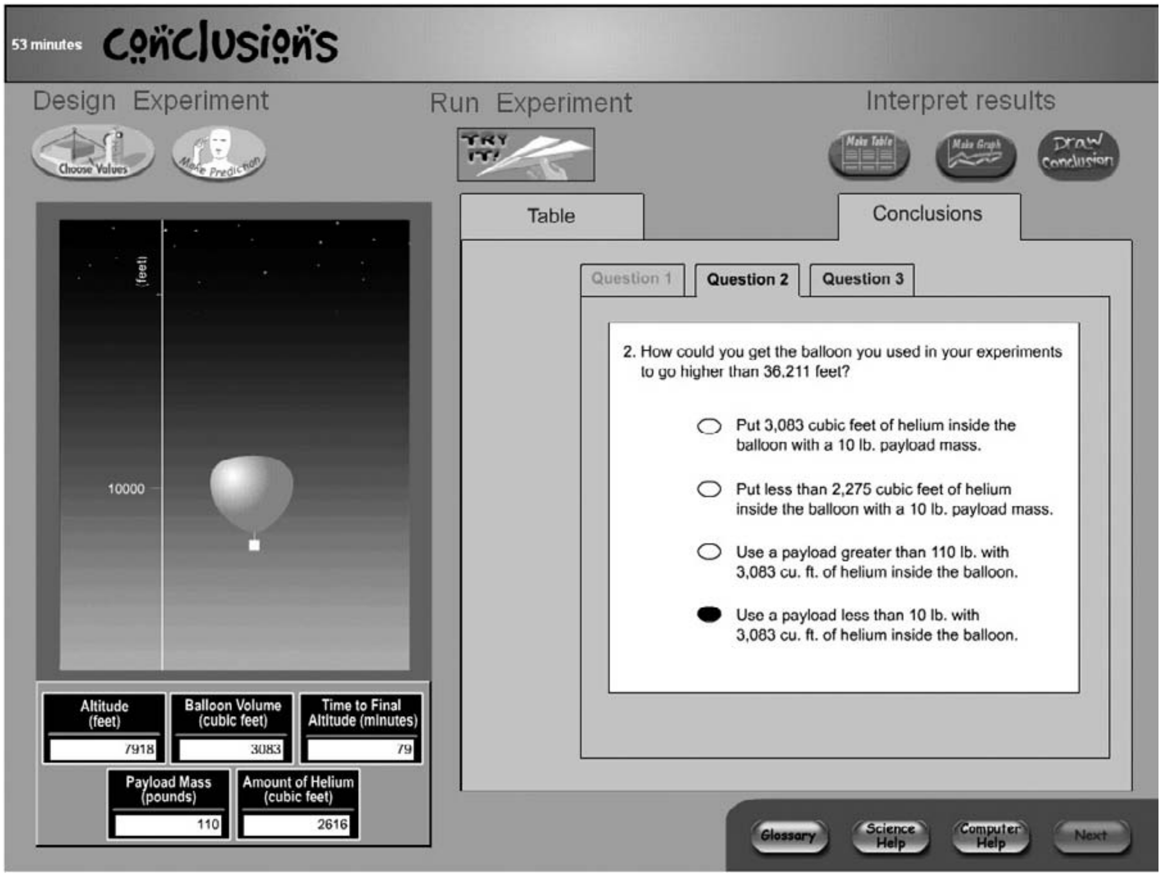
Payload Mass (pounds)	Amount of Helium (cubic feet)
10	3083

NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

When students finished problem 3, they were asked to respond to several multiple-choice questions to see how well they grasped the physics behind the overall Simulation scenario. One of the questions is shown in figure 1-24; the oval next to the correct answer is shaded. To respond to this question, students needed to have grasped that, short of increasing the size of the balloon, the only way to get the balloon to achieve a higher altitude would be to attach a payload mass smaller than any of the masses available to students in the simulation.

After completing these synthesizing questions, students could read an explanation of the physics behind helium balloons, but they could not re-enter the simulation. The explanation was included because the TRE project team believed it was important that students leave the scenario with an accurate description of the science underlying the problems they had addressed. Finally, students responded to background questionnaires, as they had done at the conclusion of the Search scenario.

Figure 1-24. Computer screen with one of the multiple-choice questions concluding the TRE Simulation scenario, grade 8: 2003



2. How could you get the balloon you used in your experiments to go higher than 36,211 feet?

○ Put 3,083 cubic feet of helium inside the balloon with a 10 lb. payload mass.

○ Put less than 2,275 cubic feet of helium inside the balloon with a 10 lb. payload mass.

○ Use a payload greater than 110 lb. with 3,083 cu. ft. of helium inside the balloon.

● Use a payload less than 10 lb. with 3,083 cu. ft. of helium inside the balloon.

Altitude (feet)	Balloon Volume (cubic feet)	Time to Final Altitude (minutes)
7918	3083	79

Payload Mass (pounds)	Amount of Helium (cubic feet)
110	2616

NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

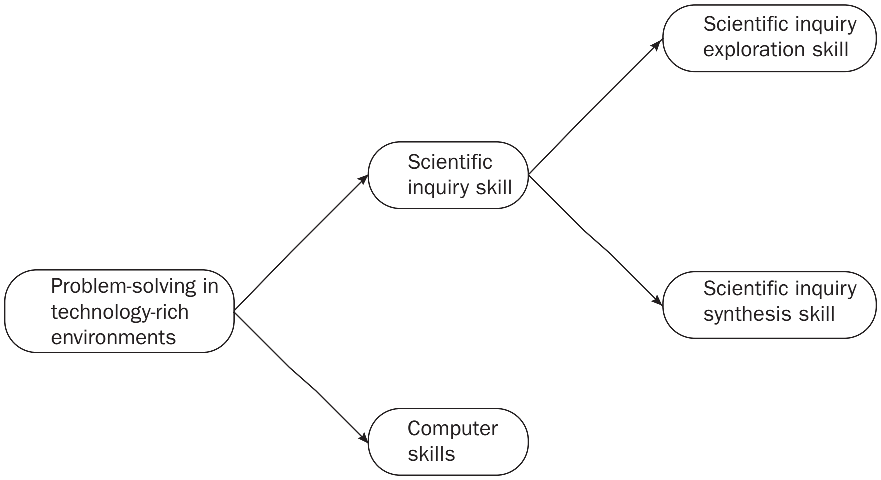
The Student and Evidence Models

While developing suitable problem-solving scenarios is a challenging task, so is interpreting the responses to such scenarios. A well-conceptualized interpretive framework is a necessity; the scenario development cost and examinee time required to perform extended problem solving on the computer can be justified only if the wealth of information that can be captured about student performance can be thoughtfully used.

In addition to the amount of data, other factors make interpretation challenging. As stated above, extended performances are typically multidimensional, relying on multiple, intertwined skills. Further, response data based on an extended scenario in which examinee actions share a common context are often locally dependent. That is, factors other than the skills of interest may influence responses to related aspects of a complex task. Such effects may arise from chance familiarity with a particular topic, personal interests, or misinterpreting directions or the intent of a question, as well as from other sources. These "context effects" are common in reading comprehension tests, where a set of items based on the same passage may share unwanted covariation for an individual because that person is (or is not) interested in the passage topic (Sireci, Thissen, and Wainer 1991; Thissen, Steinberg, and Mooney 1989). In Problem Solving in Technology-Rich Environments (TRE), an examinee's performance on the first Simulation problem relating mass to altitude may be facilitated by having recently read an article on weather balloons and the payloads they carry. However, the examinee's performance on the second problem, relating the amount of helium to altitude, may be unaffected by that contextual knowledge. The measurement models typically used in NAEP assessments do not explicitly accommodate either local dependence or multidimensionality.

The TRE team relied upon Evidence-Centered Design (ECD) to help develop the interpretive framework for the TRE scenarios (Mislevy, Almond, and Lukas 2003; Mislevy et al. 2001). ECD is a methodology for devising assessments and for using the evidence observed in complex student performances to make inferences about student proficiency. In this approach, initial specifications for scoring and interpretation are developed as part of assessment planning. These specifications take the form of student and evidence models. The student model constitutes a proposal for how the components of proficiency (or skill) are organized in the domain of problem solving in technology-rich environments. The evidence model describes how to connect student responses to these components of proficiency.[7] Figure 2-1 shows the student model.

Figure 2-1. TRE student model, grade 8: 2003


Problem-solving in technology-rich environments
Scientific inquiry skill
Computer skills
Scientific inquiry exploration skill
Scientific inquiry synthesis skill

NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

[7] In addition to student and evidence models, ECD also invokes the concept of a "task model." The task model is an abstract description of a class of situations, or tasks, intended to elicit behavior from students relevant to one or more student-model proficiencies. Because each task model defines the characteristics of a general class, such models allow test developers to generate instances of extended problem-solving exercises very efficiently. Task models are particularly useful for ongoing assessment programs that require the repeated creation of tasks. Task models were not used in the TRE study, however, because the study called for a one-time assessment.

Reading from left to right, the figure indicates that problem solving in technology-rich environments is composed of scientific inquiry skill and computer skills. Scientific inquiry skill is, in turn, composed of two subskills—exploration and synthesis. For purposes of the TRE scenarios, scientific inquiry was defined as the ability 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.

It is important to note here that the conception of scientific inquiry embodied in TRE is a partial one. The essential features of classroom scientific inquiry are acknowledged to vary along several dimensions, with some implementations considered to be full and others partial inquiry (Olson and 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. Partial inquiry was chosen for practical reasons, including limited testing time, the need to impose constraints for assessment that would be unnecessary in an instructional context, and the need to provide example scenarios for NAEP that could be taken in the direction of either a content-based assessment like science or a more general problem-solving-with-technology assessment.

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. The TRE conception of computer skills is based on the notion that, separated from all substantive knowledge, computer skill is mastery of automatized pointing, clicking, and keying. These actions become automatized through repeated practice with different software applications. The TRE scenarios build on this notion by employing common interface conventions that students knowledgeable about computers will readily recognize, such as toolbars, radio buttons, dialog boxes, and text boxes. When this mechanical computer competency is integrated with scientific inquiry, what emerges is a purposeful, nonmechanical use of the computer for scientific problem solving.

When a student takes a TRE scenario, each action is connected to one or more variables in the student model. A three-step, evidence-modeling process was used to make these connections. The three steps are feature extraction, feature evaluation, and evidence accumulation, which are described in detail in the following sections.

Feature Extraction

For each TRE scenario, all student actions are logged in a transaction record. Feature extraction involves culling particular actions from the record (e.g., the specific experiments the student ran to solve a Simulation scenario problem). These actions, called observables, are student behaviors chosen for their presumed value as evidence of a particular student-model proficiency, or skill. Observables may include both process variables (e.g., the particular experiments run) and product variables (e.g., an answer to a multiple-choice item).

Table 2-1 shows an extraction from the first minute of the record for Simulation problem 1. The extraction shows the times and values associated with given student actions. The record shows that, in designing the experiment, the student first pressed the Choose Values button and selected a payload mass of 90 for the balloon to carry. Then the student pressed Try It to launch the balloon. Next, the student created a table, with payload mass as the only variable. Finally, the student made a graph, putting altitude on the vertical axis and amount of helium on the horizontal axis.

Note that such a transaction record may contain several hundred actions for a given student, and that some of these actions may turn out to be unimportant in making inferences about what students know and can do. The challenge for the assessment designer is to identify, through theory and empirical data, which actions constitute evidence of proficiency and which can be safely ignored.

Table 2-1. A portion of the student transaction record from TRE Simulation problem 1, grade 8: 2003

Time (in seconds)[1]	Action	Action choice
137	Begin problem 1	†
150	Choose values	90
155	Select mass	†
157	Try it	†
180	Make table	†
182	Selected table variables	Payload mass
185	Make graph	†
188	Vertical axis	Altitude
190	Horizontal axis	Helium

† Not applicable.

[1] These times include 137 seconds spent interacting with introductory material presented prior to problem 1.

NOTE: TRE = Technology-Rich Environments.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Feature Evaluation

The second step in connecting observables to the student model is feature evaluation. After desired observables have been extracted, the correctness of each one is judged. Feature evaluation involves assigning scores to observables. These scoring assignments may be done by machine or by human judges. In either case, the assignments are executed in keeping with evaluation rules. The following rule describes how to evaluate the choice of experiments the student ran to solve Simulation problem 1:

- IF the list of payload masses includes the low extreme (10), the middle value (50), and the high extreme (90), with or without additional values, THEN the best experiments were run.
- IF the list omits one or more of the required values but includes at least three experiments having a range of 50 or more, THEN very good experiments were run.
- IF the list has only two experiments but the range is at least 50, OR the list has more than two experiments with a range equal to 40, THEN good experiments were run.
- IF the list has two or fewer experiments with a range less than 50, OR has more than two experiments with a range less than 40, THEN insufficient experiments were run.

This rule generates a partial-credit score that attempts to establish whether the student conducted enough experiments—and spread the values for payload mass sufficiently—to be confident that the relationship between mass and altitude was linear throughout. Too few experiments or too narrow a spread of masses would not supply sufficient evidence to support a valid inference.

Note that formulating an evaluation rule involves an iterative process in which logical challenges to the rule are posed, and, if a challenge has merit, the rule is refined. Many refinements were made to the TRE rules based on data that suggested how well the rules captured distinctions among students of varying skill levels. Even so, no rule will accurately evaluate the behavior of all performers; that is, a given rule may award too little credit to some examinees even when they know the material or too much credit even when they do not know the material. In the assessment of group proficiency, as long as these positive and negative misclassifications are not too frequent and are not systematic (e.g., do not tend to award too little credit more often than too much credit), they can be handled effectively through mechanisms that quantify uncertainty in proficiency estimates, as described below.[8]

Evidence Accumulation

The third step in connecting observables to the student model is evidence accumulation. Feature evaluations (like test items) need to be combined into summary scores that support the inferences to be made based on student performance. Evidence accumulation entails combining the feature scores in some principled manner. Item response theory (IRT) is an example of a common evidence-accumulation method.

For TRE, summary scores were created using modeling procedures that incorporate Bayesian networks (Mislevy et al. 2000; a full discussion of the Bayesian methodology used in the TRE data analysis can be found in appendix F). Bayesian models offer a formal statistical framework for reasoning about interdependent variables in the presence of uncertainty. In contrast with the procedures typically used in NAEP assessments, Bayesian (and other similarly innovative) methods are well suited to integrated tasks like those used in TRE because the methods allow the various skills that underlie performance to be modeled individually, along with the complex interrelationships that may exist among them. (See Adams, Wilson, and Wang 1997 for another suitable modeling methodology.)

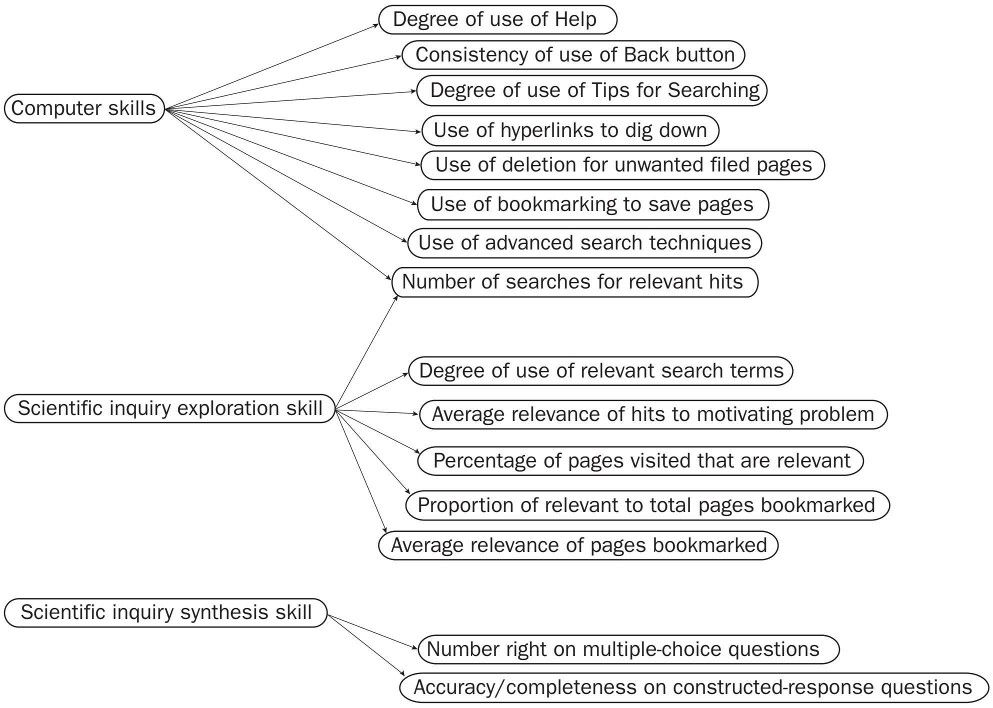
[8] Challenges were posed by advisory committee members, project team members, colleagues, and audiences hearing about the study as it progressed. Empirical evidence was gathered through several pilot tests and in the main analysis, and the rules were adjusted based on these data before the final analysis was conducted. Although they were informed by data, such revisions are ultimately judgments made by project team members. These judgments are similar to those that would be made routinely in the refinement of constructed-response rubrics during the development and scoring process for any operational assessment.

Figure 2-2 graphically depicts the evidence model for the Search senario. The model is essentially a set of hypotheses about which observables are direct evidence of the proficiencies in the student model. In the center are the student-model proficiencies—computer skills, scientific inquiry exploration skill, and scientific inquiry synthesis skill—which connect directly to the Search scenario observables. Some of the observables connected to computer skills are the use of advanced search techniques, the use of hyperlinks to drill down into web pages, and the degree of use of Tips for Searching. Some observables connected to scientific inquiry exploration skill include the degree of use of relevant search terms, the percentage of pages visited relevant to the motivating problem, and the average relevance of hits.[9] The accuracy of responses to the motivating problem and to the multiple-choice questions connect to scientific inquiry synthesis skill.

Figure 2-3 gives the evidence model for Simulation scenario problem 1. The far left of the figure shows a variable representing the context effect; that is, some local dependency among responses unrelated to the skills of interest. As stated earlier, conventional measurement models do not handle such dependency effectively. With the Bayesian methodology used in the TRE study, however, this dependency can be explicitly modeled for each problem. Note that the Search evidence model does not incorporate a context effect because the scenario contains only one main task.

The center of figure 2-3 displays the student-model proficiencies—computer skills, scientific exploration, and scientific synthesis—that connect directly to the observables. For example, how frequently Computer Help is consulted and how extensively the various components of the Simulation-tool interface are used are both connected to computer skills because they are assumed to be evidence of those skills. Some of the observables connected to scientific exploration are how frequently Science Help and the Glossary are consulted, whether the best experiments were run, whether a table or graph was used, and how

Figure 2-2. TRE Search scenario evidence model, grade 8: 2003


Computer skills
Degree of use of Help
Consistency of use of Back button
Degree of use of Tips for Searching
Use of hyperlinks to dig down
Use of deletion for unwanted filed pages
Use of bookmarking to save pages
Use of advanced search techniques
Number of searches for relevant hits
Scientific inquiry exploration skill
Degree of use of relevant search terms
Average relevance of hits to motivating problem
Percentage of pages visited that are relevant
Proportion of relevant to total pages bookmarked
Average relevance of pages bookmarked
Scientific inquiry synthesis skill
Number right on multiple-choice questions
Accuracy/completeness on constructed-response questions

NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

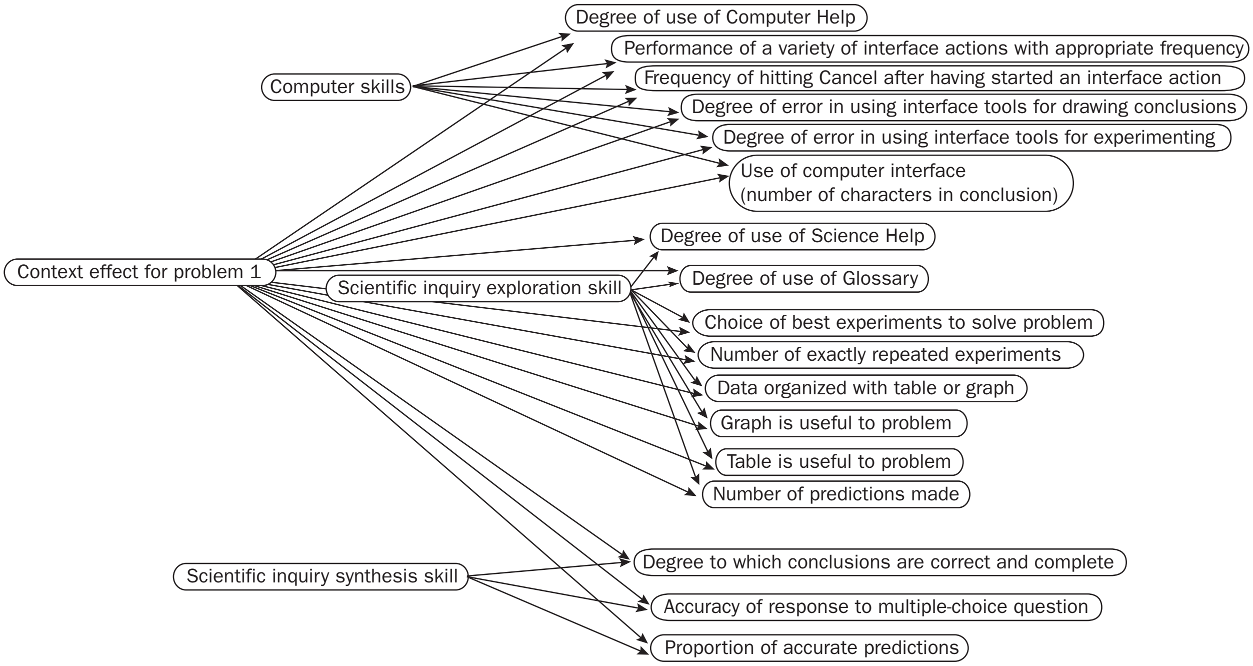
[9] Each of the approximately 5,000 pages composing the TRE Search web universe was rated independently on a scale of 1 to 4 by one staff member for its relevance to the Search motivating problem. Two additional staff members then independently rated all pages judged by the first staff member as having at least some relevance (i.e., scores of 2, 3, or 4). Disagreements between raters were resolved by consensus.

appropriate that table or graph was to the problem posed. Linked to scientific synthesis are the accuracy of answers to the constructed-response and multiple-choice questions that motivate the problem, and the proportion of accurate predictions. Some of these behaviors (such as how frequently Science Help is consulted or the same experiment is repeated) are expected to be negatively related to student proficiency. Others, like making a relevant graph, should be positively related.

How do the student and evidence models facilitate judgments about student proficiency? (Note that in the context of TRE performance, the terms "proficiency" and "proficient" denote "skill" and "skilled" and are not related to NAEP's use of "*Proficient*" as an achievement level.) As indicated by the arrows in figures 2-2 and 2-3, reasoning in the evidence model runs from left to right. That is, the likelihood of a particular level of response for an observable depends on the levels of proficiency for the variables in the student model. For example, if all other things are equal, students who are highly proficient in scientific exploration are expected to show a greater likelihood of getting the top score for running the best experiments than students who are lower in that skill. When a student responds to a scenario, the reasoning runs from right to left; the score for each observable is used to update probabilities about standing on the student-model variable to which each observable is connected. Thus, observing that a student ran the best experiments for problem 1 would increase the probability that the student is proficient in exploration skill. This increased probability would then propagate to other student-model variables linked to exploration, such as scientific inquiry and problem solving in technology-rich environments. This updating of the student model is carried out until responses to all observables are incorporated from all three Simulation problems (or from all Search scenario observables).

Note that level of standing on the student model variables constitutes a multidimensional picture of functioning that could not be generated as directly through the measurement models routinely used in main NAEP assessments. Typically, multiple skills are modeled by creating separate measurement scales, each of which is indicated by a unique set of items. With the student and evidence models implemented within a Bayesian framework, test developers can instead use integrated tasks, each of which measures a mix of skills, and attempt to model standing on each skill by connecting it to the relevant features of student responses.

Figure 2-3. TRE Simulation scenario evidence model for problem 1, grade 8: 2003


Degree of use of Computer Help
Performance of a variety of interface actions with appropriate frequency
Computer skills
Frequency of hitting Cancel after having started an interface action
Degree of error in using interface tools for drawing conclusions
Degree of error in using interface tools for experimenting
Use of computer interface
(number of characters in conclusion)
Degree of use of Science Help
Context effect for problem 1
Degree of use of Glossary
Scientific inquiry exploration skill
Choice of best experiments to solve problem
Number of exactly repeated experiments
Data organized with table or graph
Graph is useful to problem
Table is useful to problem
Number of predictions made
Degree to which conclusions are correct and complete
Scientific inquiry synthesis skill
Accuracy of response to multiple-choice question
Proportion of accurate predictions

NOTE: TRE = Technology-Rich Environments.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Chapter 3: The TRE Student Sample—Attitudes Toward and Experiences With Technology and the Nature of Science Coursework

The TRE study was conducted during the spring of 2003. The TRE student sample was a nationally representative group of 2,110 eighth-grade students from 222 schools. Students were randomly assigned to one of the two scenarios, Search or Simulation, during administrations; ultimately, 1,077 students received the Search scenario, and 1,033 received the Simulation scenario. No group of students was asked to respond to both scenarios because the time burden would have been excessive. Technical details about the methods used to obtain the student samples can be found in appendix B.

When students responded to one of the two TRE scenarios, they also responded to background questions designed to gather information about their familiarity with computers and science activities in school. Exploring the percentages of students who gave various responses to a selection of these background questions offers useful information about the kinds of knowledge, skills, and attitudes students reported bringing to the two scenarios.

For example, how familiar with computers were the participating students? Tables 3-1 through 3-4 display students' responses to computer-related background questions. Consistent with previous NAEP studies (e.g., Horkay et al. 2005), table 3-1 shows that the majority of students (88 percent for Search and 86 percent for Simulation) reported having a computer at home that they use. In addition, approximately 86 percent of students for Search and 85 percent of students for Simulation reported that they use a computer outside of school at least once a week (see table 3-2). The percentages of students who reported using a computer once a week or more at school were approximately 57 percent for Search and 59 percent for Simulation.

Table 3-1. Percentage distribution of students indicating there is a computer at home that they use, by scenario, grade 8: 2003

	Is there a computer at home that you use?	
Scenario	Yes	No
Search	88 (1.3)	12 (1.3)
Simulation	86 (2.0)	14 (2.0)

NOTE: The number of students responding was 1073 for Search and 1027 for Simulation. Detail may not sum to totals because of rounding. Standard errors of the estimates appear in parentheses.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Table 3-2. Percentage distribution of students, by frequency of computer use, and by scenario, grade 8: 2003

	How often do you use a computer at school?				
Scenario	Daily	2–3 times per week	Once a week	Once every few weeks	Never or hardly ever
Search	20 (1.6)	23 (1.7)	14 (0.9)	24 (1.7)	19 (1.6)
Simulation	23 (1.4)	21 (1.5)	15 (1.1)	23 (1.3)	18 (2.0)
	How often do you use a computer outside of school?				
Scenario	Daily	2–3 times per week	Once a week	Once every few weeks	Never or hardly ever
Search	51 (1.7)	26 (1.4)	9 (0.7)	7 (1.0)	7 (0.9)
Simulation	53 (2.2)	25 (1.1)	7 (0.8)	7 (1.0)	8 (1.1)

NOTE: The number of students responding was 1073 for Search and ranged from 1029 to 1030 for Simulation. Detail may not sum to totals because of rounding. Standard errors of the estimates appear in parentheses.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Apart from their apparent familiarity with computers, students also indicated feeling positively about using computers. Table 3-3 shows that approximately 70 percent of students for Search and 74 percent of students for Simulation reported that they agreed or strongly agreed that they are more motivated to do schoolwork on a computer. Approximately 81 percent of students for Search and 85 percent of students for Simulation agreed or strongly agreed that they have more fun learning on a computer, and about 75 percent of students for Search and 80 percent of students for Simulation agreed or strongly agreed that they get more schoolwork done when using a computer.

Table 3-3. Percentage distribution of students, by attitude statements toward computers and schoolwork, and by scenario, grade 8: 2003

	I am more motivated to do schoolwork on a computer.				
Scenario	Strongly agree	Agree	Disagree	Strongly disagree	Never use a computer
Search	18 (1.3)	52 (2.2)	22 (1.2)	4 (0.6)	3 (0.6)
Simulation	25 (1.6)	49 (1.6)	19 (1.3)	4 (0.6)	3 (0.6)
	I have more fun learning on a computer.				
Scenario	Strongly agree	Agree	Disagree	Strongly disagree	Never use a computer
Search	33 (1.5)	48 (1.8)	15 (0.9)	2 (0.4)	2 (0.4)
Simulation	35 (1.5)	50 (1.6)	11 (1.1)	2 (0.4)	1 (0.3)
	I get more done when using a computer for schoolwork.				
Scenario	Strongly agree	Agree	Disagree	Strongly disagree	Never use a computer
Search	29 (1.3)	46 (1.6)	20 (1.0)	3 (0.5)	2 (0.6)
Simulation	32 (1.2)	48 (1.4)	15 (1.0)	3 (0.5)	2 (0.4)

NOTE: The number of students responding ranged from 1060 to 1070 for Search and ranged from 1018 to 1023 for Simulation. Detail may not sum to totals because of rounding. Standard errors of the estimates appear in parentheses.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Students were also asked to what extent they used computers at home and at school: not at all, to a small extent, to a moderate extent, or to a large extent. As indicated by table 3-4, the most common pursuit was finding information on the Internet, followed by using a word processor, using e-mail, and talking in chat groups. Approximately 87 percent of students for Search and 87 percent for Simulation reported finding information on the Internet to a moderate or large extent, about 67 percent of students for both scenarios reported using word processors to a moderate or large extent, approximately 64 percent of students for both scenarios reported using e-mail to a moderate or large extent, and 55 percent of students for Search and 56 percent for Simulation reported talking in chat groups to a moderate or large extent.

Table 3-4. Percentage distribution of students, by extent of specific computer use, and by scenario, grade 8: 2003

	Play computer games			
Scenario	Not at all	Small extent	Moderate extent	Large extent
Search	8 (1.0)	44 (1.3)	36 (1.2)	12 (1.0)
Simulation	8 (1.0)	43 (2.0)	35 (1.7)	14 (1.1)
	Use a word processor			
Scenario	Not at all	Small extent	Moderate extent	Large extent
Search	10 (1.0)	23 (1.1)	40 (1.7)	27 (1.7)
Simulation	7 (0.9)	26 (1.4)	40 (1.6)	27 (1.3)
	Make drawings/art on computer			
Scenario	Not at all	Small extent	Moderate extent	Large extent
Search	25 (1.3)	48 (1.6)	18 (1.2)	8 (1.0)
Simulation	25 (1.2)	45 (1.5)	19 (1.0)	10 (1.0)
	Make tables, charts or graphs on computer			
Scenario	Not at all	Small extent	Moderate extent	Large extent
Search	26 (1.7)	46 (1.8)	22 (1.4)	7 (0.9)
Simulation	28 (1.6)	48 (1.9)	17 (1.1)	7 (0.9)
	Look up information on a CD			
Scenario	Not at all	Small extent	Moderate extent	Large extent
Search	18 (1.6)	33 (1.8)	29 (1.5)	20 (1.2)
Simulation	19 (1.1)	32 (1.4)	31 (1.3)	18 (1.1)
	Find information on the Internet			
Scenario	Not at all	Small extent	Moderate extent	Large extent
Search	2 (0.5)	10 (1.1)	32 (1.2)	55 (1.6)
Simulation	2 (0.5)	10 (1.0)	33 (1.7)	54 (1.5)
	Use e-mail			
Scenario	Not at all	Small extent	Moderate extent	Large extent
Search	19 (1.3)	17 (1.0)	23 (1.2)	41 (1.4)
Simulation	17 (2.0)	19 (1.3)	22 (1.6)	42 (2.0)
	Talk in chat groups			
Scenario	Not at all	Small extent	Moderate extent	Large extent
Search	25 (1.5)	20 (1.3)	20 (1.3)	35 (1.6)
Simulation	23 (1.7)	21 (1.6)	20 (1.5)	36 (2.2)

NOTE: The number of students responding ranged from 1068 to 1072 for Search and ranged from 1018 to 1029 for Simulation. Detail may not sum to totals because of rounding. Standard errors of the estimates appear in parentheses.

Source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Because the TRE scenarios require students to solve science problems, information was collected about students' science activities at school. Tables 3-5 and 3-6 summarize this information. Table 3-5 indicates that approximately 96 percent of students for Search and 96 percent for Simulation reported being enrolled in a science course, with most students for each scenario divided among Earth science, general science, and physical science classes.

According to table 3-6, students engaged in a variety of science activities. For instance, 68 to 77 percent of students reported that they were at least sometimes engaged in such activities as designing their own experiments, carrying out experiments, and writing up results. (The responses "sometimes, but less than once a month" and "once a month or more" were combined to derive the "at least sometimes" measure.) Further, 61 to 73 percent of students reported at least sometimes using computers for downloading data from the Internet, for analyzing data, and for collecting data. Approximately one-half of the students said they at least sometimes used computer simulations in science.

Table 3-5. Percentage distribution of students, by enrollment in particular science courses, and by scenario, grade 8: 2003

	Which best describes the science course you are taking?					
Scenario	Not taking science	Life science	Physical science	Earth science	General science	Integrated science
Search	4 (0.7)	9 (0.9)	21 (2.9)	30 (3.0)	23 (1.9)	13 (1.4)
Simulation	3 (0.7)	9 (1.3)	23 (3.0)	31 (3.4)	20 (1.8)	13 (1.6)

NOTE: The number of students responding was 1067 for Search and 1027 for Simulation. Detail may not sum to totals because of rounding. Standard errors of the estimates appear in parentheses.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Table 3-6. Percentage distribution of students, by frequency of school science activities and scenario, grade 8: 2003

	Design your own science experiment			
Scenario	Not taking science	Once a month or more	Sometimes, but less than once a month	Never
Search	3 (0.8)	26 (1.3)	44 (2.3)	26 (2.5)
Simulation	2 (0.5)	34 (1.6)	43 (2.0)	22 (1.5)
	Carry out science experiment			
Scenario	Not taking science	Once a month or more	Sometimes, but less than once a month	Never
Search	3 (0.9)	26 (1.5)	42 (2.0)	29 (2.3)
Simulation	2 (0.4)	31 (1.9)	39 (2.0)	29 (1.9)
	Write up results of science experiment			
Scenario	Not taking science	Once a month or more	Sometimes, but less than once a month	Never
Search	4 (0.7)	29 (1.8)	39 (1.6)	28 (2.1)
Simulation	2 (0.4)	35 (1.8)	39 (1.8)	24 (1.7)
	Talk to class about results of experiment			
Scenario	Not taking science	Once a month or more	Sometimes, but less than once a month	Never
Search	3 (0.8)	21 (1.7)	39 (2.0)	37 (2.5)
Simulation	2 (0.4)	25 (1.8)	38 (1.5)	36 (1.6)
	Collect data using computerized lab equipment			
Scenario	Not taking science	Once a month or more	Sometimes, but less than once a month	Never
Search	4 (0.9)	25 (1.4)	36 (1.6)	36 (1.3)
Simulation	2 (0.4)	29 (1.5)	37 (1.3)	33 (1.2)
	Download data from the Internet			
Scenario	Not taking science	Once a month or more	Sometimes, but less than once a month	Never
Search	3 (0.8)	32 (1.8)	41 (2.1)	25 (1.3)
Simulation	2 (0.5)	33 (1.7)	36 (1.6)	29 (1.5)
	Analyze data using computer			
Scenario	Not taking science	Once a month or more	Sometimes, but less than once a month	Never
Search	3 (0.8)	28 (1.2)	41 (1.5)	28 (1.6)
Simulation	2 (0.4)	28 (1.2)	38 (1.4)	32 (1.5)
	Use the Internet to exchange information with other students or scientists about experiments			
Scenario	Not taking science	Once a month or more	Sometimes, but less than once a month	Never
Search	3 (0.8)	16 (1.4)	25 (1.2)	56 (1.9)
Simulation	2 (0.4)	13 (1.2)	21 (1.3)	64 (1.7)
	Use computer simulations to perform experiments or explore science topics			
Scenario	Not taking science	Once a month or more	Sometimes, but less than once a month	Never
Search	4 (0.9)	17 (1.5)	38 (1.6)	41 (1.8)
Simulation	2 (0.4)	16 (1.3)	33 (1.3)	49 (1.5)

NOTE: The number of students responding ranged from 1059 to 1069 for Search and ranged from 1009 to 1023 for Simulation. Detail may not sum to totals because of rounding. Standard errors of the estimates appear in parentheses.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Chapter 4: Scoring TRE

As described in chapter 2, TRE employed an evidence-modeling process for scoring in which student actions were first identified, then evaluated for correctness, and finally aggregated to create scores. The evaluation portion of this process generally relied upon traditional approaches to machine scoring. In two cases, student inputs were handled differently. Students' typed responses to the open-ended motivating questions in the Search and Simulation scenarios were read and scored by human raters, and students' search queries in the Search scenario were evaluated using c-rater, an Educational Testing Service (ETS) computer program that performs automated scoring of short constructed responses. These two cases are discussed in greater detail below.

The TRE Motivating Problems

The constructed-response questions that students answered as part of the Search and Simulation scenarios are a central measure of students' scientific inquiry synthesis skills. The questions, referred to in this report as "problems" or as "motivating problems," are visible to students throughout their work on the scenarios because they were designed to inspire students' scientific inquiries in addition to serving as a measure of students' understanding at the end of the process. The Search scenario presents a single motivating problem, along with a set of multiple-choice questions, that students have 40 minutes in total to investigate and answer. The Simulation scenario uses three motivating problems, one in each of the three parts of the scenario.

Three motivating problems were originally offered in the pilot test of the TRE Search scenario; students had to respond to two of them. Two of the problems were dropped for a variety of reasons, however, including weak student performance and evidence that students did not have sufficient time to complete two problems. Having only a single motivating problem both severely limited the evidence available for estimating students' proficiency and increased the influence of problem context on performance. To increase the likelihood that enough evidence would remain to measure students' scientific inquiry synthesis skills and to reduce context effects, the second motivating problem was replaced by four multiple-choice questions. The multiple-choice questions required students to draw conclusions about topics they were likely to encounter while investigating the motivating problem. The search capability remained available in case students needed to conduct additional searches before answering the multiple-choice questions.

As is typical in National Assessment of Educational Progress (NAEP) item development, TRE staff wrote scoring guides (or evaluation rules, as they are more generally called in the Evidence-Centered Design [ECD] framework) concurrently with development of the motivating problems, and they revised those guides as the problems evolved through reviews and pilot testing. The guides contained either three or four levels, depending on how many meaningful distinctions in performance could be made reliably. In both the three- and four-level guides, the lowest level (denoted as "1") was considered to be unacceptable performance and received no credit. The top level was considered to be "best." Although responses in the highest category may have had some flaws, whatever flaws they had were considered to be minor. The scoring guides for the Search motivating problem and for Simulation motivating problem 1 used three levels, where a score of 3 was a "best" response, 2 was a "partial" response, and 1 was an "unacceptable" response. Because an additional level of response could be qualitatively distinguished, the scoring guides for Simulation motivating problems 2 and 3 used four levels. A score of 4 was a "best" response, a score of 3 was a "good" response, a score of 2 was a "partial" response, and score of 1 was an "unacceptable" response.

Scoring Procedures

Scoring for the TRE motivating problems followed procedures similar to those used in scoring other NAEP assessments, for example, mathematics and science. One member of the NAEP ETS staff was assigned to train raters for the three Simulation questions, and a second staff member trained the same raters for the Search question. Prior to scoring, the trainer read through a sample of student responses for each problem and prepared materials with which to train and guide raters. A team of six raters was assembled to score the student responses. The raters were all members of the ETS staff; most were experienced test developers well versed in scoring procedures.

Meeting as a group under the direction of the trainer, the raters read a problem and its scoring guide to understand what was expected of students. The trainer then presented and explained an "anchor set" of actual student responses chosen to illustrate the range at each score point. Next, raters independently scored two sets of practice responses. These were discussed by the group until all the raters felt comfortable applying the scoring guide. During scoring, raters generally began by working in pairs until they had scored 20 or 30 responses. The paired scoring allowed raters to discuss further the scoring guides and their application to individual student responses. Difficult issues were brought to the attention of the entire team for resolution, and scoring guides were amended as necessary to guide the scoring of similar kinds of responses that might yet appear.

Table 4-1. Interrater reliability in scoring constructed-response motivating problems, grade 8: 2003

Task	Scale	Number of second scores	Percent agreement
Search problem	1–3	268	90
Simulation problem 1	1–3	267	95
Simulation problem 2	1–4	258	89
Simulation problem 3	1–4	258	89

NOTE: The number of students responding was 1077 for Search and 1033 for Simulation.
SOURCE: U.S Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

When the raters were ready, they began to score on their own, and continued until they had read all the responses assigned to them. In all cases, scores were awarded based on the criteria that were set forth in the scoring guides and elaborated in the anchor and practice responses. As is typical in NAEP assessments, raters were concerned only with the content of a student's response, not with the quality of the prose or accuracy of the typing, except of course when poor writing and/or typing errors made it impossible to decipher what the student meant to say. Raters recorded their scores directly on the paper with the student's printed response. The scores were then compiled into a spreadsheet for analysis.

To assess the reliability of scoring, 25 percent of all student responses were read and independently scored by a second rater, who was not privy to the first rater's grade, and the degree of agreement between raters was estimated. Clean printed copies of these student responses were distributed among all six raters in such a way that each rater served as a check on all the other raters. In cases of disagreement between the first and second scores, the trainer read and assigned a resolved score to the response.

Interrater reliability was within NAEP standards for all four problems. The reliability results are shown in table 4-1. For each problem, the table presents the scale range, the number of second scores, and the percent agreement.

Scoring Guides and Sample Student Responses

This section presents the four motivating problems from the Search and Simulation scenarios. For each motivating problem, the scoring guide, the distribution of scores, and sample student responses are presented.

Search Scenario

The motivating problem and scoring guide for the TRE Search scenario are given in figure 4-1. The motivating problem requires students to present three reasons why scientists use scientific gas balloons to explore space and the atmosphere. To respond to the problem, students have to find useful web pages,

Figure 4-1. Search motivating problem and scoring guide (evaluation rule), grade 8: 2003

Some scientists study space with large helium gas balloons. These balloons are usually launched from the ground into space but can also be launched from spacecraft near other planets.

Why do scientists use these gas balloons to explore outer space and the atmosphere instead of using satellites, rockets, or other tools? Be sure to explain at least three advantages of using gas balloons.

Base your answer on more than one web page or site. Be sure to write your answer in your own words!

Scoring Guide:

3—Best: Response gives at least three advantages of using gas balloons.
Acceptable responses can include:

- Relatively cheap.
- Can be prepared in a relatively short amount of time.
- Can be launched from numerous locations.
- Payloads are recoverable and reusable (the balloons are NOT reusable).
- Can stay at a constant altitude.
- Can rise relatively slowly (making observations along the way).
- Float above much of the atmosphere, resulting in less interference.
- Can carry heavy payloads.
- Long flight duration.
- Flexibility in configuration.
- Highly reliable.
- No pollution/better for the environment.
- Vibration-free.
- Low G-forces during take-off.
- Unmanned (meaning less risk to humans, cheaper to operate).
- Safe (must explain, i.e., no explosive fuels like in rockets, no crew).

Note: If students refer to hot air balloons or weather balloons instead of properly stating "helium gas balloons," accept the answer as long as the advantages cited are true of helium gas balloons.
Do not accept (unless explained or placed in context):

- "Better."
- "Faster."
- "More efficient."
- "Easier to use."
- Scientists receive information faster.
- Safer because they won't fall on people.
- "They go high" (must explain why this is a benefit).
- "Travel long distances."

2—Partial: Response gives one or two advantages of using gas balloons.
1—Unacceptable: Response does not give any advantages of using gas balloons.

SOURCE: U.S Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

locate the necessary information within those pages, and present the information in their written answer. "Best" responses present three advantages, whereas "partial" responses present one or two advantages, as described in the scoring guide shown in figure 4-1.

The third paragraph of the Search motivating problem contains two requirements for students: "Base your answer on more than one web page or site. Be sure to write your answer in your own words!" Student compliance with these requirements was not factored into scoring; the requirements were expected to prompt better work from students.

The two requirements were the result of discussions that arose during the development of the Search scenario. The first addressed the concern that students who hit upon a web page that listed numerous advantages of scientific balloons (there were a few such pages among those available in the web universe for the Search scenario) could write their entire answers based on that page. Since the motivating problem was designed to measure synthesis skills—that is, students' abilities to gather and integrate information from more than one place—TRE staff included the suggestion that students draw upon more than one page or site in their answer.

The suggestion for students to answer the motivating problem in their own words grew from the concern, expressed by both TRE staff and the TRE Development Committee, that some students might copy their responses directly from the websites they visited. The Search scenario was designed to be as realistic as possible within the limitations of an assessment environment. Since students doing research on their computers are able to copy and paste information, it was strongly felt that students taking the Search scenario should be able to do the same. When the TRE pilot test confirmed that some students were copying, and doing so without making any effort to cite their sources or to rewrite the information in their own words, TRE staff added the new wording to the motivating problem. However, Search scenario scoring did not penalize students who might have copied text without citations.

As table 4-2 shows, 15 percent of students were able to give three advantages of using gas balloons, required for a "best" response; 35 percent could give a "partial" response with one or two advantages; and about one-half of all students received no credit on the question. For the purposes of calculating the mean, blank and off-topic responses were given the same value as an unacceptable response.

Table 4-2. Percentage distribution of student scores on Search motivating problem, grade 8: 2003

Score	Percentage
3 - "best"	15
2 - "partial"	35
1 - "unacceptable"	43
Blank or off-topic	6

NOTE: The number of students responding was 1077. Detail may not sum to totals because of rounding.
SOURCE: U.S Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

The two sample responses shown in figure 4-2 received a score of 3, "best." The main difference among answers at this score level was in the specific advantages the students listed.

Figure 4-2. Two responses to the Search motivating problem receiving a score of 3, "best," grade 8: 2003

- 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.
- Using ballons to do scientific experiments has several advantages wich I will only name a few. The first advantage is that they allow the payloads that they are carring to lift with out no vibrations or G-forces that a rocket would, and may damage the payload. Another advantage is that the ballons are quickly launched and they are quickly recoverd allowing multiple flights on the same instruments. Another advantage is that balloons offer a low-cost, quick-response method for doing scientific investigations and balloons are mobile, meaning they can be launched where the scientist needs to conduct the experiment.they are also cheap and safer for undergraduate and graduate students conducting work in scientific fields.

NOTE: Responses are the unedited, verbatim answers given by students.
SOURCE: U.S Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

The next two responses, shown in figure 4-3, received scores of 2, "partial." In the first response, the student did not provide enough detail in the second sentence for the rater to know whether there were two distinct points about human involvement being made. In the second response, no credit was awarded for saying simply that the balloon can fly high, since satellites and rockets can also fly high. To have received credit, the answer would have needed further elaboration, such as a direct comparison to earth-bound telescopes or an explanation of the advantage balloons have in taking measurements from within the stratosphere.

Figure 4-3. Two responses to the Search motivating problem receiving a score of 2, "partial," grade 8: 2003

- they use these because they are less expsenive. A human dose not have to be in one and there is no risk of loseing lives.
- Scientists use balloons for space and atmospherical experiments because they can offer cababilities that can not be made through the use of rockets or airplanes. The three advantages of using balloons for research is that balloons can be set upalmost anywere and they can be ready for flight under 6 months, and lastly they can fly real high, about 26 miles above the earth.

NOTE: Responses are the unedited, verbatim answers given by students.
SOURCE: U.S Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

The sample shown in figure 4-4, in which the response does not actually give an advantage of scientific gas balloons, is typical of many that received no credit.

Figure 4-4. A response to the Search motivating problem receiving a score of 1, "unacceptable," grade 8: 2003

You use the Balloon to go around the world and use them for Meteorology and explore outer space.

NOTE: Responses are the unedited, verbatim answers given by students.
SOURCE: U.S Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Simulation Scenario Problem 1

Figure 4-5 presents Simulation motivating problem 1 and its scoring guide.

As seen in table 4-3, about one-quarter of students received a score of "best" on the motivating problem, and 44 percent received partial credit. Almost one-third of students wrote "unacceptable" answers.

Table 4-3. Distribution of student scores on Simulation motivating problem 1, grade 8: 2003

Score	Percentage
3 - "best"	23
2 - "partial"	44
1 - "unacceptable"	31
Blank or off-topic	2

NOTE: The number of students responding was 1033. Detail may not sum to totals because of rounding.
SOURCE: U.S Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure 4-5. Simulation motivating problem 1 and scoring guide, grade 8: 2003

How do different payload masses affect the altitude of a helium balloon? Support your answer with what you saw when you experimented.

Scoring Guide:

3—Best: Response contains a correct statement summarizing the relationship between mass and altitude, i.e., "The more mass the balloon carries, the lower the balloon altitude." AND, the response refers specifically to two experiments that support the summarization in one of the following ways:

- Two masses and two altitudes.
- Two masses.
- One mass with a clear comparative statement, e.g., "I used the 50 lb. mass and then the less mass I used the higher the balloon went."

2—Partial: The response:

- Offers a comparative statement about the highest and lowest mass, e.g., "When I used the greatest mass, the balloon went lower than when I used the least mass."
- Correctly summarizes the relationship but makes no reference to any specific masses.
- Correctly summarizes the relationship with reference to one specific experiment (mass) with NO comparative statement.
- Correctly summarizes the relationship but incorrectly refers to masses and/or altitudes (without being contradictory).
- Refers to data that support correct summarization of the relationship, but offers no summary statement.
- Correctly summarizes the data, but gives a conclusion that contradicts the summary and data.

1—Unacceptable: The response:

- Offers an incorrect summary of the relationship between mass and altitude.
- Refers to data that do NOT support the correct relationship.
- Offers ONLY irrelevant information regarding volume, speed, or time.
- Offers nonsensical statements.
- Offers data and a summary statement that contradict each other.

SOURCE: U.S Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure 4-6 shows a student response that received a score of 3, “best.” The response correctly summarizes the relationship between mass and altitude and provides evidence to support it from three experiments, supplying more than the required two data points. One may argue that the student should have provided evidence from an experiment using the heaviest payload to show that the pattern continues with still greater mass. However, in the evidence model developed for the Simulation tasks, students’ choices of which experiments to run are captured separately and analyzed as part of their exploration skill rather than as part of their synthesis skill.

Figure 4-6. A response to Simulation motivating problem 1 receiving a score of 3, “best,” grade 8: 2003

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.

NOTE: Responses are the unedited, verbatim answers given by students.
SOURCE: U.S Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

The next example, given in figure 4-7, received a score of 2, “partial.” The response gives a correct summary of the relationship between mass and altitude, and refers to two experiments, but the specific data it provides are incorrect (the balloon actually reaches an altitude of 36,211 ft. with a 10 lb. payload).

Figure 4-7. A response to Simulation motivating problem 1 receiving a score of 2, “partial,” grade 8: 2003

when you put only ten pounds of payload then it will reach the height of about four thousand feet. When I put twenty pounds of pay load in the balloon it rose to a smaller height. So as the weight gets larger it will rise less and less

NOTE: Responses are the unedited, verbatim answers given by students.
SOURCE: U.S Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure 4-8 gives an example of a response that received a score of 1, “unacceptable.” As can be seen, the response gives an incorrect summary of the relationship between mass and altitude and provides no experimental data.

Figure 4-8. A response to Simulation motivating problem 1 receiving a score of 1, “unacceptable,” grade 8: 2003

The more payload mass you have the higher the baloon will go. The higher payload mass I picked the higher the balloon went.

NOTE: Responses are the unedited, verbatim answers given by students.
SOURCE: U.S Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Simulation Scenario Problem 2

Figure 4-9 shows the motivating problem and scoring guide for Simulation motivating problem 2.

Table 4-4 shows the distribution of student scores for Simulation motivating problem 2. Approximately one-third of student responses were scored either "good" or "best," and one-third were scored "partial." One-third of the responses received a score of "unacceptable."

Table 4-4. Percentage distribution of student scores on Simulation motivating problem 2, grade 8: 2003

Score	Percentage
4 - "best"	13
3 - "good"	18
2 - "partial"	33
1 - "unacceptable"	33
Blank or off-topic	2

NOTE: Number of students responding was 1033. Detail may not sum to totals because of rounding.
SOURCE: U.S Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure 4-9. Simulation motivating problem 2 and scoring guide, grade 8: 2003

How do different amounts of helium affect the altitude of a helium balloon? Support your answer with what you saw when you experimented.

Scoring Guide:

4—Best: Response contains a correct explanation of the relationship between amount of helium and balloon altitude for a payload mass of 100 lb. A correct explanation states that once enough helium is in the balloon to get the balloon off the ground, the balloon will rise to a maximum altitude and no higher, even if more helium is added.
3—Good: Response makes one of the following two points related to the step function:

- A certain amount of helium is needed to get the balloon off the ground. OR
- The response indicates that once airborne, the balloon will reach a maximum altitude no matter how much helium is added.

2—Partial: Response explains that more helium results in a higher altitude, or less helium results in a lower altitude.
1—Unacceptable: Response explains none of the points above or makes a declarative statement that the balloon does not rise.

NOTE ABOUT DESCRIBING THE BOTTOM OF THE STEP FUNCTION: For levels 3 and 4, the student must refer to more than one value of helium that fails to lift the balloon. If the student does not explicitly or implicitly state that there is a range of values for which balloon altitude is 0 and/or 2 feet and that below a certain amount of helium the balloon will remain on the ground, (e.g., "It took x amount of helium to lift the balloon..."), then the student MUST refer to more than one value of helium that fails to lift the balloon.
Examples of explicit statements or statements that imply that there is a range of values for which balloon altitude is 0 and/or 2 and that below a certain amount of helium the balloon will remain on the ground:

- It took x amount of helium to lift the balloon.
- Below x amount of helium, the balloon will not get off the ground.
- If there is not enough helium, the balloon will not go up.
- With 900 to 1500 cu. ft., it does not even move.

Do not accept answers that state that the balloon never rises.

SOURCE: U.S Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

The two student responses in figure 4-10 received scores of 4, “best.” The first response is an excellent answer that describes the step function and uses data from several experiments for support. The second response is not as good as the first—it was considered to be at the borderline between “good” and “best”—but it does meet the requirements for the top score by explaining that a minimum amount of helium is needed to lift the balloon and that once the maximum altitude is reached, additional amounts of helium have no further effect on altitude.

Figure 4-10. Two responses to Simulation motivating problem 2 receiving a score of 4, “best,” grade 8: 2003

- 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
- There has to be at least 2500 cubic feet of helium for the balloon to move. And after that point the amount of helium does not affect the height that the balloon travels

NOTE: Responses are the unedited, verbatim answers given by students.
SOURCE: U.S Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Responses that correctly described either the bottom or the top of the step function received a score of 3, “good.” The first response in figure 4-11 describes the bottom of the function well, but the phrase “the total altitude did not always change” is not a clear statement of what happens to the balloon after it lifts off the ground. The second response is written in reference to the top of the step function. The student may have wanted the first phrase to describe the bottom threshold, but unlike the description of the top, it is not clear enough to demonstrate understanding.

Figure 4-11. Two responses to Simulation motivating problem 2 receiving a score of 3, “good,” grade 8: 2003

- Different amounts of helium affect the altitude of a helium balloon greatly. The more helium that is put into the balloon the faster it rises into the air (lower time to final altitude). The total altitude did not always change when different amounts of helium were put into the balloon but when 2400 ft or less was was put into the balloon it could not support the weight of the payload mass that balloon barely liftede off of the round.
- After a certain amount of helium is used, a balloon with a the same amount of weight payload can not go past a certain altitude. It shows on the graphs after 2500 cubic feet of helium in a balloon a the ballon’s altitude levels off at 10000 feet.

NOTE: Responses are the unedited, verbatim answers given by students.
SOURCE: U.S Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

A score of 2, “partial,” was awarded to a special (and common) class of responses that offered an essentially true statement but entirely missed the nuances of the step function. The response in figure 4-12 is an example. Students seem to have arrived at this type of answer by several different paths. For example, students who ran only two experiments—one using too small an amount of helium to make the balloon rise, and the other using an amount that lifted the balloon to its maximum altitude—would have shown a straight line rising from the first data point to the second had they graphed their results. In the absence of further experiments, these students could easily, though incorrectly, conclude that a linear relationship existed, in which the greater the amount of helium the greater the altitude.

Figure 4-12. A response to Simulation motivating problem 2 receiving a score of 2, “partial,” grade 8: 2003

The more helium the higher the balloon goes up. The less helium the lower the balloon will rise.

NOTE: Responses are the unedited, verbatim answers given by students.
SOURCE: U.S Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Some anecdotal evidence from conversations with students at earlier stages of the project suggested that students simply did not want to believe the evidence in front of them; they were familiar with linear relationships but unused to seeing anything like a step function. When asked to describe the nonlinear pattern from their experiments, students questioned or ignored the information in front of them and tried to express their answers in more familiar terms.

Finally, figure 4-13 shows a response that received a score of 1, “unacceptable.” By oversimplifying and failing to distinguish between different helium volumes, it draws an incorrect conclusion for the problem as a whole.

Figure 4-13. A response to Simulation motivating problem 2 receiving a score of 1, “unacceptable,” grade 8: 2003

In my experimment I saw no matter what the volume the altitude was still the same

NOTE: Responses are the unedited, verbatim answers given by students.
SOURCE: U.S Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Simulation Scenario Problem 3

The motivating problem and scoring guide for the last of the three Simulation problems is shown in figure 4-14.

