

## Chapter 2: The TRE Interpretive Framework

### The Student and Evidence Models

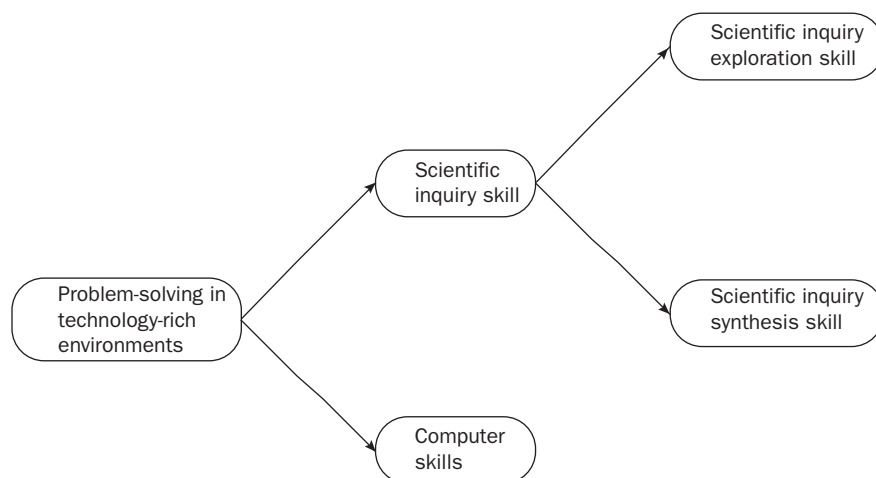
While developing suitable problem-solving scenarios is a challenging task, so is interpreting the responses to such scenarios. A well-conceptualized interpretive framework is a necessity; the scenario development cost and examinee time required to perform extended problem solving on the computer can 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.<sup>7</sup> Figure 2-1 shows the student model.

Figure 2-1. TRE student model, 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.

<sup>7</sup> 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) <sup>1</sup>	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.

<sup>1</sup> 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.<sup>8</sup>

## 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.)

<sup>8</sup> 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.