Simulation 3 was clearly the most challenging for students. To be successful, students had to manipulate two variables instead of one, run many experiments, synthesize a good deal of information, and express their complex findings in a coherent way. As can be seen in table 4-5, less than 10 percent of responses received a score of 3 or better, and 44 percent received scores of "unacceptable."

Table 4-5. Percentage distribution of student scores on Simulation motivating problem 3, grade 8: 2003

Score	Percentage
4 - "best"	2
3 - "good"	7
2 - "partial"	43
1 - "unacceptable"	44
Blank or off-topic	4

NOTE: Number of students responding was 1033. Detail may not sum to totals because of rounding.
SOURCE: U.S Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure 4-14. Simulation motivating problem 3 and scoring guide, grade 8: 2003

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.

Scoring guide

4—Best: Response contains a correct explanation of the relationship between amount of helium and balloon altitude for more than one payload mass. This explanation can be described verbally without reference to specific values, only by referring to specific values, or by a combination of the two. A correct explanation portrays the step function for multiple payload masses: The amount of helium needed to lift the balloon is greater the greater the mass the balloon carries. Once airborne, balloons will reach a maximum altitude for a given mass no matter how much helium is added. The maximum altitude decreases as mass increases.

3—Good: Response describes EITHER the bottom OR the top of the step function by making one of the following two points:·

- The amount of helium needed to lift the balloon is greater the greater the mass the balloon carries. OR
- Once airborne, balloons will reach a maximum altitude for a given mass no matter how much helium is added. The maximum altitude decreases as mass increases.

2—Partial: Response contains one of the following points that can be derived from problems 1 or 2:

- Below a certain amount of helium the balloon will not be able to get off the ground.
- The altitude the balloon reaches is lower the greater the mass.
- The balloon will reach a maximum altitude and go no higher when more helium is added.

OR

Response contains a general response that takes both variables into consideration:

- Response explains that less mass and more helium result in a higher altitude (or more mass and less helium results in a lower altitude).
- Response gives three data points with at least two different masses and volumes that suggest a linear relationship.

1—Unacceptable: Response explains none of the points above.

- General response with one or both variables in wrong direction ("less mass and more helium results in lower altitude;" "higher mass and more helium results in higher altitude").
- Response simply gives two data points.

SOURCE: U.S Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

To receive a score of 4, “best,” students had to describe the pattern of multiple step functions where, as mass increased, more helium was required to lift the payload off the ground, and the maximum altitude of the balloon decreased. Students were able to give their answers in any of three ways: by describing the pattern, by showing the pattern through the use of data, or by a combination of the two. The first example in figure 4-15 gives a good initial description, which could probably stand on its own, and then supports it with evidence. The second example succeeds through a combination of written description and data. It gives a clear description of the bottom of the step function where the “larger the payload of the balloon the more helium it takes to make the balloon take off,” whereas understanding of the top of the step function is suggested by the choice of data presented rather than by an explicit description.

Figure 4-15. Two responses to Simulation motivating problem 3 receiving a score of 4, “best,” grade 8: 2003

- 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.
- The ammount of helium and the mass of the payload affect the altitude of the balloon. The larger the payload of the balloon the more helium it takes to make the balloon take off. With 10 lbs. payload it took 910 cu. ft. of helium to make the balloon take off from the ground, and 975 cu. ft. of helium to have the balloon take off to its highest altitude. For 50 lbs. of payload mass the balloon needed 1700 cu. ft. of helium to go 2 ft. and 1875 cu. ft. of helium to go its highest altitude of 22326 ft. And for 110 lbs. of payload it took 2400 cu. ft to go 2 ft. and 2616 cu. ft. of helium to go to its highest altitude of 7918 ft.

NOTE: Responses are the unedited, verbatim answers given by students.
SOURCE: U.S Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

To earn a score of 3, “good,” responses had to demonstrate, through the use of description or data, an understanding of either the top or the bottom of the step function for multiple masses. The first sample response in figure 4-16 received credit for its description of the top, the second response for its description of the bottom of the function.

Figure 4-16. Two responses to Simulation motivating problem 3 receiving a score of 3, “good,” grade 8: 2003

- Together the helium and payload mass make up the whole experiment. The more helium, the higher the balloon flies. The higher the weight, the lower it will go. Once the weight reaches its maximum height, no mount of helium can make it go higher. With ten pounds of payload mass, the maximum altitude it could reach was 36211 feet. When I added more helium, it still stayed at 36211 feet altitude. With the 110 pound payload mass, the maximum altitude it could reach was 7918 feet. Once again, adding more helium could not change the maximum altitude for the balloon. My conclusion is that every payload mass has a maximum altitude no matter what amount of helium they are attached to.
- The amount of helium and payload mass both affect the altitude of the balloon. The more the payload the more amount of helium it is going to take to raise the balloon. The less the helium and the more the payload the balloon will not take off.

NOTE: Responses are the unedited, verbatim answers given by students.
SOURCE: U.S Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

As seen in the scoring guide, a score of 2, “partial,” was awarded to responses that fell into either of two different categories: answers that gave a correct and relevant description of the balloon’s behavior except for a single variable (i.e., a description that could have come from the experiments in Simulation problems 1 or 2), or answers that addressed two variables but were very general or only partially correct. An example of the first type is seen in the first response in figure 4-17, which somewhat vaguely describes the bottom and top of the step function for a single mass. The second response considers two variables but suggests a linear relationship between them.

Figure 4-17. Two responses to Simulation motivating problem 3 receiving a score of 2, “partial,” grade 8: 2003

- If the payload is the same and there is enough helium to lift the balloon then it will always be the same altitude.
- if you have a low helium amount and a high mass u will not be able to get it up off the ground but if u have a high helium amount and a low mass u will go very high up because the helim won’t need to pull anything very heavy up with it so it can go up very high

NOTE: Responses are the unedited, verbatim answers given by students.
SOURCE: U.S Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

“Unacceptable” responses were those giving incorrect summaries of balloon behavior or those giving no summaries and only one or two data points, as in the two examples in figure 4-18.

Figure 4-18. Two responses to Simulation motivating problem 3 receiving a score of 1, “unacceptable,” grade 8: 2003

- I saw that the lower the pounds and the amount of helium the higher it went up.
- when the mass was 110 and the helium was 700, the balloon didn’t go anywhere. when the mass was 50 and the helium was 1400, the balloon went really high

NOTE: Responses are the unedited, verbatim answers given by students.
SOURCE: U.S Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

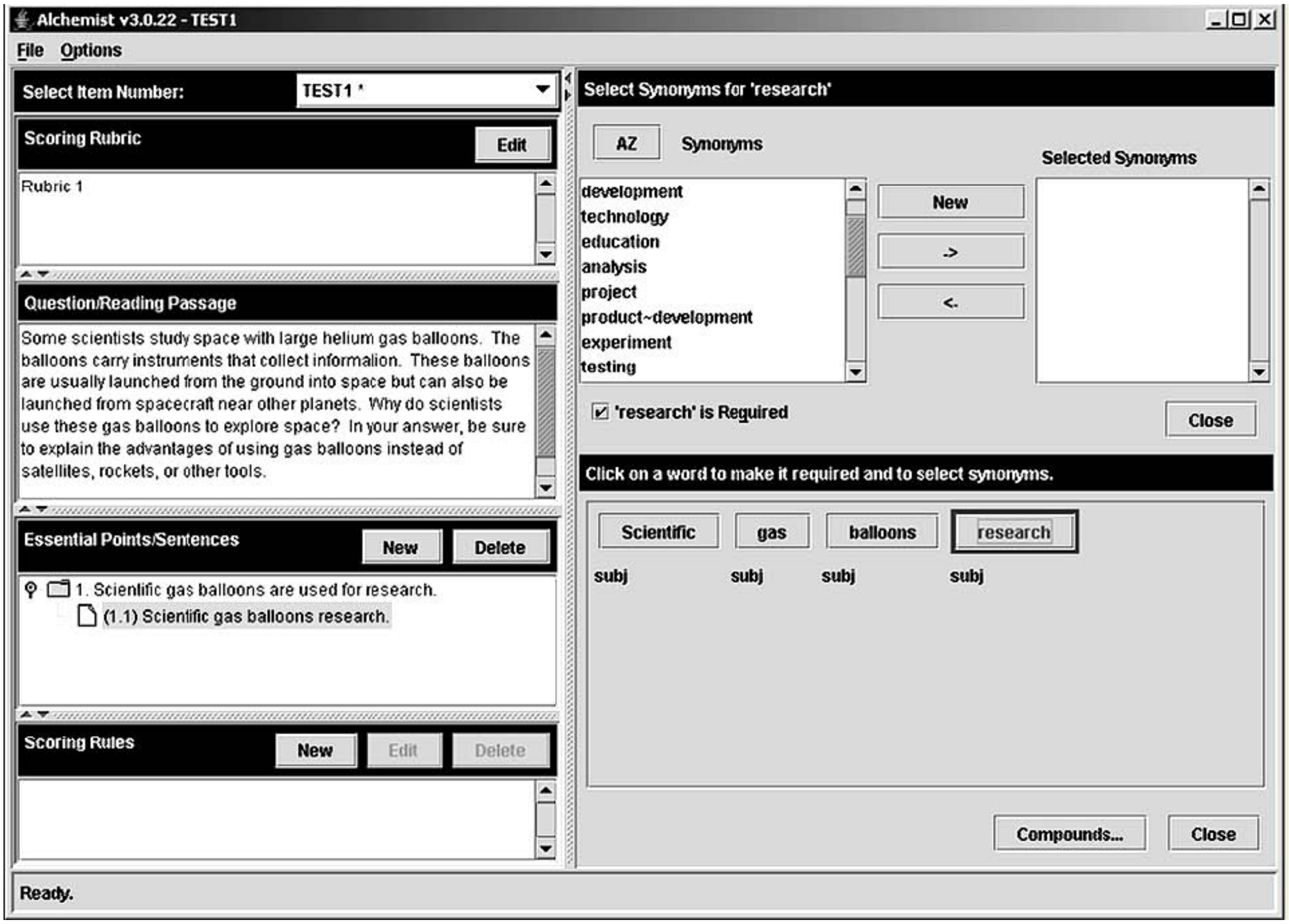
C-rater and TRE Scoring

One measure of students' exploration skill for the Search scenario was the degree to which students used terms relevant to the motivating problem in their search queries.

C-rater, a computer program developed by ETS for scoring short-answer responses, was used to score the individual search queries. To make the c-rater program usable for TRE, c-rater models—abstract descriptions of possible student queries—were manually developed. These models were developed by a computer programmer working in consultation with a NAEP assessment developer. The models implemented an evaluation rule which was established by creating queries that logical analysis, query tryout, or pilot results suggested were associated with more or less proficient searching. Proficient searching tended to employ more specific terms (e.g., scientific gas balloon), including ones taken directly from the motivating problem, whereas less proficient searching frequently relied on generic terms (e.g. balloon). The evaluation rule used seven classes of query terms and a three-point scale of "full," "partial," or "no credit" (see appendix G). The rule involved the following two steps: first, rate each search query for relevance on this three-point scale, 0-2; second, calculate the average rating for all of a student's search queries and assign a value of "high" for results above 1.4, "medium" for 0.7–1.4, and "low" for below 0.7.

C-rater models were built by entering phrases or sentences into a user interface, shown in figure 4-19. For TRE, the model developer entered a phrase, "scientific gas balloons research," query shorthand for the idea that "scientific gas balloons are used for research." Once the phrase was processed, the developer selected the term "research" as a required concept. Next, a set of

Figure 4-19. Entering concepts into a c-rater model, grade 8: 2003



Some scientists study space with large helium gas balloons. The balloons carry instruments that collect information. These balloons are usually launched from the ground into space but can also be launched from spacecraft near other planets. Why do scientists use these gas balloons to explore space? In your answer, be sure to explain the advantages of using gas balloons instead of satellites, rockets, or other tools.

1. Scientific gas balloons are used for research.

(1.1) Scientific gas balloons research.

☑ 'research' is Required

Click on a word to make it required and to select synonyms.

SOURCE: c-rater © 2003 by Educational Testing Service. All rights reserved.

words similar to “research” was presented in a scrollable window, from which the developer would then have selected acceptable synonyms (e.g., “analysis,” “study,” “exploration,” “experiment”). Additional synonyms could be entered manually.

Once all of the required concepts were entered, the developer entered the scoring rules, which indicate what scores to assign to different combinations of terms from the phrase when those terms are encountered in a student’s query.

C-rater matches phrases in the response to its rules. The program always produces the same scores for a given student response, unless its scoring rules are changed.

In processing student queries, c-rater can recognize and accept some misspelled words. For example, the system recognized the strings “baloon” and “ballon” as being “balloon.” In addition, c-rater recognizes morphological variants of words—it recognizes that “exploring” and “explored” are forms of “explore.” The test developer can also enter noun compounds, such as “space shuttle,” so that c-rater will recognize the compound “space shuttle” but not “shuttle space.”

The c-rater models constructed for scoring students’ search queries were cross-validated based on a sample of 256 queries that were independently hand scored. The agreement between c-rater and human scores for this cross-validation set was 96 percent. The 4 percent of scores that were discrepant involved students typing “outer space” as a single word and misspellings that c-rater failed to recognize. C-rater’s scoring models were adapted to account for the incorrect spelling of “outerspace” before conducting the final scoring of all student responses.

Chapter 5: The TRE Search Scenario Scales and Results

The TRE student model presented earlier proposed five proficiency scales: a TRE Search total score scale, a computer skills scale, a scientific exploration scale, a scientific synthesis scale, and a scientific inquiry scale. The scientific exploration and scientific synthesis scales were proposed as components of the scientific inquiry scale. Preliminary analysis of the TRE Search data, however, suggested that a separate scientific synthesis scale could not be empirically supported because of the number of items, or observables, associated with that scale. As a result, the scientific exploration and scientific synthesis scales were combined, resulting in three scales: a TRE Search total score scale, a scientific inquiry scale, and a computer skills scale. In addition, two observables, the degree of use of Help and the degree of use of Tips for Searching, were dropped from the analysis because they contributed little or nothing to the measurement of student performance. One observable, number of searches for relevant hits, which was originally assigned to two TRE scales, was instead assigned only to the scientific inquiry scale to simplify the analysis. Finally, one observable that had been scored on a three-point scale (use of deletion for unwanted filed pages) was recoded to dichotomous scoring.

Scores on the TRE Search total scale were estimated using a Bayesian model that combines prior information about students with student performance on the assessment instrument. Prior information about students was based on data collected on 10 variables: (1) gender, (2) race/ethnicity, (3) disability status, (4) identification as English language learner, (5) parents' highest education level, (6) number of types of reading-related items in the home, (7) eligibility for free or reduced-price lunch, (8) participation in Title I, (9) level of prior computer knowledge, and (10) whether the TRE scenario was taken on a NAEP laptop computer. Defining such priors removes bias from the estimation of TRE means for student groups (Mislevy 1991).

In keeping with the methodology employed in standard NAEP analyses (Allen, Donoghue, and Schoeps 2001), this modeling approach produces population estimates (e.g., means and standard deviations) without generating scores for individual students. Instead, population estimates are obtained by drawing five imputations, or *plausible values*, as commonly used in NAEP, for each student from the posterior distribution of proficiency, given that student's performance on the assessment instrument and the prior information described above. All means and correlations reported in this chapter employ these five imputations, except where noted. A similar process was used to determine the scale score estimates for computer skills and scientific inquiry. For convenience, all three scores were put on an arbitrary scale with a mean of 150 and a standard deviation of 35.[10] This chapter reports empirical results relating to the meaning of TRE Search scores and to student performance.

The Meaning of the TRE Search Scores

Because the TRE study used measures that are experimental, this chapter explores evidence for how well the TRE Search scenario scales captured the skills they were intended to summarize. The following sections are presented: internal consistency; the relations of student scores to students' prior knowledge; the TRE scale intercorrelations; the correlations of each observable with each of the two scales (scientific inquiry and computer skills); 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.

[10] This scale is intentionally different from the ones typically used in NAEP assessments to prevent confusion with those scales.

Internal Consistency

Internal consistency indicates the degree to which student responses to individual items (or "observables") in a scale are correlated, on average, with their responses to other items (or "observables") in the same scale. Higher values for internal consistency suggest greater similarity across items in the underlying skill being measured. For TRE, coefficient alpha, a conventional measure of internal consistency ranging from 0.00 to 1.00, was used. For the TRE Search total score, which consisted of 11 observables, the value of this statistic was .74 (data not shown). For the TRE scientific inquiry score, which had 5 observables, the comparable value was .65 (data not shown). Finally, for the TRE computer skills score, consisting of 6 observables, the value was .73 (data not shown). The values for the TRE Search total score and for the computer skills score were higher than those for the typical NAEP hands-on science block, which, although measuring skills different from the TRE Search scenario, also includes extended, problem-solving tasks. The typical NAEP hands-on science block involves a 30-minute exercise (in contrast to the approximately 40 minutes allocated to TRE Search).[11,12] For the 2000 science assessment, the mean weighted internal consistency taken across three such blocks was .62.

Correlations of TRE Search Scores With Prior Knowledge Measures

The prior knowledge measures were intended to give a rough indication of the degree of student familiarity with the science and computer-related concepts being assessed in the TRE Search scenario. The prior computer knowledge measure (which was common to all students regardless of scenario) consisted of 10 multiple-choice questions about Internet searching, word processing, spreadsheet use, and more general computer knowledge. The prior science knowledge measure (which was particular to students taking the Search scenario) comprised 10 multiple-choice questions on concepts related to the science and uses of helium gas balloons. (See appendix D for the questions included on each measure.)

Table 5-1. Weighted (disattenuated) correlations of TRE Search scores with prior knowledge measures, grade 8: 2003

TRE Search score	Prior computer knowledge measure	Prior science knowledge measure
Total	.61	.40
Computer skills	.52	.33
Scientific inquiry	.55	.39

NOTE: TRE = Technology-Rich Environments. N (number of students) = 1075. All correlations are significantly different from zero at $p < .05$. Students' scores for a particular prior knowledge measure were deleted from this analysis if they were missing seven or more questions in the scale.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Table 5-1 gives the (disattenuated) correlations of the TRE Search scores with the two prior knowledge measures—computer knowledge and science knowledge. These correlations should be considered as only suggestive because the prior knowledge measures did not consist of a sufficient number of items to be reliable or comprehensive in their coverage.[13] All of the correlations were significantly different from zero statistically. Thus, students with more prior computer knowledge and more prior science knowledge tended to perform better on each TRE Search score than did students with lower levels of prior knowledge.

[11] A NAEP hands-on block is a section of experimental tasks and constructed-response test items administered to a student.

[12] The TRE observables may not be completely independent, so the internal consistency estimates for the TRE scales may be inflated.

[13] Appendix I gives summary statistics for these measures.

Intercorrelations of the Scales

Table 5-2 gives the (disattenuated) TRE scale intercorrelations for the total sample and for gender and racial/ethnic student groups. As the table shows, in the overall sample, computer skills and scientific inquiry skill correlate about equally with the TRE Search total score (of which both computer skills and scientific inquiry skill are a part). In addition, the two scales correlate .57 with one another (as compared with values of .90 to .93 for the intercorrelations of the 1996 main NAEP eighth-grade science assessment scales [Allen, Carlson, and Zelenak 1999]).

Correlations of the Observables With the TRE Scales

Examining the correlations of the observables with each scale can also help clarify the meaning of the TRE scales. First, these correlations can suggest the degree to which the data bear out the theoretical prediction implied by assigning an observable to a particular scale. Second, the correlations indicate roughly how important each observable is to producing the score for the scale to which it is assigned.

Table 5-2. Number of students and weighted (disattenuated) intercorrelations of the TRE Search scales, by student characteristics, grade 8: 2003

Characteristic	Number of students	Computer skills with TRE Search total	Scientific inquiry with TRE Search total	Scientific inquiry with computer skills
Total	1,077	.68	.68	.57
Gender				
Male	517	.69	.68	.57
Female	560	.67	.68	.56
Race/ethnicity				
White	643	.60	.60	.46
Black	185	.69	.64	.59
Hispanic	188	.64	.60	.53

NOTE: TRE = Technology-Rich Environments. All correlations are significantly different from zero at $p < .05$. Results are shown for three mutually exclusive race/ethnicity categories. Black includes African American, and Hispanic includes Latino. Race categories exclude Hispanic origin unless specified.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Table 5-3 gives the (disattenuated) correlations of each observable with the two TRE subscales. (Correlations with the TRE Search total score scale are not shown because this scale was measured by the two subscales and not directly by the observables.) In general, each observable was intended to measure performance on one scale (that is, to measure either computer skills or scientific inquiry skill). The pattern of correlations bears out the hypotheses about which observables demonstrate which skill. That is, visual inspection suggests that the observables selected to measure computer skills and scientific inquiry correlate more highly in this student sample with the subscale to which they were assigned than they do with the other subscale.[14]

The correlations in table 5-3 also indicate the contribution of particular observables to a given scale score. It is clear from the table that, in this student sample, the scientific inquiry skill score 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 other words, students who received higher levels of credit for their performance on one or more of these observables were also likely to receive higher scientific inquiry scores.

Table 5-3. Weighted (disattenuated) correlations between score on each TRE observable and the TRE Search scales, grade 8: 2003

Observable	Computer skills	Scientific inquiry
Relevance of pages visited or bookmarked[1]	.17	**.71**
Accuracy/completeness on constructed-response question	.39	**.70**
Degree of use of relevant search terms	.33	**.51**
Number right on final multiple-choice questions	.28	**.44**
Average relevance of hits to motivating problem	.20	**.34**
Use of hyperlinks to dig down	**.69**	.37
Consistency of use of Back button	**.65**	.36
Number of searches for relevant hits[2]	**.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

[1] This observable combined the following three observables: average relevance of pages bookmarked, percentage of pages visited that are relevant, proportion of relevant to total pages bookmarked.

[2] The values for this observable were reversed (i.e., fewer searches received a higher score) to allow the correlation with scale score to be positive.

NOTE: TRE = Technology-Rich Environments. The **bold** values indicate that the scale named in the column label was the one to which an observable was assigned. All correlations are significantly different from zero at $p < .05$.

N (number of students) range = 672 to 1077. All scale scores include the observable being correlated.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

[14] Two observables were dropped from the analysis: “degree of use of Help” and “degree of use of Tips for Searching,” which related to the subscales either marginally or not at all. Also, one observable, “number of searches for relevant hits,” which was originally assigned to two TRE scales, was instead assigned only to the scientific inquiry scale to simplify the analysis.

Similarly, table 5-3 indicates that 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). Students who frequently used hyperlinks, the Back button, and bookmarking, and who found relevant information with fewer searches, were likely to receive higher computer skills scale scores. Thus, as modeled, the two scales do appear to differentiate themselves on the basis of the substantive aspects (i.e., content relevance and quality of response) versus the more technical aspects of electronic information search.

While the correlational pattern suggests a differentiation between the two scales, the data also suggest that specific computer-related behaviors were associated with higher levels of scientific problem solving with technology. Students who bookmarked, dug down with hyperlinks, employed the Back button, required fewer searches to get relevant hits, and used advanced search techniques also tended to get higher scientific inquiry scores. Further, as shown in table 5-4, students who evidenced these computer-related behaviors tended to provide better answers to the constructed-response question.

Table 5-4. Observed correlation between score on each observable and raw score on the constructed-response Search question, grade 8: 2003

Observable	Search question
Relevance of pages visited or bookmarked[1]	.55*
Use of bookmarking to save pages	.35*
Degree of use of relevant search terms	.32*
Number right on final multiple-choice questions	.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[2]	.20*
Consistency of use of Back button	.19*
Use of deletion for unwanted filed pages	.03

*Correlations are significantly different from zero at $p < .05$.

[1] This observable combined the following three observables: average relevance of pages bookmarked, percentage of pages visited that are relevant, proportion of relevant to total pages bookmarked.

[2] The values for this observable were reversed (i.e., fewer searches received a higher score) to allow the correlation with scale score to be positive.

NOTE: TRE = Technology-Rich Environments. Values are raw correlations and are not based on averages across imputations. The constructed-response Search question was scored on a 1–3 scale.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Locations of the Observables on the TRE Scales

Item maps are displays that give a context for interpreting score points on a given scale. They display the locations of observables on their respective scales by associating points on the scale with levels of correctness for particular observables, and thus describe what groups of students who attain a particular scale score on average are likely to be able to do. These maps should be interpreted carefully, however. The mapping of an observable to a point on the proficiency scale is based on an item response model and on estimated item parameters, so where an item is placed depends on the correctness of the underlying assumptions of the model and on how accurately the item parameters are estimated.[15] Also, item locations depend on the choice of a probability for correctly responding. For purposes of the TRE study, this probability was set at 65 percent, the level routinely used in NAEP assessments for the mapping of constructed-response items. With these caveats in mind, item maps can be a useful way of explicating proficiency scales.

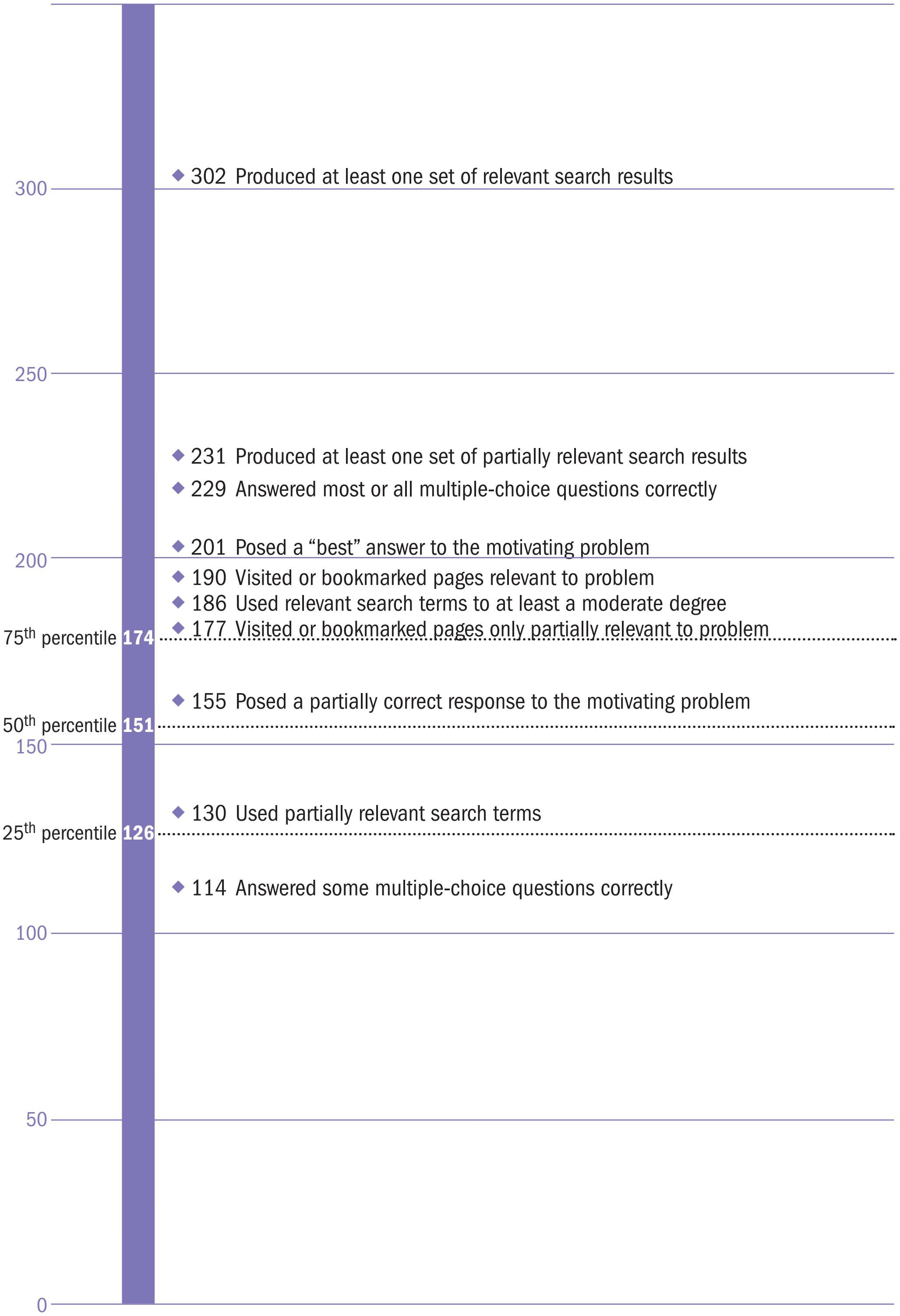
Figure 5-1 shows an item map for the scientific inquiry scale. For mapping purposes, each observable has been transformed into one or more dichotomous variables, where the number of such variables is one less than the number of levels of correctness for the observable. Thus, each location on the map represents the point on the scale at which at least 65 percent of students were likely to have achieved the indicated level of correctness for a particular observable. For example, posing a partially correct response to the motivating problem maps to a scale score of 155. This mapping means that students who received a score of 155 or more on the scientific inquiry skill scale had at least a 65 percent chance of submitting an answer achieving a score of 2 on a 1–3 scale. Full credit for responding to the motivating problem maps to a score of 201. Students with a score of 201 would have at least a 65 percent chance of submitting an answer achieving a top score of 3.

By mapping observables to the scale in this way, the scale can be described qualitatively. From the lowest mapped scale point, the ordering is as follows:

- correctly answering some (either one or two) of the four multiple-choice items that require web searching;
- using search terms that, on average, match those of proficient searchers only to a limited degree;
- constructing a response that only partially answers the motivating problem (i.e., giving only one or two advantages of using gas balloons);
- bookmarking or visiting pages that, on average, are only partially relevant to the problem posed;
- using search terms that, on average, match those of proficient searchers to at least a moderate degree;
- bookmarking or visiting pages that, on average, are relevant to the problem posed;
- constructing a "best" response that gives a complete answer to the motivating problem (i.e., gives three or more advantages of using gas balloons);
- correctly answering at least three of the four multiple-choice items that require web searching;
- producing at least one set of search results with hits that, on average, are only partially relevant to the problem posed (i.e., have relevance scores averaging between 2 and 3 on a 4-point scale, where a score of 4 denotes the most relevant hits); and
- producing at least one set of search results with hits that, on average, are relevant to the problem posed (i.e., have relevance scores averaging between 3 and 4 on a 4-point scale, where a score of 4 denotes the most relevant hit).

[15] Item mapping was done with item parameters from a scaling employing the operational, univariate NAEP IRT model as implemented by the PARSCALE program. This approach was used because no similar procedure was available within the Bayesian modeling framework. Since the two approaches do not generate equivalent item parameters, the PARSCALE item parameters were transformed so that they would estimate a proficiency with similar mean and variance as the item parameters from the Bayesian analysis.

Figure 5-1. Mapping of TRE Search observables to the scientific inquiry scale, grade 8: 2003



NOTE: TRE = Technology-Rich Environments. Each position on the map indicates the scale score at which students had a 65 percent probability of successfully attaining a given level of correctness for a particular observable. The estimated score mapping for "Produced at least one set of relevant search results" was above the scale maximum of 300 and is included in the figure for completeness.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure 5-2 is an item map for the computer skills scale. From lowest mapped scale point to the highest, the ordering is as follows:

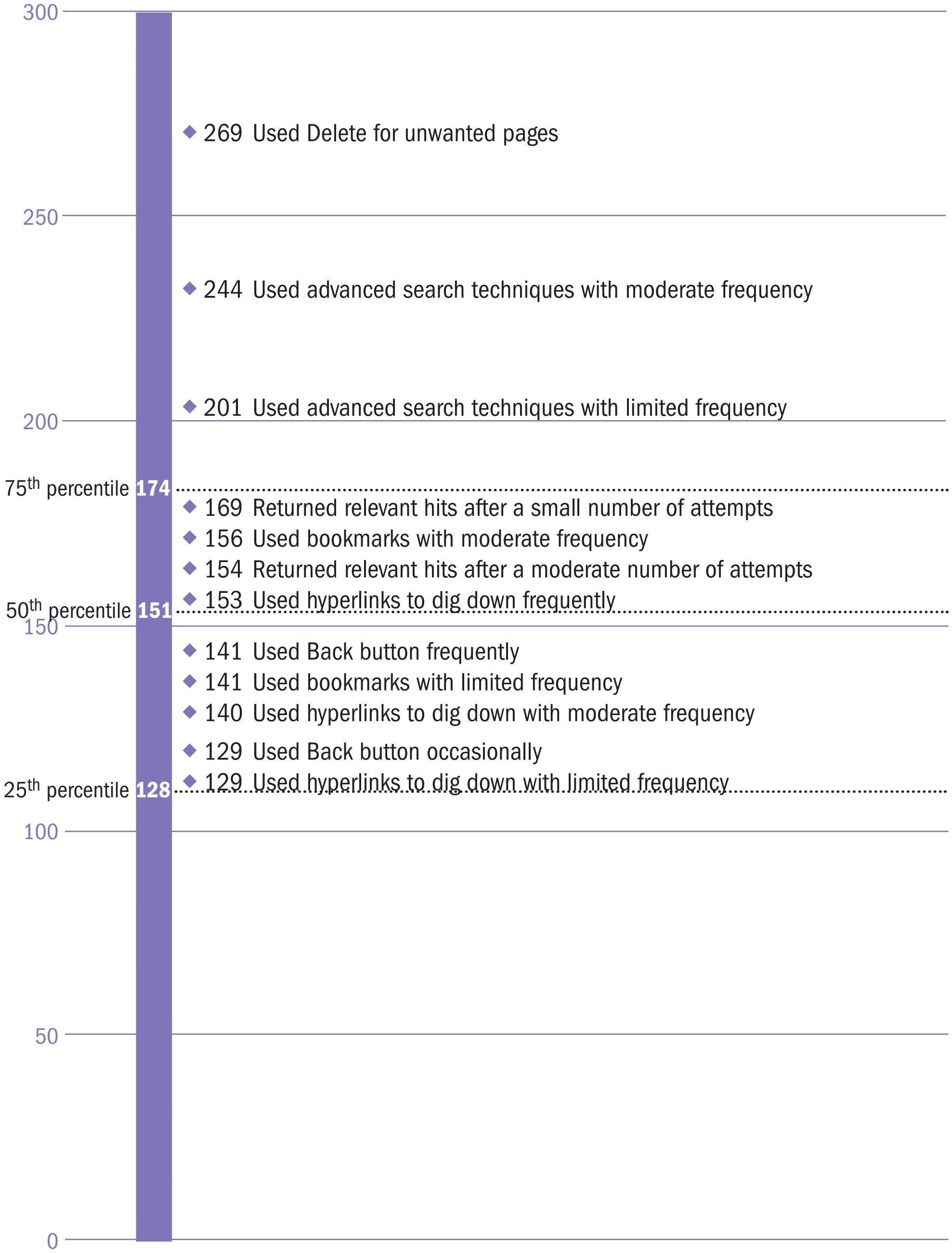
- using the Back button occasionally (3–4 times) to navigate among web pages or from web pages to the search page;
- using hyperlinks with limited frequency (1–2 times) to explore web pages linked to the page currently being viewed;
- using hyperlinks with moderate frequency (3–4 times) to explore web pages linked to the page currently being viewed;
- using the Back button frequently (at least 5 times) to navigate among web pages or from web pages to the search page;
- using bookmarks with limited frequency (1 time);
- using hyperlinks frequently (at least 5 times) to explore web pages linked to the page currently being viewed;
- returning relevant results after a moderate number of attempts (4–6);
- using bookmarks with at least moderate frequency (2 or more times);
- returning relevant results after only a small number of attempts (1–3);
- using advanced search techniques with limited frequency (1–2 searches);
- using advanced search techniques with at least moderate frequency (3 or more searches); and
- using Delete to remove a page that had been bookmarked.

Appendix J gives the percentages of students achieving each of the observable behaviors.

Response Probabilities for Prototypic Students

Examining the response probabilities for prototypic students (i.e., hypothetical students with high, medium, or low levels of proficiency) also affords a way to gain insight into the meaning of the TRE scales. The required probabilities can be generated empirically from the item response model for students with different prototypic levels of standing on the TRE proficiencies (e.g., students who are known to be at a high level of scientific inquiry as compared with those who are known to be at a medium or low level). The probability of achieving each observable can then be examined to see how prototypic students differ and if those differences are logically meaningful.

Figure 5-2. Mapping of TRE Search observables to the computer skills scale, grade 8: 2003



NOTE: TRE = Technology-Rich Environments. Two items, degree of use of Help and degree of use of Tips for Searching, are not included on the item map because they discriminated very little between high- and low-performing students, and therefore were not reliable measures of the scale. Each position on the map indicates the scale score at which students had a 65 percent probability of successfully attaining a given level of correctness for a particular observable.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Tables 5-5 and 5-6 show the response probabilities for prototypic students with different levels of scientific inquiry skill and computer skills, respectively. For these tables, the prototypic levels were defined by separately dividing in turn the scientific inquiry and computer skills score distributions into thirds and taking the midpoint in the bottom third as the prototypic low-level student, the midpoint in the center third as the prototypic middle-level student, and the midpoint in the top third as the prototypic high-level student. These values were then used to fix the proficiency level in the response model for generating the probability of achieving each of the levels of correctness on each of the observables.

The response probabilities are generally compared in the following way: First, the prototypic low-level student is described by identifying the level of correctness that student is likely to achieve on each observable. Next, the prototypic medium-level student is described in terms of only those observables that would distinguish this student from the prototypic low-level student (i.e., only those observables on which the two students would be likely to attain different degrees of correctness). Finally, the prototypic high-level student is differentiated from the prototypic medium-level student in a similar fashion.

As table 5-5 shows, the prototypic student at a low level of scientific inquiry skill was most likely to receive no credit for responses to the constructed-response question (motivating problem), the relevance of pages bookmarked, and the average relevance of hits returned from search results. This student was also most likely to receive partial credit for responses to the multiple-choice questions and for the degree of use of relevant search terms. Though the response probabilities differed, the pattern for the medium level of scientific inquiry was very similar. The main exception was that the student at this level was more likely to receive partial credit (rather than none) for answering the constructed-response question. Finally, in contrast to the low- and medium-level students, the student at a high level of scientific inquiry was most likely to get partial credit (rather than none) for bookmarking relevant pages and to get full credit (rather than partial credit) for the degree of use of relevant search terms.

Table 5-5. Probability of responding to observables on TRE Search for prototypic students, by level of scientific inquiry and level of correctness of observable response, grade 8: 2003

	Low level of scientific inquiry			Medium level of scientific inquiry			High level of scientific inquiry		
Observable	No credit[1]	Partial credit	Full credit	No credit[1]	Partial credit	Full credit	No credit[1]	Partial credit	Full credit
Accuracy/completeness on constructed-response question	**.88**	.11	.01	.44	**.50**	.05	.08	**.57**	.35
Relevance of pages visited or bookmarked[2]	**.99**	.01	.00	**.85**	.12	.03	.21	**.40**	.39
Number right on final multiple-choice questions	.30	**.64**	.05	.13	**.73**	.14	.05	**.63**	.32
Degree of use of relevant search terms	.37	**.52**	.12	.16	**.55**	.29	.06	.38	**.56**
Average relevance of hits to motivating problem	**.98**	.02	.00	**.92**	.07	.00	**.76**	.22	.01

[1] No credit, partial credit, and full credit are the levels of correctness of response specific to each observable.

[2] "Relevance of pages bookmarked" combines three observables: Average relevance of pages bookmarked, percentage of pages visited that are relevant, and proportion of relevant to total pages bookmarked.

NOTE: TRE = Technology-Rich Environments. Highest probability for each level is shown in **bold**. Detail may not sum to totals because of rounding.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

The response probabilities for computer skills, which were computed in a manner similar to that for scientific inquiry, are shown in table 5-6. As the table shows, for this scale one observable has two levels of correctness (no credit, full credit), some observables have three levels (no credit, partial credit, full credit), and one has four levels (no credit, low-partial credit, high-partial credit, full credit). The prototypic student with a low level of computer skills was likely to receive no credit for using hyperlinks, employing the Back button, getting relevant hits with few searches, bookmarking, using advanced search techniques, and deleting unwanted pages that had previously been bookmarked. The medium-level-of-computer-skills student diverged from this no-credit pattern by being likely to receive partial credit for getting relevant hits with few searches and full credit for using hyperlinks, employing the Back button, and bookmarking. Finally, the high-computer-skills student was likely to receive full credit for getting relevant hits with few searches. This hypothetical student also showed probability distributions for using hyperlinks, the Back button, and bookmarking that appeared generally more peaked at full credit than did the corresponding distributions for the medium-computer-skills student.

Table 5-6. Probability of responding to observables on TRE Search for prototypic students, by level of computer skills and level of correctness of observable response, grade 8: 2003

	Low level of computer skills				Medium level of computer skills				High level of computer skills			
Observable	No credit[1]	Low-partial credit	High-partial credit	Full credit	No credit[1]	Low-partial credit	High-partial credit	Full credit	No credit[1]	Low-partial credit	High-partial credit	Full credit
Use of hyperlinks to dig down	**.46**	.25	.17	.13	.09	.13	.23	**.55**	.01	.02	.06	**.91**

	Low level of computer skills			Medium level of computer skills			High level of computer skills		
Observable	No credit[1]	Partial credit	Full credit	No credit[1]	Partial credit	Full credit	No credit[1]	Partial credit	Full credit
Consistency of use of Back button	**.48**	.23	.29	.09	.12	**.80**	.01	.02	**.97**
Number of searches for relevant hits[2]	**.76**	.18	.06	.33	**.37**	.30	.07	.20	**.73**
Use of bookmarking to save pages	**.59**	.18	.23	.20	.17	**.62**	.04	.05	**.90**
Use of advanced search techniques	**.90**	.09	.01	**.72**	.23	.05	.43	.43	.14

	Low level of computer skills		Medium level of computer skills		High level of computer skills	
Observable	No credit[1]	Full credit	No credit[1]	Full credit	No credit[1]	Full credit
Use of deletion for unwanted filed pages	**.96**	.04	**.91**	.09	**.81**	.19

[1] No credit, partial credit (including low-partial and high-partial), and full credit are the levels of correctness of response specific to each observable.

[2] The values for this observable were such that fewer searches received higher levels of credit.

NOTE: TRE = Technology-Rich Environments. Highest probability is shown in **bold**. Detail may not sum to totals because of rounding.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

TRE Performance as a Function of Relevant Background Experience

TRE Search scores should be related in logically meaningful ways to students' reports of their background experiences. Figures 5-3 to 5-6 present data on the relationship of TRE Search score to responses to relevant computer-related background questions. (Supplementary data for these figures are available in appendix I.) In the figures, T stands for TRE Search total score, S stands for TRE Search scientific inquiry score, and C stands for TRE Search computer skills score. If the performance of students who gave the response named to the left of the row was significantly different statistically on one of the scales from that of students giving the response named at the top of a column, the cell where the row and column intersect is shaded. Which response was associated with higher TRE performance is indicated by whether the shading is light or dark. Dark shading indicates that students who gave the row response had a higher score on at least one of the three TRE measures than the students who gave the response named at the top of the column to the same question. For example, for the question "Find information on the Internet," those who indicated that they used the computer to find information on the Internet to a moderate extent had higher scores on all three scales than students who reported they used the computer in this way to a small extent. This result is indicated by the darker shading in the cell at the intersection of the moderate row and the small column, and by the letters in that cell, T, S, and C, which refer to the three TRE scores.

As a general observation, most of the statistically significant differences in performance by background question carried across all three TRE Search scales. That is, there was little evidence from the background questions that the TRE scales were functioning differently from one another. At the same time, there were differences that did seem relevant to understanding the meaning of the TRE Search scores overall. For example, as figure 5-3 shows, students who reported more frequent use of a word processor (background question 2 in appendix D) scored better on average on all three TRE scales than those who reported not using a word processor at all. Other statistically significant differences in scores associated with word processor use also appear, always in the expected direction of more use suggesting higher scores. One plausible explanation is that TRE Search requires some degree of word processing skill in order to compose an answer to the motivating problem. Another is that students who use word processors may tend to be more academically skilled in general.

TRE Search also requires students to gather relevant information from a simulated World Wide Web. Figure 5-3 indicates that students who reported using the computer to find information on the Internet (background question 6 in appendix D) to a moderate or large extent scored higher on average on all three TRE Search scales than students who reported using the Internet to a small extent for finding information.

Positive relations were also found between TRE Search performance and students' reports of the following uses of computers: e-mail (figure 5-3, background question 7 in appendix D), talking in chat groups (figure 5-3, background question 8 in appendix D), using a computer outside of school (figure 5-4, background question 11 in appendix D), and having a computer in the home that the student uses (figure 5-5, background question 12 in appendix D).

For some uses of the computer, however, more use was not associated with higher performance on the TRE Search scales. For example, students who reported using the computer to make drawings or create artwork on the computer to a large extent (figure 5-3, background question 3 in appendix D) scored 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.

Figure 5-3. Relationship between TRE Search performance and reported type of computer use, grade 8: 2003

Use a word processor

Response	Not at all	Small	Moderate	Large
Not at all	†	░ T, S, & C	░ T, S, & C	░ T, S, & C
Small	■ T, S, & C	†	░ T	░ T, S, & C
Moderate	■ T, S, & C	■ T	†	░ C
Large	■ T, S, & C	■ T, S, & C	■ C	†

Find information on the Internet

Response	Not at all	Small	Moderate	Large
Not at all	†			
Small		†	░ T, S, & C	░ T, S, & C
Moderate		■ T, S, & C	†	
Large		■ T, S, & C		†

Make drawings/art on computer

Response	Not at all	Small	Moderate	Large
Not at all	†			■ T, S, & C
Small		†		■ T, S, & C
Moderate			†	■ C
Large	░ T, S, & C	░ T, S, & C	░ C	†

Use e-mail

Response	Not at all	Small	Moderate	Large
Not at all	†		░ T, S, & C	░ T, S, & C
Small		†		░ T & C
Moderate	■ T, S, & C		†	
Large	■ T, S, & C	■ T & C		†

Make tables, charts or graphs on computer

Response	Not at all	Small	Moderate	Large
Not at all	†	░ T & C		
Small	■ T & C	†		■ T & C
Moderate			†	■ T & C
Large		░ T & C	░ T & C	†

Talk in chat groups

Response	Not at all	Small	Moderate	Large
Not at all	†			░ T, S, & C
Small		†		
Moderate			†	
Large	■ T, S, & C			†

Look up information on a CD

Response	Not at all	Small	Moderate	Large
Not at all	†			
Small		†		■ T
Moderate			†	■ T
Large		░ T	░ T	†

† Not applicable.
T = TRE Search total score.
S = TRE Search scientific inquiry score.
C = TRE Search computer skills score.
NOTE: TRE = Technology-Rich Environments. Column headings in table correspond to student questionnaire response categories as follows: Not at all = not at all; Small = small extent; Moderate = moderate extent; Large = large extent.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

■ Indicates that at least one of the three types of scores was significantly higher at the .05 level for students giving the response at the left of the row than for those giving the response at the top of the column.

□ Indicates that there was no significant difference in any of the three types of scores between students giving the response at the left of the row and those giving the response at the top of the column.

░ Indicates that at least one of the three types of scores was significantly lower at the .05 level for students giving the response at the left of the row than for those giving the response at the top of the column.

Figure 5-4. Relationship between TRE Search performance and reported frequency of computer use outside of school, grade 8: 2003

How often do you use a computer outside of school?

Response	Daily	2–3 times per week	Once a week	Once every few weeks	Never or hardly ever
Daily	†	■ T, S, & C	■ T, S, & C	■ T, S, & C	■ T, S, & C
2–3 times per week	▨ T, S, & C	†		■ T, S, & C	■ T, S, & C
Once a week	▨ T, S, & C		†	■ T, S, & C	■ T, S, & C
Once every few weeks	▨ T, S, & C	▨ T, S, & C	▨ T, S, & C	†	
Never or hardly ever	▨ T, S, & C	▨ T, S, & C	▨ T, S, & C		†

† Not applicable.
T = TRE Search total score.
S = TRE Search scientific inquiry score.
C = TRE Search computer skills score.
NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

■ Indicates that at least one of the three types of scores was significantly higher at the .05 level for students giving the response at the left of the row than for those giving the response at the top of the column.

□ Indicates that there was no significant difference in any of the three types of scores between students giving the response at the left of the row and those giving the response at the top of the column.

▨ Indicates that at least one of the three types of scores was significantly lower at the .05 level for students giving the response at the left of the row than for those giving the response at the top of the column.

Figure 5-5. Relationship between TRE Search performance and presence of a home computer that the student uses, grade 8: 2003

Is there a computer at home that you use?

Response	Yes	No
Yes	†	■ T, S, & C
No	▨ T, S, & C	†

† Not applicable.
T = TRE Search total score.
S = TRE Search scientific inquiry score.
C = TRE Search computer skills score.
NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

■ Indicates that at least one of the three types of scores was significantly higher at the .05 level for students giving the response at the left of the row than for those giving the response at the top of the column.

□ Indicates that there was no significant difference in any of the three types of scores between students giving the response at the left of the row and those giving the response at the top of the column.

▨ Indicates that at least one of the three types of scores was significantly lower at the .05 level for students giving the response at the left of the row than for those giving the response at the top of the column.

Figure 5-6. Relationship between TRE Search performance and reported use of the Internet for sharing information about science experiments, grade 8: 2003

Use the Internet to exchange information with other students or scientists about experiments

Response	Not taking science	Once a month or more	Less than once a month	Never
Not taking science	†			
Once a month or more		†		
Less than once a month			†	▨ S
Never			■ S	†

† Not applicable.
T = TRE Search total score.
S = TRE Search scientific inquiry score.
C = TRE Search computer skills score.
NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

■ Indicates that at least one of the three types of scores was significantly higher at the .05 level for students giving the response at the left of the row than for those giving the response at the top of the column.

□ Indicates that there was no significant difference in any of the three types of scores between students giving the response at the left of the row and those giving the response at the top of the column.

▨ Indicates that at least one of the three types of scores was significantly lower at the .05 level for students giving the response at the left of the row than for those giving the response at the top of the column.

Other exceptions to the general result that more computer use was associated with higher scores on the TRE Search scales are to be found in figure 5-3, background question 4, relating to using the computer to make tables, charts, or graphs; figure 5-3, background question 5, asking about using the computer to look up information on a compact disk; and figure 5-6, background question 33, which asked how often students used the Internet to exchange information with other students or scientists about experiments.

There were no statistically significant differences on the TRE scales between students who reported different levels of computer use at school, between those who reported different frequencies of downloading scientific data from the Internet, and between those who reported different frequencies of using a computer to analyze data (not shown).[16]

Finally, information was also collected about students' activities in science class, for example, the frequency of carrying out science experiments. In almost every case, the numbers of students in the various response intervals for each background question were too small for significance tests to be performed, or data based on those questions bore no statistically significant relationship to TRE Search performance (data not shown).

Performance by Student Groups

How did students perform on average? For the full sample, the mean on the TRE Search total score scale is set to an arbitrary value, that is, to a number chosen for convenience to denote the average score for the sample. However, scores can be examined for NAEP reporting groups defined by gender, race/ethnicity, parents' highest education level, students' eligibility for free or reduced-price school lunch, and school location. (See table 5-7 for performance results for student groups.) Statistically significant differences in performance were found on one or more TRE scales for all student groups except by gender. (See appendix H for graphical representations of statistically significant differences.) Notably, there was no evidence that female students were different from male students in their performance on either the scientific inquiry or computer skills components of the Search scenario.

Performance by Racial/Ethnic Group

NAEP uses school-reported data about students' race/ethnicity. For the TRE scientific inquiry scale, the performance of White students (mean scale score = 160) was significantly higher statistically than that of Black students (t, 41 = 10.59, $p < .05$), who attained a mean scale score of 125, as well as that of Hispanic students (t, 4 = 4.42, $p < .05$), who attained a mean scale score of 137.

For computer skills, too, the average performance of White students (mean scale score = 158) was significantly higher statistically than that of Hispanic students (t, 10 = 4.19, $p < .05$), who attained a mean scale score of 142, as well as that of Black students (t, 27 = 7.92, $p < .05$), who attained a mean scale score of 128. Also, the mean score for Hispanic students was higher than the mean for Black students (t, 18 = –2.87, $p < .05$).

Performance by Parents' Highest Education Level

Statistically significant performance differences were also apparent among students who reported different levels of parental education. Students who reported that a parent had graduated from college (mean scale score = 157) scored significantly higher statistically on the TRE Search total score than those students who reported that their parents did not finish high school (mean scale score = 133) (t, 45 = –5.45, $p < .05$), and also higher than those who reported that a parent had graduated from high school (mean scale score = 142) (t, 47 = –3.00, $p < .05$). Students who reported that a parent had some education after high school (mean scale score = 155) had higher mean scores than students reporting that their parents had not graduated from high school (mean scale score = 133) (t, 54 = –4.66, $p < .05$), as well as higher scores than those reporting that a parent had graduated from high school (mean scale score = 142) (t, 56 = –2.48 $p < .05$).

The scientific inquiry score of students reporting that a parent had graduated from college (mean scale score = 156) was significantly higher statistically than the score of students reporting that their parents had not finished high school (mean scale score = 135) (t, 39 = –4.22, $p < .05$), and also higher than those who reported that a parent had graduated from high school (mean scale score = 143) (t, 58 = –3.47, $p < .05$). Also, students who had a parent with some education after high school (mean scale score = 154) had statistically significantly higher scientific inquiry scores than students reporting that a parent had graduated from high school (mean scale score = 143) (t, 61 = –2.70, $p < .05$), and higher scores than students reporting that their parents had not finished high school (mean scale score = 135) (t, 43= –3.63, $p < .05$).

[16] The analyses presented in figures 5-3 to 5-6 did not control for other student background variables, such as socioeconomic status (SES). It is possible that holding such variables constant would produce a different pattern of relations between reported computer use and TRE scores from that described above.

There were also several statistically significant differences among score distributions for computer skills. Students reporting that a parent had graduated from college (mean scale score = 155) scored significantly higher statistically than students reporting that a parent had graduated from high school (mean scale score = 145) (*t*, 44 = –2.70, $p < .05$). Students with a parent who had some education after high school (mean scale score = 154) also received computer skills scores that were significantly higher statistically than those with a parent who had graduated from high school (mean scale score = 145) (*t*, 46 = –2.38, $p < .05$). Students reporting that their parents did not finish high school (mean scale score = 139) scored significantly lower statistically than those reporting that a parent had graduated from college (mean scale score = 155) (*t*, 31 = –3.11, $p < .05$), as well as lower than those reporting that a parent had some education after high school (mean scale score = 154) (*t*, 32 = –2.87, $p < .05$).

Performance by Students' Eligibility for Free or Reduced-Price School Lunch

Several statistically significant differences among score distributions were also found among students eligible and not eligible for free or reduced-price lunch, as reported by schools. Students not eligible for free or reduced-price school lunch (mean scale score = 160) received statistically significantly higher mean TRE Search total scores than students eligible for reduced-price lunch (mean scale score = 145) (*t*, 31 = 3.15, $p < .05$) and higher means than students eligible for free lunch (mean scale score = 129) (*t*, 45 = 10.33, $p < .05$). Those eligible for reduced-price lunch, in turn, received higher scores than students eligible for free lunch (*t*, 39 = 3.32, $p < .05$).

Further, students not eligible for free or reduced-price lunch received statistically significantly higher mean scientific inquiry scale scores (mean = 158) than students eligible for free lunch (mean = 131) (*t*, 40 = 8.41, $p < .05$) and those eligible for reduced-price lunch (mean = 148) (*t*, 22 = 2.59, $p < .05$). Also, students eligible for reduced-price lunch (mean = 148) performed significantly higher statistically on scientific inquiry than those eligible for free lunch (mean = 131) (*t*, 28 = 3.70, $p < .05$).

Finally, students not eligible for free or reduced-price lunch (mean scale score = 158) performed significantly higher statistically on the computer skills scale than both students eligible for free lunch (mean scale score = 133) (*t*, 37 = 7.99, $p < .05$) and students eligible for reduced-price lunch (mean scale score = 147) (*t*, 16 = 2.39, $p < .05$). Students eligible for reduced-price lunch, whose mean scale score was 147, also scored significantly higher statistically on computer skills than students eligible for free lunch, whose mean was 133 (*t*, 20 = 2.61, $p < .05$).

Performance by School Location

Students differed in their performance as a function of school location only for the TRE Search total score. On this scale, students attending central city schools (mean = 142) scored lower than students attending urban fringe/large town schools (mean = 152; *t*, 22 = –2.60, $p < .05$) and students attending rural schools (mean = 153; *t*, 26 = –2.59, $p < .05$).

Table 5-7. Mean TRE Search scores, by student characteristics, grade 8: 2003

Characteristic	Number of students	TRE Search total score	Scientific inquiry score	Computer skills score
Total	1,077	150 (2.0)	150 (2.1)	150 (1.8)
Gender				
Male	517	148 (2.4)	149 (2.7)	147 (2.5)
Female	560	151 (2.3)	150 (2.3)	152 (1.9)
Race/ethnicity				
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)
Student-reported parents' highest education level				
Did not finish high school	72	133 (3.7)	135 (4.3)	139 (4.5)
Graduated from high school	214	142 (4.4)	143 (2.9)	145 (3.1)
Some education after high school	202	155 (3.0)	154 (2.7)	154 (2.6)
Graduated from college	497	157 (2.4)	156 (2.4)	155 (2.4)
Eligibility for school lunch				
Not eligible	656	160 (1.6)	158 (2.0)	158 (1.8)
Reduced-price lunch	70	145 (4.3)	148 (3.7)	147 (4.4)
Free lunch	300	129 (2.5)	131 (2.6)	133 (2.5)
School location				
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)

NOTE: TRE = Technology-Rich Environments. Standard errors of estimate appear in parentheses. Some seemingly large differences between the performance of student groups were not statistically significant because of the large standard errors associated with those differences. Results are shown for three mutually exclusive race/ethnicity categories. Black includes African American, and Hispanic includes Latino. Race categories exclude Hispanic origin unless specified. Eligibility for free or reduced-price lunch was based on school-reported information. For details about eligibility requirements, see Eligibility for Free/Reduced-Price School Lunch in appendix K. Results are not shown for students whose eligibility status for free or reduced-price lunch was not available.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Chapter 2 of this report explained that the initial TRE student model proposed five proficiencies: (1) a total TRE scale, (2) a computer skills scale, (3) a scientific inquiry scale, (4) a scientific exploration scale, and (5) a scientific synthesis scale; the last two scales were to be components of the scientific inquiry scale. As was the case with the Search scenario data, preliminary analysis of the TRE Simulation data did not support all the proposed proficiencies; the scientific synthesis scale and the scientific exploration scale could not be effectively combined to form a scientific inquiry scale for this scenario. As a result, a separate scientific inquiry score was not estimated, leaving four scales: a total TRE Simulation scale, a computer skills scale, a scientific exploration scale, and a scientific synthesis scale.

In addition to changes in the number of scales, several Simulation scenario observables were dropped from the analysis because they contributed little or nothing to the measurement of student performance, often because they were redundant with the information provided by another observable. Table 6-1 lists the observables dropped. (See chapter 2 for preliminary versions of the evidence models.)

Table 6-1. Observables dropped from the TRE Simulation scenario analysis, grade 8: 2003

Observable	Simulation problem 1	Simulation problem 2	Simulation problem 3
Number of experiments repeated exactly	X	X	X
Number of predictions made	X	X	X
Data organized with table or graph	X	X	X
Degree of use of Science Help	X	X	X
Frequency of hitting Cancel after having started an interface action	X	X	X
Performance of a variety of interface actions with appropriate frequency	X	X	†
Proportion of accurate predictions	X	†	X
Degree of error in using interface tools for experimenting	†	X	X
Degree of use of Glossary	†	X	X
Degree of use of Computer Help	†	X	X

† Not applicable in that the observable was retained for this simulation problem.

NOTE: TRE = Technology-Rich Environments. An "X" indicates the observable was dropped from the analysis.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Finally, some of the score levels for several observables were collapsed because the performance distinctions between students at those levels did not suggest meaningful differences. Table 6-2 lists these observables.

Procedures for estimating scores on the TRE Simulation scenario were similar to those for the TRE Search scenario, discussed in chapter 5. Scores on the TRE Simulation total scale were estimated using a Bayesian model that combined prior information about students with student performance on the assessment instrument. Prior information about students was based on data collected on 10 background variables: (1) gender, (2) race/ethnicity, (3) disability status, (4) identification as English language learner, (5) student-reported parents' highest level of education, (6) number of types of reading-related items in the home, (7) participation in free or reduced-price school lunch program, (8) participation in Title I, (9) level of prior computer knowledge, and (10) whether the TRE scenario was taken on a NAEP laptop computer. Defining such priors removes bias from TRE means for student groups (Mislevy 1991).

Table 6-2. Observables for which score levels were collapsed in the TRE Simulation scenario analysis, grade 8: 2003

Observable	Simulation problem 1	Simulation problem 2	Simulation problem 3
Use of computer interface (use of various interface functions)	†	†	Collapsed from 3 levels to 2
Proportion of accurate predictions	†	Collapsed from 3 levels to 2	†
Graph is useful to problem	†	Collapsed from 3 levels to 2	Collapsed from 4 levels to 2
Table is useful to problem	Collapsed from 4 levels to 2	Collapsed from 4 levels to 3	Collapsed from 4 levels to 2
Choice of best experiments to solve problem	†	Collapsed from 4 levels to 2	Collapsed from 4 levels to 2

† Not applicable in that the original number of score levels was retained.
NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Paralleling the methodology employed in standard NAEP analyses (Allen, Donoghue, and Schoeps 2001), this modeling approach produces population estimates (e.g., means and standard deviations) without generating scores for individual students. Instead, population estimates are obtained by drawing five imputations, or "plausible values" as they are called in NAEP, for each student from the posterior distribution of proficiency given that student's performance on the assessment instrument and the prior information described above. All means and correlations reported in this chapter employ these five imputations, except where noted. A similar process was used to determine the scale score estimates for computer skills, scientific exploration, and scientific synthesis. For convenience, all four scores were put on an arbitrary scale with a mean of 150 and standard deviation of 35.[17]

[17] This scale is intentionally different from the ones typically used in NAEP assessments so as to prevent confusion with those scales.

The Meaning of TRE Simulation Scores

Because the TRE study used experimental measures, this chapter explores evidence for how well the TRE Simulation scenario scales captured the skills they were intended to summarize. The following sections are presented: internal consistency; the relations of the scores to the measures of the students' prior science and computer knowledge; the TRE scale intercorrelations; the correlations of each observable with each of the three scales (scientific exploration, scientific synthesis, and computer skills); the locations of observables on the scales; the response probabilities for prototypic students (i.e., hypothetical students with levels of low, medium, and high proficiency); and the relations of relevant student background information to performance.

Internal Consistency

As previously stated, internal consistency indicates the degree to which student responses to individual items in a scale are correlated, on average, with their responses to other items in the same scale; higher values for internal consistency suggest greater similarity across items in the underlying skill being measured. For TRE, coefficient alpha, a conventional measure of internal consistency ranging from 0.00 to 1.00, was used to represent this correlation. For the TRE Simulation total score, which consisted of 28 observables, the value of this statistic was .89 (data not shown). For the TRE Simulation scientific exploration score, which had 11 observables, the value was .78 (data not shown). The TRE Simulation scientific synthesis score had 8 observables and an internal consistency of .73 (data not shown). Finally, the TRE Simulation computer skills score had 9 observables and an internal consistency of .74 (data not shown).[18] By way of comparison, these values are higher than the average reliability for the shorter hands-on experimental-task blocks used in the 2000 NAEP science assessment, which, although measuring skills different from the TRE Simulation scenario, also include extended, problem-solving exercises. For the NAEP 2000 science assessment, the mean weighted internal consistency taken across three such blocks was .62.

Correlations of TRE Simulation Scores With Prior Knowledge Measures

The prior knowledge measures were intended to give a rough indication of the degree of student familiarity with the science and computer-related concepts being assessed in the TRE Simulation scenario.

The prior computer knowledge measure (which was common to all students regardless of scenario) consisted of 10 multiple-choice questions about Internet searching, word processing, spreadsheet use, and more general computer knowledge. The prior science knowledge measure (which was particular to students taking the Simulation scenario) comprised 10 multiple-choice questions on concepts related to the science and uses of helium gas balloons, and to the design and interpretation of science experiments. (See appendix D for the questions included on each measure.)

Table 6-3 gives the (disattenuated) correlations of the TRE Simulation scores with the two prior knowledge measures: computer knowledge and science knowledge. As with the Search scenario, these correlations should be considered only suggestive because of the limited number of items used in the prior knowledge measures. (Appendix I gives summary statistics for these measures.) All of the correlations between TRE Simulation scores and the measure of the students' prior science knowledge were significantly different from zero statistically. Thus, students with more prior science knowledge tended to receive higher TRE Simulation scores. Similarly, all of the correlations between TRE Simulation scores and the prior computer knowledge measure were significantly different from zero statistically, indicating that prior computer knowledge was also associated with better performance in the TRE Simulation scenario.

Table 6-3. Weighted (disattenuated) correlations of TRE Simulation scores with prior knowledge measures, grade 8: 2003

TRE Simulation score	Prior computer knowledge measure	Prior science knowledge measure
Total	.62	.64
Computer skills	.51	.56
Scientific exploration	.51	.58
Scientific synthesis	.60	.66

NOTE: TRE = Technology-Rich Environments. N (number of students) range from 960 to 986. All correlations are significantly different from zero at $p < .05$. Students' scores for a particular prior knowledge measure were deleted from this analysis if they were missing seven or more questions in the scale.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

[18] The TRE observables may not be completely independent, so the internal consistency estimates for the TRE scales may be inflated.

Intercorrelations of the Simulation Scales

Table 6-4 gives the (disattenuated) intercorrelations of each TRE Simulation subscale with the Simulation total score for the overall sample and for gender and racial/ethnic groups. Table 6-5 gives the (disattenuated) intercorrelations among the subscales. As the tables show, in the total sample the computer skills, scientific exploration, and scientific synthesis subscales correlate about equally with the TRE 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.

Table 6-4. Number of students and weighted (disattenuated) intercorrelations of the TRE Simulation subscales with the TRE Simulation total score, by student characteristics, grade 8: 2003

Characteristic	Number of students	Computer skills with TRE Simulation total	Scientific exploration skill with TRE Simulation total	Scientific synthesis skill with TRE Simulation total
Total	1,032	.75	.74	.76
Gender				
Male	545	.75	.74	.75
Female	487	.76	.76	.76
Race/ethnicity				
White	644	.71	.69	.71
Black	171	.66	.69	.65
Hispanic	168	.69	.70	.71

NOTE: TRE = Technology-Rich Environments. All correlations are significantly different from zero at *p* < .05. Results are shown for three mutually exclusive race/ethnicity categories. Black includes African American, and Hispanic includes Latino. Race categories exclude Hispanic origin unless specified.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Table 6-5. Number of students and weighted (disattenuated) intercorrelations among the TRE Simulation subscales, by student characteristics, grade 8: 2003

Characteristic	Number of students	Computer skills with scientific exploration skill	Scientific exploration skill with scientific synthesis skill	Scientific synthesis skill with computer skills
Total	1,032	.73	.74	.73
Gender				
Male	545	.72	.73	.74
Female	487	.74	.75	.73
Race/ethnicity				
White	644	.67	.69	.68
Black	171	.66	.65	.67
Hispanic	168	.67	.71	.66

NOTE: TRE = Technology-Rich Environments. All correlations are significantly different from zero at *p* < .05. Results are shown for three mutually exclusive race/ethnicity categories. Black includes African American, and Hispanic includes Latino. Race categories exclude Hispanic origin unless specified.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Correlations of the Observables With the TRE Simulation Scales

Examining the correlations of the observables with each scale can suggest the degree to which the data bear out the theoretical prediction implied by assigning an observable to a particular scale. Also, the correlations indicate roughly how important each observable is to producing the score for the scale to which it is assigned.

Table 6-6 gives the (disattenuated) correlations of each observable with the three TRE subscales. Each observable was intended to measure proficiency on one scale (that is, computer skills, scientific exploration skill, or scientific synthesis skill). Although the distinctions between the scales are not as sharp as they were for TRE Search, in general, visual inspection suggests that the Simulation observables correlate in this student sample more with the scale they were intended to measure than with the other scales. That is, the observables selected to measure computer skills generally appear to correlate more with the computer skills subscale than with the scientific exploration or scientific synthesis subscale, and the same is true for the other scales.

The correlations in table 6-6 also indicate the impact of particular observables on a given scale score. 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. The correlations between these particular observables and the scientific exploration scale score ranged from .49 to .74.

For the scientific synthesis scale, table 6-6 indicates that, in this student sample, the observable most highly associated with this scale score 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). In other words, students who wrote longer responses to the constructed-response question that concluded each Simulation problem tended to receive higher computer skills scale scores than students who wrote shorter answers.

As noted, a correct and complete response to the constructed-response question concluding each Simulation problem is key to achieving a high scientific synthesis score in the TRE Simulation scenario. The scoring guides for Simulation motivating problem 1 used three levels, where a score of 3 was a "best" answer, 2 was a "partial" answer, and 1 was an "unacceptable" answer. Because an additional level could be distinguished, the scoring guides for Simulation problems 2 and 3 used four levels. A score of 4 was a "best" answer, a score of 3 was a "good" answer, a score of 2 was a "partial" answer, and a score of 1 was an "unacceptable" answer.

Table 6-6. Weighted (disattenuated) correlations between score on each observable and TRE Simulation scales, grade 8: 2003

Observable	Computer skills	Scientific exploration	Scientific synthesis
Simulation 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
Use of computer interface (number of characters in conclusion)	**.72**	.49	.54
Degree of error in using interface tools for drawing conclusions	**–.32**	–.25	–.28
Degree of error in using interface tools for experimenting	**–.28**	–.24	–.27
Degree of use of Computer Help	**–.26**	–.22	–.24
Simulation 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
Use of computer interface (number of characters in conclusion)	**.78**	.52	.55
Degree of error in using interface tools for drawing conclusions	**–.27**	–.21	–.23
Simulation 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
Use of computer interface (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
Degree of error in using interface tools for drawing conclusions	**–.21**	–.19	–.20
Conclusion			
Degree of correctness of responses to multiple-choice items	.47	.48	**.58**

NOTE: TRE = Technology-Rich Environments. The **bold** values indicate the scale to which an observable was assigned. All correlations are significantly different from zero at *p* < .05. N (number of students) range = 221 to 1032. All scale scores include the observable being correlated.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

What student behaviors were associated with providing successful responses to the Simulation motivating problems? Table 6-7 indicates that 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 the length of the response, the results show statistically significant positive relationships between 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 tended to receive higher scores for responses to the concluding question than did students who chose a less adequate set of experiments. Further, students who made graphs and tables appropriate to Simulation problems 1 and 2 tended to receive higher scores for their conclusions to those problems than students who did not make such graphs and tables. Finally, table 6-7 shows that students who controlled for one variable in their experiments for Simulation problem 3 tended to attain higher scores on the constructed-response question.

Table 6-7. Observed correlation between score on each observable and raw score on the constructed-response questions for each of three Simulation problems, grade 8: 2003

Observable	Correlation
Simulation problem 1	
Use of computer interface (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
Degree of error in using interface tools for drawing conclusions	–.23
Degree of error in using interface tools for experimenting	–.18
Degree of use of Computer Help	–.15
Degree of use of glossary	–.14
Simulation problem 2	
Use of computer interface (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
Degree of error in using interface tools for drawing conclusions	–.16
Proportion of accurate predictions	.15
Simulation problem 3	
Proportion of experiments controlled for one variable	.45
Use of computer interface (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
Degree of error in using interface tools for drawing conclusions	–.11

NOTE: TRE = Technology-Rich Environments. All correlations are significantly different from zero at $p < .05$. Values are raw correlations and not based on averages across imputations. 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.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Locations of the Observables on the TRE Simulation Scales

Item maps are displays that give a context for interpreting score points on a given scale. They display the locations of items (in the TRE context, observables) on their respective scales by associating points on the scale with levels of correctness for particular observables, and thus describe what students who attain a particular score on each scale are likely to be able to do. As noted in the previous chapter, item maps should be interpreted carefully because an item's location is dependent on the extent to which the underlying assumptions of the response model are met and on the accuracy with which item parameters are estimated. Also, item locations depend on the choice of a probability for correctly responding. For purposes of the TRE study, this probability was set at 65 percent, the level routinely used in NAEP assessments for the mapping of constructed-response items.

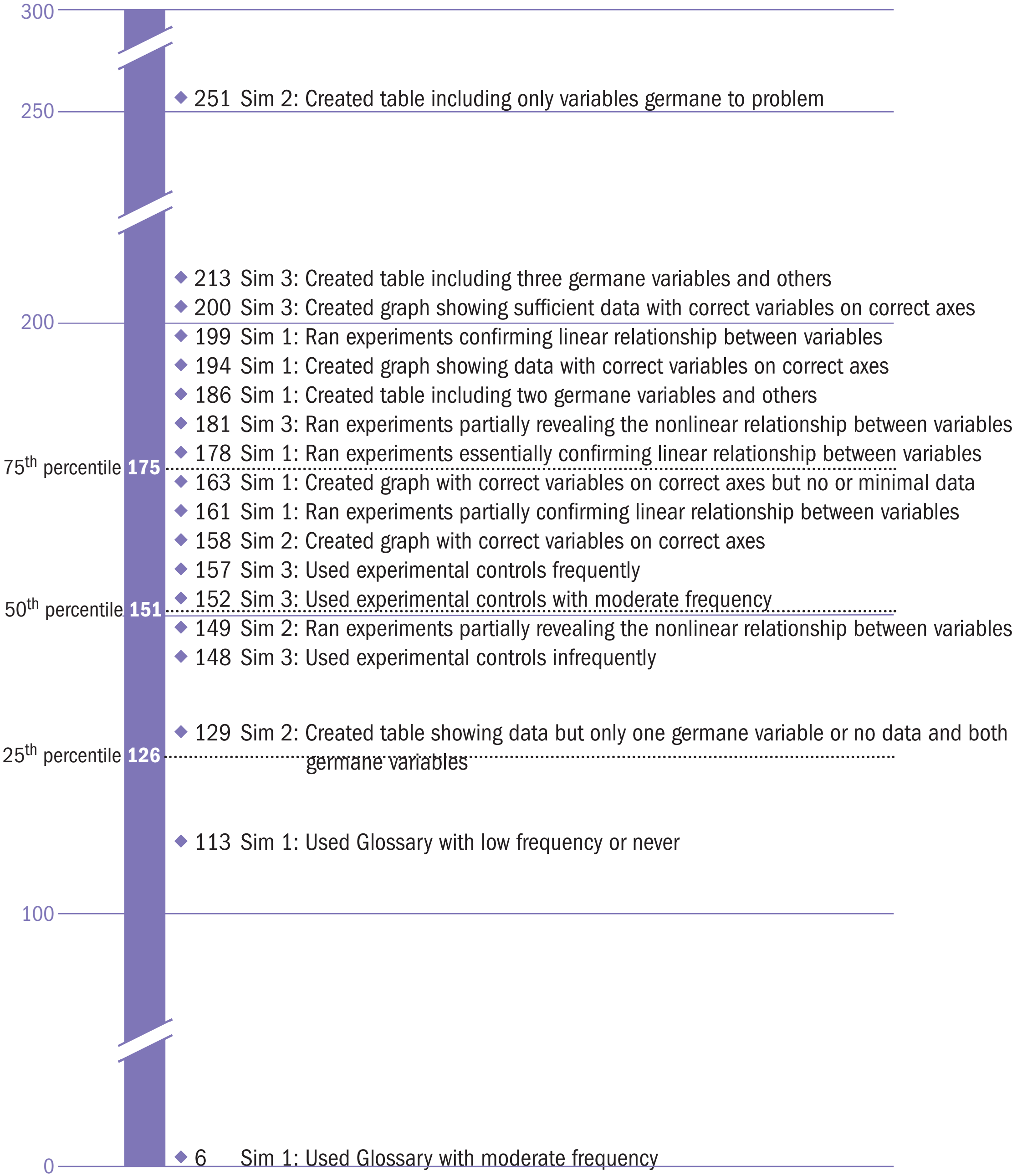
Figure 6-1 shows an item map for the scientific exploration scale. For mapping purposes, each observable has been transformed into one or more dichotomous variables, where the number of such variables is one less than the number of levels of correctness for the observable. Thus, each location on the map represents the point on the scale at which at least 65 percent of students were likely to have achieved the indicated level of correctness for a particular observable. For example, the lowest level of partial credit for running the best experiments for Simulation problem 1 maps to a scale score of 161. This mapping means that students who received a mean score of 161 or more on the scientific exploration scale had at least a 65 percent chance of running experiments that partially confirmed the negative linear relationship between variables for Simulation problem 1. Full credit for running the best experiments for Simulation problem 1 maps to a score of 199; students with this mean score had at least a 65 percent chance of running experiments for Simulation problem 1 that were sufficient to confirm the negative linear relationship between variables.

As shown in chapter 5, mapping observables to the scale enables the scale to be qualitatively described. For the Simulation scientific exploration scale, the scale is defined by the following ordering, from the lowest mapped scale point to the highest:

- using the glossary of science terms in Simulation problem 1 with moderate frequency (note that using the glossary is hypothesized as suggesting a lower level of skill than not using it);
- using the glossary of science terms in Simulation problem 1 with low frequency or never;
- creating a table for Simulation problem 2 that either includes one of the variables relevant to solving the problem with experimental data, or includes both relevant variables without data;
- controlling for one variable in less than 40 percent of the experiments run for Simulation problem 3;
- running a set of experiments that partially reveals the nonlinear relationship between altitude and amount of helium for Simulation problem 2;
- controlling for one variable in 40 to 65 percent of the experiments run for Simulation problem 3;
- controlling for one variable in at least 66 percent of the experiments run for Simulation problem 3;
- creating a graph for Simulation problem 2 with the correct variables on the correct axes, with or without data;
- running experiments sufficient either in number or in range to confirm the negative linear relationship between altitude and mass for Simulation problem 1;
- creating a graph for Simulation problem 1 with the correct variables on the correct axes but showing no data or only one data point;
- running experiments in Simulation problem 1 sufficient in number and range, but not in distribution, to confirm the negative linear relationship between mass and altitude;
- running experiments in Simulation problem 3 for at least one value of mass and conducting a set of experiments with amounts of helium that partially reveals a nonlinear relationship between altitude and volume;
- creating a table for Simulation problem 1 that includes the variables relevant to the problem as well as other variables;

- creating a graph for Simulation problem 1 that has the correct variables on the correct axes and shows at least two data points;
- running a set of experiments in Simulation problem 1 sufficient in number, range, and distribution to confirm the negative linear relationship between altitude and mass;
- creating a graph for Simulation problem 3 that has the correct variables on the correct axes and shows data for at least four experiments (two experiments for each of at least two values of mass);
- creating a table for Simulation problem 3 that includes the three variables relevant to the problem as well as other variables; and

Figure 6-1. Mapping of TRE Simulation observables to the scientific exploration scale, grade 8: 2003



NOTE: TRE = Technology-Rich Environments. Sim 1 = Simulation problem 1; Sim 2 = Simulation problem 2; Sim 3 = Simulation problem 3. Each position on the map indicates the scale score at which students had a 65 percent probability of successfully attaining a given level of correctness for a particular observable.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

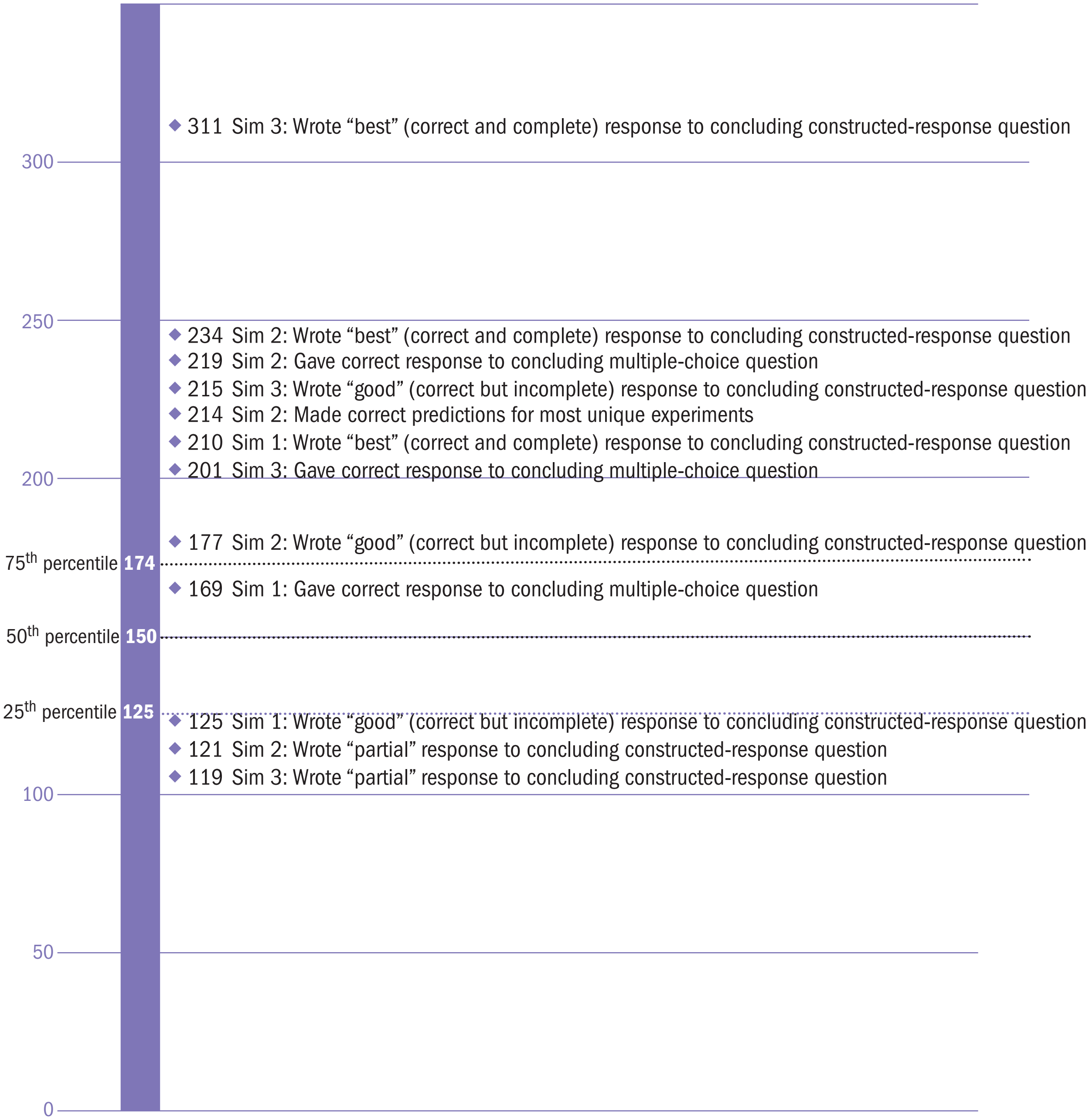
- creating a table for Simulation problem 2 that includes only the dependent and independent variables germane to the problem.

Appendix J gives the percentages of students achieving each of these observable behaviors.

Figure 6-2 shows the locations of the levels of correctness for the observables on the scientific synthesis scale. From the lowest scale point, the ordering is as follows:

- offering "partial" responses to the concluding question for Simulation problem 3 that could be derived from the experiments conducted for Simulations 1 or 2 (e.g., "Below a certain amount of helium the balloon cannot get off the ground");
- offering "partial" responses to the concluding question for Simulation problem 2 that incorrectly describe the relationship between altitude and amount of helium as a positive linear one (e.g., "More helium inside the balloon will make the balloon go higher");
- offering "good" responses to the concluding question for Simulation problem 1 that correctly express the negative linear relationship between mass and altitude (e.g., "A smaller mass will make the balloon go higher"), but do not make specific references to experiments;
- correctly answering the concluding multiple-choice question about the relationship between altitude and mass in Simulation problem 1;
- offering "good" responses to the concluding question for Simulation problem 2 that correctly describe either the top or the bottom segments (but not both) of the step function (e.g., "Once in the air, the balloon will reach a maximum altitude no matter how much helium is added");
- correctly answering the concluding multiple-choice question about the relationships among altitude, mass, and amount of helium in Simulation problem 3;
- offering "best" responses to the concluding question for Simulation problem 1 that correctly express the negative linear function and refer to at least two specific experiments;
- making correct predictions for more than one-half of the unique experiments run for Simulation problem 2;
- offering "good" responses to the concluding question for Simulation problem 3 that correctly describe either the top or the bottom segments of the step function (but not both) in terms of various values of mass (e.g., "Once in the air, the balloon will reach a maximum altitude no matter how much helium is added, and the maximum altitude the balloon can reach decreases as payload mass increases");
- correctly answering the concluding multiple-choice question about the relationship between altitude and amount of helium in Simulation problem 2;
- offering "best" responses to the concluding question for Simulation problem 2 that correctly describe both the top and the bottom segments of the step function (e.g., "Once the balloon has enough helium to rise into the air, the balloon will rise to a maximum height and go no higher no matter how much helium is added"); and
- offering "best" responses to the concluding question for Simulation problem 3 that correctly and completely describe both the top and the bottom segments of the step function in terms of various values of mass (e.g., "The amount of helium needed to lift the balloon increases as mass increases. Once the balloon has enough helium to rise into the air, the balloon will rise to a maximum height for a given mass no matter how much helium is added. This maximum altitude decreases as mass increases.")

Figure 6-2. Mapping of TRE Simulation observables to the scientific synthesis skill scale, grade 8: 2003



NOTE: TRE = Technology-Rich Environments. Sim 1 = Simulation problem 1; Sim 2 = Simulation problem 2; Sim 3 = Simulation problem 3. Each position on the map indicates the scale score at which students had a 65 percent probability of successfully attaining a given level of correctness for a particular observable. The estimated score mapping for “Sim 3: Wrote ‘best’ (correct and complete) response to concluding constructed-response questions” was above the scale maximum of 300 and is included on the figure for completeness.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

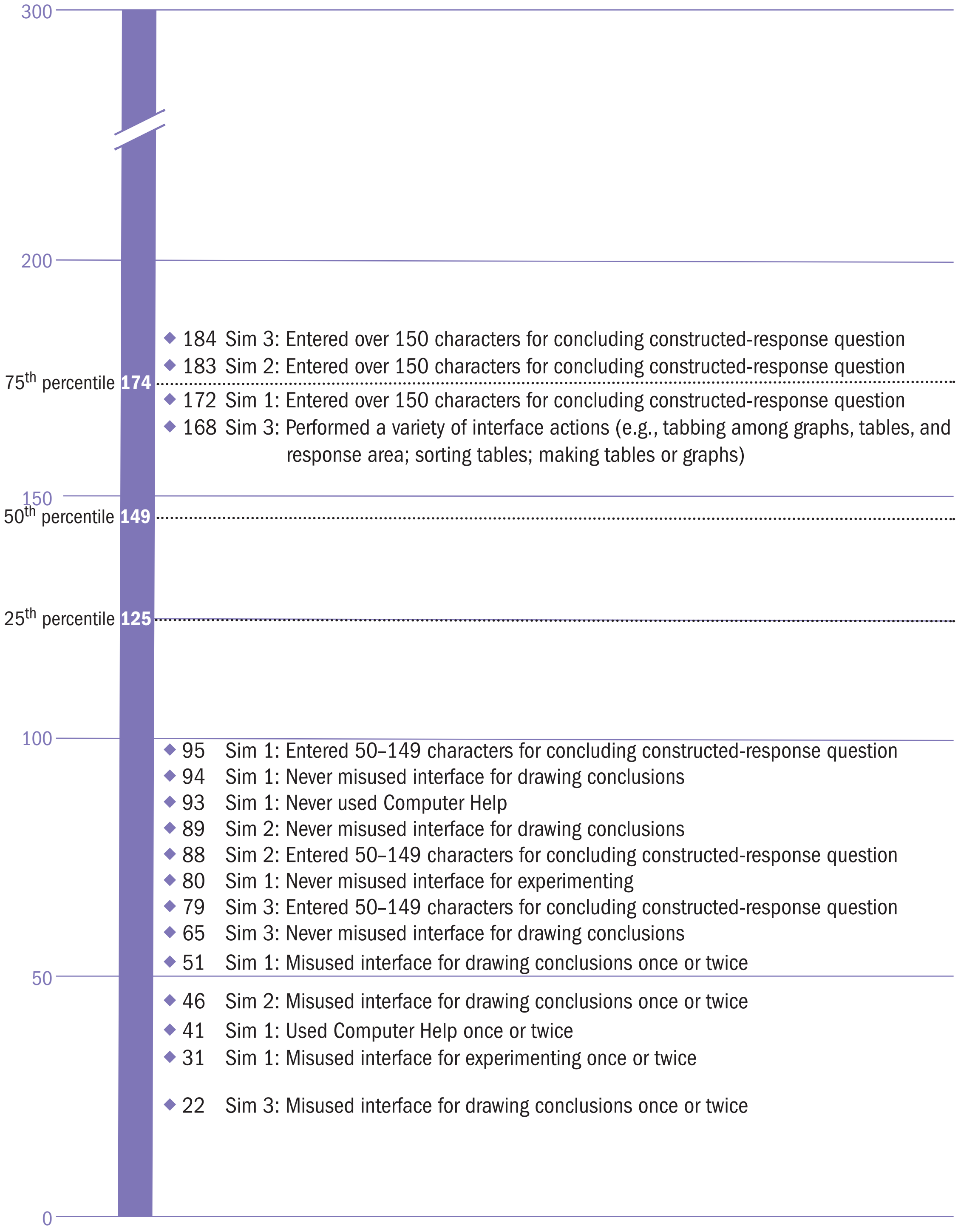
Figure 6-3 shows the locations of the levels of correctness for the observables on the computer skills scale. From the lowest scale point to the highest, the ordering is as follows:

- using interface tools in the wrong order for drawing conclusions once or twice in Simulation problem 3 (e.g., clicking on the Draw Conclusions button before running any experiments);[19]
- using interface tools in the wrong order for experimenting once or twice in Simulation problem 1 (e.g., clicking on the Make Predictions button without having chosen any values with which to experiment);
- using Computer Help once or twice in Simulation problem 1 (note that using Computer Help is proposed as suggesting a lower level of skill than not using it);
- using interface tools in the wrong order for drawing conclusions once or twice in Simulation problem 2 (e.g., clicking on Next without responding to the concluding multiple-choice question);
- using interface tools in the wrong order for drawing conclusions once or twice in Simulation problem 1 (e.g., clicking on the concluding multiple-choice question without first responding to the concluding constructed-response question);
- never using interface tools in the wrong order for drawing conclusions in Simulation problem 3 (e.g., clicking on the Draw Conclusions button before running any experiments);
- key-entering a response of 50 to 149 characters to the constructed-response question concluding Simulation problem 3;
- never using interface tools in the wrong order for experimenting in Simulation problem 1 (e.g., clicking on Try It before choosing a value for a first experiment);
- key-entering a response of 50 to 149 characters to the constructed-response question concluding Simulation problem 2;
- never using interface tools in the wrong order for drawing conclusions in Simulation problem 2 (e.g., clicking on Next without responding to the concluding multiple-choice question);
- never using Computer Help in Simulation problem 1;
- never using interface tools in the wrong order for drawing conclusions in Simulation problem 1 (e.g., clicking on Next without responding to the concluding multiple-choice question);
- key-entering a response of 50 to 149 characters to the constructed-response question concluding Simulation problem 1;
- performing a variety of interface actions (e.g., tabbing among graphs, tables, and the response area; sorting tables; making tables or graphs) in Simulation problem 3;
- key-entering a response of over 150 characters to the constructed-response question concluding Simulation problem 1;
- key-entering a response of over 150 characters to the constructed-response question concluding Simulation problem 2; and
- key-entering a response of over 150 characters to the constructed-response question concluding Simulation problem 3.

Appendix J gives the percentages of students achieving each of these observable behaviors.

[19] The rule for determining whether students used interface tools in the wrong order did not account for students who purposively clicked on each tool to find out what the tool did. However, relatively few students could have taken this approach because, as the item map shows, all of the observables associated with using interface tools in the wrong order fell at the low end of the computer skills scale.

Figure 6-3. Mapping of TRE Simulation observables to the computer skills scale, grade 8: 2003



NOTE: TRE = Technology-Rich Environments. Sim 1 = Simulation problem 1; Sim 2 = Simulation problem 2; Sim 3 = Simulation problem 3. Each position on the map indicates the scale score at which students had a 65 percent probability of successfully attaining a given level of correctness for a particular observable.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Response Probabilities for Prototypic Students

As discussed in chapter 5, examining the response probabilities for prototypic students (that is, hypothetical students with low, medium, or high levels of proficiency) also affords a way to gain insight into the meaning of the TRE scales. The required probabilities can be generated empirically from the item response model for students with different prototypic levels of standing on the TRE proficiencies (e.g., students who are known to be high on scientific exploration as compared with those who are known to be medium or low). The probability of achieving each observable can then be examined to see how prototypic students differ and if those differences are logically meaningful.

Tables 6-8, 6-9, and 6-10 show the response probabilities for prototypic students with different levels of scientific exploration, scientific synthesis, and computer skills. For these tables, the prototypic levels were defined by separately dividing in turn the scientific exploration, scientific synthesis, and computer skills score distributions into thirds and taking the middle value in the bottom third as the prototypic low student, the middle value in the center third as the prototypic medium student, and the middle value in the top third as the prototypic high student. These values were then used to fix the proficiency level in the response model for generating the probability of achieving each of the levels of correctness on each of the observables.[20]

The response probabilities are generally compared in the following way: First, the prototypic low-level student is described by identifying the level of correctness that student is likely to achieve on each observable. Next, the prototypic medium-level student is described in terms of only those observables that would distinguish this student from the prototypic low-level student (i.e., only those observables on which the two students would be likely to attain different degrees of correctness). Finally, the prototypic high-level student is differentiated from the prototypic medium-level student in a similar fashion.

As table 6-8 shows, the low-scientific-exploration student was most likely to receive no credit (i.e., “low” in terms of level of correctness) for a large number of observables:

- running the best experiments for Simulation problem 1,
- controlling variables in experiments for Simulation problem 3,
- creating a useful graph for Simulation problem 1,
- creating a useful table for Simulation problem 1,
- running the best experiments for Simulation problem 2,
- creating a useful graph for Simulation problem 2,
- running the best experiments for Simulation problem 3,
- creating a useful graph for Simulation problem 3, and
- creating a useful table for Simulation problem 3.

The low-scientific-exploration student was also most likely to receive partial credit for creating a useful table for Simulation problem 2 and full credit for degree of use of the glossary in Simulation problem 1, meaning that this student was *unlikely* to make frequent use of the glossary.

The pattern for the medium-scientific-exploration student differed from the low-scientific-exploration student in that the medium-scientific exploration student was more likely to achieve *full* credit, rather than *no* credit, for the following observables:

- controlling variables in experiments for Simulation problem 3,
- running the best experiments for Simulation problem 2, and
- creating a useful graph for Simulation problem 2.

Finally, in contrast to the medium-scientific-exploration student, the high-scientific-exploration student was most likely to get *full*, rather than *no*, credit for the following observables:

- running the best experiments for Simulation problem 1,
- creating a useful graph for Simulation problem 1,
- creating a useful table for Simulation problem 1,
- running the best experiments for Simulation problem 3, and
- creating a useful graph for Simulation problem 3.

[20] Note that some observables have two levels of correctness (no credit, full credit), some have three levels (no credit, partial credit, and full credit), and some have four levels (no credit, low-partial credit, high-partial credit, and full credit).

Table 6-8. Probability of responding to observables on TRE Simulation for prototypic students, by level of scientific exploration skill and level of correctness of observable response, grade 8: 2003

	Low level of scientific exploration				Medium level of scientific exploration				High level of scientific exploration			
Observables	No credit[1]	Low-partial credit	High-partial credit	Full credit	No credit[1]	Low-partial credit	High-partial credit	Full credit	No credit[1]	Low-partial credit	High-partial credit	Full credit
Sim1 Ran best experiments	**.67**	.17	.09	.07	**.37**	.23	.19	.20	.16	.17	.23	**.45**
Sim3 Proportion of experiments controlled for 1 variable	**.87**	.06	.03	.04	.31	.13	.13	**.43**	.02	.02	.04	**.92**

	Low level of scientific exploration			Medium level of scientific exploration			High level of scientific exploration		
Observables	No credit[1]	Partial credit	Full credit	No credit[1]	Partial credit	Full credit	No credit[1]	Partial credit	Full credit
Sim1 Degree of use of Glossary[2]	.04	.31	**.65**	.02	.19	**.79**	.01	.11	**.88**
Sim1 Usefulness of graph	**.77**	.17	.06	**.45**	.34	.21	.18	.31	**.51**

	Low level of scientific exploration		Medium level of scientific exploration		High level of scientific exploration	
Observables	No credit[1]	Full credit	No credit[1]	Full credit	No credit[1]	Full credit
Sim1 Usefulness of table	**.86**	.14	**.64**	.36	.35	**.65**
Sim2 Ran best experiments	**.81**	.19	.32	**.68**	.05	**.95**
Sim2 Usefulness of graph	**.68**	.32	.39	**.61**	.16	**.84**
Sim3 Ran best experiments	**.99**	.01	**.85**	.15	.33	**.67**
Sim3 Usefulness of graph	**.81**	.19	**.64**	.36	.44	**.56**
Sim3 Usefulness of table	**.60**	.40	**.50**	**.50**	.41	**.59**

[1] No credit, partial credit (including low-partial and high-partial), and full credit are the levels of correctness of response specific to each observable.

[2] The values for this observable were such that less glossary use received a higher score.

NOTE: TRE = Technology-Rich Environments. Sim1 = Simulation problem 1; Sim2 = Simulation problem 2; Sim 3 = Simulation problem 3. Highest probability for each level is shown in **bold**. Detail may not sum to totals because of rounding.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Table 6-9 gives the response probabilities for the prototypic students with different levels of scientific synthesis skill, which were computed in a manner similar to that for scientific exploration. The low-scientific-synthesis student was most likely to get no credit for every observable except for the accuracy of the responses to the concluding multiple-choice synthesizing questions, for which this student would more likely receive partial credit. By contrast, the medium-scientific-synthesis student was likely to receive partial credit, instead of no credit, for the accuracy of the responses to the final constructed-response questions for Simulation problems 1, 2, and 3, and for the accuracy of the response to the final multiple-choice question for Simulation problem 1.

Compared with the student with medium proficiency on scientific synthesis, the high-scientific-synthesis student was likely to receive full instead of partial credit for the accuracy of the response to the final constructed-response question for Simulation problem 1, the accuracy of the responses to the concluding multiple-choice synthesizing questions, the proportion of accurate predictions for experimental results for Simulation problem 2, and the accuracy of the response to the final multiple-choice question for Simulation problem 3.

Table 6-9. Probability of responding to observables on TRE Simulation for prototypic students, by level of scientific synthesis skill and level of correctness of observable response, grade 8: 2003

	Low level of scientific synthesis				Medium level of scientific synthesis				High level of scientific synthesis			
Observables	No credit[1]	Low-partial credit	High-partial credit	Full credit	No credit[1]	Low-partial credit	High-partial credit	Full credit	No credit[1]	Low-partial credit	High-partial credit	Full credit
Sim2 Accuracy of response to constructed-response question	**.74**	.22	.03	.00	.26	**.50**	.21	.03	.04	.24	**.50**	.22
Sim3 Accuracy of response to constructed-response question	**.86**	.14	.00	.00	.46	**.50**	.03	.00	.11	**.69**	.19	.01

	Low level of scientific synthesis			Medium level of scientific synthesis			High level of scientific synthesis		
Observables	No credit[1]	Partial credit	Full credit	No credit[1]	Partial credit	Full credit	No credit[1]	Partial credit	Full credit
Sim1 Accuracy of response to constructed-response question	**.68**	.31	.02	.22	**.66**	.12	.03	.47	**.50**
Accuracy of responses to concluding multiple-choice synthesizing questions	.35	**.58**	.07	.12	**.64**	.24	.03	.40	**.56**

	Low level of scientific synthesis		Medium level of scientific synthesis		High level of scientific synthesis	
Observables	No credit[1]	Full credit	No credit[1]	Full credit	No credit[1]	Full credit
Sim1 Accuracy of response to multiple-choice question	**.57**	.43	.41	**.59**	.26	**.74**
Sim2 Accuracy of response to multiple-choice question	**.92**	.08	**.81**	.19	**.61**	.39
Sim2 Proportion of accurate predictions made	**.77**	.23	**.63**	.37	.47	**.53**
Sim3 Accuracy of response to multiple-choice question	**.89**	.11	**.73**	.27	.48	**.52**

[1] No credit, partial credit (including low-partial and high-partial), and full credit are the levels of correctness of response specific to each observable.

NOTE: TRE = Technology-Rich Environments. Sim1 = Simulation problem 1; Sim2 = Simulation problem 2; Sim3 = Simulation problem 3. Highest probability for each level is shown in **bold**. Detail may not sum to totals because of rounding.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Finally, the high-synthesis student was also more likely to receive a higher degree of partial credit than the medium-synthesis student for the accuracy of the response to the final constructed-response question for Simulation problem 2.

Table 6-10 gives the response probabilities for computer skills. The prototypic low-computer-skills student was likely to receive no credit for performing a variety of interface actions with appropriate frequency (e.g., tabbing among graphs, tables, and the response area; sorting tables; and making tables or graphs) in Simulation problem 3, and partial credit for the number of characters used in the final constructed-response questions for Simulation problems 1, 2, and 3. The low-computer-skills student was likely to receive the full score for making errors in using interface tools to draw conclusions in Simulation problems 1, 2, and 3; for making errors in using interface tools for experimenting in Simulation problem 1; and for frequency of use of the Computer Help tool in Simulation problem 1, meaning that this student was not very likely to make such errors or to frequently use Computer Help.

Table 6-10. Probability of responding to observables on TRE Simulation for prototypic students, by level of computer skills and level of correctness of observable response, grade 8: 2003

	Low level of computer skills			Medium level of computer skills			High level of computer skills		
Observables	No credit[1]	Partial credit	Full credit	No credit[1]	Partial credit	Full credit	No credit[1]	Partial credit	Full credit
Sim1 Interface errors in drawing conclusions[2]	.02	.35	**.63**	.01	.21	**.78**	.00	.11	**.88**
Sim1 Interface errors in running experiments[2]	.02	.28	**.70**	.01	.18	**.81**	.01	.11	**.89**
Sim1 Degree of use of Computer Help[2]	.02	.25	**.73**	.01	.15	**.84**	.01	.09	**.91**
Sim1 Number of characters used in response to constructed-response question	.19	**.73**	.07	.01	.46	**.52**	.00	.06	**.94**
Sim2 Interface errors in drawing conclusions[2]	.01	.15	**.83**	.01	.07	**.93**	.00	.03	**.97**
Sim2 Number of characters used in response to constructed-response question	.28	**.69**	.03	.01	**.53**	.45	.00	.05	**.95**
Sim3 Interface errors in drawing conclusions[2]	.01	.10	**.89**	.00	.05	**.95**	.00	.02	**.98**
Sim3 Number of characters used in response to constructed-response question	.19	**.74**	.07	.01	.44	**.55**	.00	.04	**.96**

	Low level of computer skills		Medium level of computer skills		High level of computer skills	
Observables	No credit[1]	Full credit	No credit[1]	Full credit	No credit[1]	Full credit
Sim3 Performing a variety of interface actions with appropriate frequency (e.g., tabbing among graphs and tables)	**.76**	.24	**.54**	.46	.28	**.72**

[1] No credit, partial credit, and full credit are the levels of correctness of response specific to each observable.

[2] The values for these observables were such that fewer errors or less use received higher levels of credit.

NOTE: TRE = Technology-Rich Environments. Sim1 = Simulation problem 1; Sim2 = Simulation problem 2; Sim3 = Simulation problem 3. Highest probability for each level is shown in **bold**. Detail may not sum to totals because of rounding.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

The medium-computer-skills student differed from the low-computer-skills student most obviously by being likely to receive full credit for the number of characters used in the constructed-response questions concluding Simulation problems 1 and 3.

Finally, the high-computer-skills student was likely to receive full credit for the number of characters used in the constructed-response question concluding Simulation problem 2, and for performing a variety of interface actions with appropriate frequency (e.g., tabbing among graphs, tables, and the response area; sorting tables; and making tables or graphs) in Simulation problem 3. In contrast, the medium-computer-skills student was likely to get partial credit for the first observable and no credit for the second observable.

TRE Performance as a Function of Relevant Background Experience

As previously discussed, students responded to sets of background questions when they took the TRE scenarios. One set of questions asked students about their experiences with computers in and out of school, as well as their activities in science class. Figures 6-4 to 6-6 show the relationship of students' TRE Simulation scenario scores with some kinds of experience with computers that students reported. For each background question in the tables, statistically significant differences in student performance and the directions of those differences are indicated; T denotes the TRE Simulation total score, E denotes the TRE Simulation scientific exploration score, S denotes the TRE Simulation scientific synthesis score, and C denotes the TRE Simulation computer skills score.

As shown in figure 6-4, and as might be expected, students who reported using computers more frequently for a variety of activities, ranging from using a word processor to making tables and graphs, outperformed their peers who reported using computers less frequently for these activities. While some activities—for example, using computers to make art (data not shown)—were not associated with any statistically significant score differences, in no case were computer-based activities negatively associated with student performance.

Students reporting using a word processor to a small, moderate, or large extent performed better on all four scales than students reporting not using a word processor at all. Further, students reporting using a word processor to a moderate or large extent outperformed students reporting using one to a small extent; and, finally, students reporting using a word processor to a large extent outperformed students reporting using one to a moderate extent. These results make sense as the TRE Simulation scenario requires students to use their word processing skills to compose responses to the constructed-response questions concluding each section of the scenario.

Also notable in figure 6-4 is that students who reported using a computer to make charts, tables, and graphs to a small or moderate extent performed better on all four TRE scales than students who reported that they did not do so at all. Although they did not have to, students could choose to make tables and graphs in the TRE Simulation scenario to keep track of experiments they had run and to help them interpret the results of their experiments; students who reported using charts, tables, and graphs outside of the TRE experience to a small or moderate extent received higher scale scores than students who did not report such use. One possible explanation for this association is that experience with making tables and graphs on the computer was helpful to students taking the TRE Simulation scenario.

Figure 6-4 indicates that students who reported finding information on the Internet to a large extent had higher scale scores for all four TRE Simulation scales than their peers who reported doing so to a small extent, and also higher scientific synthesis scale scores than students who reported finding information on the Internet to a moderate extent. A possible explanation for this association is that, while the TRE Simulation scenario did not require web searching, its interface conventions (for example, arrows to move forward and backward among pages and functions activated by clicking) would all likely be very familiar to students who spend time navigating on the Web.

Finally, figures 6-5 and 6-6 show results consistent with those from figure 6-4, as they indicate that the frequency of using a computer outside of school (figure 6-5) and the presence of a computer at home (figure 6-6) are both positively associated with student performance. On all four TRE Simulation scale scores, students who reported using a computer outside of school daily outperformed students who reported doing so 2 to 3 times per week, once every few weeks, and never or hardly ever. On the TRE Simulation total, exploration, and computer skills scales, students who reported using a computer outside of school daily outperformed students who reported doing so once a week. Additionally, students who reported using a computer outside of school 2–3 times a week outperformed those who reported doing so once every few weeks on the scientific exploration scale and on the total score scale, and outperformed those who reported doing so never or hardly ever on all four TRE Simulation scales.

The overall positive pattern of relationships between student performance and computer use generally held true for all four TRE Simulation scales, indicating that the TRE scales were functioning similarly with respect to these background indicators. There was one notable exception, however: 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. There were no statistically significant relationships between student reports about this variable and their scores on the other three TRE Simulation scales. This result may reflect the fact that the TRE Simulation observables assigned to the TRE exploration scale resemble the activities involved in some complex computer games; manipulating conditions, keeping track of choices made and their outcomes, observing and interpreting animated displays, and creating and manipulating tables and graphs are effective strategies for solving problems in a variety of computer-based environments.

Information was also collected about students' activities in science class, for example, the frequency of carrying out science experiments. In almost every case, the numbers of students in the various response intervals for each background question were too small for significance tests to be performed, or data based on these questions bore no statistically significant relationships to student performance. In no instance were reported science activities negatively associated with student performance (data not shown).[21]

Figure 6-4. Relationship between TRE Simulation performance and reported type of computer use, grade 8: 2003

Play computer games

Response	Not at all	Small	Moderate	Large
Not at all	†		E	E
Small		†		
Moderate	E		†	
Large	E			†

Use a word processor

Response	Not at all	Small	Moderate	Large
Not at all	†	T, E, S, & C	T, E, S, & C	T, E, S, & C
Small	T, E, S, & C	†	T, E, S, & C	T, E, S, & C
Moderate	T, E, S, & C	T, E, S, & C	†	T, E, S, & C
Large	T, E, S, & C	T, E, S, & C	T, E, S, & C	†

Make tables, charts, or graphs on computer

Response	Not at all	Small	Moderate	Large
Not at all	†	T, E, S, & C	T, E, S, & C	
Small	T, E, S, & C	†		
Moderate	T, E, S, & C		†	
Large				†

Find information on the Internet

Response	Not at all	Small	Moderate	Large
Not at all	†			
Small		†	T, S, & C	T, E, S, & C
Moderate		T, S, & C	†	S
Large		T, E, S, & C	S	†

† Not applicable.
T = TRE Simulation total score.
E = TRE Simulation scientific exploration score.
S = TRE Simulation scientific synthesis score.
C = TRE Simulation computer skills score.
NOTE: TRE = Technology-Rich Environments. Column headings in table correspond to student questionnaire response categories as follows: Not at all = not at all; Small = small extent; Moderate = moderate extent; Large = large extent.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

■ Indicates that at least one of the four types of scores was significantly higher at the .05 level for students giving the response at the left of the row than for those giving the response at the top of the column.

□ Indicates that there was no significant difference in any of the four types of scores between students giving the response at the left of the row and those giving the response at the top of the column.

▒ Indicates that at least one of the four types of scores was significantly lower at the .05 level for students giving the response at the left of the row than for those giving the response at the top of the column.

[21] The analyses presented in figures 6-5 to 6-6 did not control for other background variables, such as socioeconomic status (SES). It is possible that holding such variables constant would produce a different pattern of relations between reported computer use and TRE scores from that described above.

Figure 6-5. Relationship between TRE Simulation performance and reported frequency of computer use outside of school, grade 8: 2003

How often do you use a computer outside of school?

Response	Daily	2–3 times per week	Once a week	Once every few weeks	Never or hardly ever
Daily	†	T, E, S, & C	T, E, & C	T, E, S, & C	T, E, S, & C
2–3 times per week	T, E, S, & C	†		T, E	T, E, S, & C
Once a week	T, E, & C		†		
Once every few weeks	T, E, S, & C	T, E		†	
Never or hardly ever	T, E, S, & C	T, E, S, & C			†

† Not applicable.
T = TRE Simulation total score.
E = TRE Simulation scientific exploration score.
S = TRE Simulation scientific synthesis score.
C = TRE Simulation computer skills score.
NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

■ Indicates that at least one of the four types of scores was significantly higher at the .05 level for students giving the response at the left of the row than for those giving the response at the top of the column.

□ Indicates that there was no significant difference in any of the four types of scores between students giving the response at the left of the row and those giving the response at the top of the column.

▒ Indicates that at least one of the four types of scores was significantly lower at the .05 level for students giving the response at the left of the row than for those giving the response at the top of the column.

Figure 6-6. Relationship between TRE Simulation performance and presence of a home computer that the student uses, grade 8: 2003

Is there a computer at home that you use?

Response	Yes	No
Yes	†	T, E, S, & C
No	T, E, S, & C	†

† Not applicable.
T = TRE Simulation total score.
E = TRE Simulation scientific exploration score.
S = TRE Simulation scientific synthesis score.
C = TRE Simulation computer skills score.
NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

■ Indicates that at least one of the four types of scores was significantly higher at the .05 level for students giving the response at the left of the row than for those giving the response at the top of the column.

□ Indicates that there was no significant difference in any of the four types of scores between students giving the response at the left of the row and those giving the response at the top of the column.

▒ Indicates that at least one of the four types of scores was significantly lower at the .05 level for students giving the response at the left of the row than for those giving the response at the top of the column.

Performance by Student Groups

Analyses were carried out for average scores for NAEP reporting groups defined by gender, race/ethnicity, parents' education level, eligibility for free or reduced-price school lunch, and school location. (See table 6-11 for performance results for student groups.) Statistically significant differences in student performance were found on one or more TRE Simulation scales for all groups except gender and school location, and are discussed below. (More details on TRE scale scores and percentiles by student groups are available in appendix H for those groups and scales on which statistically significant differences were observed.) It is notable that no difference was found between the average scores of male and female students in the Simulation scenario.

Performance by Racial/Ethnic Group

NAEP uses school-reported data to identify students' race/ethnicity. For each of the four TRE Simulation score scales, there were statistically significant differences among the racial/ethnic groups: White students received higher scores on all four TRE scales than their Black and Hispanic peers. On the TRE Simulation total score, White students scored higher (mean scale score = 161) than Black students (mean scale score = 127) (*t*, 15 = 8.21, $p < .05$) and Hispanic students (mean scale score = 128) (*t*, 5 = 6.68, $p < .05$).

On the scientific exploration scale, White students (mean scale score = 160) had higher scores than did Black students (mean scale score = 131) (*t*, 12 = 6.97, $p < .05$) and Hispanic students (mean scale score = 130) (*t*, 6 = 6.72, $p < .05$).

For scientific synthesis, too, the average performance of White students (mean scale score = 161) was higher than that of Hispanic students (*t*, 10 = 7.14, $p < .05$), who received a mean scale score of 130, as well as that of Black students (*t*, 13 = 6.73, $p < .05$), who received a mean scale score of 128.

Finally, for the computer skills scale score, White students (mean scale score = 159) received higher scale scores than did Hispanic students (mean scale score = 132) (*t*, 18 = 5.04, $p < .05$) and Black students (mean scale score = 132) (*t*, 31 = 5.09, $p < .05$).

Performance by Parents' Education Level

Statistically significant performance differences were also present for groups of students reporting different levels of parental education. NAEP asks how far the student's mother went in school and how far the student's father went in school and uses the higher level for this category. As is typical for NAEP results, students who reported higher levels of parental education outperformed their peers who reported lower levels. For the TRE Simulation total score, students reporting that a parent graduated from college (mean scale score = 161) outperformed students reporting that a parent graduated from high school (mean scale score = 141) (*t*, 37 = –5.02, $p < .05$), and outperformed students reporting that their parents did not finish high school (mean scale score = 121) (*t*, 20 = –7.19, $p < .05$). In addition, students reporting that a parent had some education after high school (mean scale score = 150) outperformed those reporting that a parent graduated from high school (*t*, 41 = –2.18, $p < .05$) and those reporting that their parents did not finish high school (*t*, 22 = –5.05, $p < .05$).

On the scientific exploration scale, the performance of students reporting that a parent had graduated from college (mean scale score = 159) was higher than the performance of students reporting that a parent had graduated from high school (mean scale score = 142) (*t*, 38 = –4.18, $p < .05$) and higher than the performance of students whose parents did not finish high school (mean scale score = 127) (*t*, 32 = –7.02, $p = <.05$). Additionally, students whose parents had some education after high school (mean scale score = 151) also outperformed students whose parents did not finish high school (*t*, 32 = –4.79, $p < .05$).

For the scientific synthesis scale, students who reported that a parent graduated from college (mean scale score = 160) scored higher than students with a parent who had some education after high school (mean scale score = 150) (*t*, 37 = –2.22, $p < .05$); than students who reported a parent who graduated from high school (mean scale score = 142) (*t*, 27 = –4.87, $p < .05$); and than students whose parents did not finish high school (mean scale score = 125) (*t*, 35 = –7.48, $p < .05$). Further, students with a parent who had some education after high school (mean scale score = 150) scored higher than students whose parents did not finish high school (mean scale score = 125) (*t*, 48 = –4.38, $p < .05$).

There were also several statistically significant differences among the groups for computer skills. Students reporting that a parent graduated from college (mean scale score = 160) scored higher on the computer skills scale than students with a parent whose highest level of education was graduation from high school (mean scale score = 143) (*t*, 52 = –3.32, $p < .05$), and than students whose parents did not finish high school (mean scale score = 125) (*t*, 45 = –6.54, $p < .05$).

Performance by Eligibility for Free or Reduced-Price School Lunch

Performance can also be analyzed for groups defined according to eligibility for free or reduced-price school lunch, as reported by schools. Eligibility is based on family income and is thus related to socioeconomic status. Those students not eligible for free or reduced-price lunch received higher mean TRE Simulation total scores (mean scale score = 160) than students eligible for reduced-price lunch (mean scale score = 143) (*t*, 36 = 3.25, $p < .05$) and students eligible for free lunch (mean scale score = 127) (*t*, 22 = 8.67, $p < .05$). Students eligible for reduced-price lunch, in turn, performed better (mean scale score = 143) than students eligible for free lunch (mean scale score = 127)(*t*, 37 = 2.94, $p < .05$).

For the scientific exploration scale, students who were not eligible for free or reduced-price lunch received higher scores (mean scale score = 158) than students who were eligible for free lunch (mean scale score = 131) (*t*, 12 = 6.61, $p < .05$).

For the scientific synthesis scale, students who were not eligible for free or reduced-price lunch performed better (mean scale score = 159) than students who were eligible for reduced-price lunch (mean scale score = 146) (*t*, 22 = 2.17, $p < .05$) and students who were eligible for free lunch (mean scale score = 130) (*t*, 21 = 7.31, $p < .05$). Additionally, students who were eligible for reduced-price lunch (mean scale score = 146) had higher scores than those who were eligible for free lunch (mean scale score = 130) (*t*, 30 = 2.53, $p < .05$).

For the computer skills scale, students who were not eligible for free or reduced-price lunch (mean scale score = 158) performed better than those who were eligible for free lunch (mean scale score = 131) (*t*, 25 = 5.29, $p < .05$).

Table 6-11. Mean TRE Simulation scores, by student characteristics and number of students, grade 8: 2003

Characteristic	Number of students	TRE Simulation total score	Scientific exploration score	Scientific synthesis score	Computer skills score
Total	1,032	150 (2.4)	150 (2.3)	150 (2.3)	150 (3.4)
Gender					
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)
Race/ethnicity					
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)
Student-reported parents' highest education level					
Did not finish high school	66	121 (5.1)	127 (3.8)	125 (4.1)	125 (3.7)
Graduated from high school	199	141 (3.3)	142 (3.1)	142 (3.1)	143 (3.5)
Some education after high school	180	150 (2.8)	151 (3.3)	150 (3.9)	149 (4.4)
Graduated from college	493	161 (2.4)	159 (2.6)	160 (2.2)	160 (3.7)
Eligibility for school lunch					
Not eligible	625	160 (2.1)	158 (1.4)	159 (1.7)	158 (3.2)
Reduced-price lunch	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)
School location					
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)

NOTE: TRE = Technology-Rich Environments. Standard errors of the estimates appear in parentheses. Some seemingly large differences between the performance of student groups were not statistically significant because of the large standard errors associated with those differences. Results are shown for three mutually exclusive race/ethnicity categories. Black includes African American, and Hispanic includes Latino. Race categories exclude Hispanic origin unless specified. Eligibility for free or reduced-price lunch was based on school-reported information. For details about eligibility requirements, see Eligibility for Free/Reduced-Price School Lunch in Appendix K. Results are not shown for students whose eligibility status for free or reduced-price lunch was not available.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Chapter 7: Summary of Results

The Problem Solving in Technology-Rich Environments (TRE) study was designed to demonstrate and explore an innovative use of computers for developing, administering, scoring, and analyzing the results of NAEP assessments. To accomplish this exploration, researchers developed two sample scenarios focused on using computers for problem solving. Because the TRE project was intended as an exploratory study involving only two scenarios in one domain of science, results cannot be generalized to problem solving in technology-rich environments as a whole. However, by reflecting eighth-graders' performance in a narrow domain, the study illustrates the kinds of tasks, analyses, and results that scenario-based technology assessment can provide in NAEP.

TRE Search Scenario Results

TRE Search consisted of 11 observables and produced a total score and two subscores: scientific inquiry and computer skills. The internal consistency of the three TRE Search scores ranged from .65 to .74. These values compare favorably to those for the NAEP grade 8 hands-on science blocks, which, although measuring skills different from TRE, also include extended exercises. The hands-on science blocks usually feature 30-minute extended exercises (in contrast to the approximately 40 minutes allocated to TRE Search). For the 2000 NAEP science assessment, the mean weighted internal consistency, taken across three such hands-on blocks, was .62 (O'Sullivan et al. 2003).

The Search subscores provided overlapping but not redundant information; the (disattenuated) intercorrelation of the scores was .57. The scientific inquiry skill scale score was most related in the student sample 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 (disattenuated correlations between performance on the observable and scale score = .51 to .71). In contrast, the computer skills scale score was most related in the student sample to the following factors: 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 (disattenuated correlation range = .60 to .69). Although the Search scenario required more time than the typical NAEP science assessment block, the scenario produced more score information because performance was evaluated along three dimensions instead of one.

Some of the differences observed among the performances of major NAEP reporting groups on NAEP assessments were also observed on TRE Search. On the total score, White students scored higher than Black and Hispanic students, and Hispanic students scored higher than Black students. Students who reported that at least one parent graduated from college scored higher than students who reported that their parents did not finish high school and higher than those who reported that at least one parent graduated from high school. Students who were not eligible for free or reduced-price lunch scored higher than eligible students. Overall, similar patterns of difference were also evident for the two Search subscales.

TRE Simulation Scenario Results

The TRE Simulation scenario consisted of 28 observables and produced a total score and three subscores: scientific exploration, scientific synthesis, and computer skills. The internal consistency of the four scales ranged from .73 to .89. Like the Search scenario, Simulation compared favorably to the NAEP hands-on science blocks, which measure skills different from TRE but which employ extended tasks. TRE Simulation required more time than the typical NAEP science block, and Simulation appeared to be somewhat more reliable and produced more score information than NAEP science blocks.

As with the Search scenario, the Simulation subscores provided overlapping but not redundant information; the (disattenuated) intercorrelations of the scores ranged from .73 to .74. The scientific exploration skill scale score was most related in the student sample to three factors—which experiments students chose to run to solve 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 in Simulation problem 3. The scientific synthesis scale score was most related in the student sample to the degree of correctness and completeness of conclusions drawn for each Simulation problem. Finally, performance on the computer skills scale was most associated in the student sample with the number of characters in the conclusions students constructed for each of the three Simulation problems.

Also, as with the Search scenario, many of the performance differences observed among student groups on NAEP assessments held true for TRE Simulation. On the TRE Simulation total score, White students scored significantly higher statistically than Black and Hispanic students. Students who reported that at least one parent graduated from college scored higher than students who reported that their parents did not finish high school and higher than those who reported that at least one parent graduated from high school. Finally, students who were not eligible for free or reduced-price lunch scored higher than eligible students. Similar patterns of difference were also evident for the three Simulation subscales.

References

Adams, R.J., Wilson, M., and Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*(1), 1–23.

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

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

Almond, R.G. (forthcoming). *Learning and Revising Models From Data.* Princeton, NJ: Educational Testing Service.

Almond, R.G., DiBello, L., Jenkins, F., Mislevy, R.J., Senturk, D., Steinberg, L.S., and Yan, D. (2001). Models for Conditional Probability Tables in Educational Assessment. In T. Jaakkola and T. Richardson (Eds.), *Artificial Intelligence and Statistics 2001: Proceedings of the 8th International Workshop on Artificial Intelligence and Statistics* (pp. 137–143). San Mateo, CA: Morgan Kaufmann.

Baxter, G.P., and Glaser, R. (1998). Investigating the Cognitive Complexity of Science Assessments. *Educational Measurement: Issues and Practice, 17*(3): 37–45.

c-rater. Princeton, NJ: Educational Testing Service [computer software].

Fidel, R., Davies, R.K., Douglass, M.H., Holder, J.K., Hopkins, C.J., Kushner, E.J., Miyagishima, B.K., and 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.

Gelman, A., Carlin, J., Stern, H., and Rubin, D.B. (1995). *Bayesian Data Analysis.* London: Chapman & Hall.

Gelman, A., and Rubin, D.B. (1992). Inference from Iterative Simulation Using Multiple Sequences (with discussion and rejoinder). *Statistical Science, 7:* 457–472.

Geman, S., and Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 6:* 721–741.

Horkay, N., Bennett, R., Allen, N., and Kaplan, B. (2005). Online Assessment in Writing. In B. Sandene, N. Horkay, R. Bennett, N. Allen, J. Braswell, B. Kaplan, and A. Oranje. *Online Assessment in Mathematics and Writing: Reports From the NAEP Technology-Based Assessment Project, Research and Development Series* (NCES 2005–457) (Part II). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

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

Johnson, M.S., and Jenkins, F. (2005). *A Bayesian Hierarchical Model for Large-Scale Educational Surveys: An Application to the National Assessment of Educational Progress.* ETS Research Report, RR-04-38. Princeton, NJ: Educational Testing Service.

Klein, D.C.D., Yarnall, L., and 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., and 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: Lawrence Erlbaum Associates, Inc.

Kleiner, A., and Lewis, L. (2003). *Internet Access in U.S. Public Schools and Classrooms: 1994–2002* (NCES 2004-111). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Knowledge Integration Environment. (KIE) (1997). *Deformed Frogs! Web KIE Project, KIE-Roosevelt Curriculum Development Partnership.* Retrieved January 25, 2005, from http://kie.berkeley.edu/roosevelt/frogs.html.

Lauritzen, S.L., and Spiegelhalter, D.J. (1988). Local Computations With Probabilities on Graphical Structures and Their Application to Expert Systems (with discussion). *Journal of the Royal Statistical Society, Series B, 50:* 157–224.

Learning Technology Center, Vanderbilt University. (1992). *The Adventures of Jasper Woodbury.* Retrieved January 25, 2005, from http://peabody.vanderbilt.edu/projects/funded/jasper/intro/jasperintro.html.

Massachusetts Department of Education. (2001). *Massachusetts Science and Technology Engineering Framework.* Retrieved April 11, 2005, from http://www.doe.mass.edu/frameworks.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. (1953). Equations of State Calculations by Fast Computing Machines. *Journal of Chemical Physics, 21*(6): 1087–1092.

Mislevy, R.J. (1991). Randomization-Based Inference About Latent Variables From Complex Samples. *Psychometrika, 56*(2): 177–196.

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.

Mislevy, R.J., Steinberg, L.S., Almond, R.G., Breyer, F.J., and Johnson, L. (2001, March). *Making Sense of Data From Complex Assessments* (CSE Technical Report 538). Retrieved January 25, 2005, from http://www.cse.ucla.edu/CRESST/Reports/RML%20TR%20538.pdf.

Muraki, E., and Bock, R.D. (1997). *PARSCALE: IRT Item Analysis and Test Scoring for Rating Scale Data* [computer software]. Chicago, IL: Scientific Software International.

National Academy of Sciences. (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.

Neal, R.M. (2003). Slice Sampling (with discussion). *Annals of Statistics. 31*(3): 705–767.

Nichols, P., and Sugrue, B. (1999). The Lack of Fidelity Between Cognitively Complex Constructs and Conventional Test Development Practice. *Educational Measurement: Issues and Practice, 18*(2): 18–29.

Noetic Systems, Inc. (2001). ERGO [computer software]. Baltimore, MD: Author.

North Carolina State Department of Education (2004). *Science Standard Course of Studies and Grade Level Competencies.* Retrieved April 11, 2005, from http://www.ncpublicschools.org/curriculum/science.

O'Sullivan, C.Y., Lauko, M.A., Grigg, W.S., Qian, J., and Zhang, J. (2003). *The Nation's Report Card: Science 2000* (NCES 2003–453). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics.

Olson, S., and 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/.

Patz, R.J., and Junker, B.W. (1999). Applications and Extensions of MCMC in IRT: Multiple Item Types, Missing Data, and Rated Responses: *Journal of Educational and Behavioral Statistics, 24*(4):342–366.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* San Mateo, CA: Morgan Kaufmann.

Pellegrino, J.W., Jones, L.R., and Mitchell, K.J. (Eds.). (1999). *Grading the Nation's Report Card: Evaluating NAEP and Transforming the Assessment of Educational Progress.* Washington, DC: National Academy Press.

Raghavan, K., Sartoris, M.L., and 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*(5): 547–567.

Riley, R.W., Holleman, F.S., and 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/offices/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.

Samejima, F. (1969). Estimation of Latent Ability Using a Response Pattern of Graded Scores. *Psychometrika, Monograph No. 17, 34*(4, Part 2).

Sandene, B., Bennett, R., Braswell, J., and Oranje, A. (2005). Online Assessment in Mathematics. In B. Sandene, N. Horkay, R. Bennett, N. Allen, J. Braswell, B. Kaplan, and A. Oranje. *Online Assessment in Mathematics and Writing: Reports From the NAEP Technology-Based Assessment Project, Research and Development Series* (NCES 2005–457) (Part I). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Schacter, J., Chung, G.K.W.K., and 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., Duschl, R.A., Schulze, S., and John, J. (1995). Students' Understanding of the Objectives and Procedures of Experimentation in the Science Classroom. *The Journal of the Learning Sciences, 4*(2): 131–166.

Schauble, L., Glaser, R., Raghavan, K., and 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., and Reiner, M. (1992). The Integration of Knowledge and Experimentation Strategies in Understanding a Physical System. *Applied Cognitive Psychology, 6:* 321–343.

Scott, S.L., and Ip, E.H. (2002). Empirical Bayes and Item-Clustering Effects in a Latent Variable Hierarchical Model: A Case Study From the National Assessment of Educational Progress. *Journal of the American Statistical Association, 97:* 409–419.

Shute, V.J., and Glaser, R. (1990). A Large Scale Evaluation of an Intelligent Discovery World: Smithtown. *Interactive Learning Environments, 1*(March): 51–77.

Shute, V.J., and 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., and Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28*(3): 237–247.

Spiegelhalter, D.J., Thomas, A., Best, N.G., and Gilks, W.R. (2004). *BUGS: Bayesian Inference Using Gibbs Sampling*. Version 1.4 [computer software]. Cambridge, UK: MRC Biostatistics Unit.

Thissen, D., Steinberg, L., and Mooney, J. (1989). Trace lines for testlets: A use of multiple-categorical response models. *Journal of Educational Measurement, 26:* 247–260.

Tierney, L. (1994). Markov Chains for Exploring Posterior Distributions. *Annals of Statistics, 22:* 1701–1762.

U.S. Department of Education, National Center for Education Statistics (2005, June). *Issue Brief: Rates of Computer and Internet Use by Children in Nursery School and Students in Kindergarten Through Twelfth Grade: 2003* (NCES 2005–111).

White, B.Y., and Frederiksen, J.R. (1998). Inquiry, Modeling, and Metacognition: Making Science Accessible to All Students. *Cognition and Instruction, 16*(1): 3–118.

THIS PAGE INTENTIONALLY LEFT BLANK.

Appendix A: Development Committee for the Problem Solving in Technology-Rich Environments (TRE) Study

List of Committee Members

Paul Cohen	Director of Curriculum Pascack Valley Regional High School District, New Jersey
Karen Cooper	Librarian and Technology Coordinator Montgomery Middle School, New Jersey
Lamont Fuchs	Technology Director Buncombe County Schools, North Carolina
Kathleen Gibbs	Middle School Science Teacher John Witherspoon Middle School, New Jersey
Cheryl Lemke	President and CEO, The Metiri Group, California
Christopher Manno	Assistant Superintendent Montgomery Township Schools, New Jersey
Kevin Mattingly	Middle School Science Teacher The Lawrenceville School, New Jersey
Joan Mazur	Assistant Professor University of Kentucky Department of Curriculum & Instruction, Kentucky
Diane Reed	Technology Teacher in Residence Portage Path School of Technology Akron, Ohio
Ismael Salas	Career and Technology Instructor Fabens Independent School District, Texas
Martha Veale	Director Technology/Career Education Canutillo Independent School District, Texas
Randy Bennett	ETS Staff
Bob Evans	NCES Project Monitor
Vonda Kiplinger	NCES Project Monitor
Hilary Persky	ETS Staff
Holly Spurlock	NCES Project Monitor

Appendix B: Sample Selection

The TRE 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.[1] The third and final stage involved selection of students within schools and their assignment to either the Search scenario or the Simulation scenario.

Fifty-two primary sampling units (PSUs) were included in the first stage, with the 10 largest PSUs being certainty PSUs and the remaining 42 noncertainty PSUs. The schools were selected systematically from a sorted list with probabilities proportional to assigned measures of size. To increase cost-efficiency in sampling, samples were designed to include more relatively large schools. Also, because the TRE administration was so different from the traditional NAEP assessment, school selection probabilities were adjusted so that the TRE sample overlapped as little as possible with the main 2003 NAEP assessment. The selection procedure resulted in a sample of 270 schools, 222 of which participated in the assessment, for a weighted cooperation rate of 85.1 percent.

From the 222 participating schools, 2,409 students were selected to participate in the study. Of these students, 150 were nonrespondents. An additional 125 students were excluded who could not participate in the assessment as it was normally conducted. The weighted exclusion rate for such students was 4.8 percent. After accounting for excluded students and nonrespondents, the total number of students assessed was 2,134, resulting in a weighted student participation rate of 93.5 percent. 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.

When resulting data files were examined, it was found that, for unknown reasons, 25 students did not have scenario data and that 1 student, who was mistakenly coded as a nonrespondent, actually did have scenario data but no sampling weights. This resulted in a total number of students with data of 2,110 but sampling weights for only 2,109. Results reported in chapters 5 and 6 used the sample of 2,109.

Assignment to the Search and Simulation scenarios within schools was random. The number of students taking the Search scenario was 1,077. The number taking the Simulation scenario was 1,033, including the student without a sampling weight.

[1] County equivalents refer to the Anchorage Municipality and all Boroughs and Census Areas in Alaska, the District of Columbia, all Parishes in Louisiana, and all Independent Cities in Virginia, as well as Baltimore City, Maryland; St. Louis, Missouri; and Carson City, Nevada.

Appendix C: Technical Specifications for Participating Schools

Hardware

- Internet connection: Dedicated line (non-dial up) 200Kb per second or greater
- Computers: PC with Pentium Class 266 megahertz microprocessor or better (Macintosh computers were not acceptable.)
- Memory: 32MB or greater for Windows 95 and 98; 64MB or greater for other operating systems
- Operating system: Windows 95, Windows 98, Windows ME, Windows NT, Windows 2000, or Windows XP
- Hard drives: 10MB free disk space
- Graphics capabilities: SVGA support – 1024 x 768 resolution with minimum 65536 (16 bit) colors

Software

- Web browser: Microsoft's Internet Explorer Version 5.0 or later.[1]

[1] Some minor enhancements to Internet Explorer were required. These were installed during the certification process if they were not already present. The enhancements included:

- Macromedia Flash 5.0 Player
- Microsoft Virtual Machine (Java)

Appendix D: Prior Knowledge and Background Questions for Search and Simulation Scenarios

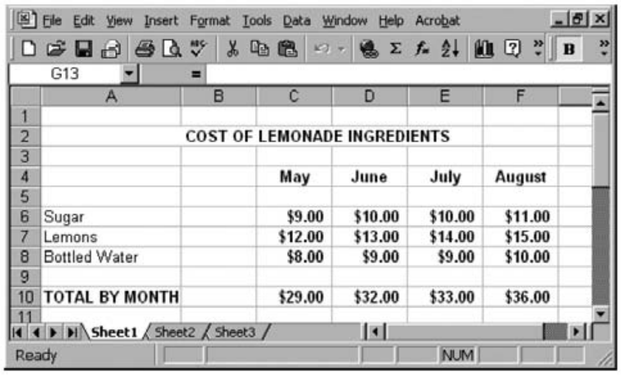
The correct answers to prior knowledge questions in this appendix are shown in bold.

Problem Solving in Technology-Rich Environments (TRE) Search Scenario and Simulation Scenario Prior Computer Knowledge Questions

1. What is the main role of a computer program?
   - A. To put data into the computer
   - B. To give the computer a memory
   - **C. To tell the computer what to do**
   - D. To let the computer know if it is doing a good job

> Put dough in a pie dish. Grease pie dish. Open can of cherry pie filling and pour it in pie dish. Bake at 350 degrees for 45 minutes and let cool.

2. In the recipe above, the words "Grease pie dish" should go before "Put dough in a pie dish." What is the best way to fix this problem using your word processor?
   - A. Search and Replace
   - **B. Move (or Cut and Paste)**
   - C. Insert
   - D. Delete



	A	B	C	D	E	F
1						
2		COST OF LEMONADE INGREDIENTS				
3						
4			May	June	July	August
5						
6	Sugar		$9.00	$10.00	$10.00	$11.00
7	Lemons		$12.00	$13.00	$14.00	$15.00
8	Bottled Water		$8.00	$9.00	$9.00	$10.00
9						
10	TOTAL BY MONTH		$29.00	$32.00	$33.00	$36.00
11						

3. Pat has made the spreadsheet above to calculate the cost of supplies for a lemonade stand open from May through August.

   What should Pat do to calculate the total cost of lemons for all four months?
   - A. Calculate the sum of cells F6 through F10.
   - B. Calculate the sum of cells A7 through C7.
   - C. Calculate the sum of cells C6 through F6.
   - **D. Calculate the sum of cells C7 through F7**.

4. Your teacher has asked you to do a web search to find out about what African elephants eat. Which of the following search terms would likely return the most relevant pages?
   - A. African elephant
   - B. Elephant diet
   - C. Elephant
   - **D. Diet African elephant**

5. What does the web search query elephant OR tiger mean?
   - A. Find pages with references to both elephants and tigers.
   - **B. Find pages with references to either elephants or tigers**.
   - C. Find pages with references to elephants or tigers, but not both.
   - D. Find pages with elephant and tiger in the page title.

6. When talking about the Internet, what is a "link"?

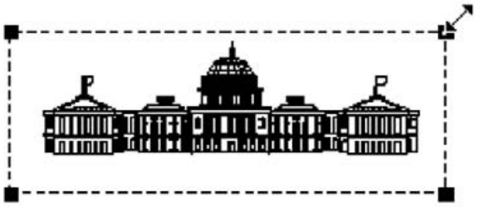
A. The cables connecting computers together
B. The missing information in a document
C. A connection between web pages
D. A kind of email message

7. After you enter a search query, you get a list of hits. Where in the list of hits are you likely to find information most related to your query?

A. At the bottom of the list
B. In the middle of the list
C. Anywhere on the list
D. At the top of the list

8. In order to automatically repeat the same text at the bottom of each page of a multipage report you need to

A. use a footer
B. use a header
C. place it in a table
D. type in outline mode



9. By clicking and dragging on the point indicated by the arrow, the user will be able to

A. change the color
B. cut the graphic
C. resize the graphic
D. paste the graphic

10. What is a "URL"?

A. A computer processor
B. A security password
C. An internet address
D. A computer monitor

Problem Solving in Technology-Rich Environments (TRE) Simulation Scenario Prior Science Knowledge Questions

1. Which of the following is the best example of the concept of mass?

A. The amount of space that a liquid takes up
B. The energy it takes a person to carry an object
C. The amount of material in an object
D. The length of a piece of material

2. Which statement best describes what happens to a specific amount of gas when it is moved from a larger to a smaller closed container?

A. The mass of the gas decreases.
B. The temperature of the gas decreases.
C. The density of the gas increases.
D. The volume of the gas increases.

3. A rubber gas balloon can hold 10 cubic feet of helium. Ellen puts 5 cubic feet of helium inside the balloon, so its starting volume is 5 cubic feet. The balloon rises and expands. When the balloon stops rising, its final volume is 10 cubic feet.

Why did the balloon volume change from start to finish?

A. As the balloon rises, decreasing air pressure allows the amount of helium gas inside the balloon to increase.
B. As the balloon rises, decreasing air pressure allows the helium inside the balloon to expand and push out the sides of the balloon.
C. As the balloon rises, increasing air pressure makes the helium gas inside the balloon denser and therefore heavier.
D. As the balloon rises, increasing air pressure makes the helium gas inside the balloon less dense so it expands.

4. Brad thinks that water will evaporate at different rates depending on the temperature of a room. If he wants to do an experiment to test his idea, what would be the best experimental set up?

   A. Put equal amounts of water at the same temperature in bowls of different sizes, each in a different room with each room having a different temperature and a different humidity.

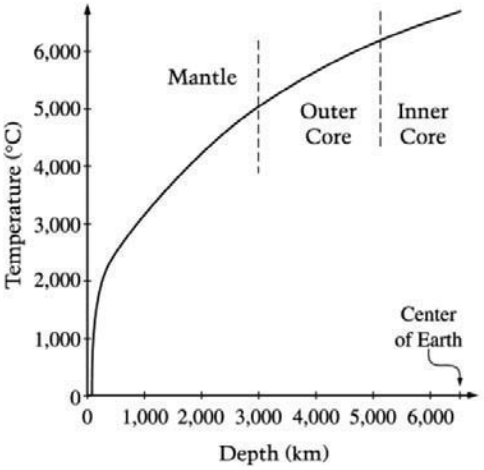
   B. Put equal amounts of water at the same temperature in bowls of equal size, each in a different room with each room having a different temperature but the same humidity.

   C. Put equal amounts of water at the same temperature in bowls of equal size, each in a different room with each room having the same temperature but different humidity.

   D. Put equal amounts of water at the same temperature in bowls of different sizes, each in a different room with each room having the same temperature and the same humidity.

The graph below shows the change in temperature inside the Earth as the depth below the surface increases.

Graph 1: Change in Temperature with Increasing Depth Below Earth's Surface



5. Which of the following is true of the temperature inside the Earth?

   A. It increases rapidly with depth near the surface, then remains constant.

   B. It increases rapidly with depth near the surface, then it increases more slowly in the inner layers.

   C. It increases slowly with depth near the surface, then it increases more rapidly in the inner layers.

   D. It increases with depth at a constant rate.

6. Which statement best describes what makes a gas balloon rise into the air?

   A. The gas inside the balloon decreases in volume as the balloon rises into the air.

   B. The temperature of the air increases as the balloon rises into the air.

   C. The mass of the balloon material is greater than the mass of the gas inside the balloon.

   D. The density of the air surrounding the balloon is greater than the density of the gas inside the balloon.

Questions 7–9 refer to the description below.

A scientist questioned the ability of fish raised in a hatchery (farm) to survive in the wild. She believed the fish raised in hatcheries had lost their fear of predators.

To test her idea, she placed 15 hatchery salmon and 15 wild salmon of the same age into two separate but identical tanks. She then placed a clear piece of plastic into each tank. In each tank, she put the salmon on one side of the plastic and a large predatory fish, the cod, on the other side of the plastic. She then recorded the amount of time it took the salmon in each tank to move to the back of the tank away from the cod.

She found that the hatchery fish were much slower in moving away than the wild fish. This led her to believe that the hatchery fish have less fear of predators than do wild fish.

7. What is a control in the experiment?

A. The hatchery salmon
B. The wild salmon
C. The time it took the wild salmon to move away from the cod
D. The time it took the hatchery salmon to move away from the cod

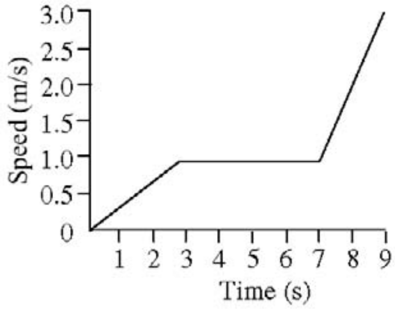
8. What is the hypothesis in the experiment?

A. Wild fish have less fear of predators than hatchery fish.
B. Hatchery fish have lost their fear of predators.
C. Hatchery fish will move rapidly away from predators placed in their tanks.
D. Wild fish will survive attacks from predators more often than hatchery fish.

9. What is the conclusion of the experiment?

A. Wild fish swim more rapidly than do hatchery fish.
B. Wild fish take more time to move away from predators than do hatchery fish.
C. Hatchery fish have less fear of predators than do wild fish.
D. Hatchery fish will be able to survive in a wild environment.

The graph below contains information about the movement of a bicycle.



10. At which time is the bicycle's speed constant?

A. At 1 second
B. At 2 seconds
C. At 4 seconds
D. At 8 seconds

Problem Solving in Technology-Rich Environments (TRE) Search Scenario Prior Science Knowledge Questions

1. Which statement best describes what happens to a specific amount of gas when it is moved from a larger to a smaller closed container?

A. The mass of the gas decreases.
B. The temperature of the gas decreases.
C. The density of the gas increases.
D. The volume of the gas increases.

2. What kind of gas would most likely be used to lift a balloon 10 miles into the sky?

A. Helium
B. Oxygen
C. Hot Air
D. Nitrogen

3. The main reason a scientist might prefer to observe distant stars from high above earth than from on the ground is because

A. the force of gravity is weaker
B. it is always nighttime high above earth
C. there is less interference from the atmosphere
D. it shortens the distance to the stars being observed

4. Which of the following physical forces is mostly responsible for pulling a balloon toward the ground?

A. Air resistance
B. Gravity
C. Atomic force
D. Magnetic force

5. A rubber balloon filled with air will sink to the ground. Which of the following actions would make the balloon rise?

A. Release the balloon from the top of a mountain.
B. Make the balloon out of lighter material.
C. Put more air into the balloon.
D. Heat the air in the balloon.

6. Which statement best describes what makes a gas balloon rise into the air?

A. The gas inside the balloon decreases in volume as the balloon rises into the air.
B. The temperature of the air increases as the balloon rises into the air.
C. The mass of the balloon material is greater than the mass of the gas inside the balloon.
D. The density of the air surrounding the balloon is greater than the density of the gas inside the balloon.

7. What will likely happen to a rubber balloon filled with gas as it rises into the air?

A. It will remain the same size.
B. It will shrink in size until it collapses.
C. It will expand in size until it bursts.
D. It will expand and then shrink.

8. Scientists interested in studying weather would most likely send a weather balloon into which part of the atmosphere?

A. Mesosphere
B. Stratosphere
C. Thermosphere
D. Troposphere

9. Scientists currently use gas balloons to collect information on which of the following?

A. Condition of the ozone layer
B. Effects of gravity on humans
C. Contents of craters on the Moon
D. Patterns of airplane traffic

10. One problem with using hydrogen gas in scientific balloons is that hydrogen gas

A. gives less lift than most other gases
B. is a rare and expensive gas
C. is highly explosive
D. turns to liquid as the balloon rises

Background Questions Used in the Search and Simulation Scenarios

Questions 1–8. To what extent do you do the following on a computer? Include things you do in school and things you do outside of school.

1. Play computer games

A. Not at all
B. Small extent
C. Moderate extent
D. Large extent

2. Write using a word processing program

A. Not at all
B. Small extent
C. Moderate extent
D. Large extent

3. Make drawings or art projects on the computer

A. Not at all
B. Small extent
C. Moderate extent
D. Large extent

4. Make tables, charts, and graphs on the computer

A. Not at all
B. Small extent
C. Moderate extent
D. Large extent

5. Look up information on a CD

A. Not at all
B. Small extent
C. Moderate extent
D. Large extent

6. Find information on the Internet for a project or report for school

A. Not at all
B. Small extent
C. Moderate extent
D. Large extent

7. Use e-mail to communicate with others

A. Not at all
B. Small extent
C. Moderate extent
D. Large extent

8. Talk in chat groups or with other people who are logged on at the same time

A. Not at all
B. Small extent
C. Moderate extent
D. Large extent

9. Who taught you the most about how to use a computer?

A. I learned the most on my own.
B. I learned the most from my friends.
C. I learned the most from my teachers.
D. I learned the most from my family.
E. I don't really know how to use a computer.

10. How often do you use a computer at school? Include use anywhere in the school and at any time of day.

A. Every day
B. Two or three times a week
C. About once a week
D. Once every few weeks
E. Never or hardly ever

11. How often do you use a computer outside of school?

A. Every day
B. Two or three times a week
C. About once a week
D. Once every few weeks
E. Never or hardly ever

12. Is there a computer at home that you use?

A. Yes
B. No

Questions 13–15. Please indicate the extent to which you AGREE or DISAGREE with the following statements.

13. I am more motivated to get started doing my schoolwork when I use a computer

A. Strongly agree
B. Agree
C. Disagree
D. Strongly disagree
E. I never use a computer.

14. I have more fun learning when I use a computer

A. Strongly agree
B. Agree
C. Disagree
D. Strongly disagree
E. I never use a computer.

15. I get more done when I use a computer for schoolwork

A. Strongly agree
B. Agree
C. Disagree
D. Strongly disagree
E. I never use a computer.

16. Which best describes you?

A. White (not Hispanic)
B. Black (not Hispanic)
C. Hispanic ("Hispanic" means someone who is from a Mexicano, Mexican America, Chicano, Puerto Rican, Cuban, or other Spanish or Hispanic background
D. Asian ("Asian" means someone who is from a Chinese, Japanese, Vietnamese, or other Asian background)
E. Pacific Islander ("Pacific Islander" means someone who is from a Filipino, Hawaiian, or other Pacific Islander background)
F. American Indian or Alaskan Native ("American Indian or Alaskan Native" means someone who is from one of the American Indian tribes, or one of the original people of Alaska)
G. Other

17. If you are Hispanic, what is your Hispanic background?

A. I am not Hispanic.
B. Mexican, Mexican America, or Chicano
C. Puerto Rican
D. Cuban
E. Other Spanish or Hispanic background

18. How far in school did your mother go?

A. She did not finish high school.
B. She graduated from high school.
C. She had some education after high school.
D. She graduated from college.
E. I don't know.

19. How far in school did your father go?

A. He did not finish high school.
B. He graduated from high school.
C. He had some education after high school.
D. He graduated from college.
E. I don't know.

20. About how many books are there in your home?

A. Few (0–10)
B. Enough to fill one shelf (11–25)
C. Enough to fill one bookcase (26–100)
D. Enough to fill several bookcases (more than 100)

21. Does your family get a newspaper at least four times a week?

A. Yes
B. No
C. I don't know.

22. Does your family get any magazines regularly?

A. Yes
B. No
C. I don't know.

23. Is there an encyclopedia in your home? It could be a set of books, or it could be on the computer.

A. Yes
B. No
C. I don't know.

24. On a school day, about how many hours do you usually watch TV or videotapes outside of school?

A. None
B. I hour or less
C. 2 or 3 hours
D. 4 or 5 hours
E. 6 hours or more

25. Which best describes the science course you are taking this year?

A. I am not taking a science course this year.
B. Life science (for example, biology)
C. Physical science (for example, physics or chemistry)
D. Earth science (for example, geology or astronomy)
E. General science (several content areas of science taught separately)
F. Integrated science (several content areas of science combined and taught throughout the year)

Questions 26–29. About how often do you do each of the following in your science class?

26. Design your own science experiment or investigation

A. I am not taking science.
B. Once a month or more
C. Sometimes, but less than once a month
D. Never

27. Carry out the science experiment or investigation you designed

A. I am not taking science.
B. Once a month or more
C. Sometimes, but less than once a month
D. Never

28. Write up results of the experiment or investigation you designed

A. I am not taking science.
B. Once a month or more
C. Sometimes, but less than once a month
D. Never

29. Talk to class about the results of your experiment or investigation

A. I am not taking science.
B. Once a month or more
C. Sometimes, but less than once a month
D. Never

Questions 30–34. If you are taking a science class this year, about how often do you use a computer to do the following?

30. Collect data using lab equipment that interfaces with computers (for example, probes)

A. I am not taking science.
B. Once a month or more
C. Sometimes, but less than once a month
D. Never

31. Download data and related information from the Internet

A. I am not taking science.
B. Once a month or more
C. Sometimes, but less than once a month
D. Never

32. Analyze data using the computer

A. I am not taking science.
B. Once a month or more
C. Sometimes, but less than once a month
D. Never

33. Use the Internet to exchange information with other students or scientists about science experiments or investigations

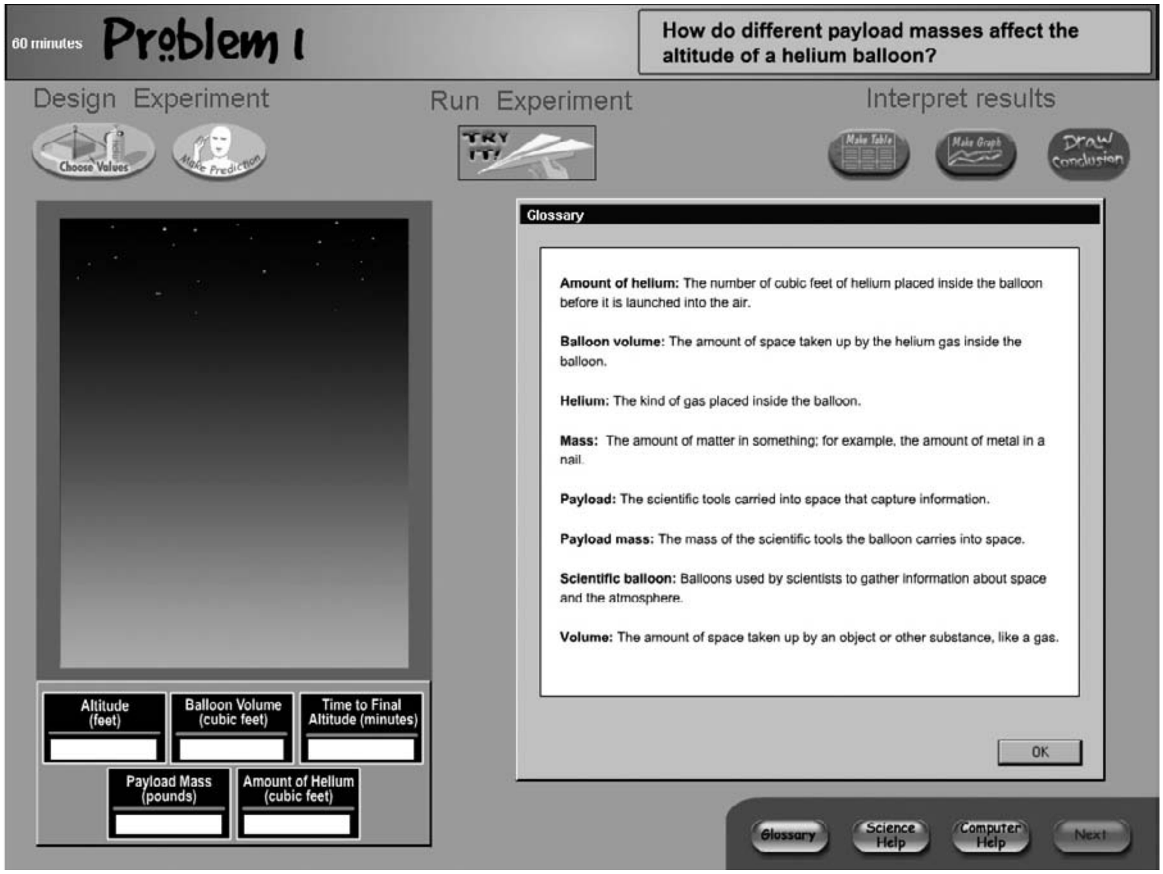
A. I am not taking science.
B. Once a month or more
C. Sometimes, but less than once a month
D. Never

34. Use computer simulations to perform experiments or explore science topics

A. I am not taking science.
B. Once a month or more
C. Sometimes, but less than once a month
D. Never

Appendix E: TRE Simulation Glossary, Help, and Tutorial Screens

Figure E-1. Computer screen showing the TRE Simulation glossary, grade 8: 2003



Amount of helium: The number of cubic feet of helium placed inside the balloon before it is launched into the air.

Balloon volume: The amount of space taken up by the helium gas inside the balloon.

Helium: The kind of gas placed inside the balloon.

Mass: The amount of matter in something; for example, the amount of metal in a nail.

Payload: The scientific tools carried into space that capture information.

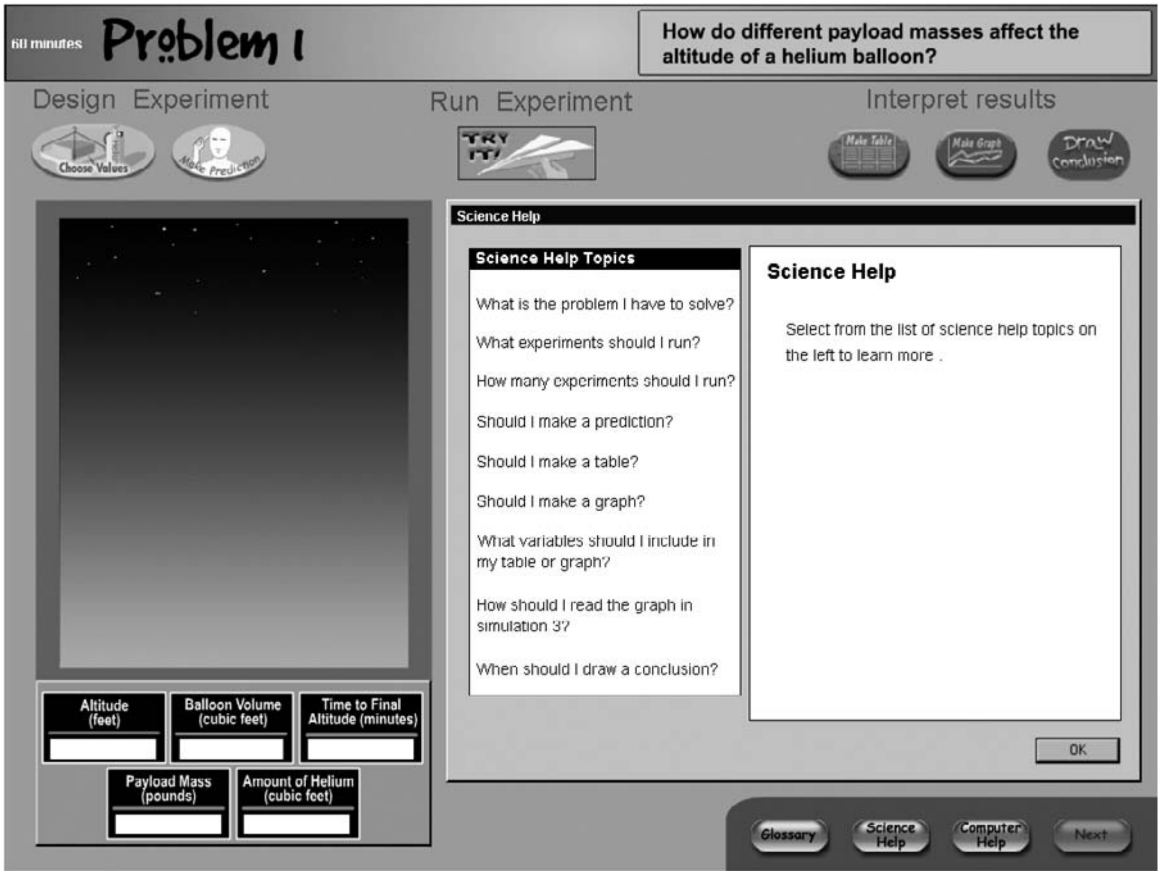
Payload mass: The mass of the scientific tools the balloon carries into space.

Scientific balloon: Balloons used by scientists to gather information about space and the atmosphere.

Volume: The amount of space taken up by an object or other substance, like a gas.

NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure E-2. Computer screen showing the TRE Simulation Science Help topics menu, grade 8: 2003



How do different payload masses affect the altitude of a helium balloon?

Science Help Topics

What is the problem I have to solve?

What experiments should I run?

How many experiments should I run?

Should I make a prediction?

Should I make a table?

Should I make a graph?

What variables should I include in my table or graph?

How should I read the graph in simulation 3?

When should I draw a conclusion?

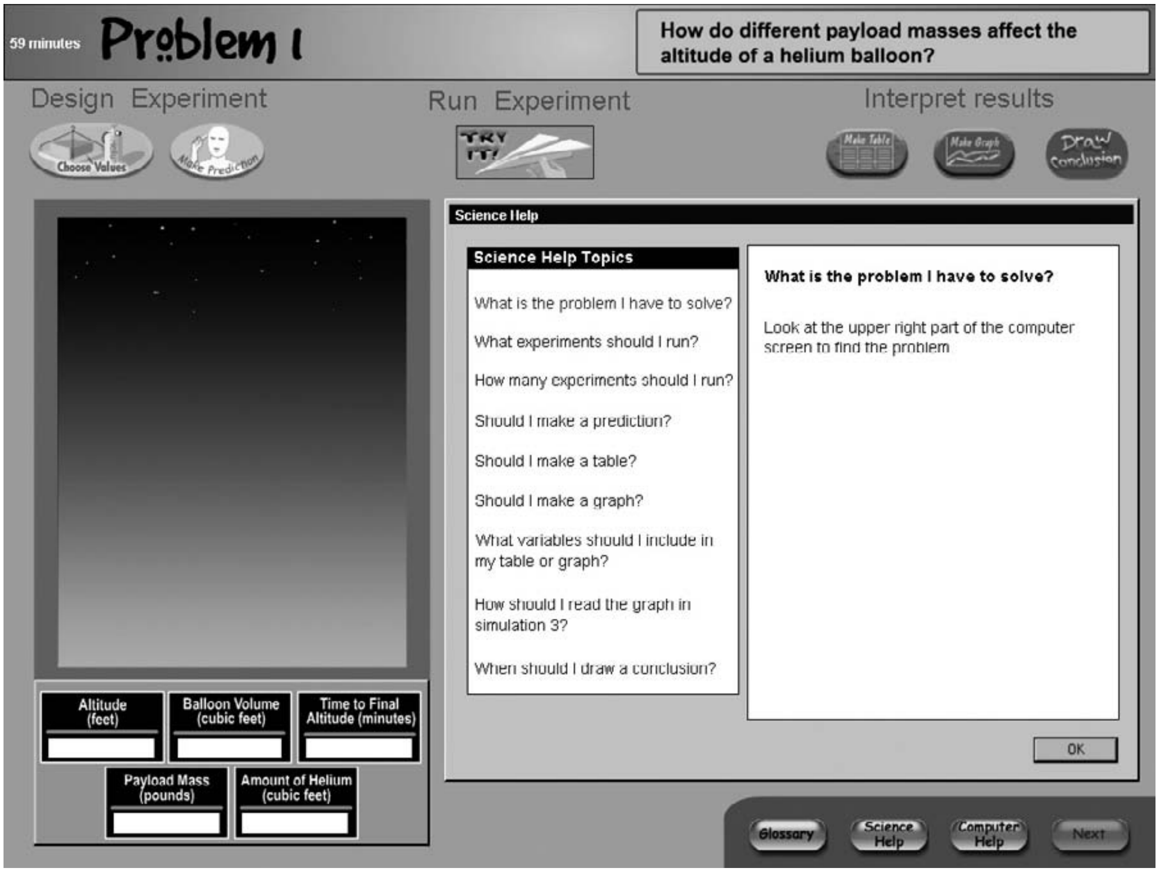
Science Help

Select from the list of science help topics on the left to learn more .

NOTE: TRE = Technology-Rich Environments.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure E-3. Computer screen showing help for the first TRE Simulation Science Help topic, grade 8: 2003



How do different payload masses affect the altitude of a helium balloon?

Science Help

Science Help Topics

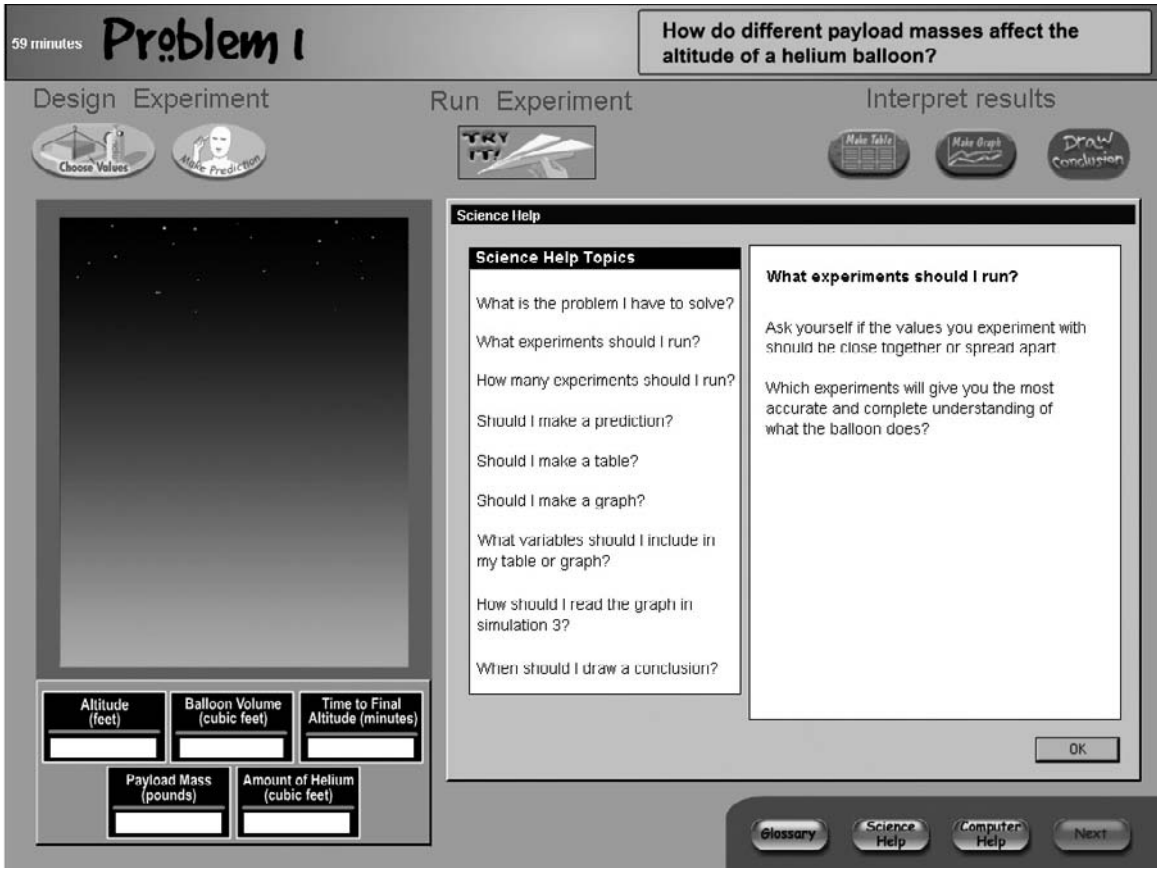
- What is the problem I have to solve?
- What experiments should I run?
- How many experiments should I run?
- Should I make a prediction?
- Should I make a table?
- Should I make a graph?
- What variables should I include in my table or graph?
- How should I read the graph in simulation 3?
- When should I draw a conclusion?

What is the problem I have to solve?

Look at the upper right part of the computer screen to find the problem.

NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure E-4. Computer screen showing help for the second TRE Simulation Science Help topic, grade 8: 2003



Science Help

Science Help Topics

- What is the problem I have to solve?
- What experiments should I run?
- How many experiments should I run?
- Should I make a prediction?
- Should I make a table?
- Should I make a graph?
- What variables should I include in my table or graph?
- How should I read the graph in simulation 3?
- When should I draw a conclusion?

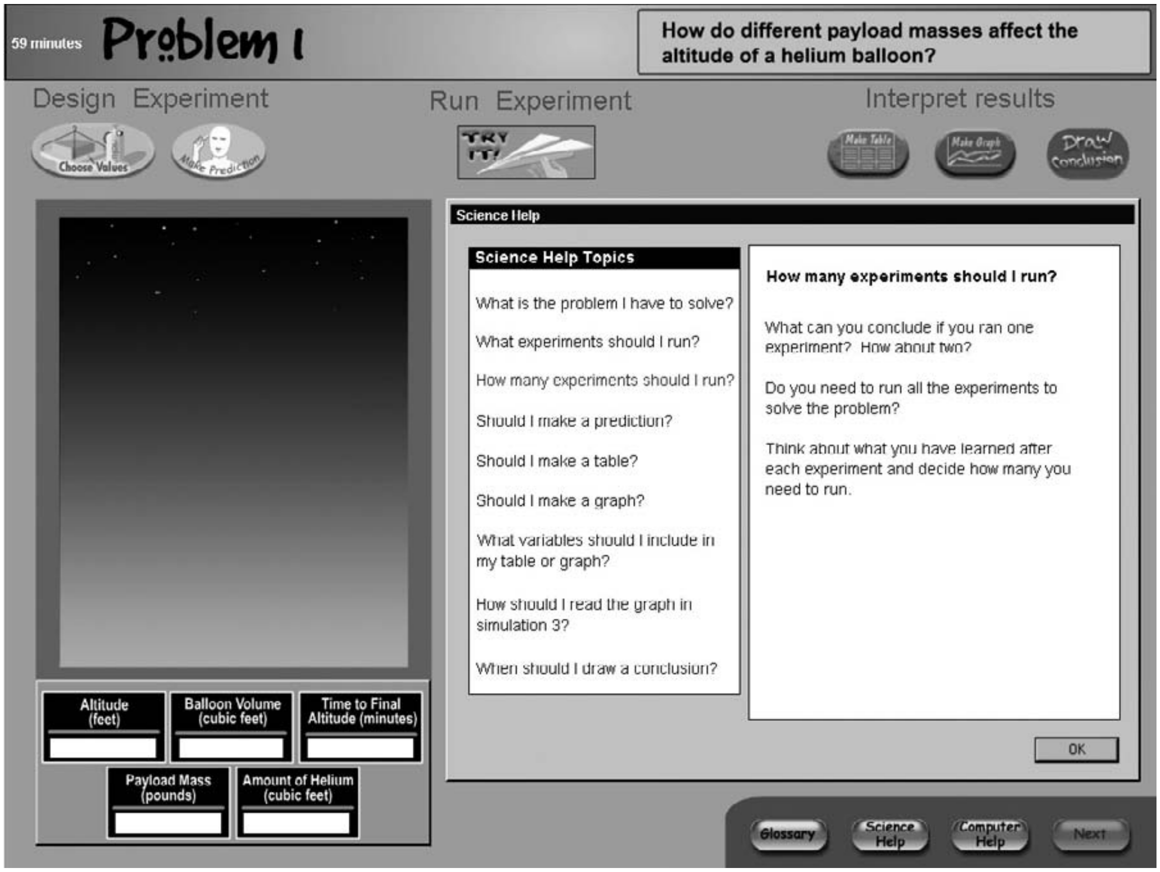
What experiments should I run?

Ask yourself if the values you experiment with should be close together or spread apart.

Which experiments will give you the most accurate and complete understanding of what the balloon does?

NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure E-5. Computer screen showing help for the third TRE Simulation Science Help topic, grade 8: 2003



What is the problem I have to solve?

What experiments should I run?

How many experiments should I run?

Should I make a prediction?

Should I make a table?

Should I make a graph?

What variables should I include in my table or graph?

How should I read the graph in simulation 3?

When should I draw a conclusion?

How many experiments should I run?

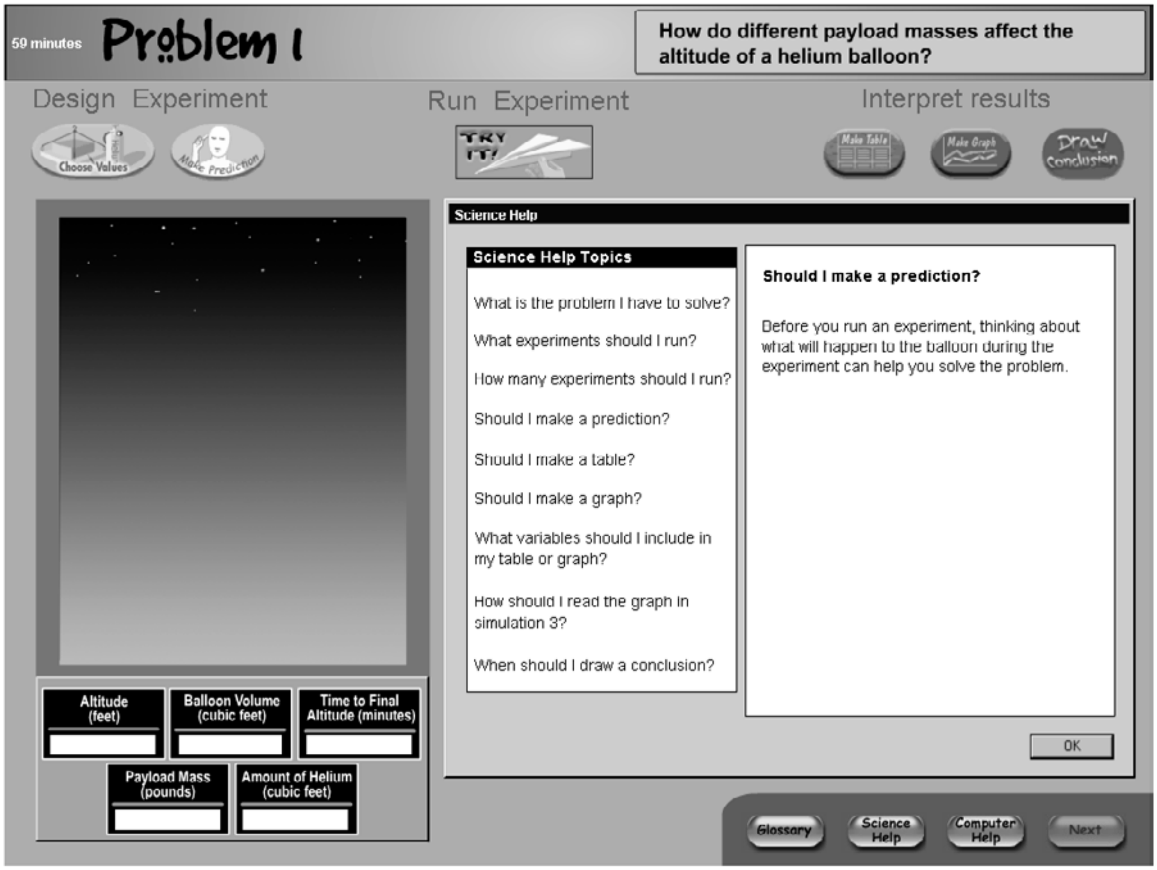
What can you conclude if you ran one experiment? How about two?

Do you need to run all the experiments to solve the problem?

Think about what you have learned after each experiment and decide how many you need to run.

NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure E-6. Computer screen showing help for the fourth TRE Simulation Science Help topic, grade 8: 2003



How do different payload masses affect the altitude of a helium balloon?

Science Help

Science Help Topics

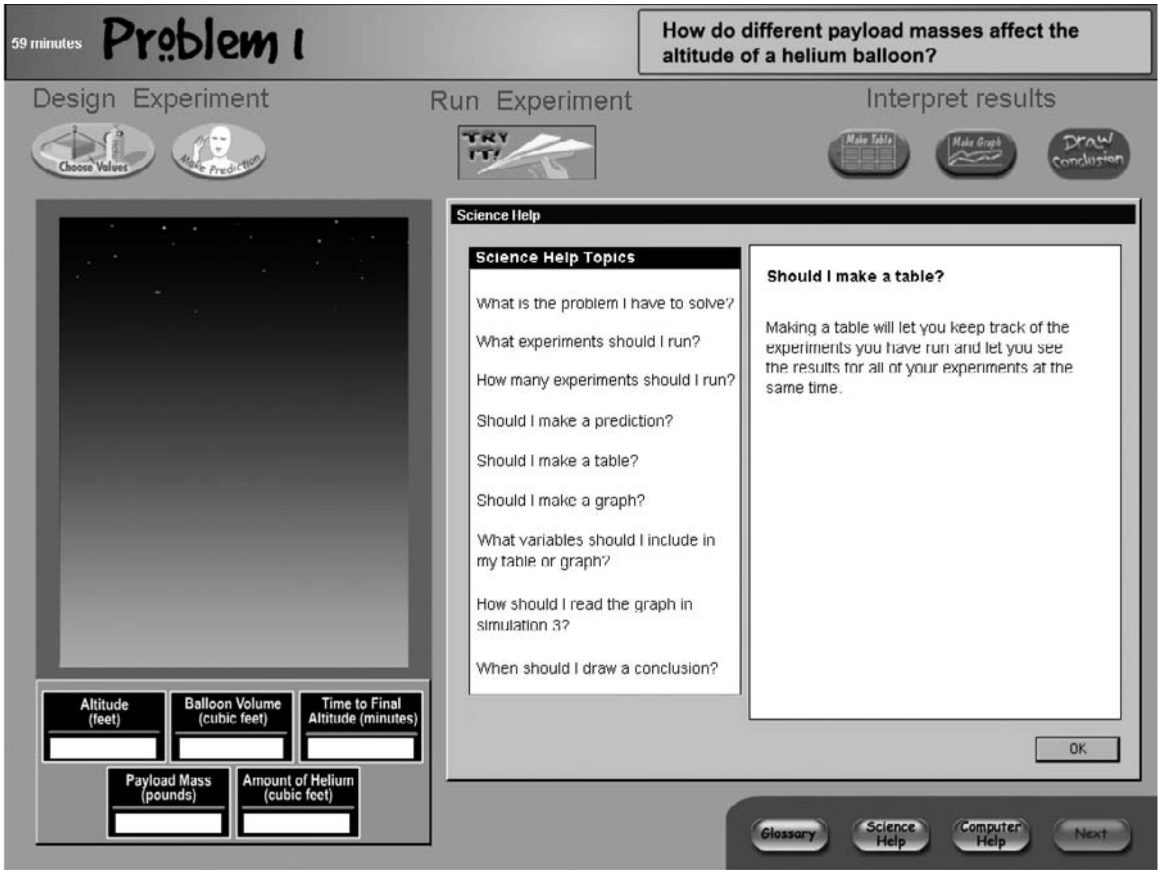
- What is the problem I have to solve?
- What experiments should I run?
- How many experiments should I run?
- Should I make a prediction?
- Should I make a table?
- Should I make a graph?
- What variables should I include in my table or graph?
- How should I read the graph in simulation 3?
- When should I draw a conclusion?

Should I make a prediction?

Before you run an experiment, thinking about what will happen to the balloon during the experiment can help you solve the problem.

NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure E-7. Computer screen showing help for the fifth TRE Simulation Science Help topic, grade 8: 2003



How do different payload masses affect the altitude of a helium balloon?

Science Help

Science Help Topics

What is the problem I have to solve?

What experiments should I run?

How many experiments should I run?

Should I make a prediction?

Should I make a table?

Should I make a graph?

What variables should I include in my table or graph?

How should I read the graph in simulation 3?

When should I draw a conclusion?

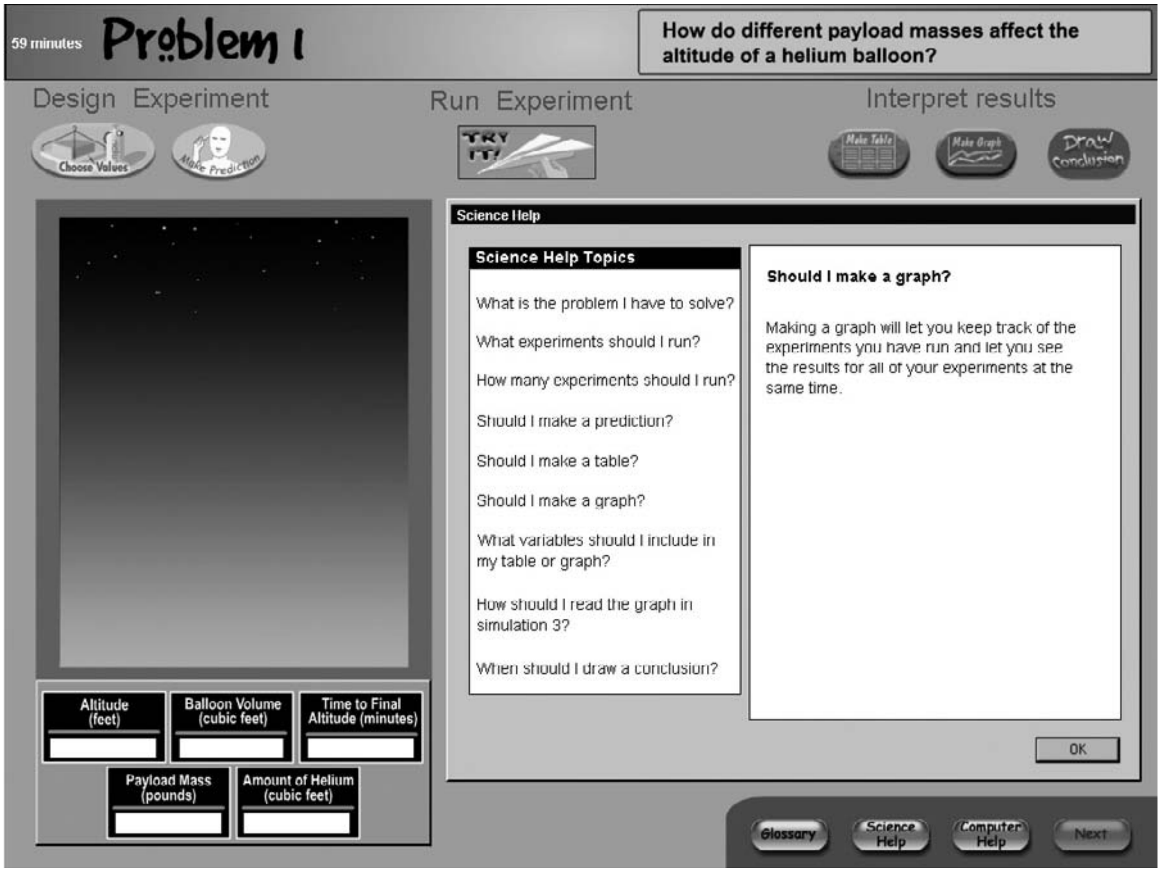
Should I make a table?

Making a table will let you keep track of the experiments you have run and let you see the results for all of your experiments at the same time.

NOTE: TRE = Technology-Rich Environments.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure E-8. Computer screen showing help for the sixth TRE Simulation Science Help topic, grade 8: 2003



How do different payload masses affect the altitude of a helium balloon?

Science Help

Science Help Topics

What is the problem I have to solve?

What experiments should I run?

How many experiments should I run?

Should I make a prediction?

Should I make a table?

Should I make a graph?

What variables should I include in my table or graph?

How should I read the graph in simulation 3?

When should I draw a conclusion?

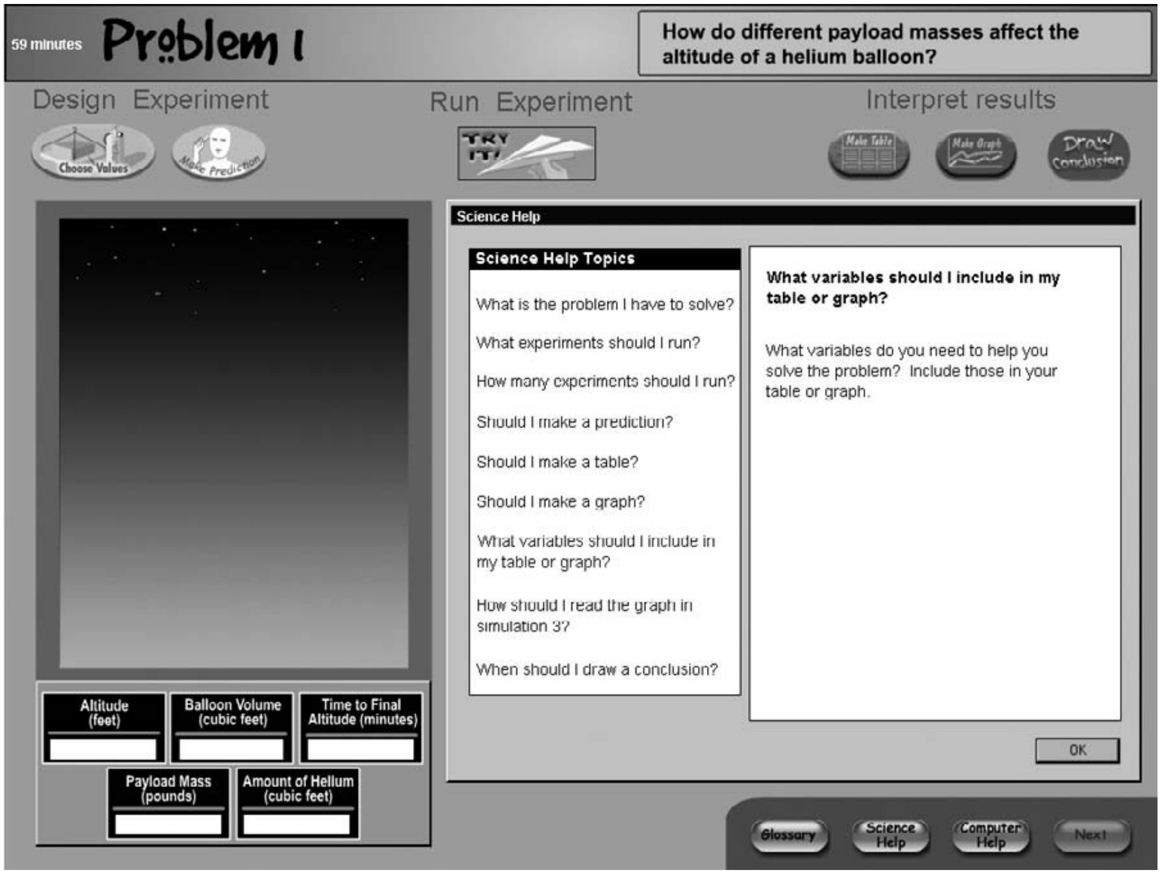
Should I make a graph?

Making a graph will let you keep track of the experiments you have run and let you see the results for all of your experiments at the same time.

NOTE: TRE = Technology-Rich Environments.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure E-9. Computer screen showing help for the seventh TRE Simulation Science Help topic, grade 8: 2003



How do different payload masses affect the altitude of a helium balloon?

Science Help Topics

What is the problem I have to solve?

What experiments should I run?

How many experiments should I run?

Should I make a prediction?

Should I make a table?

Should I make a graph?

What variables should I include in my table or graph?

How should I read the graph in simulation 3?

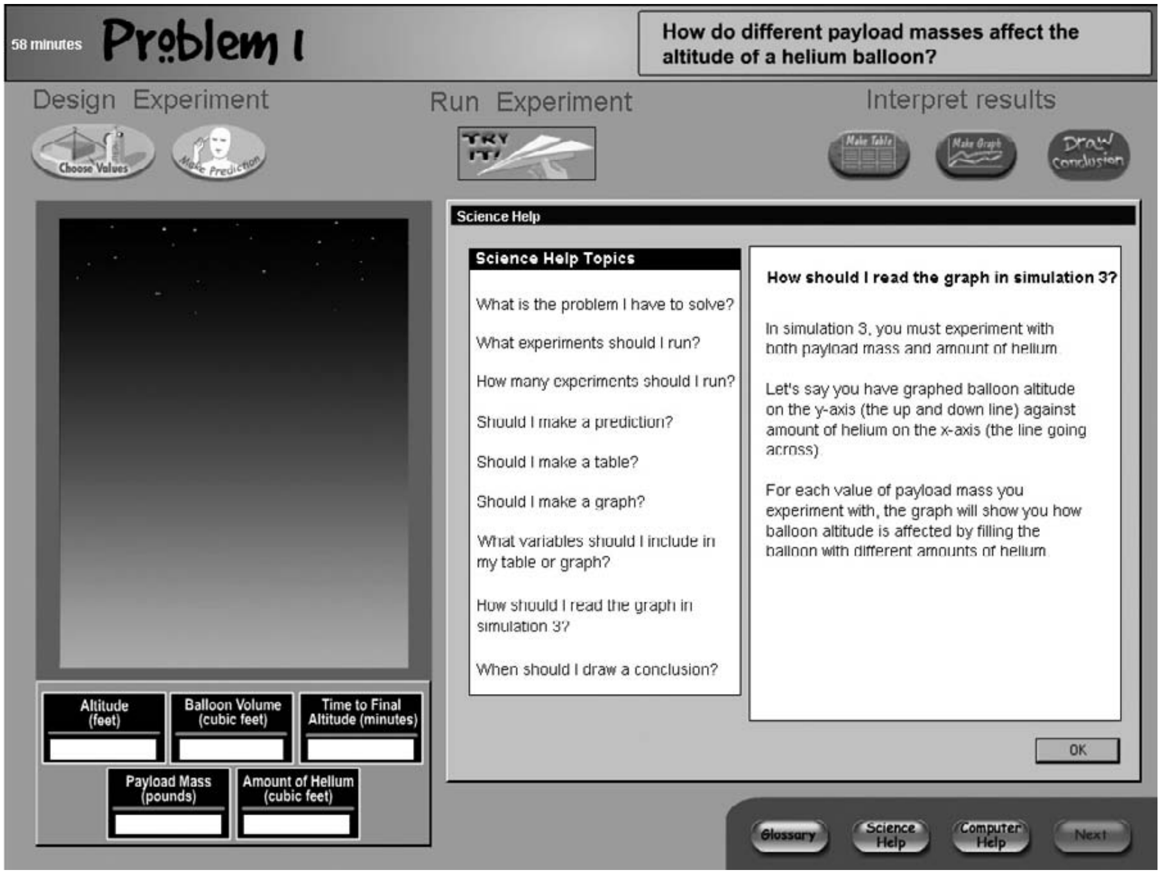
When should I draw a conclusion?

What variables should I include in my table or graph?

What variables do you need to help you solve the problem? Include those in your table or graph.

NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure E-10. Computer screen showing help for the eighth TRE Simulation Science Help topic, grade 8: 2003



How do different payload masses affect the altitude of a helium balloon?

Science Help

Science Help Topics

What is the problem I have to solve?

What experiments should I run?

How many experiments should I run?

Should I make a prediction?

Should I make a table?

Should I make a graph?

What variables should I include in my table or graph?

How should I read the graph in simulation 3?

When should I draw a conclusion?

How should I read the graph in simulation 3?

In simulation 3, you must experiment with both payload mass and amount of helium.

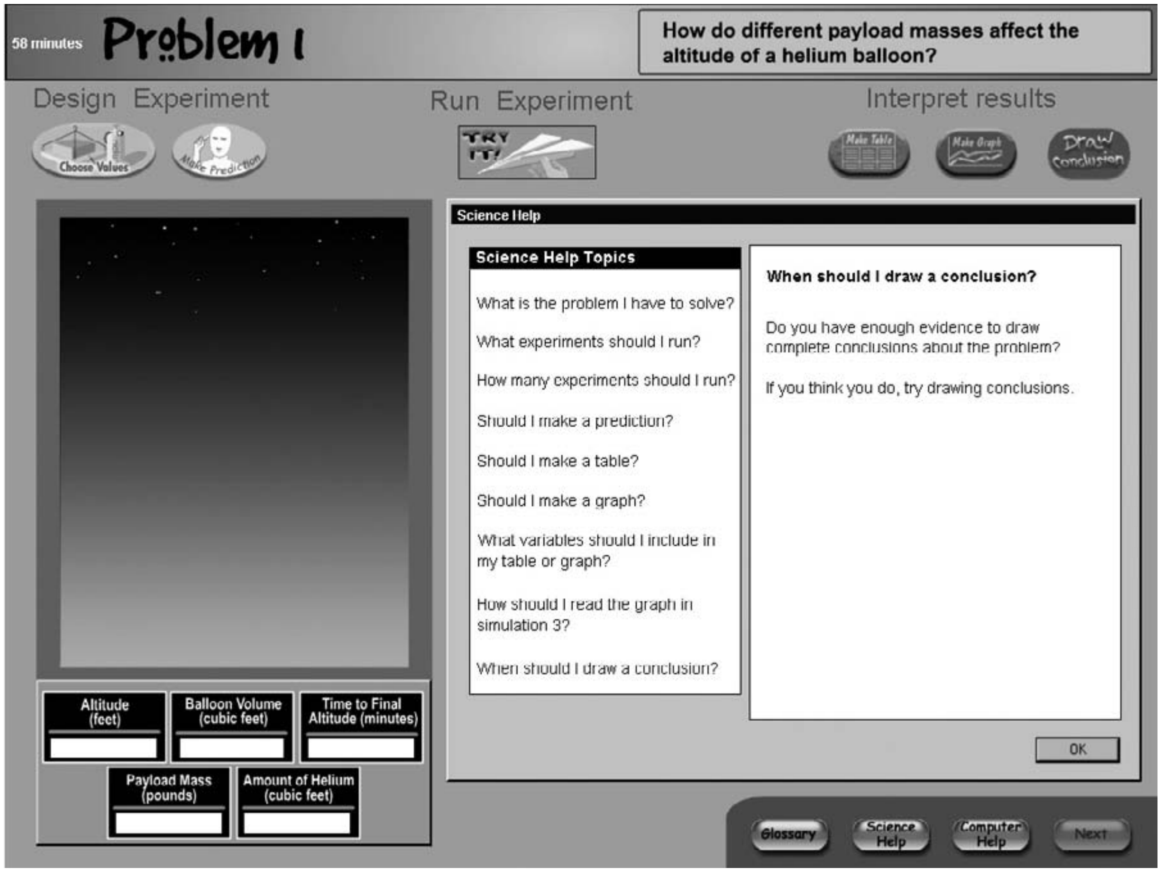
Let's say you have graphed balloon altitude on the y-axis (the up and down line) against amount of helium on the x-axis (the line going across).

For each value of payload mass you experiment with, the graph will show you how balloon altitude is affected by filling the balloon with different amounts of helium.

NOTE: TRE = Technology-Rich Environments.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure E-11. Computer screen showing help for the ninth TRE Simulation Science Help topic, grade 8: 2003



58 minutes
Problem 1
How do different payload masses affect the altitude of a helium balloon?
Design Experiment
Choose Values
Make Prediction
Run Experiment
TRY IT!
Interpret results
Make Table
Make Graph
Draw Conclusion
Altitude (feet)
Balloon Volume (cubic feet)
Time to Final Altitude (minutes)
Payload Mass (pounds)
Amount of Helium (cubic feet)
Science Help
Science Help Topics
What is the problem I have to solve?
What experiments should I run?
How many experiments should I run?
Should I make a prediction?
Should I make a table?
Should I make a graph?
What variables should I include in my table or graph?
How should I read the graph in simulation 3?
When should I draw a conclusion?
OK
Glossary
Science Help
Computer Help
Next

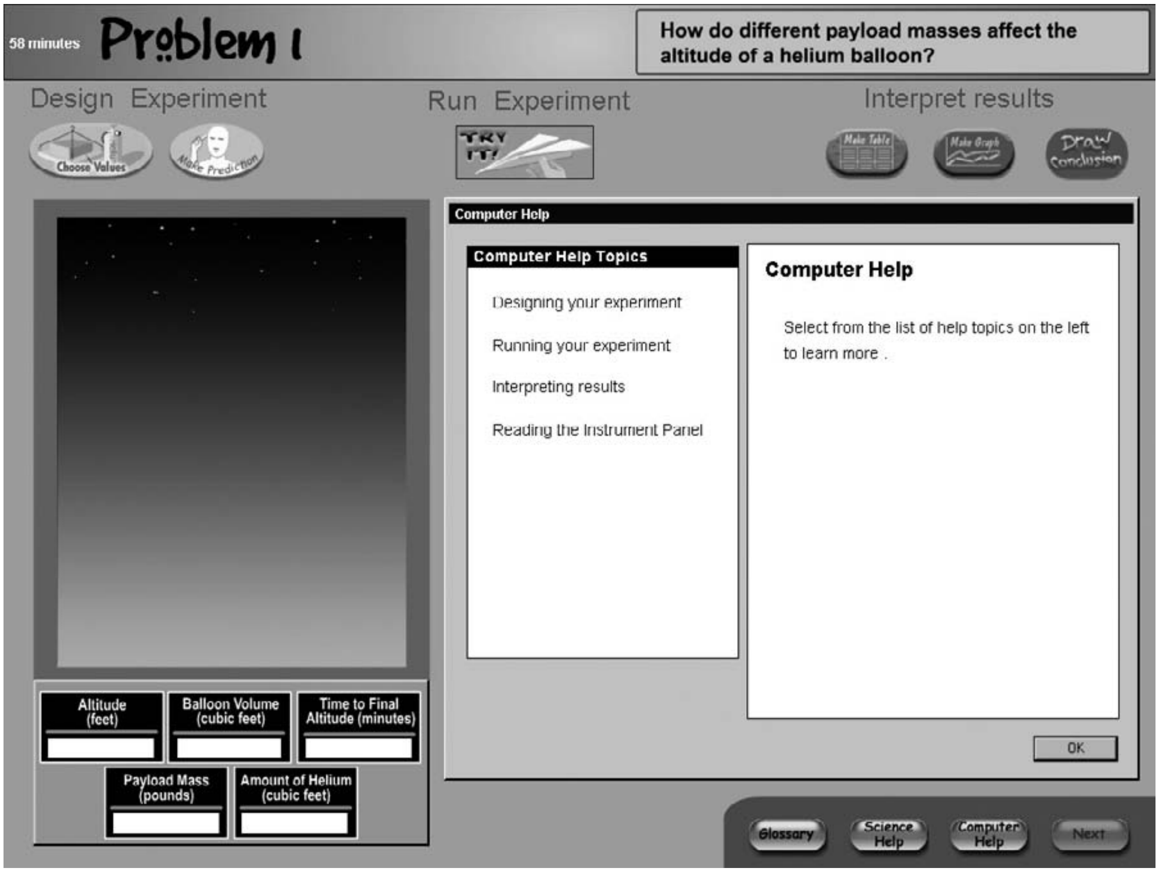
When should I draw a conclusion?

Do you have enough evidence to draw complete conclusions about the problem?

If you think you do, try drawing conclusions.

NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure E-12. Computer screen showing the TRE Simulation Computer Help topics menu, grade 8: 2003



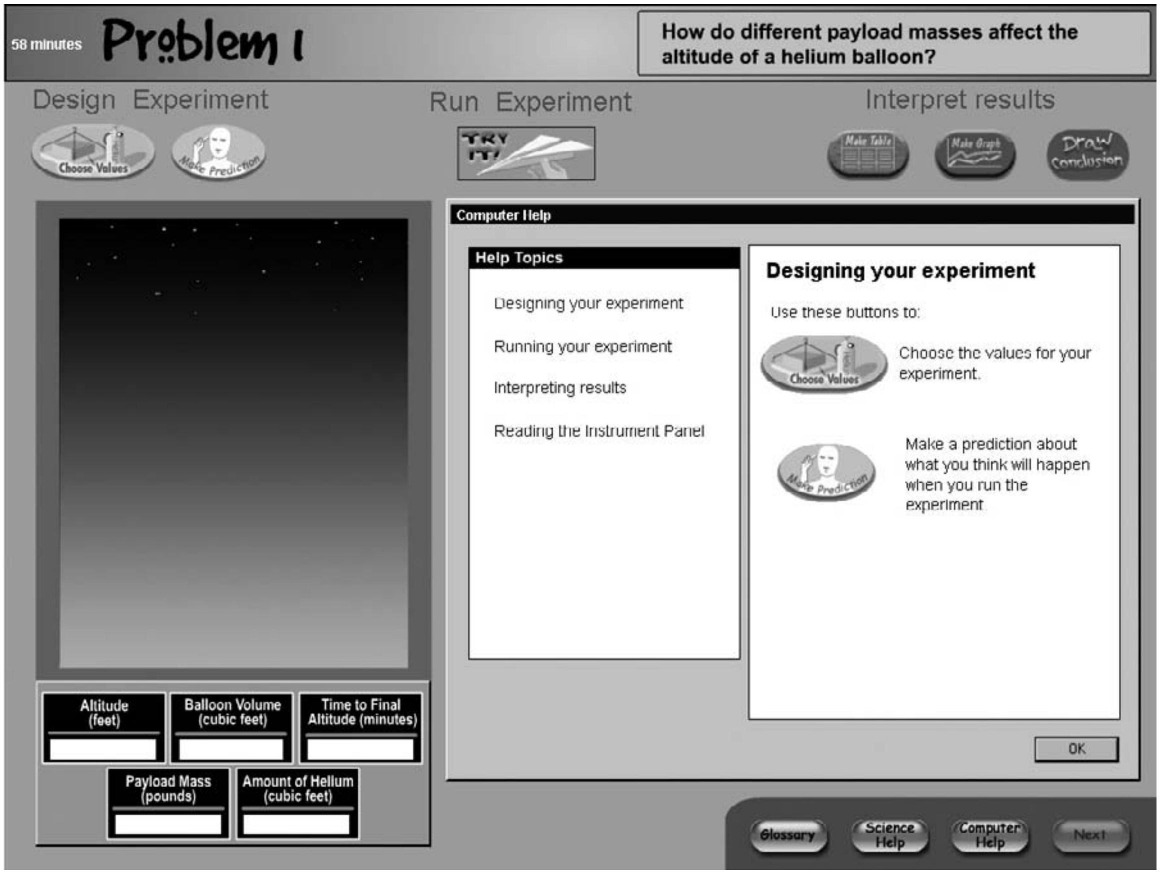
How do different payload masses affect the altitude of a helium balloon?

Computer Help

Select from the list of help topics on the left to learn more .

NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure E-13. Computer screen showing help for the first TRE Simulation Computer Help topic, grade 8: 2003



How do different payload masses affect the altitude of a helium balloon?

Designing your experiment

Running your experiment

Interpreting results

Reading the Instrument Panel

Designing your experiment

Use these buttons to:

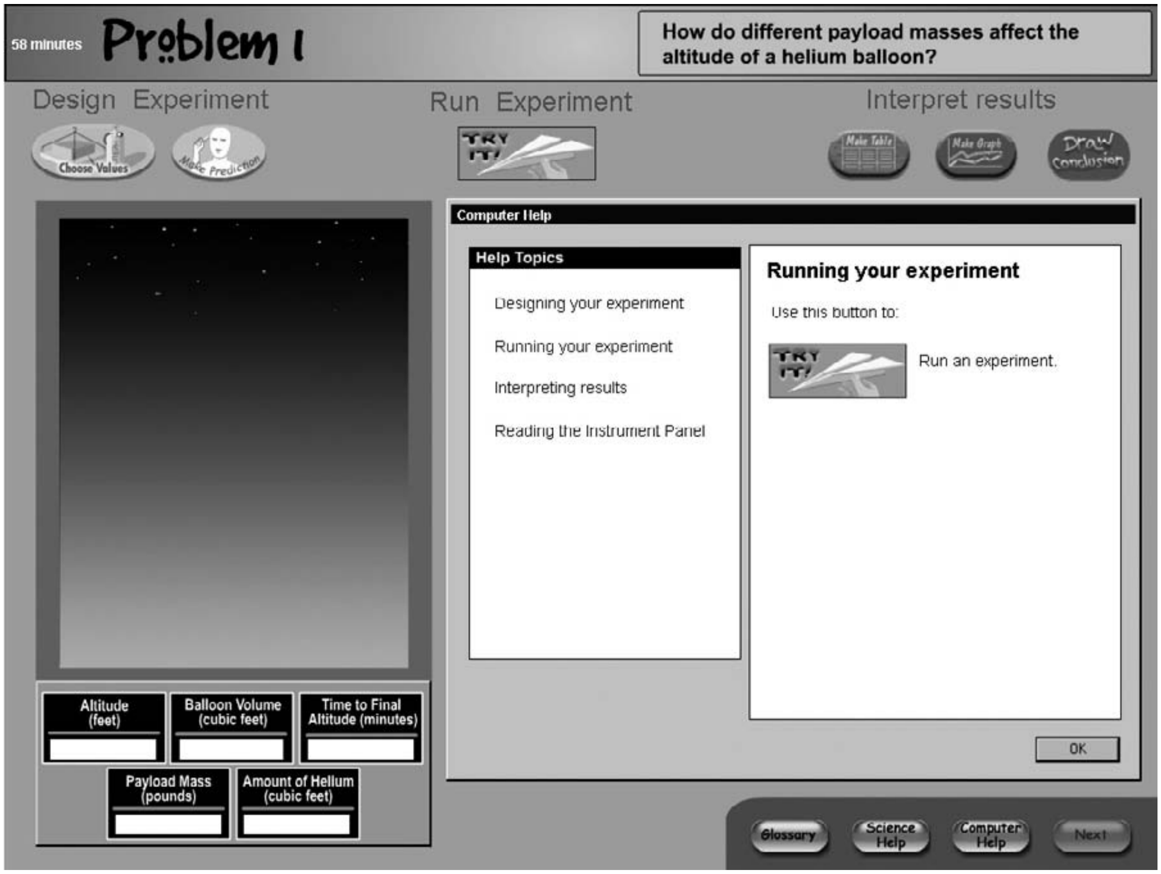
Choose the values for your experiment.

Make a prediction about what you think will happen when you run the experiment.

NOTE: TRE = Technology-Rich Environments.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure E-14. Computer screen showing help for the second TRE Simulation Computer Help topic, grade 8: 2003



How do different payload masses affect the altitude of a helium balloon?

Designing your experiment

Running your experiment

Interpreting results

Reading the Instrument Panel

Running your experiment

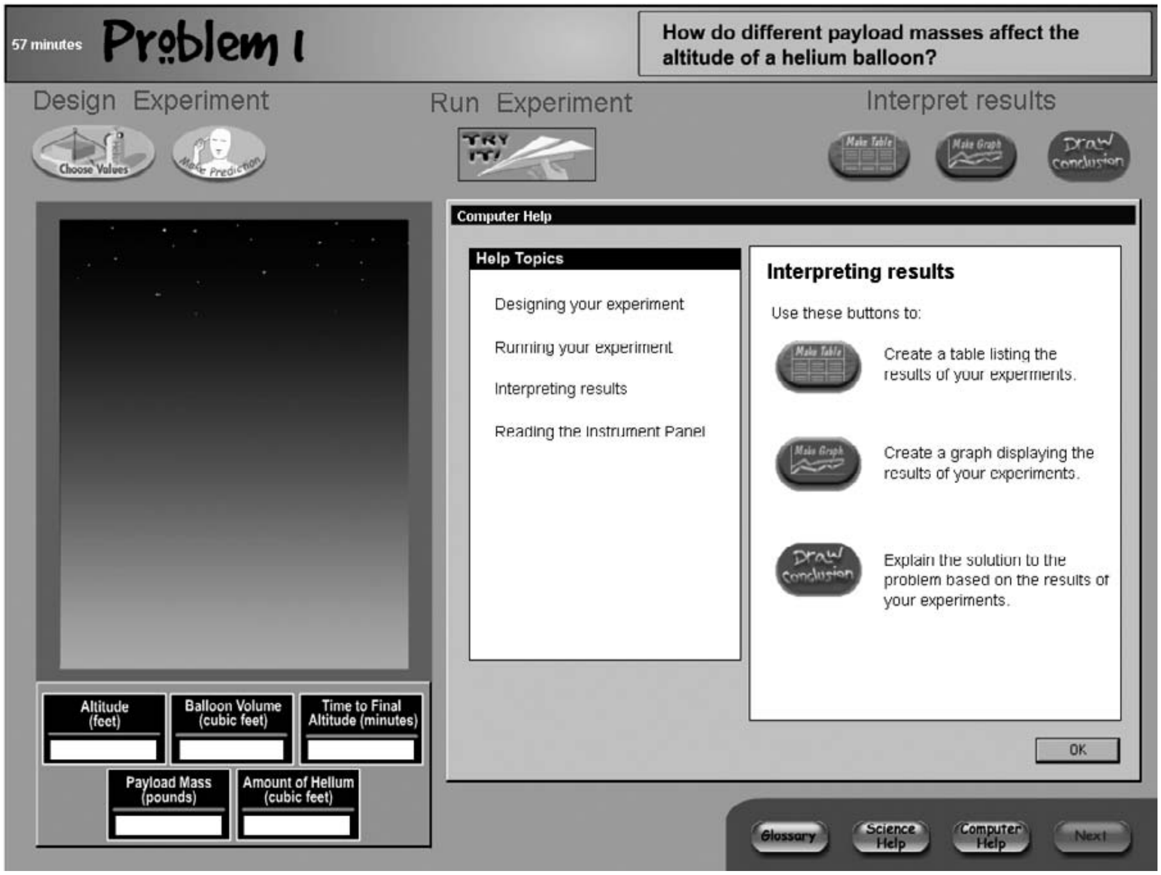
Use this button to:

Run an experiment.

NOTE: TRE = Technology-Rich Environments.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure E-15. Computer screen showing help for the third TRE Simulation Computer Help topic, grade 8: 2003



How do different payload masses affect the altitude of a helium balloon?

Computer Help

Help Topics

Designing your experiment

Running your experiment

Interpreting results

Reading the Instrument Panel

Interpreting results

Use these buttons to:

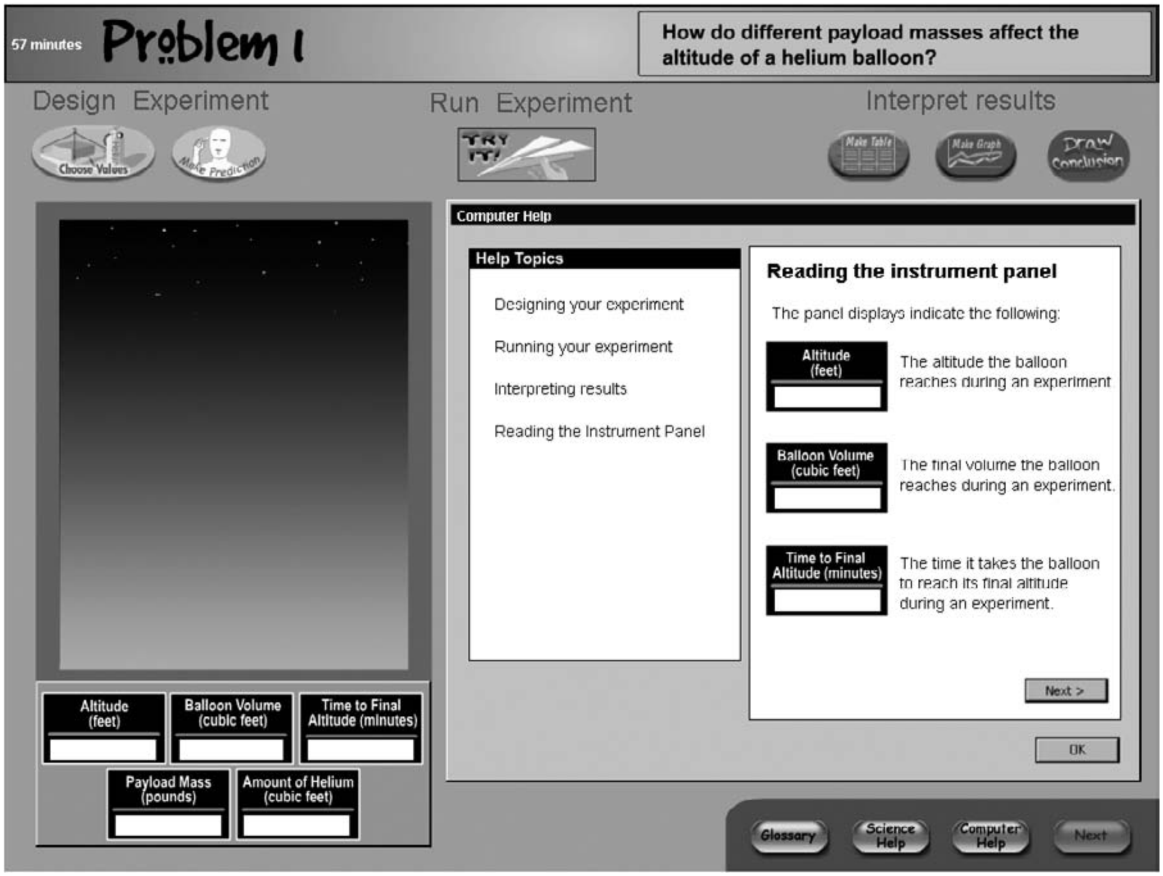
Make Table — Create a table listing the results of your experiments.

Make Graph — Create a graph displaying the results of your experiments.

Draw Conclusion — Explain the solution to the problem based on the results of your experiments.

NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure E-16. Computer screen showing help for the first part of the fourth TRE Simulation Computer Help topic, grade 8: 2003



How do different payload masses affect the altitude of a helium balloon?

Help Topics

- Designing your experiment
- Running your experiment
- Interpreting results
- Reading the Instrument Panel

Reading the instrument panel

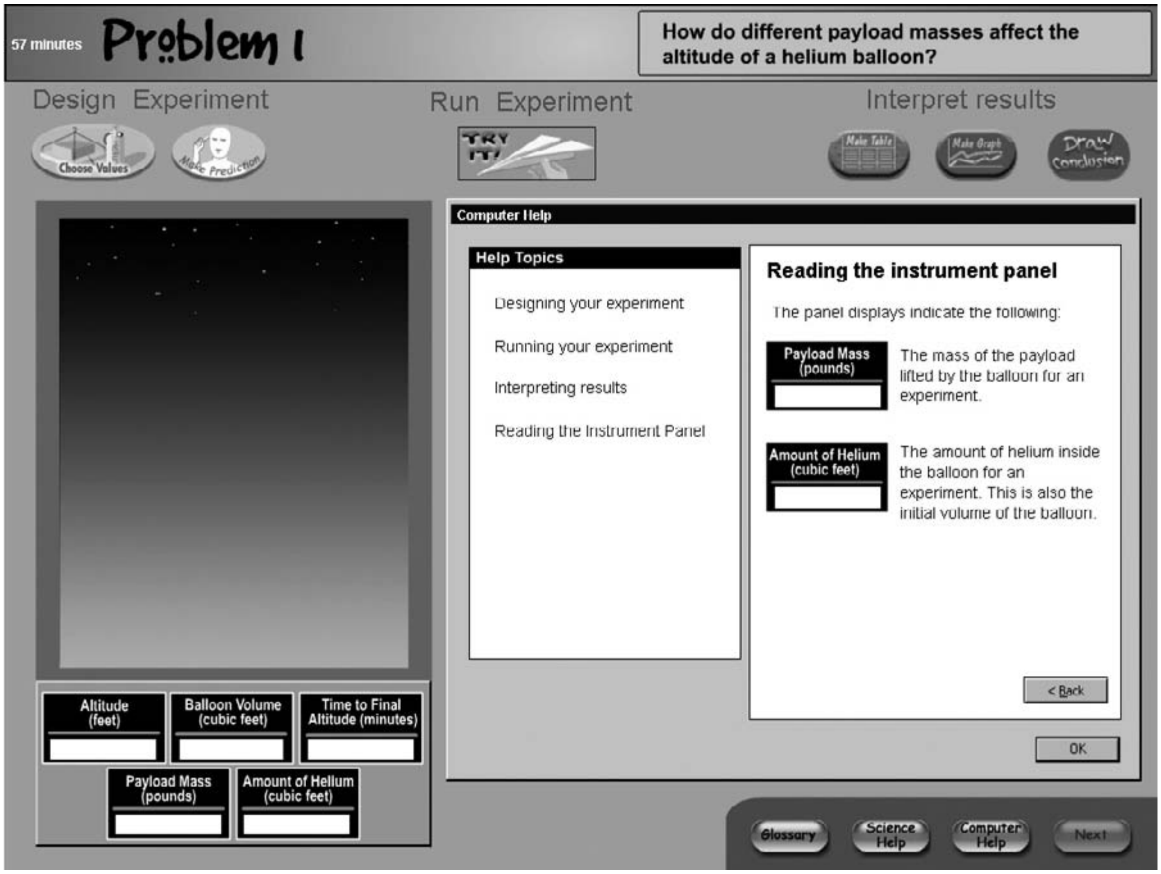
The panel displays indicate the following:

Altitude (feet)	The altitude the balloon reaches during an experiment
Balloon Volume (cubic feet)	The final volume the balloon reaches during an experiment.
Time to Final Altitude (minutes)	The time it takes the balloon to reach its final altitude during an experiment.

NOTE: TRE = Technology-Rich Environments.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure E-17. Computer screen showing help for the second part of the fourth TRE Simulation Computer Help topic, grade 8: 2003



How do different payload masses affect the altitude of a helium balloon?

Computer Help

Help Topics

- Designing your experiment
- Running your experiment
- Interpreting results
- Reading the Instrument Panel

Reading the instrument panel

The panel displays indicate the following:

Payload Mass (pounds) — The mass of the payload lifted by the balloon for an experiment.

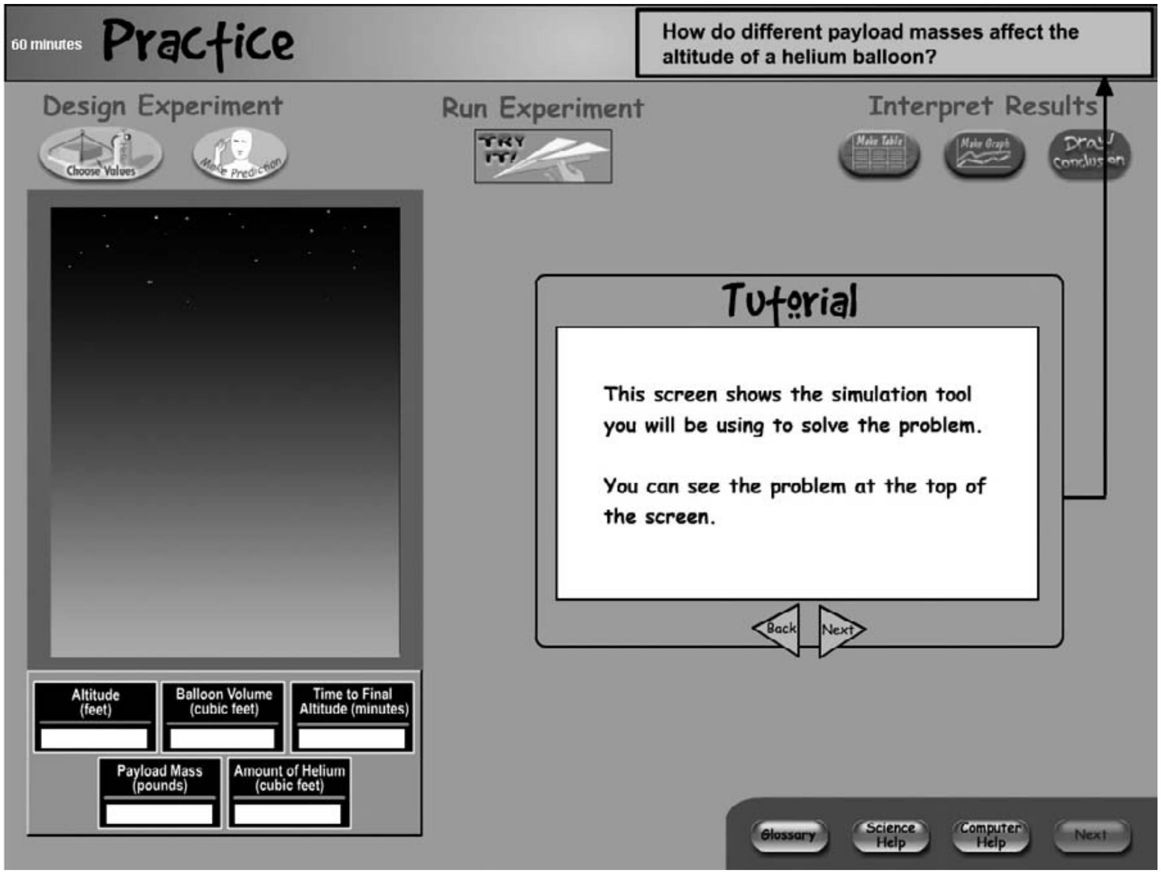
Amount of Helium (cubic feet) — The amount of helium inside the balloon for an experiment. This is also the initial volume of the balloon.

Altitude (feet)	Balloon Volume (cubic feet)	Time to Final Altitude (minutes)

Payload Mass (pounds)	Amount of Helium (cubic feet)

NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure E-18. TRE Simulation tutorial screen 1 showing the problem to be solved, grade 8: 2003



How do different payload masses affect the altitude of a helium balloon?

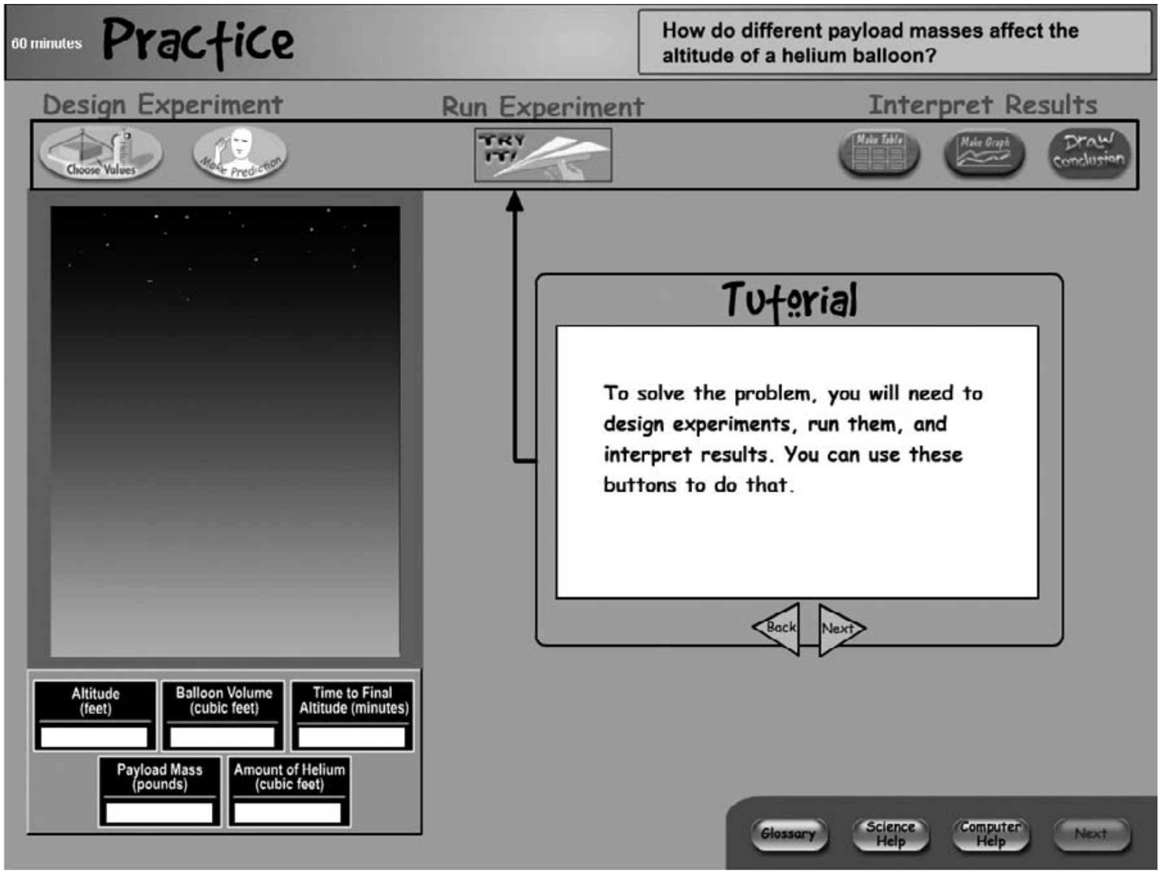
This screen shows the simulation tool you will be using to solve the problem.

You can see the problem at the top of the screen.

NOTE: TRE = Technology-Rich Environments.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure E-19. TRE Simulation tutorial screen 2 showing the task bar for solving the problem, grade 8: 2003

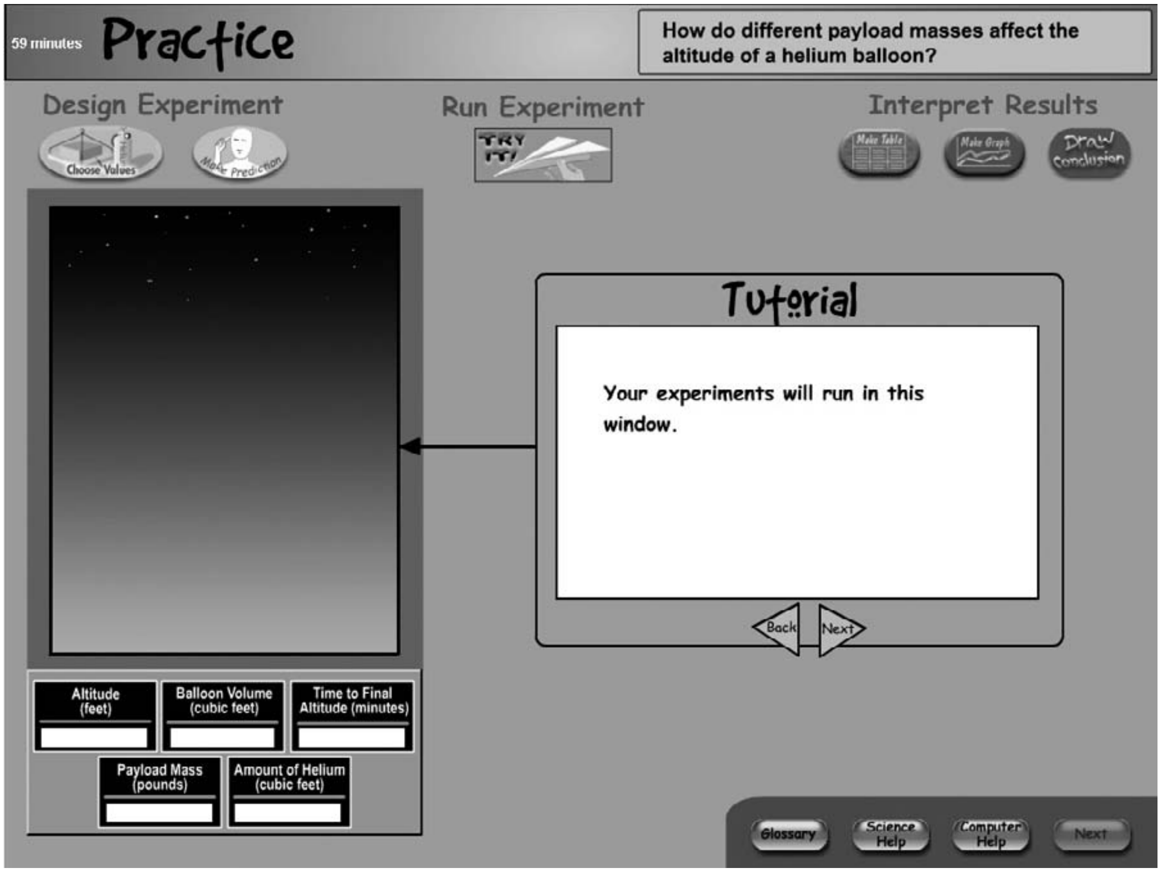


How do different payload masses affect the altitude of a helium balloon?

To solve the problem, you will need to design experiments, run them, and interpret results. You can use these buttons to do that.

NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure E-20. TRE Simulation tutorial screen 3 showing the experiment display window, grade 8: 2003

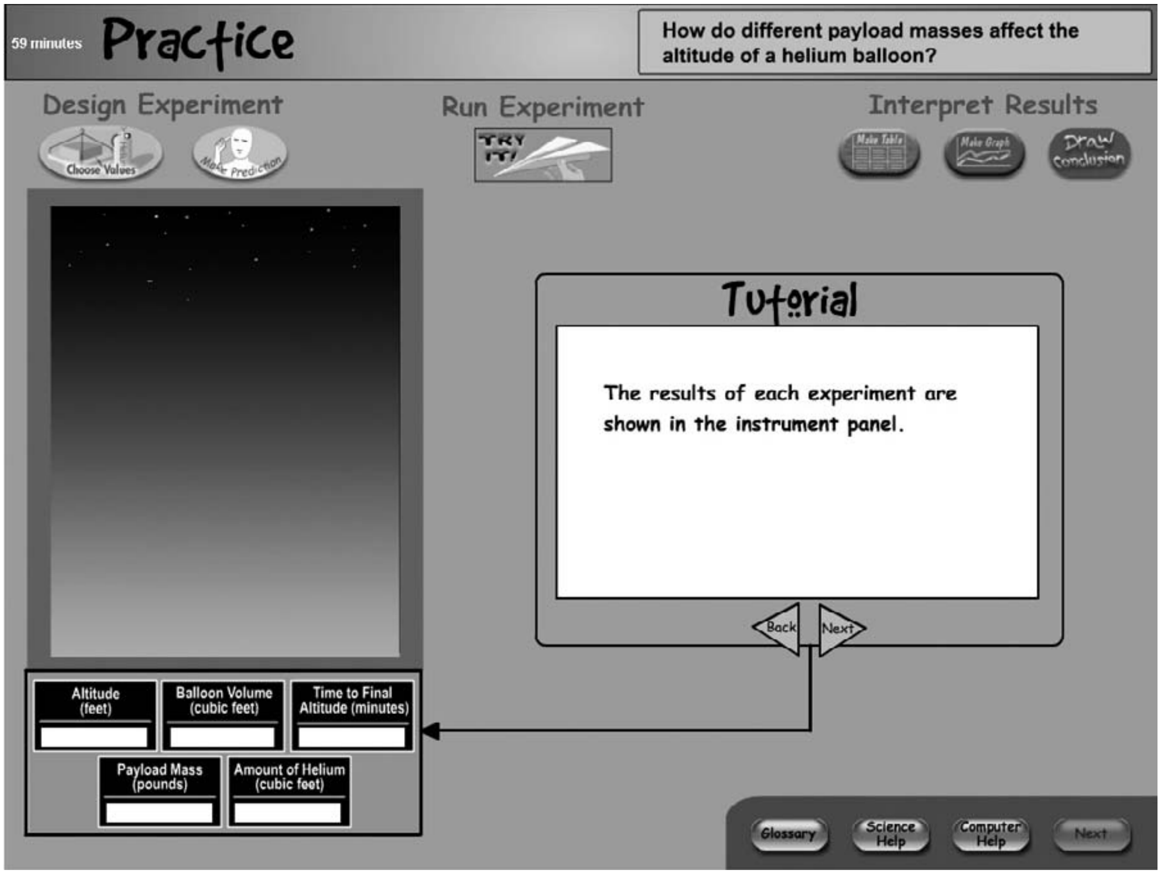


How do different payload masses affect the altitude of a helium balloon?

Your experiments will run in this window.

NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure E-21. TRE Simulation tutorial screen 4 showing the instrument panel to show data, grade 8: 2003



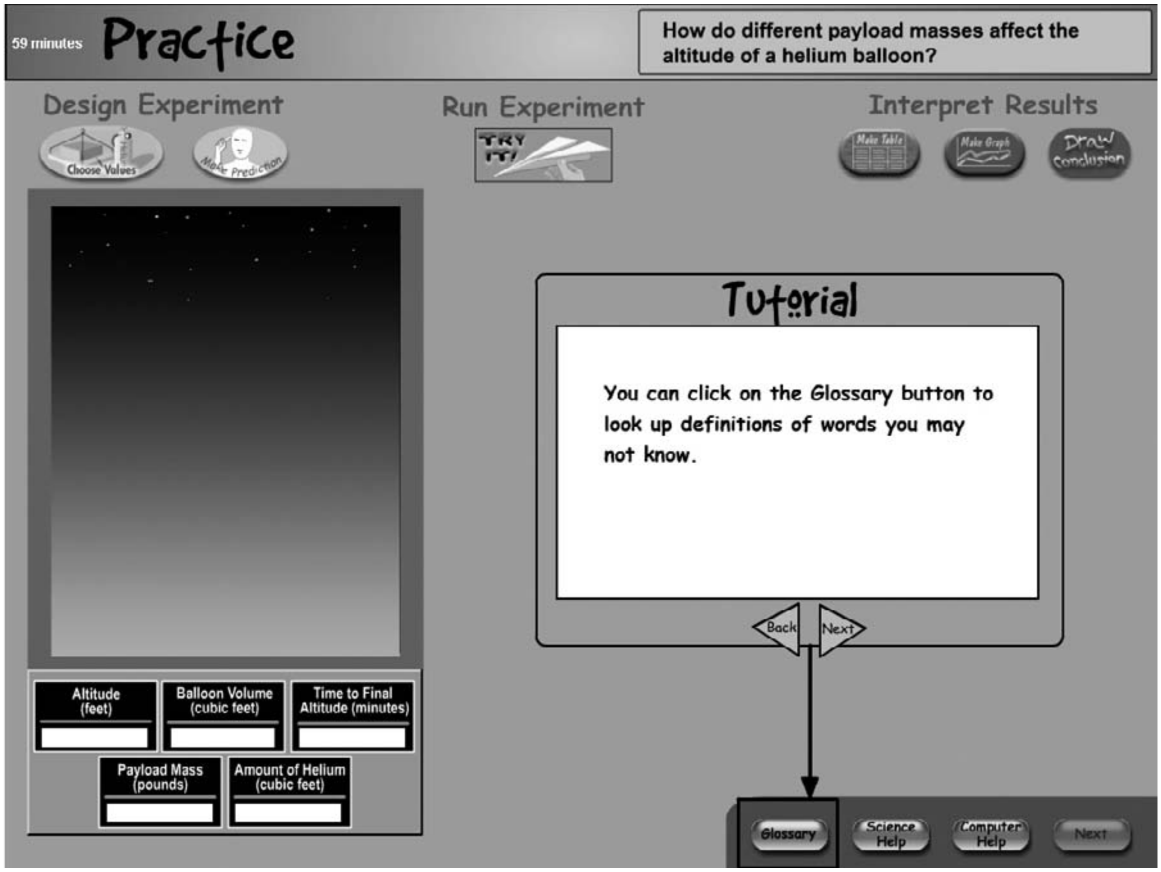
How do different payload masses affect the altitude of a helium balloon?

The results of each experiment are shown in the instrument panel.

NOTE: TRE = Technology-Rich Environments.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure E-22. TRE Simulation tutorial screen 5 showing the Glossary tool button, grade 8: 2003



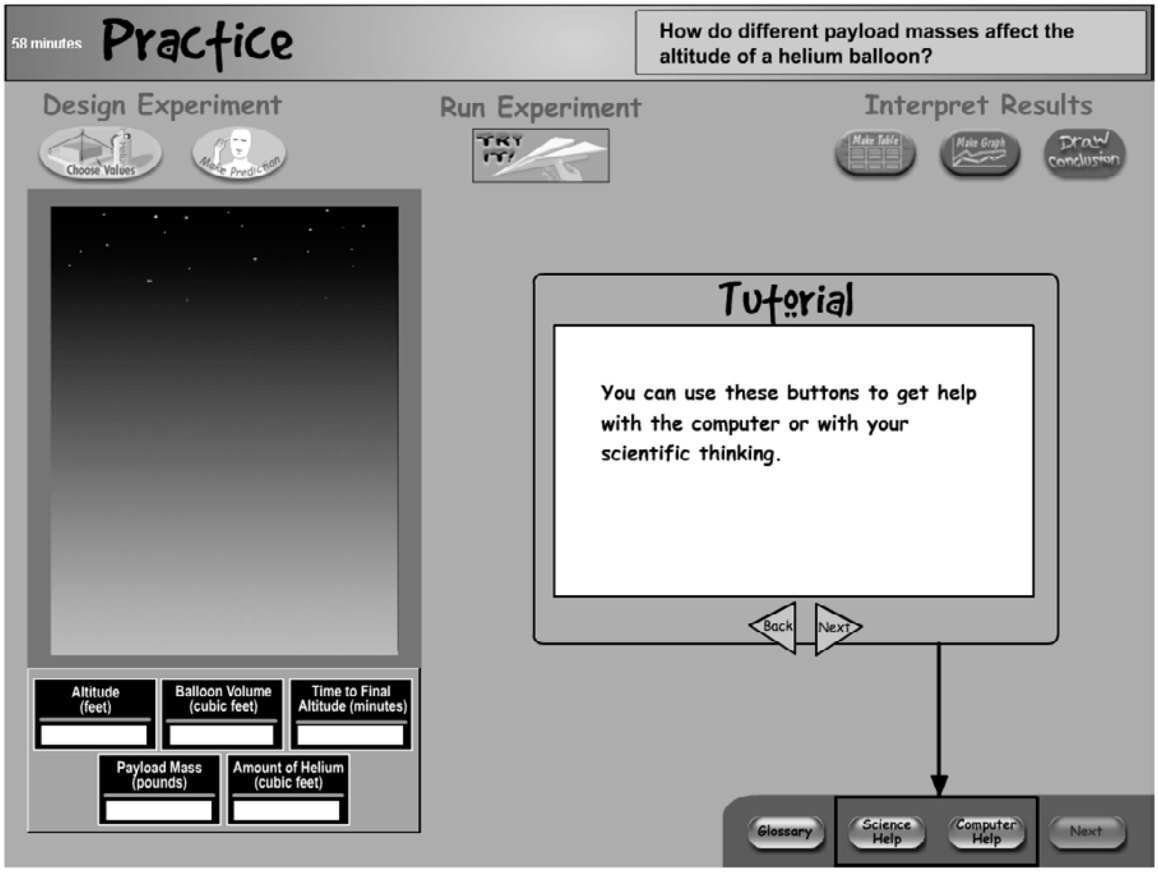
How do different payload masses affect the altitude of a helium balloon?

You can click on the Glossary button to look up definitions of words you may not know.

NOTE: TRE = Technology-Rich Environments.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure E-23. TRE Simulation tutorial screen 6 showing the Science and Computer Help tool buttons, grade 8: 2003



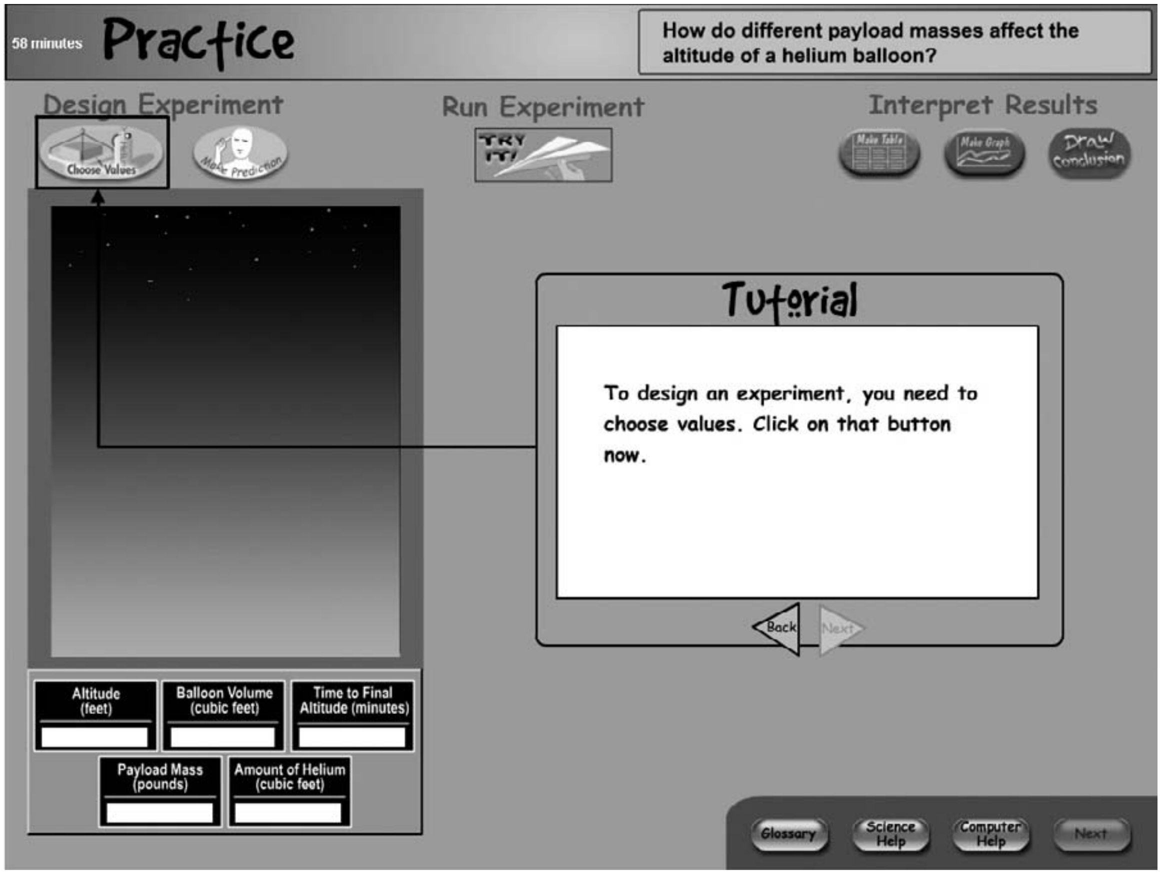
How do different payload masses affect the altitude of a helium balloon?

You can use these buttons to get help with the computer or with your scientific thinking.

NOTE: TRE = Technology-Rich Environments.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure E-24. TRE Simulation tutorial screen 7 showing the Choose Values (for experiments) button, grade 8: 2003

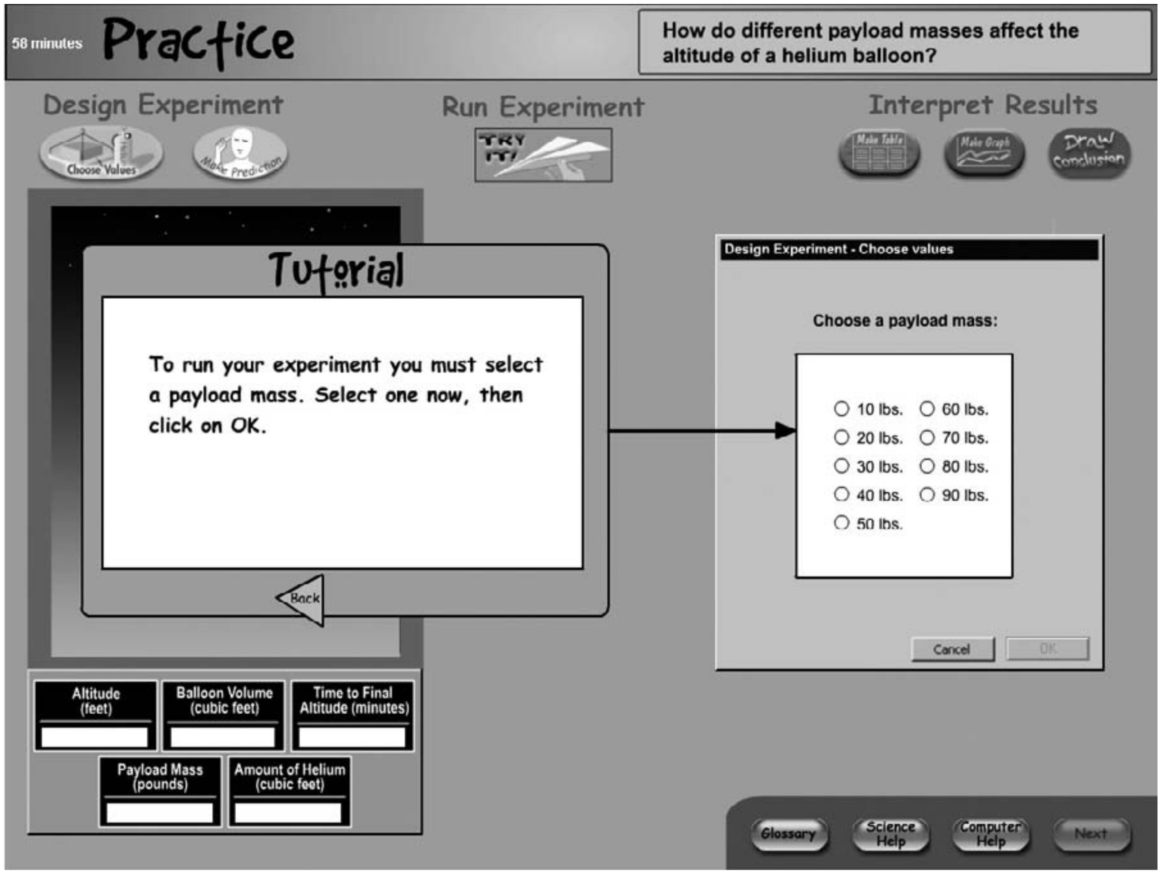


How do different payload masses affect the altitude of a helium balloon?

To design an experiment, you need to choose values. Click on that button now.

NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure E-25. TRE Simulation tutorial screen 8 showing the payload mass menu, grade 8: 2003



How do different payload masses affect the altitude of a helium balloon?

Tutorial

To run your experiment you must select a payload mass. Select one now, then click on OK.

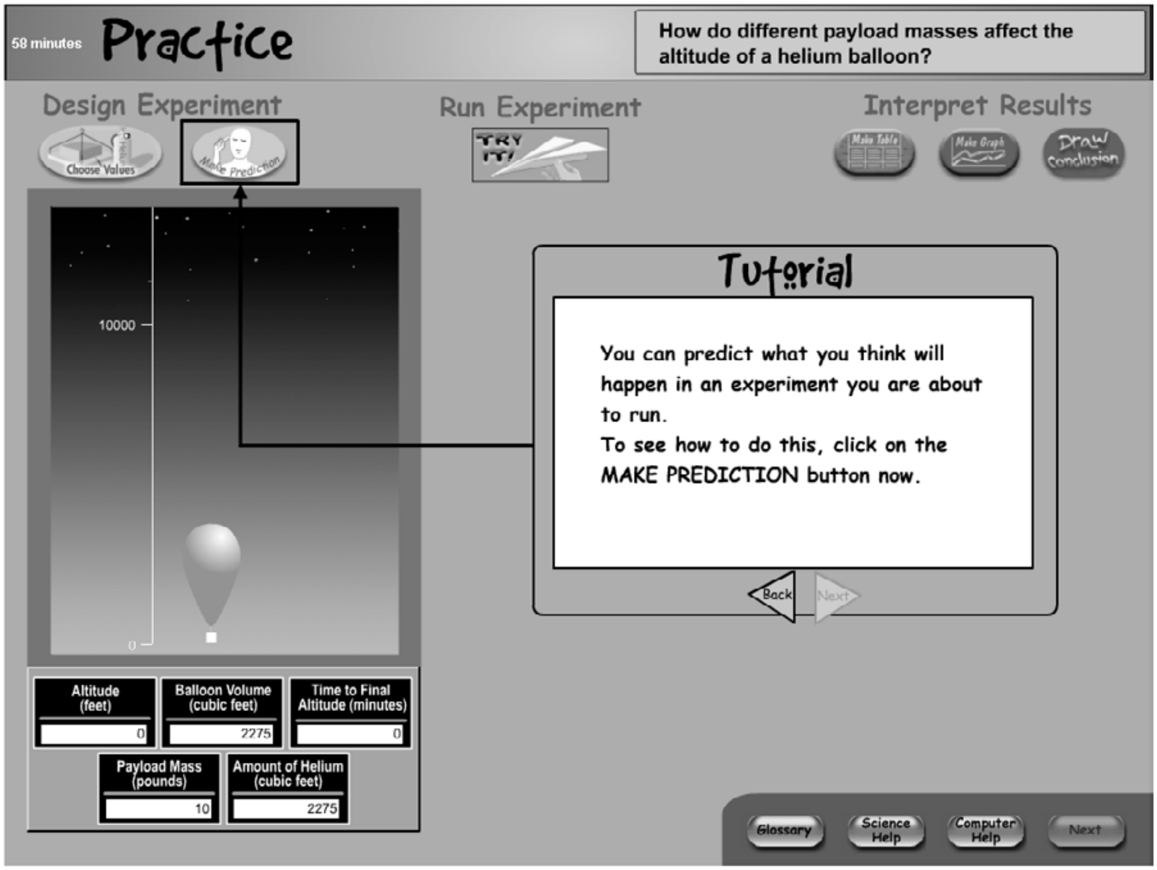
Design Experiment - Choose values

Choose a payload mass:

○ 10 lbs. ○ 60 lbs.
○ 20 lbs. ○ 70 lbs.
○ 30 lbs. ○ 80 lbs.
○ 40 lbs. ○ 90 lbs.
○ 50 lbs.

NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure E-26. TRE Simulation tutorial screen 9 showing the Make Prediction button, grade 8: 2003



How do different payload masses affect the altitude of a helium balloon?

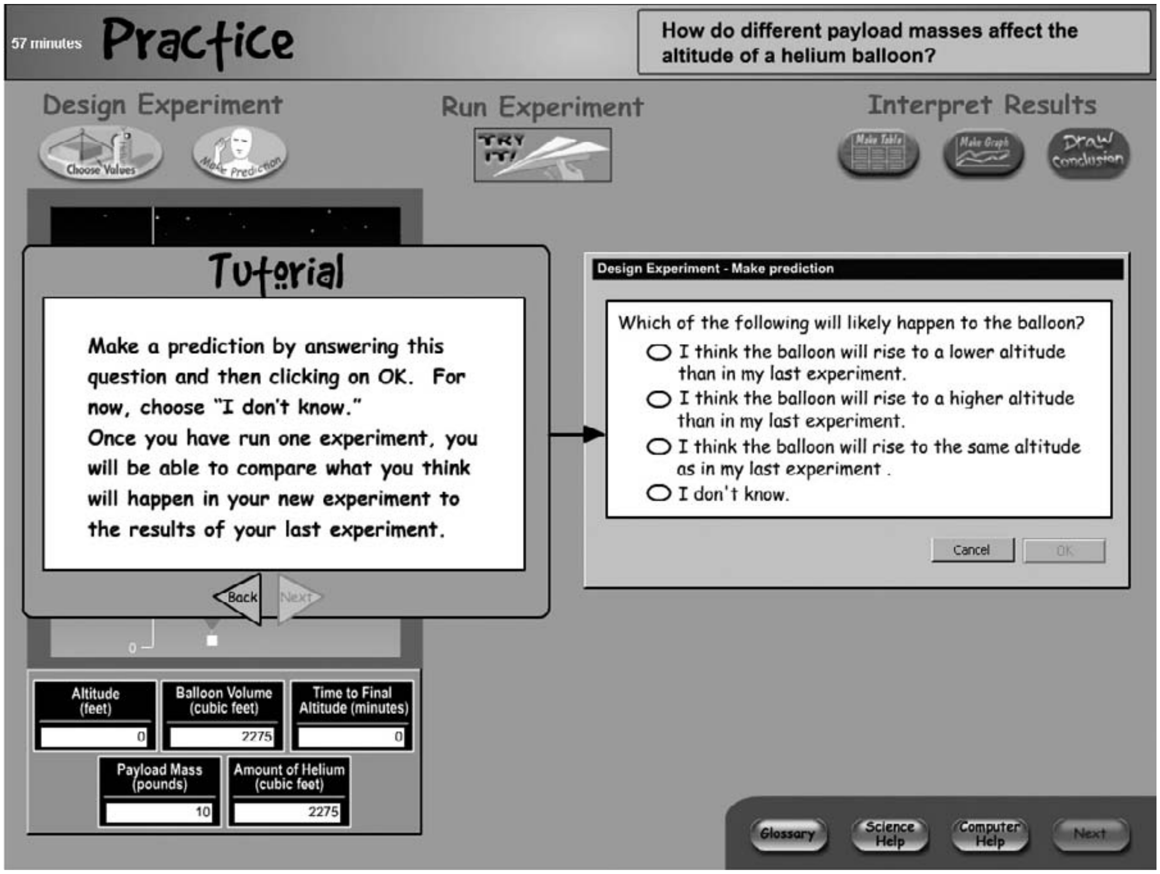
You can predict what you think will happen in an experiment you are about to run.
To see how to do this, click on the MAKE PREDICTION button now.

Altitude (feet)	Balloon Volume (cubic feet)	Time to Final Altitude (minutes)
0	2275	0

Payload Mass (pounds)	Amount of Helium (cubic feet)
10	2275

NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure E-27. TRE Simulation tutorial screen 10 showing the Make Prediction question and response choices, grade 8: 2003



How do different payload masses affect the altitude of a helium balloon?

Make a prediction by answering this question and then clicking on OK. For now, choose "I don't know."
Once you have run one experiment, you will be able to compare what you think will happen in your new experiment to the results of your last experiment.

Design Experiment - Make prediction

Which of the following will likely happen to the balloon?

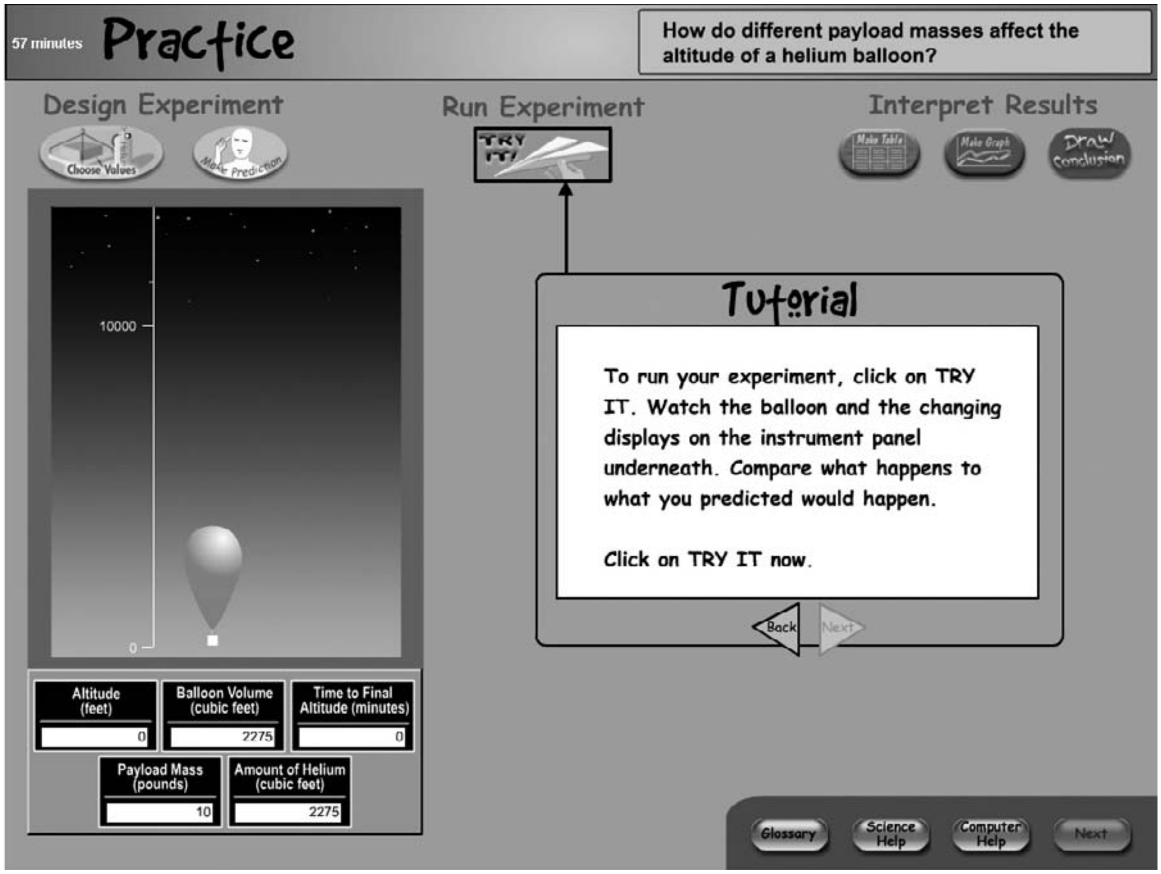
○ I think the balloon will rise to a lower altitude than in my last experiment.
○ I think the balloon will rise to a higher altitude than in my last experiment.
○ I think the balloon will rise to the same altitude as in my last experiment .
○ I don't know.

Altitude (feet)	Balloon Volume (cubic feet)	Time to Final Altitude (minutes)
0	2275	0

Payload Mass (pounds)	Amount of Helium (cubic feet)
10	2275

NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure E-28. TRE Simulation tutorial screen 11 showing the Try It button for running an experiment, grade 8: 2003



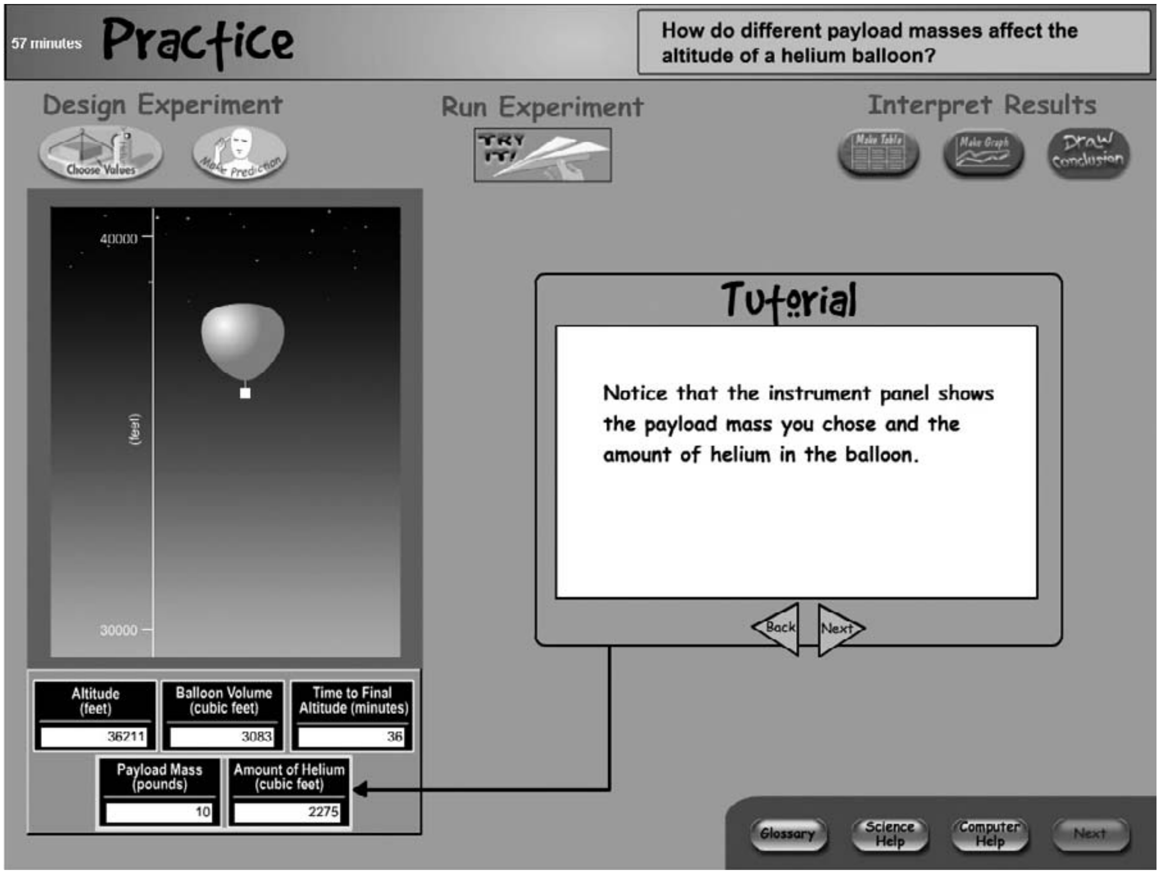
How do different payload masses affect the altitude of a helium balloon?

To run your experiment, click on TRY IT. Watch the balloon and the changing displays on the instrument panel underneath. Compare what happens to what you predicted would happen.

Click on TRY IT now.

NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure E-29. TRE Simulation tutorial screen 12 showing the instrument panel in detail, grade 8: 2003



How do different payload masses affect the altitude of a helium balloon?

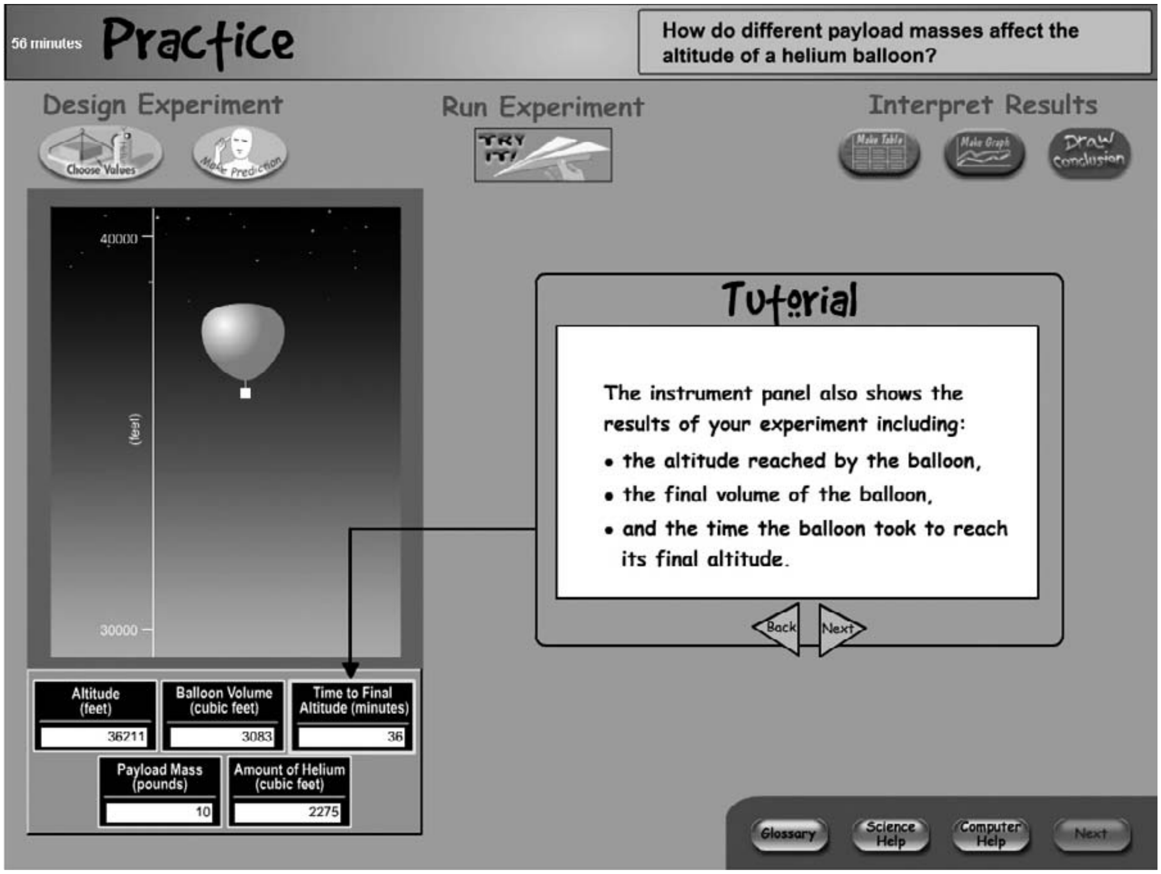
Notice that the instrument panel shows the payload mass you chose and the amount of helium in the balloon.

Altitude (feet)	Balloon Volume (cubic feet)	Time to Final Altitude (minutes)
36211	3083	36

Payload Mass (pounds)	Amount of Helium (cubic feet)
10	2275

NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure E-30. TRE Simulation tutorial screen 13 showing the instrument panel in detail, grade 8: 2003



How do different payload masses affect the altitude of a helium balloon?

Tutorial

The instrument panel also shows the results of your experiment including:

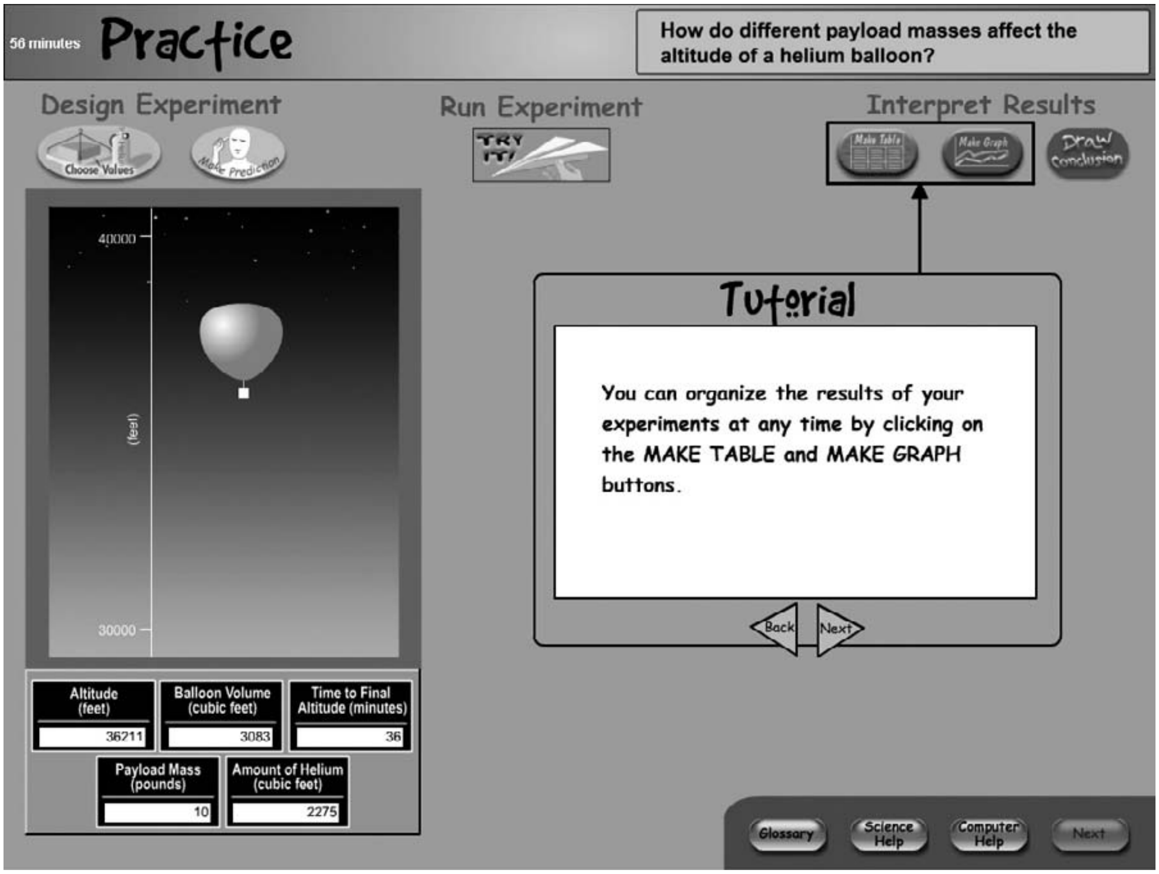
- the altitude reached by the balloon,
- the final volume of the balloon,
- and the time the balloon took to reach its final altitude.

Altitude (feet)	Balloon Volume (cubic feet)	Time to Final Altitude (minutes)
36211	3083	36

Payload Mass (pounds)	Amount of Helium (cubic feet)
10	2275

NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure E-31. TRE Simulation tutorial screen 14 showing the buttons for making tables and graphs, grade 8: 2003



How do different payload masses affect the altitude of a helium balloon?

You can organize the results of your experiments at any time by clicking on the MAKE TABLE and MAKE GRAPH buttons.

Altitude (feet): 36211

Balloon Volume (cubic feet): 3083

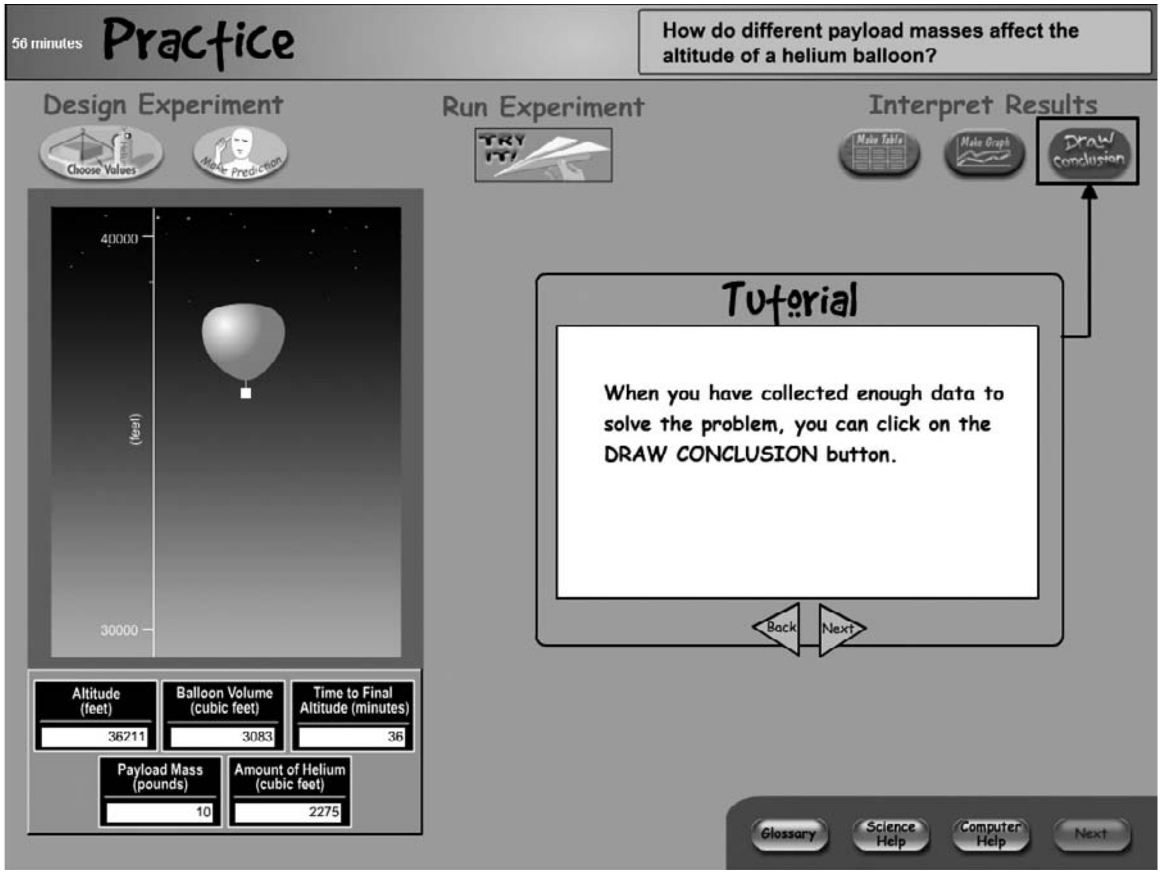
Time to Final Altitude (minutes): 36

Payload Mass (pounds): 10

Amount of Helium (cubic feet): 2275

NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure E-32. TRE Simulation tutorial screen 15 showing the button for drawing conclusions, grade 8: 2003



How do different payload masses affect the altitude of a helium balloon?

When you have collected enough data to solve the problem, you can click on the DRAW CONCLUSION button.

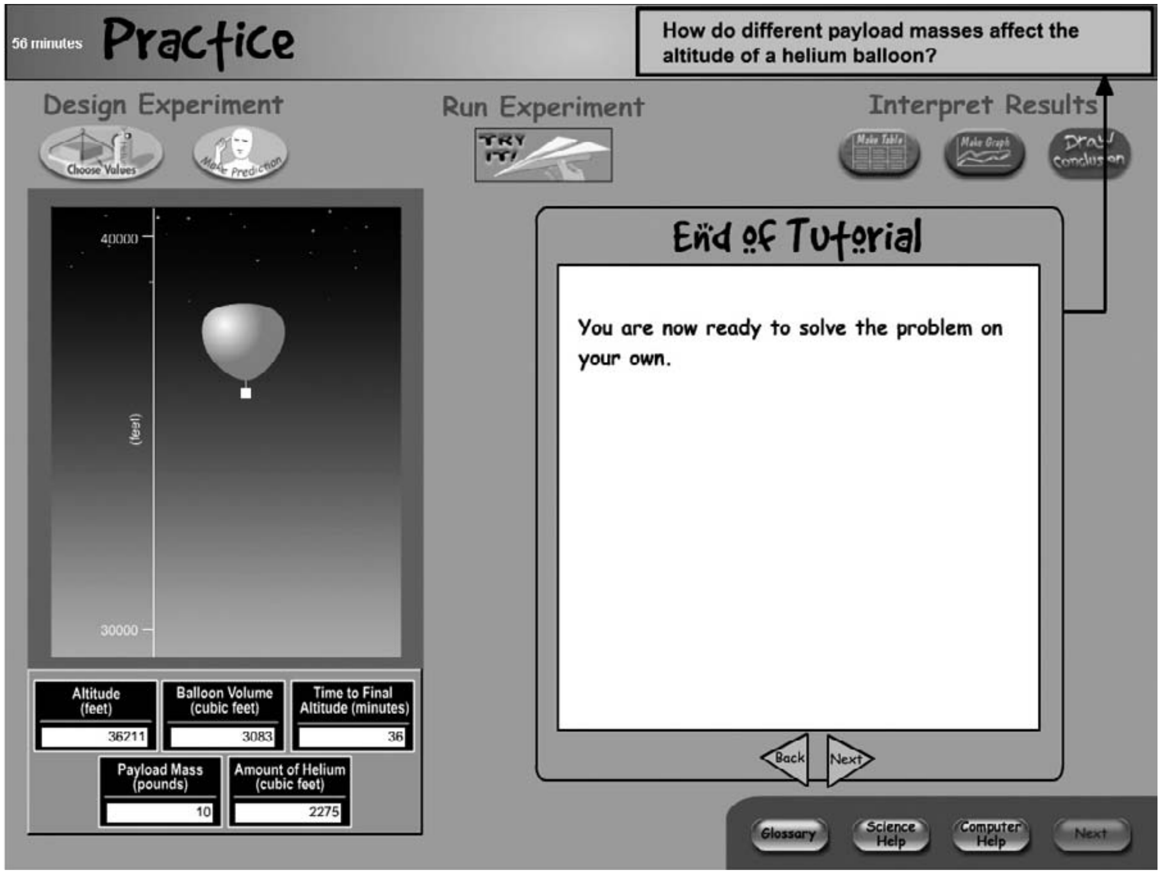
Altitude (feet)	Balloon Volume (cubic feet)	Time to Final Altitude (minutes)
36211	3083	36

Payload Mass (pounds)	Amount of Helium (cubic feet)
10	2275

NOTE: TRE = Technology-Rich Environments.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure E-33. TRE Simulation tutorial screen 16 showing the problem to be solved, grade 8: 2003



How do different payload masses affect the altitude of a helium balloon?

End of Tutorial

You are now ready to solve the problem on your own.

Altitude (feet): 36211

Balloon Volume (cubic feet): 3083

Time to Final Altitude (minutes): 36

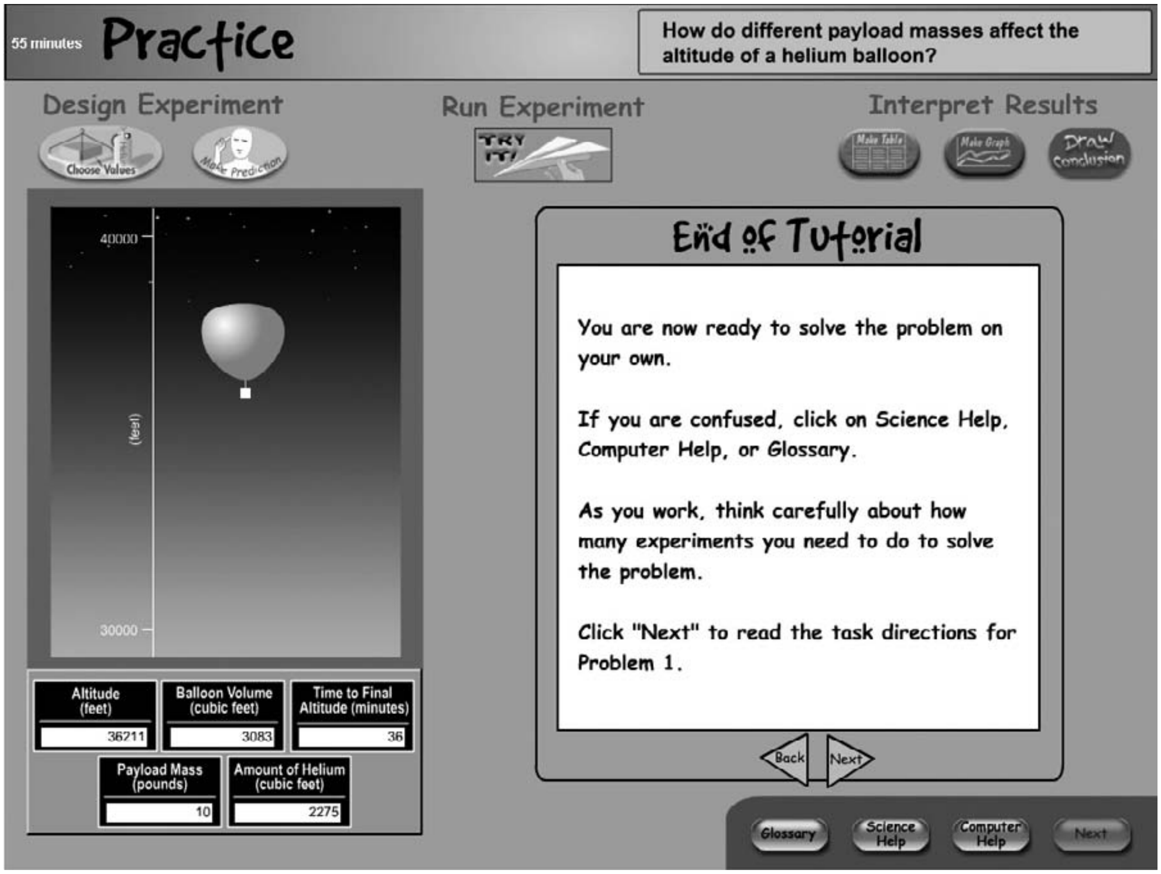
Payload Mass (pounds): 10

Amount of Helium (cubic feet): 2275

NOTE: TRE = Technology-Rich Environments.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure E-34. TRE Simulation tutorial screen 17 showing entry into the Simulation scenario, grade 8: 2003



How do different payload masses affect the altitude of a helium balloon?

Altitude (feet)	Balloon Volume (cubic feet)	Time to Final Altitude (minutes)	Payload Mass (pounds)	Amount of Helium (cubic feet)
36211	3083	36	10	2275

End of Tutorial

You are now ready to solve the problem on your own.

If you are confused, click on Science Help, Computer Help, or Glossary.

As you work, think carefully about how many experiments you need to do to solve the problem.

Click "Next" to read the task directions for Problem 1.

NOTE: TRE = Technology-Rich Environments.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Appendix F: Bayesian Estimation in the Problem Solving in Technology-Rich Environments Study

Introduction

The Problem Solving in Technology-Rich Environments (TRE) study incorporates several design features that are not found in standard NAEP analysis. These features include

- an a priori hypothesized structure of the relationship among the set of latent proficiency variables,
- the potential to accommodate multivariate items (i.e., items that measure more than one latent proficiency), and
- inclusion of context effects; items sharing a context are related to each other more strongly than to other items.

All three of these features are beyond the scope of measurement models used in operational NAEP. Operational NAEP employs a univariate Item Response Theory (IRT) model that uses a simple structure, i.e., each item measures only one latent proficiency. Since the IRT model is univariate, there can be no structural relations among latent proficiencies, there can be no item that measures more than one proficiency, and there can be no context effect in addition to the latent proficiency.

This appendix outlines the cognitive models that were used in the TRE study. (The term, cognitive model, is used here to refer to the union of the student and evidence models described in chapter 2 of this report.) These are represented by directed graphs showing latent proficiency, observable, and context variables, with arrows showing direction of influence. Note that two scenarios, or separate computer tasks, were delivered. One was the Search scenario, in which students used a simulated web search to answer questions about scientific balloons. They conducted searches, gathered information, and then summarized results. The second scenario was Simulation. In this activity, students used a simulation tool to conduct a series of experiments in order to discover relationships among variables related to the physics of balloon behavior in the atmosphere.

This appendix also presents the Bayesian models used to analyze the data and estimate item parameters. These consist of the IRT model for items; the structural model for representing relationships among the latent proficiencies; the conditioning model, which describes the structured prior distribution of the latent problem-solving in TRE proficiency; and finally the population model for deriving estimates of population means, percents, and associated standard errors.

Finally, this appendix discusses the construction of a real-time inference engine for the Search scenario. Model parameters estimated from the Bayesian IRT analysis are imported as fixed quantities into an inference engine (ERGO 2001 by Noetic Systems, Inc.), enabling sensitivity testing of the model and scoring of student responses. Profiles of proficiencies can be selected to see what response probabilities of the observables will result. Also, a vector of observed responses can be selected, and the resulting proficiency scores can be estimated. The inference engine can also be used as a stand-alone application to get real-time estimates of proficiency as an examinee responds to the assessment. This aspect of the Bayesian inference engine demonstrates the feasibility of using a computer to assess and immediately provide proficiency estimates over the Web.

The Cognitive Models

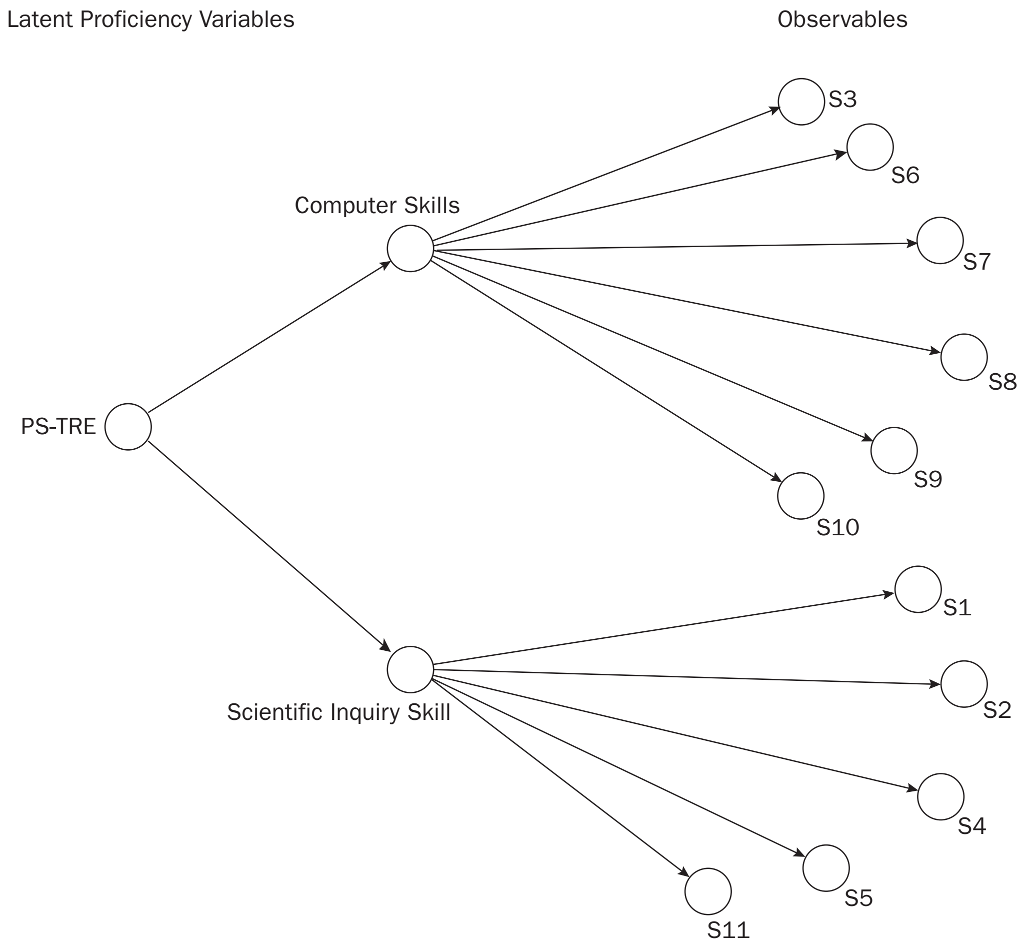
Two somewhat different cognitive models were fitted to the two TRE scenarios. First, consider the directed graph in figure F-1, which depicts the relationships among variables for the Search scenario. Two classes of variables are shown. To the left are latent proficiencies, and to the right are observables, representing observed scores on performance tasks.

This discussion of latent proficiencies follows customary usage in calling precursor variables “parents” and other latent variables “children” (to avoid use of causal language). In this model, the parent proficiency is problem solving in technology-rich environments (PS-TRE), which has computer skills and scientific inquiry skill as resultant or “child” proficiencies. Arrows between the latent skills indicate the direction of influence.[1]

[1] Note that scientific inquiry skill was originally proposed as having two component skills: scientific inquiry exploration skill and scientific inquiry synthesis skill. With the Search scenario, it was found that there were too few observables to reliably measure these constructs. As a result, they were combined into a single scientific inquiry proficiency in the final model.

To the right of figure F-1 are observables. These are summaries of observed behaviors that can be mapped onto several levels of partial credit (from two to four levels). The probability that a student will score at a specific level is a function of that student's latent skill. The nature of this function is defined by an IRT model. According to the model, computer skill contributes to a student's propensity to respond correctly to observables requiring computer-related abilities such as keyboarding, using menus correctly, and not needing to use the help function. Similarly, scientific inquiry skill contributes to a student's propensity to explore content and draw conclusions about scientific questions correctly.

Figure F-1. The TRE Search cognitive model, grade 8: 2003



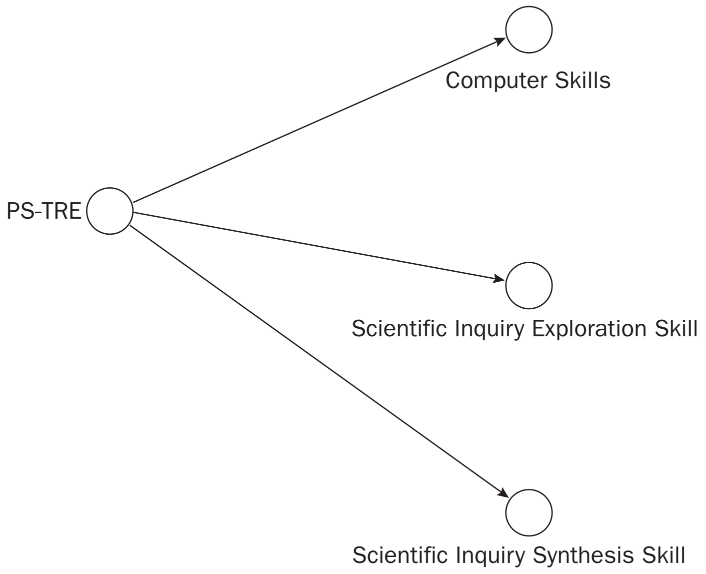
NOTE: PS-TRE = Problem solving in technology-rich environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure F-2 shows the directed graph depicting a structural (or student) model for the latent proficiencies in the Simulation scenario. In this model, PS-TRE is the parent of three other latent skills: computer skills, scientific inquiry exploration skill, and scientific inquiry synthesis skill. These latter three are proficiencies that contribute to the propensity to respond correctly to observables.[2]

Figure F-3 shows the cognitive model for the Simulation scenario. The variables on the left, PS-TRE, computer skills, scientific inquiry exploration skill, and scientific inquiry synthesis skill, are latent proficiencies. These are the direct precursors of observables, which are found in the middle of the diagram. Each observable measured (was the child of) just one latent proficiency. This simple structure was confirmed to fit the data best. On the far right of figure F-3 are three other latent variables, which define the effect of context.

The three context effects correspond to the three Simulation problems in the scenario. The context variables represent any knowledge, skill, or other factor that is specific to one Simulation task but not another. Students with a higher level of task-specific skill will tend to do better on all the items in the task. As a result, items sharing a common task tend to be more highly correlated than items in different tasks. The context effect can be thought of as controlling for a type of nuisance variation. With context effects in the model, conditional independence of observables, given a student's latent skills, holds. The assumption of conditional independence is a basic tenet of any explanatory model. This assumption also underlies all conventional IRT estimation.

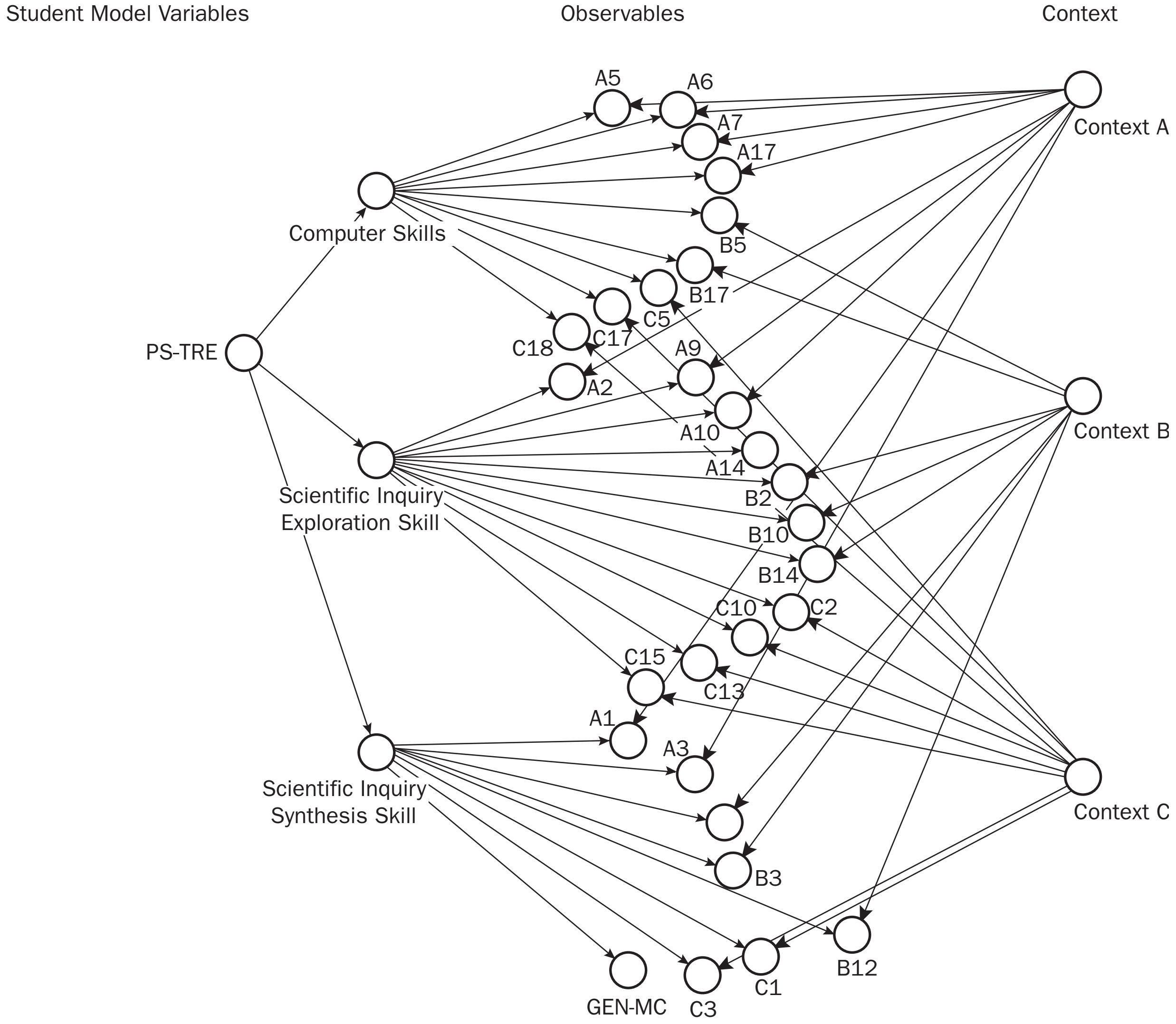
Figure F-2. Student model for TRE Simulation scenario, grade 8: 2003


Computer Skills
PS-TRE
Scientific Inquiry Exploration Skill
Scientific Inquiry Synthesis Skill

NOTE: PS-TRE = Problem solving in technology-rich environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

[2] Unlike the Search scenario, Simulation had a sufficient number of observables to reliably measure exploration and synthesis as separate skills. However, scientific inquiry skill was dropped as a precursor to the latter two proficiencies, because scientific inquiry skill was not reliably measured by its component skills.

Figure F-3. Cognitive model for TRE Simulation scenario, grade 8: 2003



NOTE: PS-TRE = Problem solving in technology-rich environments. GEN-MC = Synthesizing multiple-choice items.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

General Description of the Bayesian Model

The previous section gave an outline of the cognitive model behind this analysis. This section presents a detailed description of the models used to analyze data and estimate item parameters.

Item Response Model

In the TRE study, the item-category responses (i.e., the probability of responding correctly to a category of an observable) are modeled as dichotomous item responses. In the Simulation scenario, a student's behavior on an observable is influenced by a latent proficiency skill (student model variable) and a context effect. As a result, the item response is multivariate in form. In the present form, this is a compensatory model, with equal slopes for Θ_{ij}, the value for a student on the latent proficiency, and Φ_{im}, the value for a student on the latent context effect. This model is compensatory in that the two latent variables have an additive effect on item response. Other types of relationships (e.g., disjunctive) could have been modeled to represent different sorts of relationships between the latent variables (Almond et al. 2001).

For observables with a dichotomous response (i.e., that can either be correct or incorrect), the multivariate item response takes the form

$$p_{ij}(x_{ij}=1|a_j,b_j,\Theta_{ij},\Phi_{im}) = \frac{1}{1+\exp[-K*a_j(\Theta_{ij}+\Phi_{im}-b_j)]} \qquad (1)$$

where

- K is a scaling constant,
- P_{ij} is the probability of student i correctly responding to item j,
- Θ_{ij} is the value of student i on the parent proficiency j,
- Φ_{im} is the value of student i on latent context effect m,
- a_j is the slope of the item response function for item j, and
- b_j is the difficulty of the item response function for item j.

The probability of responding incorrectly to the observable is the complement of success, 1-p_{ij}.

As previously explained, the context effect represents the correlation among responses to observables having a common context. In the Simulation scenario, there are three problems of increasing complexity. Each problem forms a context. Any task-specific skills contribute to a latent context propensity in the student. This parameterization of the context effect follows the item cluster effect model of Scott and Ip (2002). In the Bayesian IRT model, the context effect has prior

$$\Phi_{im} \sim N(0,\tau_m)$$

for task m. The precision of the context is given a gamma prior:

$$\frac{1}{\tau_m} \sim Gamma(.01,.01).$$

Gelman and colleagues (1995) point out that a gamma distribution with parameter values approaching zero constitutes a noninformative prior. In this case, the sampled values would be very dispersed, approaching a uniform distribution.

For observables with polytomous responses, i.e., that can be responded to in two or more categories of partial credit, the item response is more complicated. The probability of responding to each category of partial credit, or higher, is modeled as a compensatory multivariate item response as above, but with an additional item-category parameter, d_{jk}, for item j and category k. Since the probability is for a given category, *or higher,* it is referred to here as p^{cum}. Such a formulation follows Samejima (1969).

$$p^{CUM}(x_{ij}=k|a_j,b_j,\Theta_{ij},\Phi_{im}) = \frac{1}{1+\exp[-K*a_j(\Theta_{ij}+\Phi_{im}-b_j-d_{jk})]} \qquad (2)$$

where $p^{cum}_{i,j,k}$ is the probability of responding in categories k, k+1, …Q, where Q is the highest category of partial credit.

Although these parameters will be estimated by Bayesian techniques using a Markov Chain Monte Carlo (MCMC) algorithm, constraints to assure identifiability of item-category parameters were employed. This was accomplished by stipulating that the item-category associated with the first category, d_{j0} , is zero, and setting $\sum_{m=1}^{M} d_{jm} = 0$. In practice, only a single item category parameter was estimated. For three-category items, d_{j1} had a positive prior, $N(1{,}1000)$, and $d_{j2} = - d_{j1}$. For four-category items, $d_{j1} \sim N(1{,}1000)$, $d_{j2} = 0$ and $d_{j3} = -d_{j1}$. The positive prior means that d_{j1} will likely be associated with more difficult levels of item response.

Since the response probabilities are cumulative in that they are the probability of responding in category k or higher, the item-category probabilities (except for the last one) must be calculated by subtraction:

$$p_{\mathrm{i,j,0}} = 1 - p_{\mathrm{i,j,1}}^{\mathrm{cum}}$$

$$p_{\mathrm{i,j,1}} = p_{\mathrm{i,j,1}}^{\mathrm{cum}} - p_{\mathrm{i,j,2}}^{\mathrm{cum}}$$

$$\vdots$$

$$p_{\mathrm{i,j,Q}} = p_{\mathrm{i,j,Q}}^{\mathrm{cum}}$$

Determining the Scale of the Latent Proficiencies

The scale of the latent proficiencies is indeterminate. This indeterminacy can be resolved in a Bayesian model either by specifying strong informative priors or by constraining the item parameters. The latter course was taken. The scale for each of the measured latent proficiencies was determined by setting the following constraints on the item parameters corresponding to the observables that measure that scale:

$$\sum_{j=1}^{\mathrm{J_p}} b_{\mathrm{jp}} = 0$$

where $\mathrm{J_p}$ is the number of items in proficiency p, and b_{jp} is the difficulty parameter for item j in proficiency p; and

$$\prod_{j=1}^{J_p} a_{\mathrm{jp}} = 1 \;,$$

where a_{jp} is the slope associated with item j in proficiency p.

Structural Equation Model

There is a network of relations among the student model variables. These structural relations are modeled as simple linear regressions:

$$\Theta_{\mathrm{i}}^{\mathrm{child}} = B_0 + B_1 * \Theta_{\mathrm{i}}^{\mathrm{parent}} + e_{\mathrm{i}}$$

$$\text{with,} \qquad (3)$$

$$\mathrm{VAR}(e_{\mathrm{i}}) = \sigma_{\text{child-parent}}^{2}$$

In the Simulation scenario, for example, these describe how PS-TRE influences computer skills, how PS-TRE influences scientific inquiry exploration skill, and how PS-TRE influences scientific inquiry synthesis skill.

Because of the complexity of the overall model, the structural equations were constrained to a limiting case with slopes fixed to 1.0. An informative prior was set for B_0 , at N(0,1). Finally, Var(e_{i}) was set to 1.0, as a way to control the overall variance of the proficiency estimates.

Structured Prior for the Summary Proficiency, Problem Solving in Technology-Rich Environments

With all NAEP assessments, the average number of items measuring each subproficiency for an examinee is small. Such sparseness of measurement can lead to biased estimates of group quantities. A way to remedy this problem is to use auxiliary information related to an examinee's ability in the estimation of group means and percents. This is accomplished by regressing latent proficiency scores on student background information. In operational NAEP, a Bayesian estimation procedure is employed in which item response information is combined with student background information to get posterior distributions of proficiency for each examinee (Mislevy 1991). In the present application, background information is introduced by defining a structured prior on the unmeasured summary proficiency, PS-TRE.

Auxiliary information is introduced by assuming that an examinee's prior ability is structured (i.e., derived from a regression of proficiency on background variables),

$$\text{PS}-\text{TRE}_{\text{i}} \sim N(\mathbf{\Gamma}'\mathbf{y}_{\text{i}}, \sigma^2), \tag{4}$$

where $\mathbf{y}_\mathbf{i}$ is a vector of background variables for examinee i, Γ is a vector of regression effects, and σ^2 is a common variance for all examinees. In the present application, there are 10 categorical background variables that are recoded into 21 dummy variables. These variables consist of gender, race/ethnicity, whether the student had disabilities or was an English language learner, whether the scenario was administered to the student on a laptop computer, prior computer knowledge level, and socioeconomic status (SES), including parents' education level, number of reading-related materials in the home, whether the student was eligible for free/reduced-price school lunch, and whether the student was in the Title I program.

In order to control the contribution to proficiency variance made by the structured prior, two conditions were imposed. First, regression parameters were given informative priors with high precision,

$$\Gamma_{\text{p}} \sim N(1,1), \tag{5}$$

for regression weight p (p = 1 to 21). Next, the predictors, $\mathbf{y}_{\text{i}}$, were standardized and weighted by approximately $\frac{1}{\sqrt{21}}$ (the square root of the inverse of the number of predictors), so that the variance would not increase as the number of predictors increased. The R-squares of the conditioning models for the Search and Simulation scenarios were modest, between .34 and .41, but within the range of operational NAEP assessments.

In the present application, regression parameters, variance components, and the prior proficiency distribution of PS-TRE are estimated by using an MCMC algorithm, in which all model parameters are jointly estimated, conditional on the data. A general outline of the MCMC algorithm will be given in the next section.

General Description of MCMC Estimation Techniques

In operational NAEP procedures, item parameters are estimated using a marginal maximum likelihood approach (Muraki and Bock 1997). Multivariate proficiencies with a structured prior distribution are estimated in a conditioning phase in which item parameters in the first phase are introduced as fixed parameters (Mislevy 1991). In TRE, an MCMC algorithm to estimate all parameters simultaneously was employed. For item parameter estimates, the MCMC approach has been shown to produce point estimates and standard errors that are similar to those in operational NAEP estimates (Patz and Junker 1999). Further, if the scope is extended to include item parameters, conditioning parameters, and sampling variances, MCMC estimation produces results similar to those produced by operational NAEP techniques, when models are parallel (Johnson and Jenkins 2005). In the present research, MCMC estimation is applied to a model that is unlike an operational NAEP model in several key aspects (e.g., multivariate items and structured relationships among latent proficiencies). Also, unlike that in Johnson and Jenkins, the present model does not incorporate estimates of sampling variances. These are estimated by a separate jackknife procedure, which is an approach similar to that of Scott and Ip (2002).

A Markov chain is a sequence of random variables,

$$\psi^{1}, \psi^{2}, \ldots, \psi^{\mathrm{T}},$$

such that the probability of observing ψ^{t} is the transition probability,

$$p(\psi^{t} \mid \psi^{t-1}). \tag{6}$$

So ψ^{t} depends only on the previous state of the chain.

Under certain regularity conditions (Tierney 1994, section 3.1), the Markov chain converges to a stationary distribution (i.e., is invariant over time *t*). The general idea behind MCMC estimation is to set up a chain, which converges to a stationary distribution that equals the joint conditional distribution of model parameters, given data:

$$p(\psi \mid \mathrm{X}).$$

The procedure for deriving statistical estimates from a Markov chain is the following: Simulate a series of "burn in" observations from the chain until it is judged that the chain has converged to its stationary distribution,

$$\psi^{\text{-M}}, \psi^{\text{-(M-1)}}, \ldots, \psi^{0}.$$

The Gelman-Rubin diagnostic gives one test for convergence (Gelman and Rubin 1992). The M iterations till convergence are called "burn in iterations." For the burn-in phase, 5000 iterations were required. These were then tested for convergence.

After convergence, a series of T further observations are drawn from the joint distribution of the model parameters:

$$\psi^{1}, \psi^{2}, \ldots, \psi^{\mathrm{T}}.$$

Typically, between 5,000 and 10,000 samples of each parameter were drawn from the joint posterior.

Point estimates of model parameters are calculated from sample averages:

$$\hat{\psi}_{\mathrm{p}} = \frac{1}{\mathrm{T}} \sum_{t=1}^{\mathrm{T}} \psi^{t}_{\mathrm{p}}, \tag{7}$$

where T is the number of MCMC iterations.

This procedure would yield a point estimate of parameter p, such as an item difficulty or the proficiency score for examinee i. However, for more complex parameters, such as "percent above achievement-level cut-point K," estimates are averages of functions of parameters:

$$\begin{aligned} \hat{\Theta}_{\mathrm{p}} &= \sum_{t=1}^{\mathrm{T}} f(\psi^{t}_{\mathrm{p}}) \\ &= \frac{1}{\mathrm{T}} \sum_{t=1}^{\mathrm{T}} \frac{1}{\mathrm{N}} \sum_{i=1}^{\mathrm{N}} I(\Theta^{t}_{i}), \end{aligned} \tag{8}$$

where $I(\Theta^{t}_{i}),$ is an indicator of whether proficiency Θ for examinee i is at or above achievement-level cutscore K, and N is the sample size.

It is often difficult to simulate multivariate draws from the joint conditional distribution. A way to simplify the process is to take univariate draws from a distribution conditional on the data and all other model parameters. This has been shown to approximate draws from the joint posterior distribution (Geman and Geman 1984). By this approach, one draw of the parameters at iteration t, ψ^{t}, would consist of P univariate draws, each draw conditioned on the data and the rest of the parameters. If a set of parameters is symbolized by Ω, then the sequential set of draws for iteration t is described by:

$$\begin{aligned} &\psi^{t+1}_{1} \sim \pi(\psi_{1} \mid X, \psi^{t}_{1\notin\Omega}) \\ &\psi^{t+1}_{2} \sim \pi(\psi_{2} \mid X, \psi^{t}_{2\notin\Omega}) \\ &\vdots \\ &\psi^{t+1}_{\mathrm{P}} \sim \pi(\psi_{\mathrm{P}} \mid X, \psi^{t}_{\mathrm{P}\notin\Omega}) \end{aligned},$$

where $\pi(* \mid *)$ is the stationary distribution of a parameter, and $\psi^{t}_{p\notin\Omega}$ is the most current vector of parameters with parameter p excluded.

The MCMC simulating package BUGS (Spiegelhalter et al. 2004) was used to get Bayesian estimates of parameters. When posterior distributions can be explicitly defined, BUGS uses a Gibbs sampler. When posterior distributions of a particular parameter are not explicitly available, it uses two types of approximation for the univariate draw: Metropolis Hastings (Metropolis et al. 1953) and slice sampling (Neal 2003). In the present research, BUGS employed all three types of sampling.

Estimation of Population Parameters

Point estimates for most model parameters (e.g., item parameters and regression coefficients) were calculated from MCMC sample averages as described in equation 7. However, for estimates of mean proficiencies of student groups and their associated standard errors, approximation procedures from operational NAEP were employed.

Plausible Values of Latent Proficiencies

Plausible values consist of a set of M independent draws from each examinee's posterior proficiency distribution. With MCMC estimation, drawing plausible values consists of systematically selecting 5 values from the thousands of MCMC draws, taking care that each draw has a minimum of 50 draws between them. Equation 6 implies that each MCMC draw is dependent on the previous draw. As a result, the MCMC series of parameter draws are autocorrelated. Diagnostics indicated that it took about 25 to 50 draws for the autocorrelation to fall to zero. In practice, the 5 independent draws were separated by several hundred iterations. Following NAEP terminology, these 5 independent draws will be called plausible values (Allen, Carlson, and Zelenak 1999, section 12.3.3).

Calculating Student Group Means

The Bayesian model did not contain a model for the population. Such a model would have to include proficiency distributions corresponding to all primary sampling units and schools in the sampling frame. This would have been impractical for the present analysis. As a result, sampling weights are used to approximate population estimates.

The targets of reporting are student group means and standard errors. Student group means are calculated on each of the 5 plausible values and then averaged:

$$\hat{\mu}_{\mathrm{kG}} = \frac{1}{N_{\mathrm{G}}} \sum_{i \in \mathrm{G}} w_i PV_{\mathrm{k}i}, \tag{9}$$

where $\hat{\mu}_{\mathrm{kG}}$ is the estimated population mean of student group G, for the k^th^ set of plausible values, w_i is a sampling weight for examinee I, N_G is the weighted size (sum of sample weights) of student group G, and PV_{ki} is the plausible value k for examinee i.

Point estimates are averages over plausible values (Allen, Carlson, and Zelenak 1999, section 12.4.1),

$$\hat{\mu}_{\mathrm{G}} = \frac{1}{\mathrm{M}} \sum_{k=1}^{\mathrm{M}} \hat{\mu}_{k\mathrm{G}}, \tag{10}$$

where M is the number of plausible values (which is 5 in this application).

Estimating Standard Errors

Measurement variance

Measurement variance is the variance across plausible values of the target statistic. The first step in the procedure is to calculate t_{im}, a sample statistic, based on the m$^{\mathrm{th}}$ plausible value. It is equal to either a student group mean or a student group percent above achievement level. The variance over plausible values is:

$$U_{\mathrm{G}} = \frac{1}{\mathrm{M}-1} \sum_{m=1}^{\mathrm{M}} (t_{\mathrm{G}}^{m} - \bar{t}_{\mathrm{G}})^2, \tag{11}$$

where U_G is the measurement variance, $t_{\mathrm{G}}^{\mathrm{m}}$ is the value of the statistic over all examinees in group G for plausible value m, and $\bar{t}_{\mathrm{G}}$ is the mean value of the statistic averaged over plausible values.

Sampling variance

The procedure used to estimate sampling variance followed operational NAEP procedures. Typically, schools are grouped into 2P primary sampling units (PSUs). These are stratified into P pairs of PSUs, where the PSUs within a pair are similar on various SES measures. The procedure of the jackknife is to work through the P pairs one by one. Each time a PSU pair is selected, a single PSU is dropped from the pair, the data are suitably reweighted, and an estimated sample statistic (called a pseudoestimate), $t_{\mathrm{G}}^{\mathrm{p}}$, is calculated on the remaining sample. In the present case, this statistic is a group mean. This process is followed till a series of P sample statistics is estimated, $t_{\mathrm{G}}^{1}, t_{\mathrm{G}}^{2}, \ldots, t_{\mathrm{G}}^{\mathrm{P}}$. The sampling variance is calculated as

$$V_{\mathrm{G}} = \sum_{p=1}^{\mathrm{P}} (t_{\mathrm{G}}^{p} - \bar{t}_{\mathrm{G}})^2, \tag{12}$$

where $\bar{t}_{\mathrm{G}}$ is the average statistic over P pseudoestimates.

Note that the proper estimate of V_G is the average of the estimate calculated over the k set of plausible values. Practice in NAEP has shown that using an estimate based on one plausible value is sufficiently accurate.

Standard errors

The total variance of a sample statistic is a weighted combination of measurement and sampling variances (Mislevy 1991). As a result, the standard error for a sample statistic for group *G* is

$$SE_G = \sqrt{V_G + (1+\frac{1}{M})U_G}, \qquad (13)$$

where M is the number of plausible values (Allen, Carlson, and Zelenak 1999, section 12.4.1).

Creation of a Real-Time Inference Engine for the Search Scenario

As part of the demonstration of the feasibility of delivering an assessment that uses the full potential of the computer, a Bayesian inference engine for the Search scenario was developed. A Bayesian inference engine is a system of variables like those depicted in figures F-1 and F-3. It is assumed that beliefs about the system, i.e., the conditional probability of any variable given the values of any precursor (parent) variables, can be defined. These conditional probabilities may come from the judgments of experts or from parameters estimated from the Bayesian analysis of data (as is the case with the present research). The goal of using an inference engine is to be able to estimate the probability distribution of any variable in the system given the observed or hypothesized value of any other variables in the system. On one hand, there is interest in being able to score an examinee; that is, given that a certain pattern of responses on the observables is obtained, it is desirable to estimate the distribution of the latent variables. On the other hand, there might be interest, given a certain profile of scores on the latent variables, in gauging the sensitivity of the model by estimating the probability of responding correctly on the observables.

Estimating probabilities in an inference engine is not straightforward. This is because often some variables in a network are not conditionally independent. As a result, information about observed values of variables may be redundantly accounted for when updating the system. To avoid such overcounting of evidence, a Bayes net has to be transformed into a structure that can propagate information throughout the network without redundancy. To accomplish this, a directed graph (such as the ones in figures F-1 and F-3) and conditional probabilities are translated into a linear inference tree, or clique tree. For details, see Lauritzen and Spiegelhalter (1988) and Pearl (1988). To make calculations in such a system tractable, all variables have to be defined as categorical. A program package called ERGO (Noetic Systems, Inc. 2001) automatically accomplishes the task of compiling a Bayes net into a linear inference tree.

There were several steps in defining an inference engine from the results of the Bayesian MCMC analysis.

1. Point estimates for all model parameters had to be extracted from the MCMC estimation.
2. The estimated sample distributions of the latent proficiency variables had to be made discrete. This was done by partitioning the distribution into 15 equal-probability regions. The values associated with these were the inverse normal probability functions of the midpoints.
3. Conditional probability tables that represent the relationship between the variables had to be constructed. The structural relations between latent proficiencies are represented with a normal translation model (Almond forthcoming), where the discrete values of the child variable are a linear function of the parent variable. This representation reflects the structural regression estimated in the MCMC phase. For the observables, the conditional probabilities of each observable are a function of the parent latent proficiency. This procedure employs an IRT model using item parameters from the Bayesian estimation.
4. The conditional probability tables were then imported into the ERGO program and compiled into a linear inference tree.

With the inference engine, it was possible to input profiles of latent proficiencies and see what probabilities of response resulted for the observables. For example, if a high level of computer skills was stipulated, there should be a high probability of a high score on all of the computer observables.

The inference engine was confirmed with the MCMC algorithm. This was done in the following way. The data were augmented by a few dozen dummy cases which had profiles of latent proficiencies fixed. This data set, which included some 1,100 *real* cases, was input into a run of the MCMC estimation program.[3] Average response probabilities of the observables corresponding to the dummy cases were then estimated. In a parallel analysis, the same profiles of latent proficiencies were input into the

[3] The n of ~1,100 was the number of students responding to the TRE Search scenario. This sample size was based on the minimum assumed for scaling in main NAEP and for detecting mean differences among reporting groups of interest.

inference engine, and the resulting response probabilities for observables were noted. It was found that the response probabilities derived from the MCMC algorithm almost exactly matched with those derived from the inference engine.

The ultimate utility of such a Bayes net would be to score results immediately from a computer-delivered assessment. It could also be part of a tailored test, in which the interim proficiency estimates would be used as a basis for deciding how to branch the assessment to more or less challenging activities.

In the current research, the inference engine provided a proof of concept for an approach to Bayesian IRT estimation. In an assessment using an inference engine, the model to estimate parameters from data could involve continuous latent-proficiency variables. It has been demonstrated that parameters from such a model can be translated into a discrete system.

Appendix G: C-rater Rules for Scoring Students' Search Queries

Terms are assigned to the following seven categories:

1. *Comparative terms:* better, advantages, disadvantages, prefer, more, over, worse
2. *Relevant terms:* weather, atmosphere, space, outer space, cost, helium, science, scientist, astronomer, astronomy, astrophysics, NASA, study, research, explore, learn, experiment
3. *Tool terms:* satellite, rocket, telescope, space shuttle
4. *Weak balloon terms:* balloon, air balloon, hot air balloon
5. *Good balloon terms:* gas balloon, helium balloon, helium gas balloon, weather balloon
6. *Special balloon terms:* scientific balloon, scientific gas balloon, scientific helium balloon, super pressure balloon, long duration balloon, zero pressure balloon
7. *Explore terms:* study, research, explore, learn, experiment

Scoring rules (numbers represent categories):

SCORE = 2

1. 1 & 3 & 4
2. 1 & 2 & 7
3. 1 & 3 & 7
4. 2 & 5
5. 3 & 5
6. 6
7. 2 & 3 & 4
8. 4 & 2 (at least two from 2)
9. 4 & 3 (at least two from 3)

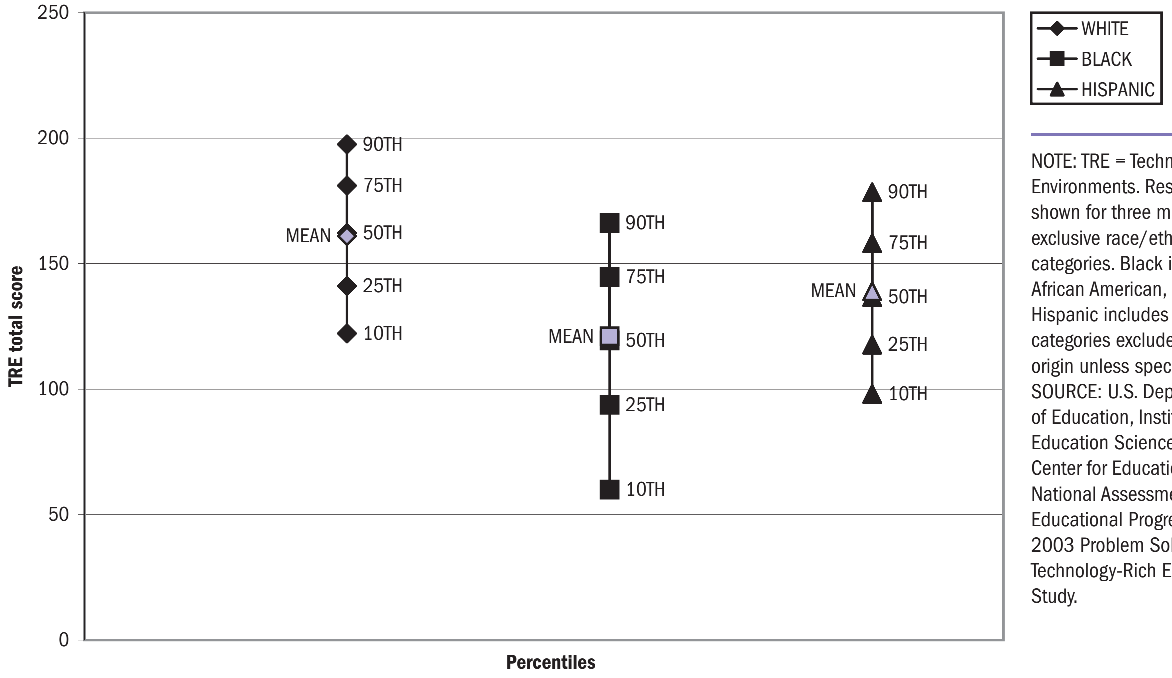
SCORE = 1

10. 2 & 3
11. 2 & 4
12. 3 & 4
13. 5

SCORE = 0 if no rules are met.

Appendix H: TRE Search and Simulation Scale Scores and Percentiles by Student Reporting Groups for Scales on Which Statistically Significant Group Differences Were Observed

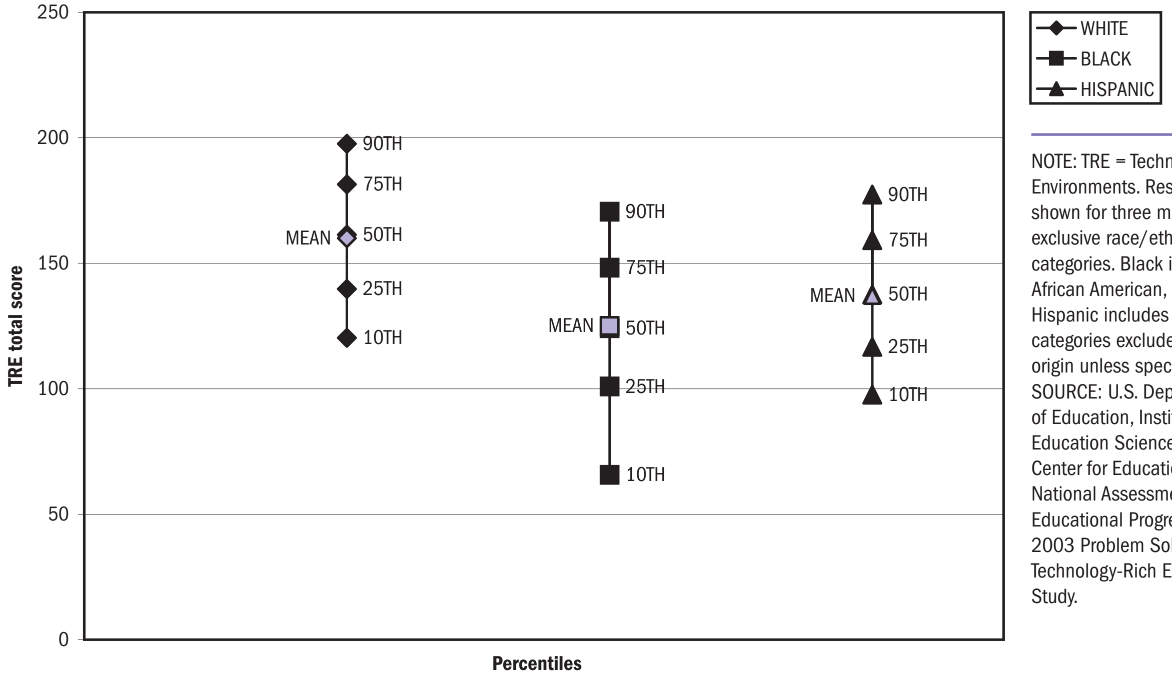
Figure H-1. TRE Search total score distribution, by race/ethnicity, grade 8: 2003



NOTE: TRE = Technology-Rich Environments. Results are shown for three mutually exclusive race/ethnicity categories. Black includes African American, and Hispanic includes Latino. Race categories exclude Hispanic origin unless specified.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

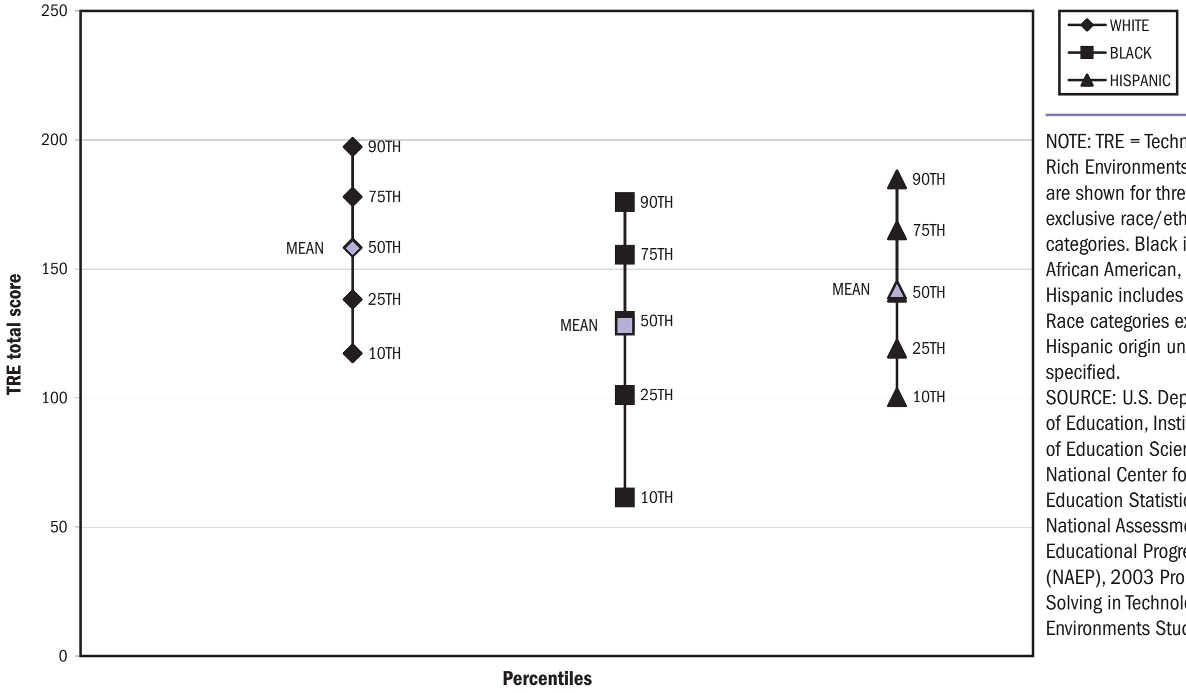
Figure H-2. TRE Search scientific inquiry skill score distribution, by race/ethnicity, grade 8: 2003



NOTE: TRE = Technology-Rich Environments. Results are shown for three mutually exclusive race/ethnicity categories. Black includes African American, and Hispanic includes Latino. Race categories exclude Hispanic origin unless specified.

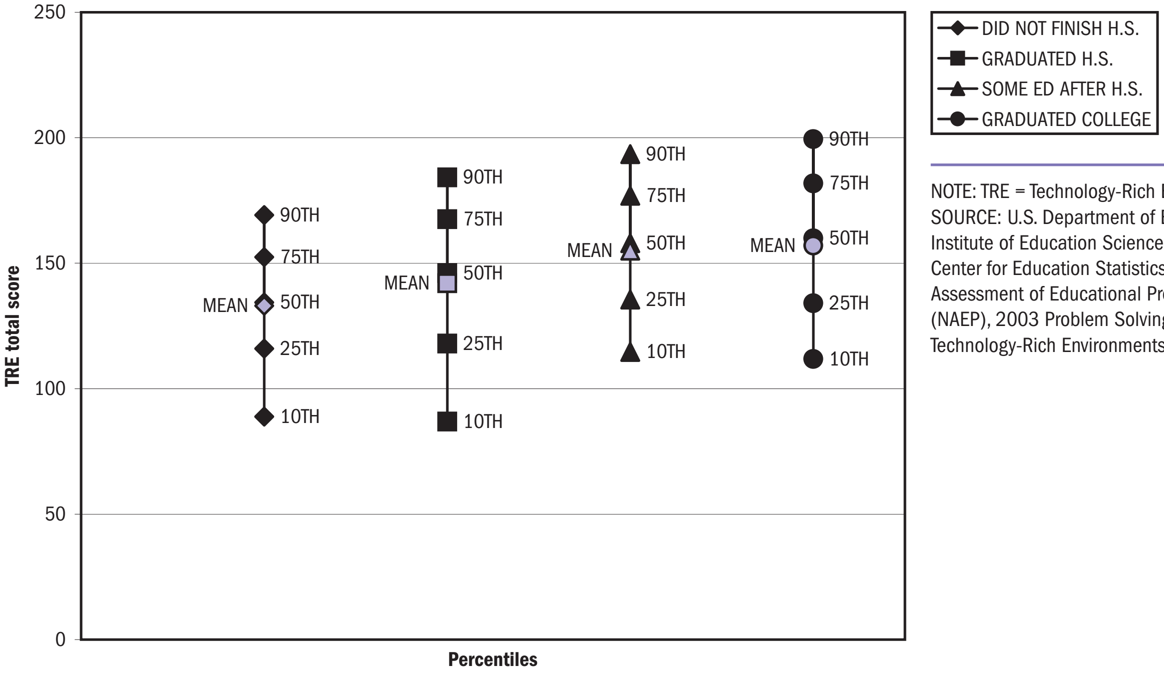
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure H-3. TRE Search computer skills score distribution, by race/ethnicity, grade 8: 2003



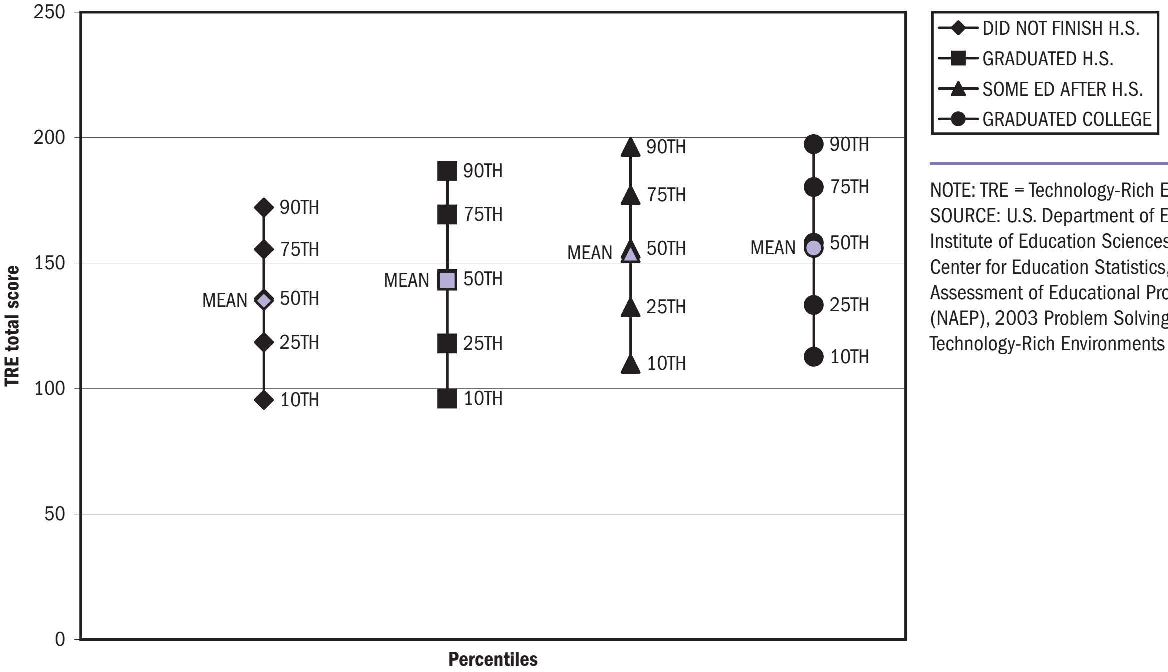
NOTE: TRE = Technology-Rich Environments. Results are shown for three mutually exclusive race/ethnicity categories. Black includes African American, and Hispanic includes Latino. Race categories exclude Hispanic origin unless specified.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure H-4. TRE Search total score distribution, by student-reported parents' highest education level, grade 8: 2003



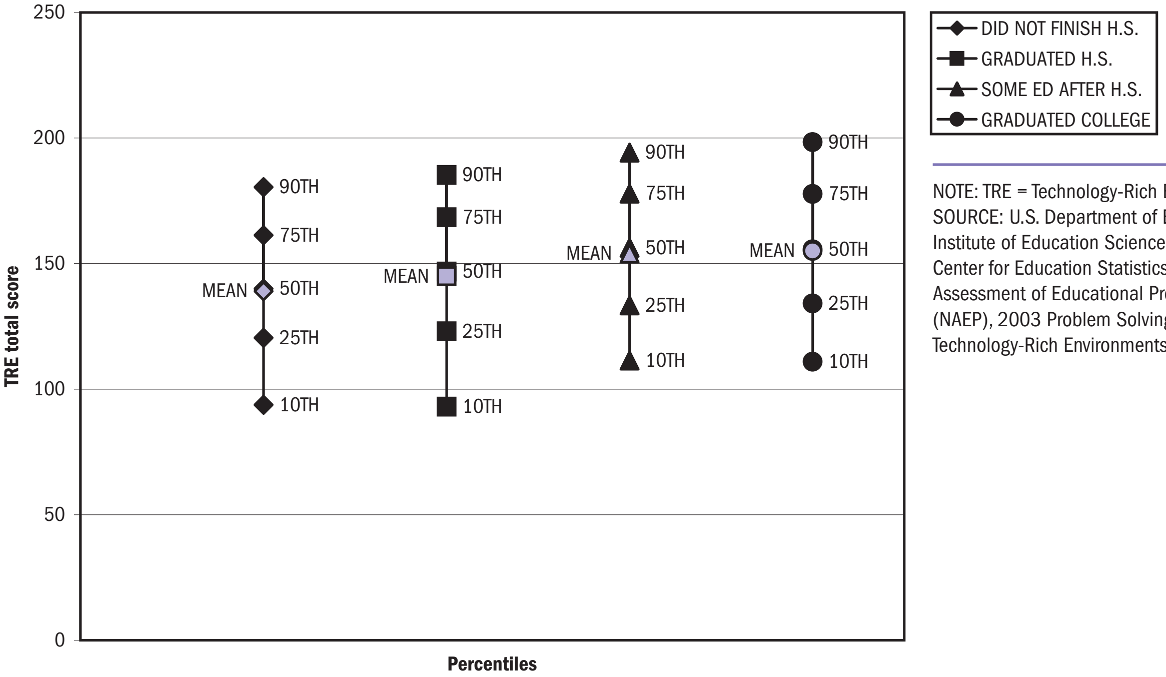
NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure H-5. TRE Search scientific inquiry skill score distribution, by student-reported parents' highest education level, grade 8: 2003



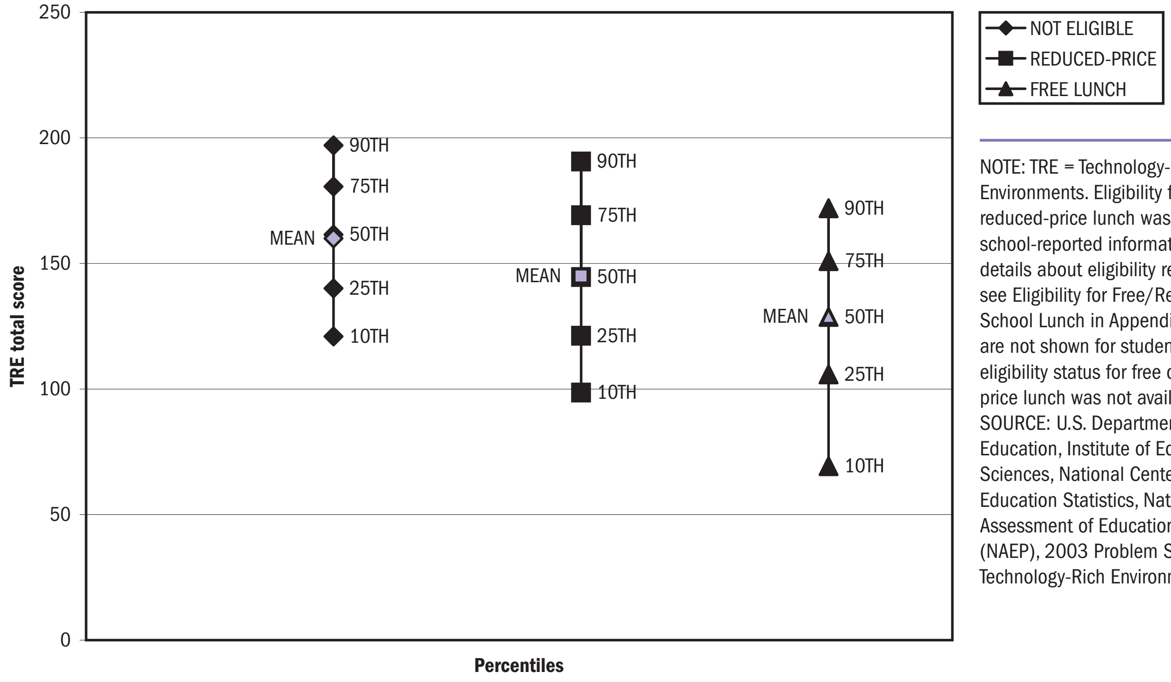
NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure H-6. TRE Search computer skills score distribution, by student-reported parents' highest education level, grade 8: 2003



NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

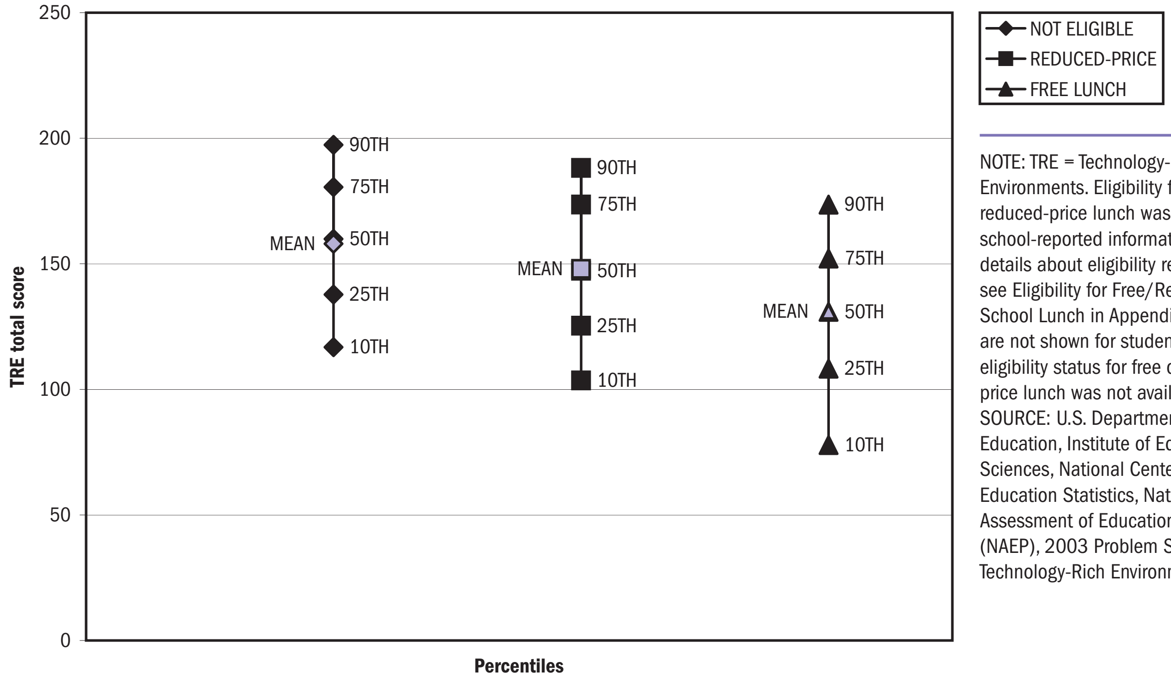
Figure H-7. TRE Search total score distribution, by eligibility for free or reduced-price school lunch, grade 8: 2003



NOTE: TRE = Technology-Rich Environments. Eligibility for free or reduced-price lunch was based on school-reported information. For details about eligibility requirements, see Eligibility for Free/Reduced-Price School Lunch in Appendix K. Results are not shown for students whose eligibility status for free or reduced-price lunch was not available.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

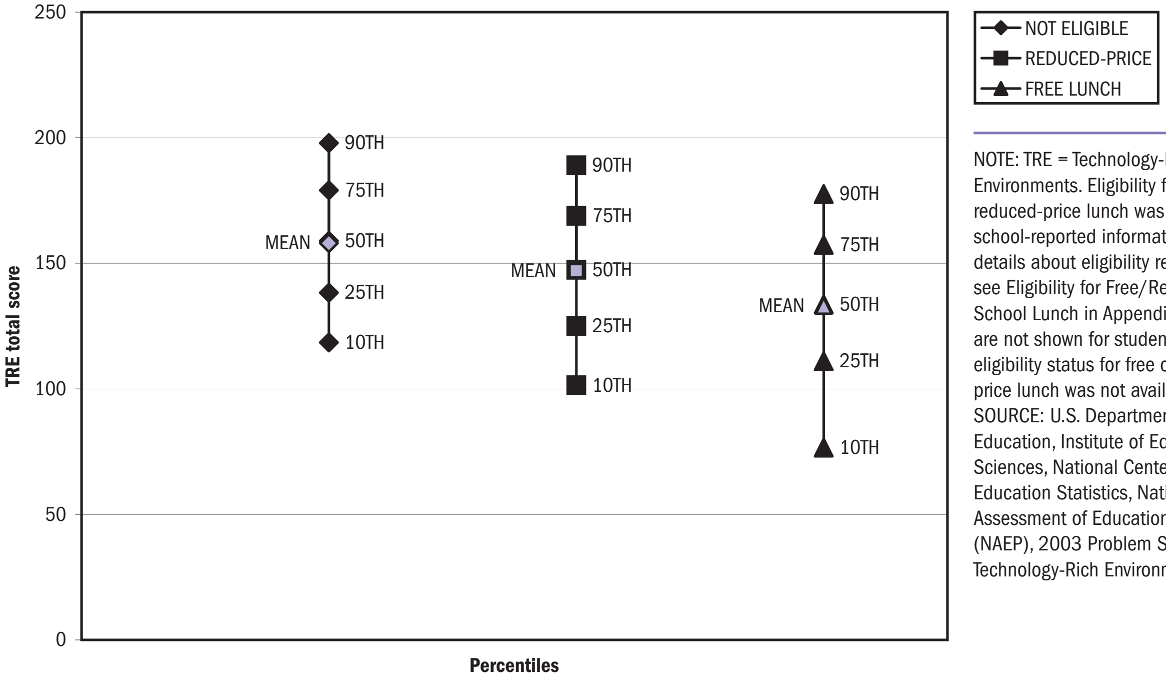
Figure H-8. TRE Search scientific inquiry skill score distribution, by eligibility for free or reduced-price school lunch, grade 8: 2003



NOTE: TRE = Technology-Rich Environments. Eligibility for free or reduced-price lunch was based on school-reported information. For details about eligibility requirements, see Eligibility for Free/Reduced-Price School Lunch in Appendix K. Results are not shown for students whose eligibility status for free or reduced-price lunch was not available.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

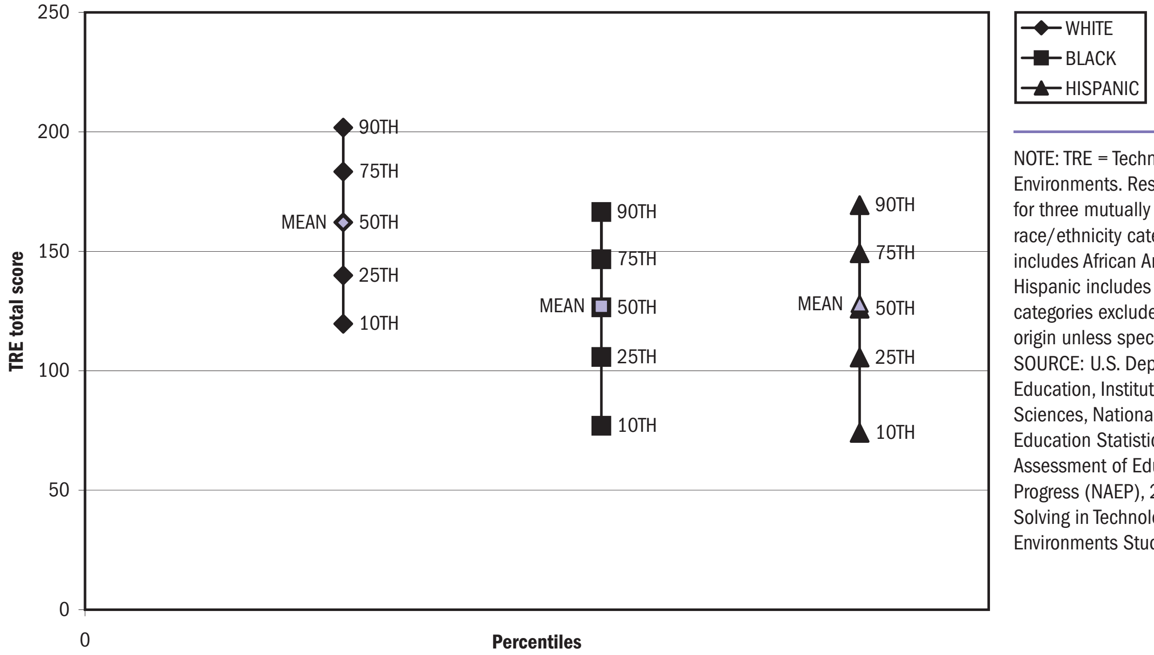
Figure H-9. TRE Search computer skills score distribution, by eligibility for free or reduced-price school lunch, grade 8: 2003



NOTE: TRE = Technology-Rich Environments. Eligibility for free or reduced-price lunch was based on school-reported information. For details about eligibility requirements, see Eligibility for Free/Reduced-Price School Lunch in Appendix K. Results are not shown for students whose eligibility status for free or reduced-price lunch was not available.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

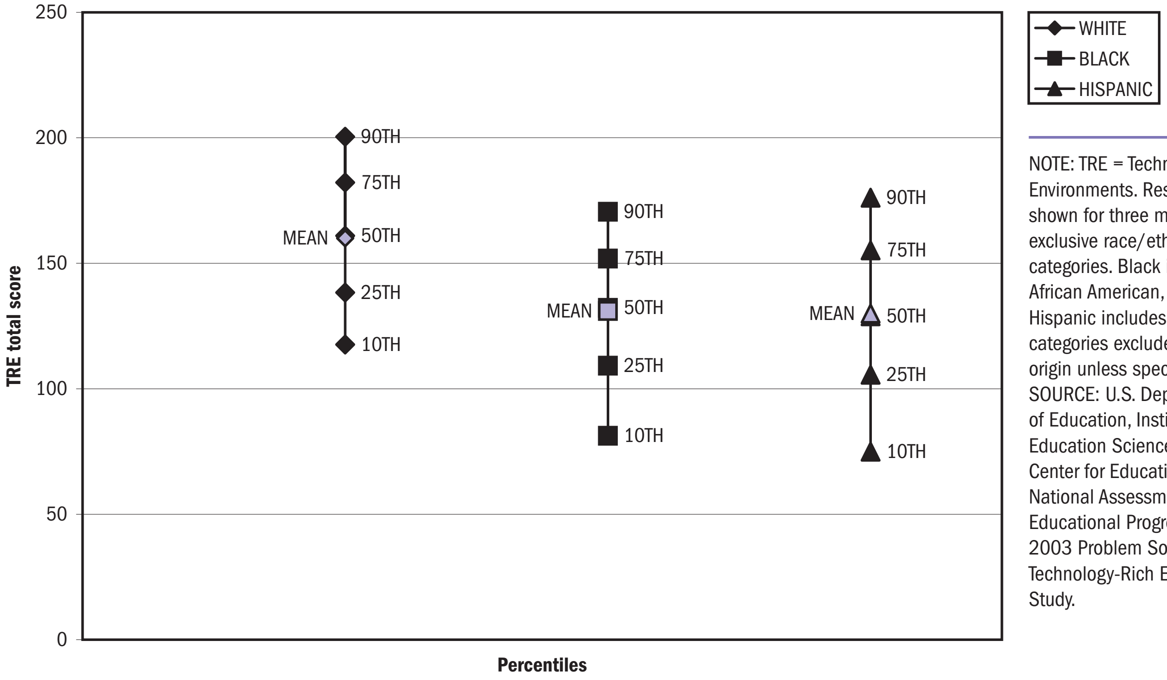
Figure H-10. TRE Simulation total score distribution, by race/ethnicity, grade 8: 2003



NOTE: TRE = Technology-Rich Environments. Results are shown for three mutually exclusive race/ethnicity categories. Black includes African American, and Hispanic includes Latino. Race categories exclude Hispanic origin unless specified.

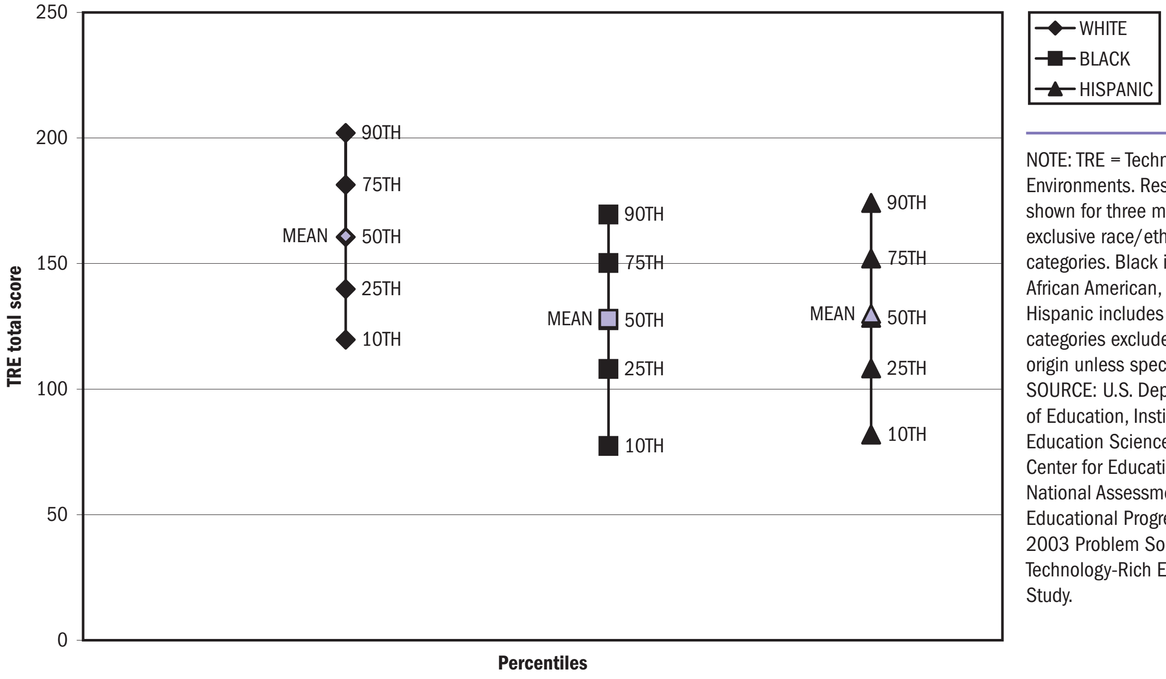
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure H-11. TRE Simulation scientific exploration skill score distribution, by race/ethnicity, grade 8: 2003



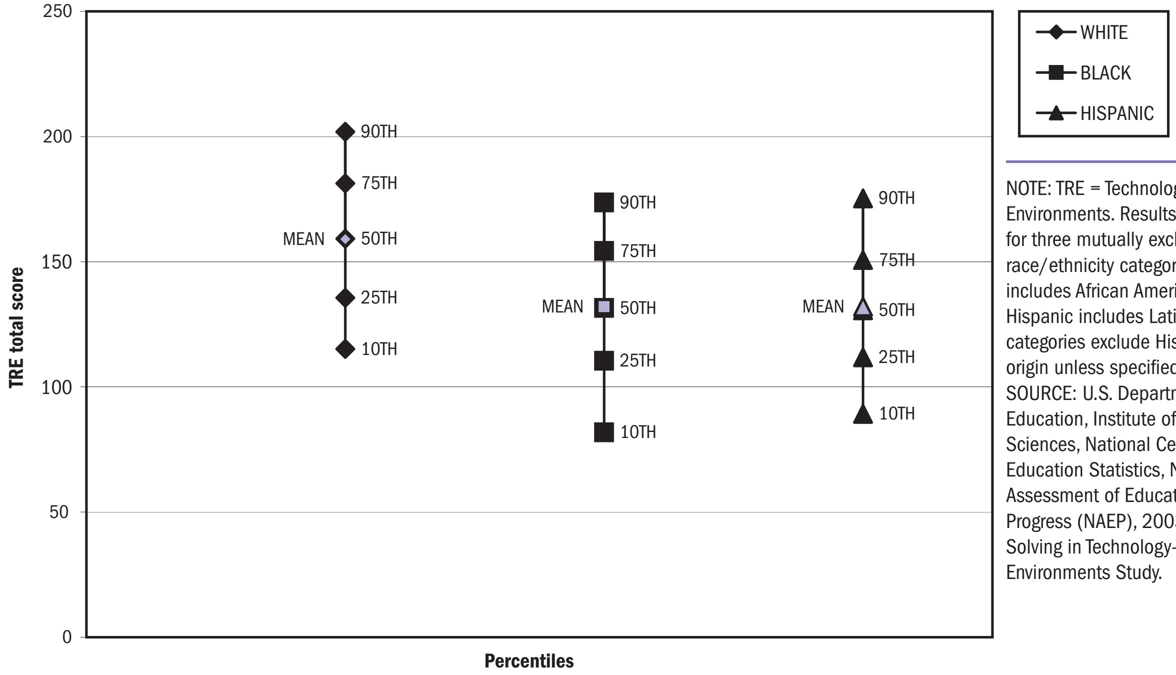
NOTE: TRE = Technology-Rich Environments. Results are shown for three mutually exclusive race/ethnicity categories. Black includes African American, and Hispanic includes Latino. Race categories exclude Hispanic origin unless specified.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure H-12. TRE Simulation scientific synthesis score distribution, by race/ethnicity, grade 8: 2003



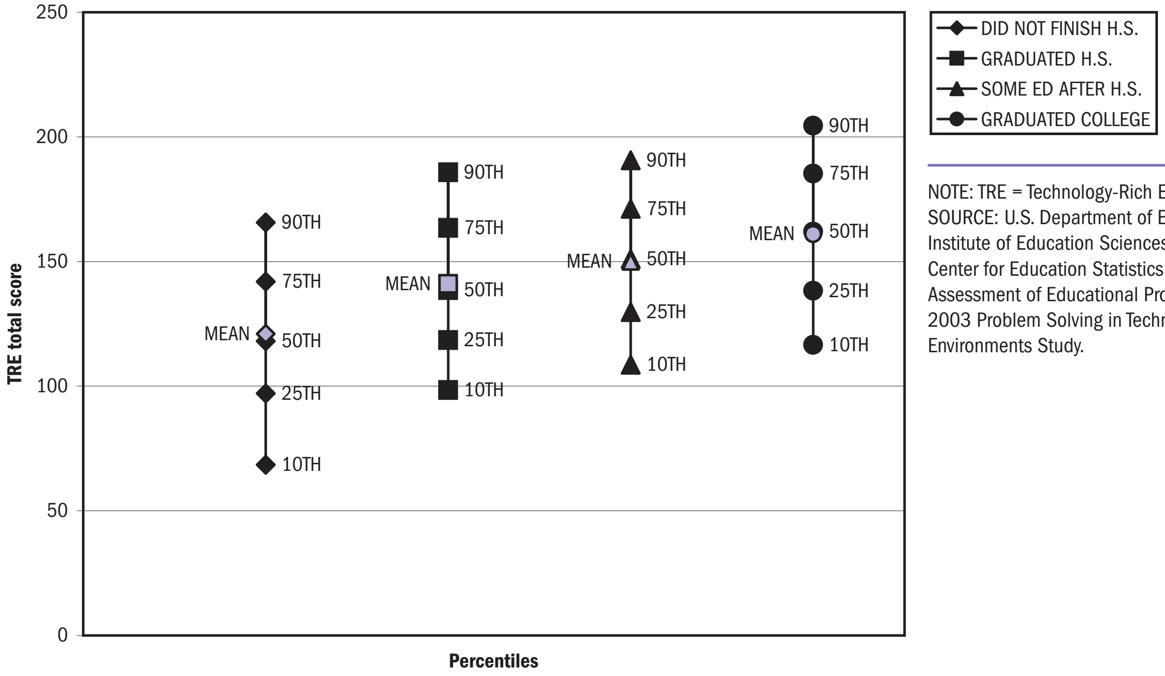
NOTE: TRE = Technology-Rich Environments. Results are shown for three mutually exclusive race/ethnicity categories. Black includes African American, and Hispanic includes Latino. Race categories exclude Hispanic origin unless specified.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure H-13. TRE Simulation computer skills score distribution, by race/ethnicity, grade 8: 2003



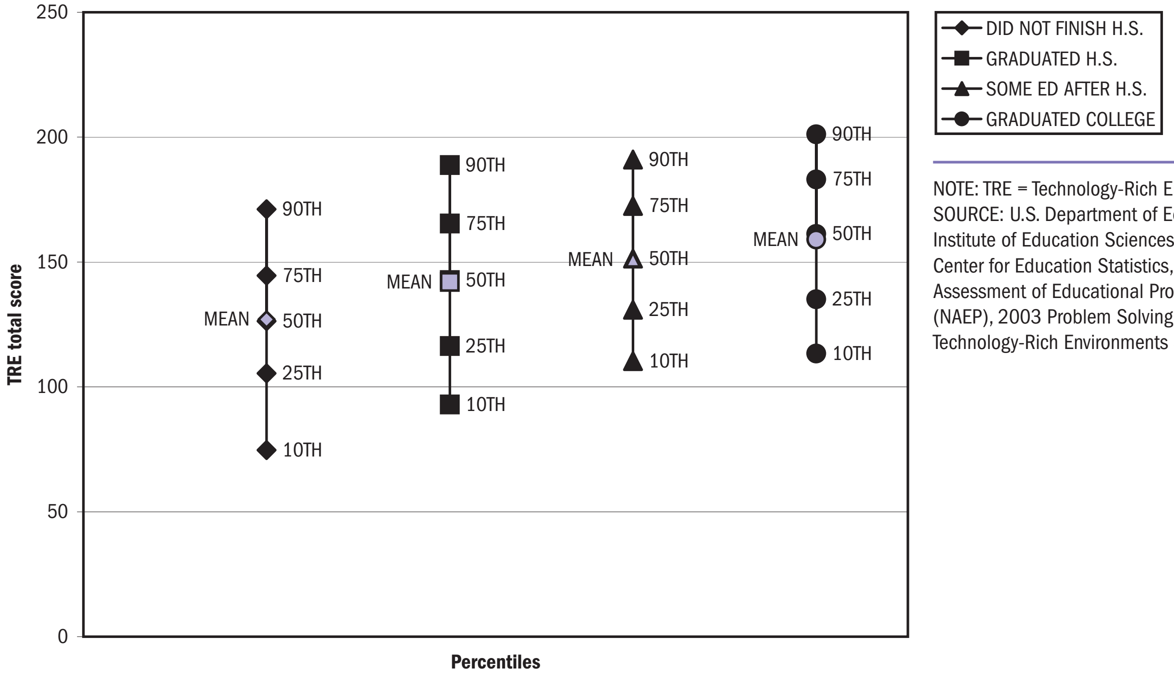
NOTE: TRE = Technology-Rich Environments. Results are shown for three mutually exclusive race/ethnicity categories. Black includes African American, and Hispanic includes Latino. Race categories exclude Hispanic origin unless specified.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure H-14. TRE Simulation total score distribution, by student-reported parents' highest education level, grade 8: 2003



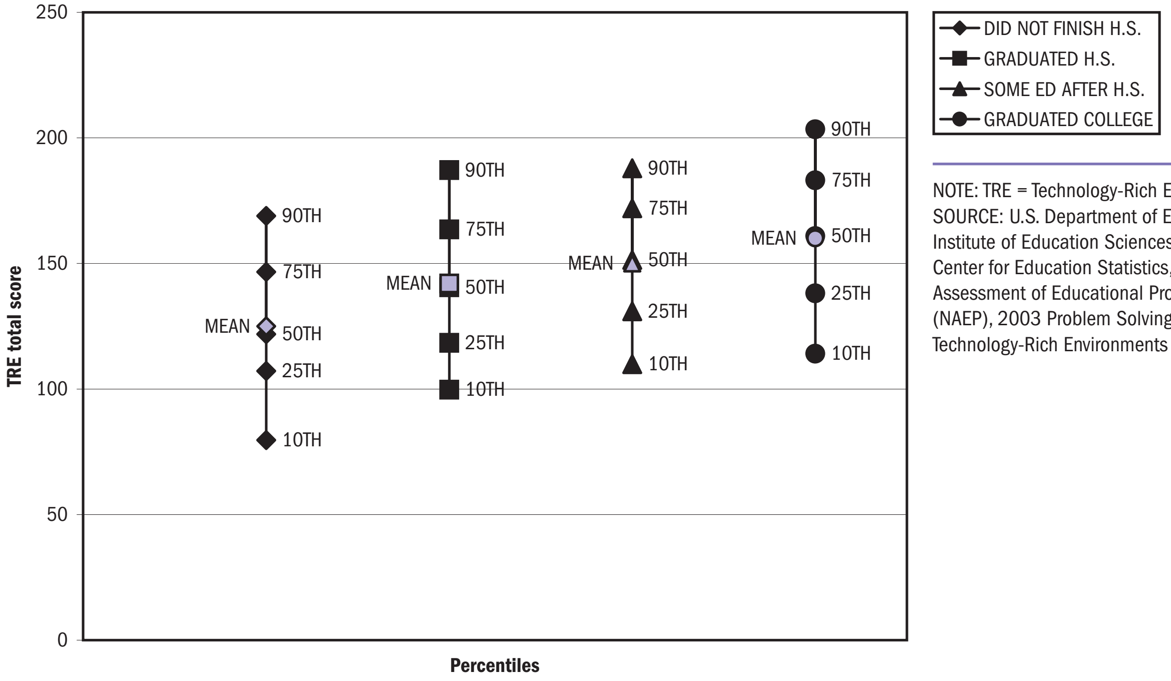
NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure H-15. TRE Simulation scientific exploration skill score distribution, by student-reported parents' highest education level, grade 8: 2003



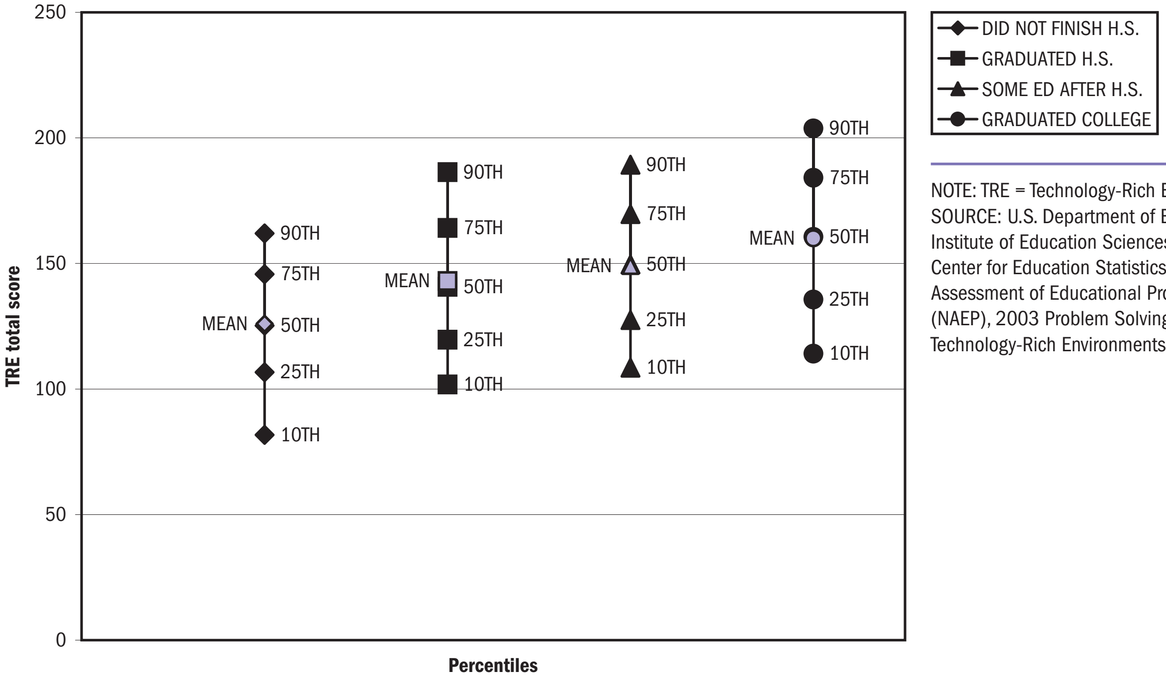
NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure H-16. TRE Simulation scientific synthesis score distribution, by student-reported parents' highest education level, grade 8: 2003



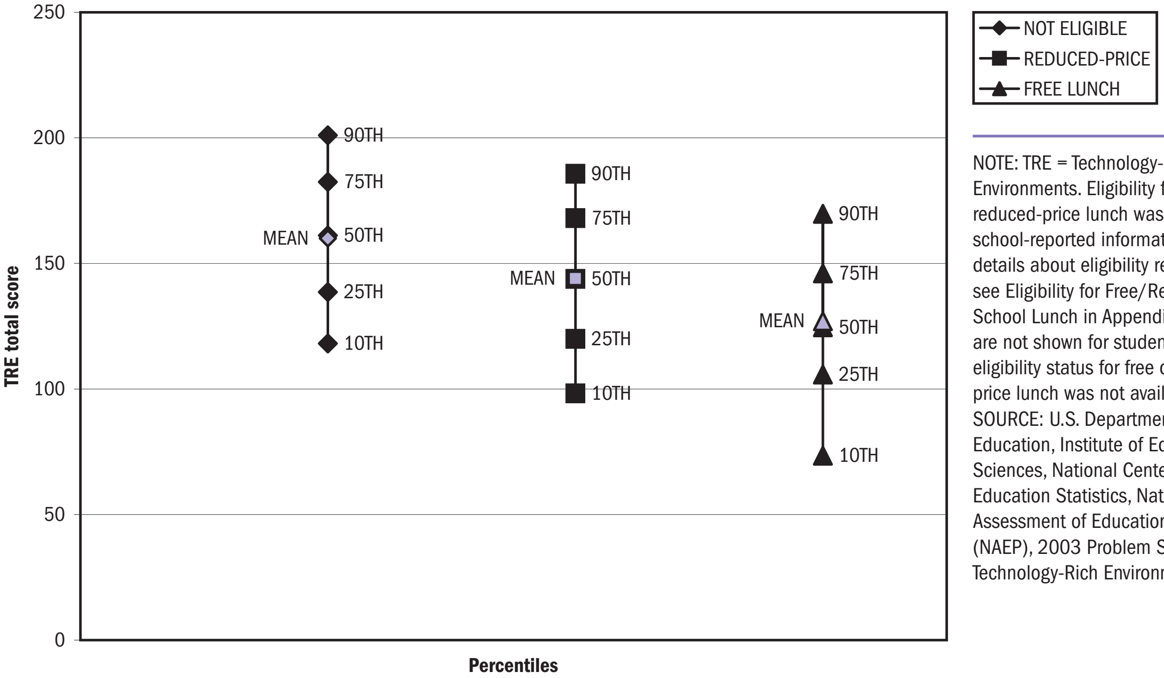
NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure H-17. TRE Simulation computer skills score distribution, by student-reported parents' highest education level, grade 8: 2003.



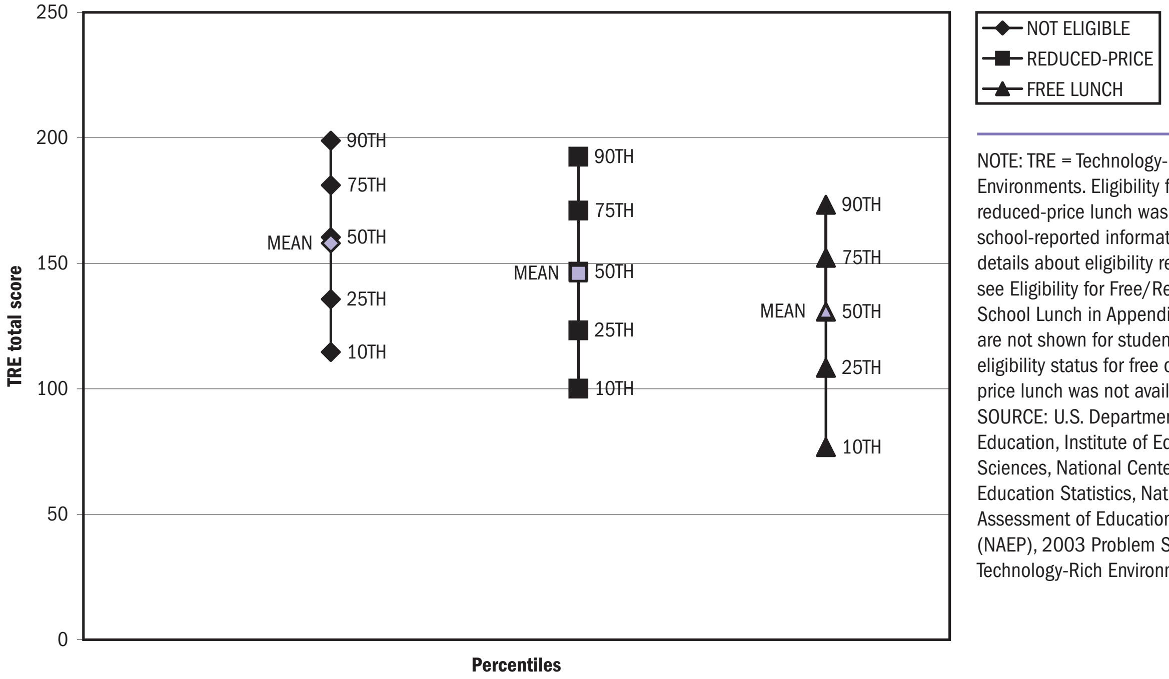
NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure H-18. TRE Simulation total score distribution, by eligibility for free or reduced-price school lunch, grade 8: 2003



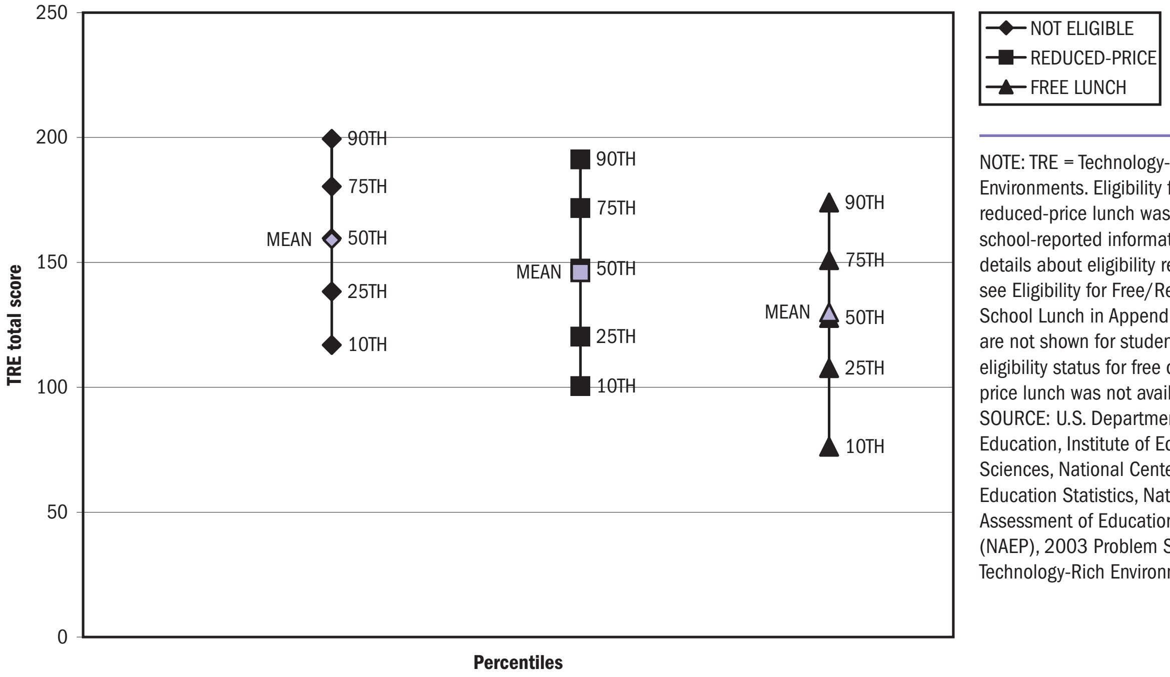
NOTE: TRE = Technology-Rich Environments. Eligibility for free or reduced-price lunch was based on school-reported information. For details about eligibility requirements, see Eligibility for Free/Reduced-Price School Lunch in Appendix K. Results are not shown for students whose eligibility status for free or reduced-price lunch was not available.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure H-19. TRE Simulation scientific exploration skill score distribution, by eligibility for free or reduced-price school lunch, grade 8: 2003



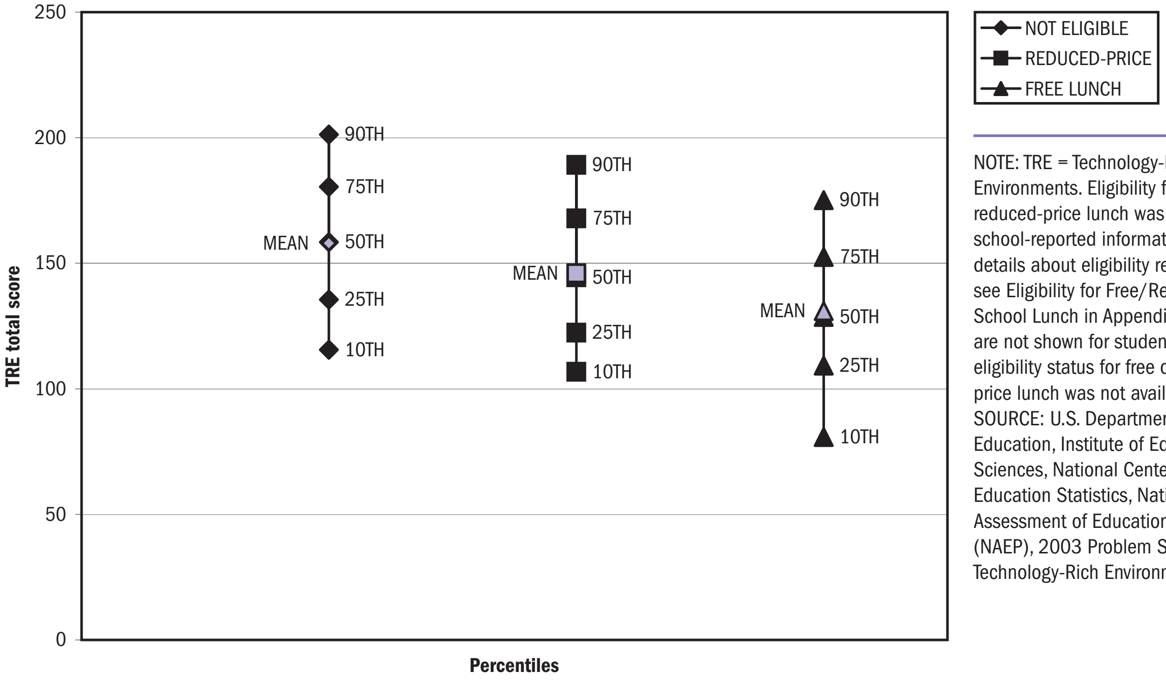
NOTE: TRE = Technology-Rich Environments. Eligibility for free or reduced-price lunch was based on school-reported information. For details about eligibility requirements, see Eligibility for Free/Reduced-Price School Lunch in Appendix K. Results are not shown for students whose eligibility status for free or reduced-price lunch was not available.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure H-20. TRE Simulation scientific synthesis score distribution, by eligibility for free or reduced-price school lunch, grade 8: 2003



NOTE: TRE = Technology-Rich Environments. Eligibility for free or reduced-price lunch was based on school-reported information. For details about eligibility requirements, see Eligibility for Free/Reduced-Price School Lunch in Appendix K. Results are not shown for students whose eligibility status for free or reduced-price lunch was not available.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure H-21. TRE Simulation computer skills score distribution, by eligibility for free or reduced-price school lunch, grade 8: 2003



NOTE: TRE = Technology-Rich Environments. Eligibility for free or reduced-price lunch was based on school-reported information. For details about eligibility requirements, see Eligibility for Free/Reduced-Price School Lunch in Appendix K. Results are not shown for students whose eligibility status for free or reduced-price lunch was not available.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Appendix I: Summary Statistics for Prior Knowledge Measures and Mean Scale Scores for Background-Question Response Options[1]

Table I-1. Unweighted summary statistics for Search scenario prior knowledge measures, grade 8: 2003

Statistic	Prior computer knowledge	Prior science knowledge
Number of students	1,059	1,062
Mean score	5.6	5.0
Standard deviation	2.1	1.8
Scale range	0–10	0–10
Coefficient alpha reliability	.58	.39

NOTE: Students' scores for a particular prior knowledge measure were deleted from this analysis if they did not answer all 10 questions in a scale.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Table I-2. Unweighted summary statistics for Simulation scenario prior knowledge measures, grade 8: 2003

Statistic	Prior computer knowledge	Prior science knowledge
Number of students	960	986
Mean score	5.5	5.3
Standard deviation	2.0	2.4
Scale range	0–10	0–10
Coefficient alpha reliability	.51	.67

NOTE: Students' scores for a particular prior knowledge measure were deleted from this analysis if they did not answer all 10 questions in a scale.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

[1] The items composing the Prior Computer Knowledge measure were the same for the Search and Simulation scenarios. For the Prior Science Knowledge measure, different items were used for each scenario.

Table I-3. Data for figure 5-3, mean scale scores, by extent of specific computer use and scale for Search scenario, grade 8: 2003

	Use a word processor			
Scale	Not at all	Small extent	Moderate extent	Large extent
Search total score	130 (4.1)	145 (2.8)	153 (2.2)	157 (2.5)
Search scientific inquiry score	132 (3.9)	145 (3.4)	153 (2.7)	156 (2.3)
Search computer skills score	133 (3.9)	146 (2.8)	151 (2.4)	159 (2.6)
	Make drawings/art on computer			
Scale	Not at all	Small extent	Moderate extent	Large extent
Search total score	151 (3.3)	152 (2.2)	149 (2.7)	138 (3.7)
Search scientific inquiry score	151 (3.2)	152 (2.2)	149 (3.3)	137 (4.0)
Search computer skills score	151 (2.4)	151 (2.3)	151 (2.5)	139 (4.2)
	Make tables, charts, or graphs on computer			
Scale	Not at all	Small extent	Moderate extent	Large extent
Search total score	145 (2.8)	155 (2.1)	150 (3.4)	134 (5.8)
Search scientific inquiry score	146 (2.9)	154 (2.7)	149 (2.8)	136 (5.8)
Search computer skills score	145 (2.8)	154 (1.8)	151 (3.7)	137 (5.6)
	Look up information on a CD			
Scale	Not at all	Small extent	Moderate extent	Large extent
Search total score	148 (2.8)	154 (2.7)	152 (2.7)	141 (3.4)
Search scientific inquiry score	149 (3.0)	154 (3.2)	151 (3.1)	143 (3.0)
Search computer skills score	148 (3.2)	153 (2.5)	152 (2.5)	144 (3.1)
	Find information on the Internet			
Scale	Not at all	Small extent	Moderate extent	Large extent
Search total score	‡	136 (3.8)	149 (2.7)	154 (2.2)
Search scientific inquiry score	‡	137 (4.4)	150 (3.4)	153 (2.3)
Search computer skills score	‡	134 (4.0)	149 (2.6)	154 (2.5)
	Use e-mail			
Scale	Not at all	Small extent	Moderate extent	Large extent
Search total score	138 (3.1)	146 (3.1)	151 (3.8)	156 (2.2)
Search scientific inquiry score	139 (3.8)	147 (3.7)	152 (2.6)	155 (2.2)
Search computer skills score	141 (3.3)	145 (3.0)	151 (2.7)	155 (2.1)
	Talk in chat groups			
Scale	Not at all	Small extent	Moderate extent	Large extent
Search total score	142 (2.6)	147 (3.6)	149 (3.3)	157 (2.3)
Search scientific inquiry score	143 (3.4)	147 (2.9)	149 (2.5)	156 (2.6)
Search computer skills score	143 (2.8)	147 (3.4)	149 (3.3)	157 (2.0)

‡ Reporting standards not met. Sample size was insufficient to permit a reliable estimate.

NOTE: The range of scores for each scale is 0–300. Standard errors of the estimated scores appear in parentheses.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Table I-4. Data for figure 5-4, mean scale scores, by frequency of computer use and scale for Search scenario, grade 8: 2003

	How often do you use a computer outside of school?				
Scale	Daily	2–3 times per week	Once a week	Once every few weeks	Never or hardly ever
Search total score	158 (2.4)	146 (2.2)	147 (3.6)	130 (5.8)	126 (5.1)
Search scientific inquiry score	157 (2.3)	147 (2.0)	147 (3.7)	131 (6.1)	129 (4.5)
Search computer skills score	157 (2.1)	148 (2.4)	147 (3.8)	129 (4.7)	131 (3.1)

NOTE: The range of scores for each scale is 0–300. Standard errors of the estimated scores appear in parentheses.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Table I-5. Data for figure 5-5, mean scale scores, by students indicating there is a computer at home that they use and scale for Search scenario, grade 8: 2003

	Is there a computer at home that you use?	
Scale	Yes	No
Search total score	153 (1.9)	125 (3.4)
Search scientific inquiry score	152 (1.9)	129 (3.3)
Search computer skills score	152 (1.9)	131 (3.5)

NOTE: The range of scores for each scale is 0–300. Standard errors of the estimated scores appear in parentheses.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Table I-6. Data for figure 5-6, mean scale scores, by frequency of school science activity and scale for Search scenario, grade 8: 2003

	Use the Internet to exchange information with other students or scientists about experiments			
Scale	Not taking science	Once a month or more	Sometimes, but less than once a month	Never
Search total score	‡	146 (3.3)	145 (3.5)	154 (2.2)
Search scientific inquiry score	‡	145 (3.4)	144 (3.3)	154 (1.8)
Search computer skills score	‡	147 (2.7)	147 (3.1)	153 (2.0)

‡ Reporting standards not met. Sample size was insufficient to permit a reliable estimate.
NOTE: The range of scores for each scale is 0–300. Standard errors of the estimated scores appear in parentheses.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Table I-7. Data for figure 6-4, mean scale scores, by extent of specific computer use and scale for Simulation scenario, grade 8: 2003

	Play computer games			
Scale	Not at all	Small extent	Moderate extent	Large extent
Simulation total score	140 (5.8)	149 (3.1)	153 (2.6)	152 (3.2)
Simulation scientific exploration score	137 (4.9)	149 (2.7)	153 (2.4)	154 (3.7)
Simulation scientific synthesis score	141 (4.8)	148 (3.4)	153 (2.2)	151 (3.3)
Simulation computer skills score	143 (6.0)	150 (3.7)	152 (3.7)	148 (4.0)

	Use a word processor			
Scale	Not at all	Small extent	Moderate extent	Large extent
Simulation total score	121 (4.3)	140 (3.6)	153 (2.6)	163 (2.7)
Simulation scientific exploration score	125 (5.3)	141 (4.0)	153 (2.3)	161 (2.3)
Simulation scientific synthesis score	124 (4.2)	141 (3.8)	153 (2.5)	161 (2.0)
Simulation computer skills score	123 (4.4)	138 (4.5)	152 (3.2)	165 (4.4)

	Make tables, charts, or graphs on computer			
Scale	Not at all	Small extent	Moderate extent	Large extent
Simulation total score	136 (2.9)	157 (2.4)	154 (3.7)	148 (5.3)
Simulation scientific exploration score	138 (3.2)	156 (2.1)	153 (3.4)	147 (5.4)
Simulation scientific synthesis score	136 (3.5)	156 (2.2)	154 (3.1)	149 (5.9)
Simulation computer skills score	135 (3.7)	156 (3.2)	155 (4.9)	151 (6.5)

	Find information on the Internet			
Scale	Not at all	Small extent	Moderate extent	Large extent
Simulation total score	‡	133 (4.4)	147 (3.5)	156 (2.5)
Simulation scientific exploration score	‡	137 (3.7)	147 (3.3)	155 (2.2)
Simulation scientific synthesis score	‡	136 (4.5)	147 (3.1)	155 (2.2)
Simulation computer skills score	‡	131 (4.4)	148 (4.0)	156 (3.6)

‡ Reporting standards not met. Sample size was insufficient to permit a reliable estimate.
NOTE: The range of scores for each scale is 0–300. Standard errors of the estimated scores appear in parentheses.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Table I-8. Data for figure 6-5, mean scale scores, by frequency of computer use and scale for Simulation scenario, grade 8: 2003

	How often do you use a computer outside of school?				
Scale	Daily	2–3 times per week	Once a week	Once every few weeks	Never or hardly ever
Simulation total score	160 (2.1)	147 (2.9)	134 (7.1)	130 (6.1)	118 (3.0)
Simulation scientific exploration score	159 (2.3)	148 (2.5)	136 (7.3)	134 (5.1)	119 (5.3)
Simulation scientific synthesis score	159 (2.0)	148 (2.4)	136 (9.1)	135 (5.9)	119 (2.7)
Simulation computer skills score	159 (3.4)	147 (3.6)	135 (7.3)	134 (6.7)	121 (3.7)

NOTE: The range of scores for each scale is 0–300. Standard errors of the estimated scores appear in parentheses.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Table I-9. Data for figure 6-6, mean scale scores, by students indicating there is a computer at home that they use and scale for Simulation scenario, grade 8: 2003

	Is there a computer at home that you use?	
Scale	Yes	No
Simulation total score	154 (2.1)	123 (4.4)
Simulation scientific exploration score	154 (1.8)	125 (5.2)
Simulation scientific synthesis score	154 (2.0)	125 (4.4)
Simulation computer skills score	153 (3.3)	128 (4.7)

NOTE: The range of scores for each scale is 0–300. Standard errors of the estimated scores appear in parentheses.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Appendix J: Performance on Problem Solving in Technology-Rich Environments (TRE) Observables

Table J-1. Weighted percentage of students achieving each level of correctness on each Search scenario scientific inquiry observable in order of first appearance on item map (figure 5-1), grade 8: 2003

Observable and level of correctness	Weighted percent
Correctly answering most, if not all (three or four), of the four multiple-choice items that require web searching.	18
Correctly answering some (one or two) of the four multiple-choice items that require web searching.	64
Correctly answering none of the four multiple-choice items that require web searching.	18
Using search terms that, on average, match those of proficient searchers to at least a moderate degree.	33
Using search terms that, on average, match those of proficient searchers only to a limited degree.	46
Using search terms that, on average, did not match those of proficient searchers.	21
Constructing a response that gives a reasonably complete answer to the motivating Search problem (i.e., three or more advantages of using gas balloons).	15
Constructing a response that only partially answers the motivating Search problem (i.e., giving only one or two advantages of using gas balloons).	35
Constructing a response that fails to answer the motivating Search problem (i.e., giving no advantages of using gas balloons).	43
Did not construct a response.	7
Bookmarking or visiting pages that are, on average, relevant to the question posed.	14
Bookmarking or visiting pages that are, on average, partially relevant to the question posed.	12
Bookmarking or visiting pages that are, on average, irrelevant to the question posed.	36
Did not bookmark, did not visit pages, did not search, or produced otherwise unscorable response for this observable.	38
Producing at least one set of search results with hits that are, on average, relevant to the question posed (i.e., have relevance scores averaging between 3 and 4 on a four-point scale, where a score of 4 denotes the most relevant hits).	1
Producing at least one set of search results with hits that are, on average, partially relevant to the question posed (i.e., have relevance scores averaging between 2 and 3 on a four-point scale, where a score of 4 denotes the most relevant hits).	11
Producing search results with hits that are, on average, irrelevant to the question posed (i.e., have relevance scores below 2 on a four-point scale, where a score of 4 denotes the most relevant hits).	83
Did not run any searches.	5

NOTE: Detail may not sum to totals because of rounding. Evaluation levels for certain observables were collapsed during analysis; hence, not all the levels for these observables shown in this table appear in the item map.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Table J-2. Weighted percentage of students achieving each level of correctness on each Search scenario computer skills observable in order of first appearance on item map (figure 5-2), grade 8: 2003

Observable and level of correctness	Weighted percent
Using the Back button frequently (at least five times) to navigate among web pages or from web pages to the search page.	69
Using the Back button occasionally (three or four times) to navigate among web pages or from web pages to the search page.	10
Using the Back button rarely (two times or less) to navigate among web pages or from web pages to the search page.	21
Using hyperlinks frequently (at least 5 times) to explore web pages linked to the page currently being viewed.	55
Using hyperlinks with moderate frequency (3 to 4 times) to explore web pages linked to the page currently being viewed.	11
Using hyperlinks with limited frequency (1 to 2 times) to explore web pages linked to the page currently being viewed.	15
Did not use hyperlinks to explore web pages linked to the page currently being viewed.	20
Using bookmarks with at least moderate frequency (two or more times).	58
Using bookmarks with limited frequency (one time).	13
Did not use bookmarks.	29
Returning relevant results after only a small number of attempts (1–3).	37
Returning relevant results after a moderate number of attempts (4–6).	24
Returning relevant results after many attempts (more than 6) or does not return relevant results at all.	34
Did not attempt any searches.	5
Using advanced search techniques with at least moderate frequency (3 or more searches).	8
Using advanced search techniques with limited frequency (1–2 searches).	24
Did not use advanced search techniques.	68
Using Delete with at least moderate frequency (2 or more times) to remove a page that had been bookmarked.	3
Using Delete with limited frequency (1 time) to remove a page that had been bookmarked.	8
Did not to use Delete to remove a page that had been bookmarked.	89

NOTE: Detail may not sum to totals because of rounding. Evaluation levels for certain observables were collapsed during analysis; hence, not all the levels for these observables shown in this table appear in the item map.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Table J-3. Weighted percentage of students achieving each level of correctness on each Simulation scenario scientific exploration observable in order of first appearance on item map (figure 6-1), grade 8: 2003

Observable and level of correctness	Weighted percent
Using the glossary of science terms in Simulation problem 1 with low frequency or never.	80
Using the glossary of science terms in Simulation problem 1 with moderate frequency.	17
Using the glossary of science terms in Simulation problem 1 with high frequency.	2
Did not produce a scorable response for this observable.	1
Creating a table for Simulation problem 2 that includes only the dependent and independent variables germane to the problem.	9
Creating a table for Simulation problem 2 that includes both of the variables germane to solving the problem along with other variables.	19
Creating a table for Simulation problem 2 that either includes one of the variables germane to solving the problem along with experimental data, or both germane variables without data.	17
Creating a table for Simulation problem 2 that does not include either of the variables germane to solving the problem, or includes one germane variable without experimental data.	13
Did not create a table for Simulation problem 2.	42
Controlling for one variable in at least 66 percent of the experiments run for Simulation problem 3.	46
Controlling for one variable in 40 to 65 percent of the experiments run for Simulation problem 3.	9
Controlling for one variable in less than 40 percent of the experiments run for Simulation problem 3.	3
Running an insufficient number of experiments for controlled experimentation to be evaluated for Simulation problem 3.	40
Did not produce scorable response for this observable.	1
Running a set of experiments sufficient in number, range, and distribution to confirm that the relationship between altitude and amount of helium takes the form of a step function for Simulation problem 2.	#
Running a set of experiments sufficient in number, range, and distribution to confirm that the relationship between altitude and amount of helium is nonlinear for Simulation problem 2.	51
Running a set of experiments that suggests that the relationship between altitude and amount of helium takes the form of a two-piece linear one for Simulation problem 2.	9
Running a set of experiments that suggests that the relationship between altitude and amount of helium is linear for Simulation problem 2.	40
Running a set of experiments sufficient in number, range, and distribution to reveal the linear relationship between altitude and mass for Simulation problem 1.	24
Running experiments sufficient in number and range but not in distribution to confirm the linear relationship between mass and altitude for Simulation problem 1.	24
Running experiments either sufficient in number or in range to confirm the linear relationship between altitude and mass for Simulation problem 1.	10
Running experiments insufficient in number, range, or distribution to confirm the linear relationship between altitude and mass for Simulation problem 1.	42
Did not produce scorable response for this observable.	1
Creating a graph for Simulation problem 2 with the correct variables on the correct axes, with experimental data.	22
Creating a graph for Simulation problem 2 with the correct variables on the correct axes, with minimal experimental data or without data.	13
Creating a graph for Simulation problem 2 with only one or neither of the correct variables on the correct axes.	22
Did not create a graph for Simulation problem 2.	42

See notes at end of table.

Table J-3. Weighted percentage of students achieving each level of correctness on each Simulation scenario scientific exploration observable in order of first appearance on item map (figure 6-1), grade 8: 2003—Continued

Observable and level of correctness	Weighted percent
Creating a graph for Simulation problem 1 with the correct variables on the correct axes that shows at least two data points.	19
Creating a graph for Simulation problem 1 with the correct variables on the correct axes but that shows no experimental data or only one data point.	16
Creating a graph for Simulation problem 1 with only one or neither of the correct variables on the correct axes.	27
Did not create a graph for Simulation problem 1.	38
Running experiments for at least two values of mass and, for at least one of those values, conducting a set of experiments with amounts of helium sufficient in number and in range to confirm that the relationship between altitude and volume takes the form of a step function for Simulation problem 3.	9
Running experiments for at least one value of mass and conducting a set of experiments with amounts of helium sufficient in number and in range to confirm that the relationship between altitude and volume is nonlinear for Simulation problem 3.	4
Running experiments for at least one value of mass and conducting a set of experiments with amounts of helium that suggest that the relationship between altitude and volume takes the form of a two-piece linear function for Simulation problem 3.	15
Running experiments for at least one value of mass and conducting a set of experiments that suggest that the relationship between altitude and volume takes the form of a linear function for Simulation problem 3.	71
Creating a table for Simulation problem 1 that includes only the dependent and independent variables most germane to the problem.	8
Creating a table for Simulation problem 1 that includes the dependent and independent variables most germane to the problem as well as other variables.	18
Creating a table for Simulation problem 1 that includes the dependent OR independent variable most germane to the problem along with experimental data, OR that includes the dependent and independent variables most germane to the problem as well as other variables, but no data.	16
Creating a table for Simulation problem 1 that includes neither the dependent nor independent variable most germane to the problem, OR that includes either the dependent OR the independent variable most germane to the problem but no experimental data.	20
Did not create a table for Simulation problem 1.	37
Creating a graph for Simulation problem 3 with the correct variables on the correct axes that shows data for at least four experiments (two experiments for each of at least two values of mass).	20
Creating a graph for Simulation problem 3 with the correct variables on the correct axes that shows data for at least one experiment for each of two masses.	3
Creating a graph for Simulation problem 3 with the correct variables on the correct axes that shows data for one or no experiments.	27
Creating a graph for Simulation problem 3 that does not have the correct variables on the correct axes.	#
Did not create a graph for Simulation problem 3.	50
Creating a table for Simulation problem 3 that includes only the three variables most germane to the problem.	4
Creating a table for Simulation problem 3 that includes the three variables most germane to the problem along with other variables.	26
Creating a table for Simulation problem 3 that includes the three variables most germane to the problem along with other variables but no experimental data, OR any two of the most germane variables with data.	26
Creating a table for Simulation problem 3 that includes only one of the three variables most germane to the problem with experimental data, OR any two of the most germane variables without data.	#
Did not create a table for Simulation problem 3.	44

# The estimate rounds to zero.

NOTE: Detail may not sum to totals because of rounding. Evaluation levels for certain observables were collapsed during analysis; hence, not all the levels for these observables shown in this table appear in the item map.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Table J-4. Weighted percentage of students achieving each level of correctness on each Simulation scenario scientific synthesis observable in order of first appearance on item map (figure 6-2), grade 8: 2003

Observable and level of correctness	Weighted percent
Offering correct and complete ("best") responses to the constructed-response question that concludes Simulation problem 3 that explain how the relationship between amount of helium and balloon altitude for more than one payload mass takes the form of a series of step functions (e.g., "Once the balloon has enough helium to rise into the air, the balloon will rise to a maximum height and go no higher no matter how much helium is added.").	2
Offering correct but incomplete ("good") responses to the constructed-response question that concludes Simulation problem 3 by explaining either the top or the bottom of the step function (e.g., "Once in the air, the balloon will reach a maximum altitude no matter how much helium is added, and the maximum altitude the balloon can reach decreases as payload mass increases.").	7
Offering partially correct responses that can be derived from Simulation problems 1 or 2 to the concluding question for Simulation problem 3 (e.g., "Below a certain amount of helium the balloon cannot get off the ground.").	43
Offering wholly inaccurate responses to the concluding question for Simulation problem 3.	45
Did not produce scorable response for this observable.	4
Offering correct and complete ("best") responses to the constructed-response question that concludes Simulation problem 2 that explain how the relationship between amount of helium and balloon altitude for a payload mass of 100 lb. takes the form of a step function (e.g., "Once the balloon has enough helium to rise into the air, the balloon will rise to a maximum height and go no higher matter how much helium is added.").	13
Offering correct but incomplete ("good") responses referring either to the top or the bottom of the step function to the concluding question for Simulation problem 2 (e.g., "Once in the air, the balloon will reach a maximum altitude no matter how much helium is added.").	18
Offering partially correct responses that express a linear relationship between altitude and amount of helium to the concluding question for problem 2 (e.g., "More helium inside the balloon will make the balloon go higher.").	33
Offering wholly inaccurate responses to the concluding question for Simulation problem 2.	34
Did not produce scorable response for this observable.	2
Offering correct and complete ("best") responses to the constructed-response question that concludes Simulation problem 1 with specific references to experiments (e.g., "As the payload mass increases, the balloon's altitude decreases. For example, when I put 90 lb. of payload on the balloon, it only went to 10,000 feet. But when I put 50 lb. of payload mass on the balloon, it went to 22,326, and when I put 10 lb., it went to 36,211 feet.")	23
Offering correct but incomplete ("partial") responses that express the linear relationship between mass and altitude to the concluding question for Simulation problem 1 (e.g., "As the payload mass increases, the balloon's altitude decreases") with no specific references to experiments.	44
Offering wholly inaccurate response to the concluding question for Simulation problem 1.	31
Did not produce scorable response for this observable.	2
Correctly answering the multiple-choice question about the relationship between variables concluding Simulation problem 1.	59
Incorrectly answering the multiple-choice question about the relationship between variables concluding Simulation problem 1.	41
Correctly answering the multiple-choice question about the relationship among variables concluding Simulation problem 3.	31
Incorrectly answering the multiple-choice question about the relationship among variables concluding Simulation problem 3.	68
Did not produce scorable response for this observable.	1

See notes at end of table.

Table J-4. Weighted percentage of students achieving each level of correctness on each Simulation scenario scientific synthesis observable in order of first appearance on item map (figure 6-2), grade 8: 2003—Continued

Observable and level of correctness	Weighted percent
Making correct predictions for more than one half of unique experiments run for Simulation problem 2.	9
Making correct predictions for one half to one third of unique experiments run for Simulation problem 2.	6
Making correct predictions for less than one third of unique experiments run for Simulation problem 2.	6
Did not make predictions for Simulation problem 2.	79
Correctly answering the multiple-choice question about the relationship between variables concluding Simulation problem 2.	23
Incorrectly answering the multiple-choice question about the relationship between variables concluding Simulation problem 2.	77

NOTE: Detail may not sum to totals because of rounding. Evaluation levels for certain observables were collapsed during analysis; hence, not all the levels for these observables shown in this table appear in the item map.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Table J-5. Weighted percentage of students achieving each level of correctness on each Simulation scenario computer skills observable in order of first appearance on item map (figure 6-3), grade 8: 2003

Observable and level of correctness	Weighted percent
Never using the interface tools in the wrong order for drawing conclusions in Simulation problem 3 (e.g., clicking on the Draw Conclusions button without having run any experiments).	93
Using the interface tools in the wrong order for drawing conclusions once or twice in Simulation problem 3 (e.g., clicking on the Draw Conclusions button without having run any experiments).	6
Using the interface tools in the wrong order for drawing conclusions at least 3 times in Simulation problem 3 (e.g., clicking on the Draw Conclusions button without having run any experiments).	#
Never using the interface tools in the wrong order for experimenting in Simulation problem 1 (e.g., clicking on the Make Predictions button without having chosen any values with which to experiment).	79
Using the interface tools in the wrong order for experimenting once or twice in Simulation problem 1 (e.g., clicking on the Make Predictions button without having chosen any values with which to experiment).	20
Using the interface tools in the wrong order for experimenting at least 3 times in Simulation problem 1 (e.g., clicking on the Make Predictions button without having chosen any values with which to experiment).	1
Did not produce scorable response for this observable.	1
Never using Computer Help in Simulation problem 1.	81
Using Computer Help once or twice in Simulation problem 1.	17
Using Computer Help at least 3 times in Simulation problem 1.	1
Did not produce scorable response for this observable.	1
Never using the interface tools in the wrong order for drawing conclusions in Simulation problem 2 (e.g., clicking on the Draw Conclusions button without having run any experiments).	90
Using the interface tools in the wrong order for drawing conclusions once or twice in Simulation problem 2 (e.g., clicking on the Draw Conclusions button without having run any experiments).	9
Using the interface tools in the wrong order for drawing conclusions at least 3 times in Simulation problem 2 (e.g., clicking on the Draw Conclusions button without having run any experiments).	1
Never using the interface tools in the wrong order for drawing conclusions in Simulation problem 1 (e.g., clicking on the Draw Conclusions button without having run any experiments).	75
Using the interface tools in the wrong order for drawing conclusions once or twice in Simulation problem 1 (e.g., clicking on the Draw Conclusions button without having run any experiments).	23
Using the interface tools in the wrong order for drawing conclusions at least 3 times in Simulation problem 1 (e.g., clicking on the Draw Conclusions button without having run any experiments).	1
Did not produce scorable response for this observable.	1
Key-entering a response of over 150 characters to the constructed-response question concluding Simulation problem 3.	51
Key-entering a response of 50 to 149 characters to the constructed-response question concluding Simulation problem 3.	37
Key-entering a response of less than 50 characters to the constructed-response question concluding Simulation problem 3.	11
Did not produce scorable response for this category.	1

See notes at end of table.

Table J-5. Weighted percentage of students achieving each level of correctness on each Simulation scenario computer skills observable in order of first appearance on item map (figure 6-3), grade 8: 2003—Continued

Observable and level of correctness	Weighted percent
Key-entering a response of over 150 characters to the constructed-response question concluding Simulation problem 2.	47
Key-entering a response of 50 to 149 characters to the constructed-response question concluding Simulation problem 2.	39
Key-entering a response of less than 50 characters to the constructed-response question concluding Simulation problem 2.	13
Did not produce scorable response for this observable.	#
Key-entering a response of over 150 characters to the constructed-response question concluding Simulation problem 1.	51
Key-entering a response of 50 to 149 characters to the constructed-response question concluding Simulation problem 1.	38
Key-entering a response of less than 50 characters to the constructed-response question concluding Simulation problem 1.	10
Did not produce scorable response for this observable.	1
Performing a variety of interface actions (e.g., tabbing among graphs, tables, and the response area; sorting tables) in Simulation problem 3.	47
Performing some interface actions (e.g., tabbing among graphs, tables, and the response area; sorting tables) in Simulation problem 3.	28
Performing few interface actions (e.g., tabbing among graphs, tables, and the response area; sorting tables) in Simulation problem 3.	25

# The estimate rounds to zero.

NOTE: Detail may not sum to totals because of rounding. Evaluation levels for certain observables were collapsed during analysis; hence, not all the levels for these observables shown in this table appear in the item map.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Appendix K: Understanding NAEP Reporting Groups

NAEP results are provided for groups of students defined by shared characteristics—gender, race/ethnicity, parental education, and eligibility for free/reduced-price school lunch. Based on participation rate criteria, results are reported for subpopulations only when sufficient numbers of students and adequate school representation are present. The minimum requirement is at least 62 students in a particular subgroup from at least five primary sampling units (PSUs).[1] However, the data for all students, regardless of whether their subgroup was reported separately, were included in computing overall results. Definitions of the subpopulations are presented below.

Gender

Results are reported separately for male students and female students.

Race/Ethnicity

In all NAEP assessments, data about student race/ethnicity is collected from two sources: school records and student self-reports. Prior to 2002, NAEP used students' self-reported race as the primary race/ethnicity reporting variable. As of 2002, the race/ethnicity variable presented in NAEP reports is based on the race reported by the school. When school-recorded information is missing, student-reported data are used to determine race/ethnicity. The mutually exclusive racial/ethnic categories are White, Black, Hispanic, Asian/Pacific Islander, American Indian (including Alaska Native), and Other. Information based on student self-reported race/ethnicity is available on the NAEP Data Explorer (http://nces.ed.gov/nationsreportcard/nde/).

Parental Education

Eighth-graders were asked the following two questions, the responses to which were combined to derive the parental education variable.

How far in school did your mother go?

A. She did not finish high school.

B. She graduated from high school.

C. She had some education after high school.

D. She graduated from college.

E. I don't know.

Students were also asked

How far in school did your father go?

A. He did not finish high school.

B. He graduated from high school.

C. He had some education after high school.

D. He graduated from college.

E. I don't know.

The information was combined into one parental education reporting variable in the following way: If a student indicated the extent of education for only one parent, that level was included in the data. If a student indicated the extent of education for both parents, the higher of the two levels was included in the data. If a student responded "I don't know" for both parents, or responded "I don't know" for one parent and did not respond for the other, the parental education level was classified as "I don't know." If the student did not respond for either parent, the student was recorded as having provided no response.

Eligibility for Free/Reduced-Price School Lunch

As part of the Department of Agriculture's National School Lunch Program, schools can receive cash subsidies and donated commodities in turn for offering free or reduced-price lunches to eligible children. Based on available school records, students were classified as either currently eligible for free/reduced-price school lunch or not eligible. Eligibility for the program is determined by students' family income in relation to the federally established poverty level. Free lunch qualification is set at 130 percent of the poverty level, and reduced-price lunch qualification is set at between 130 and 185 percent of the poverty level. Additional information on eligibility may be found at the Department of Agriculture website (http://www.fns.usda.gov/cnd/lunch/). The classification applies only to the school year when the TRE scenarios were administered (i.e., the 2002–2003 school year) and is not based on eligibility in previous years. If school records were not available, the student's information was recorded as "Unavailable." If the school did not participate in the program, all students in that school were classified as "Unavailable."

[1] A PSU is a selected geographic region (a county, group of counties, or metropolitan statistical area).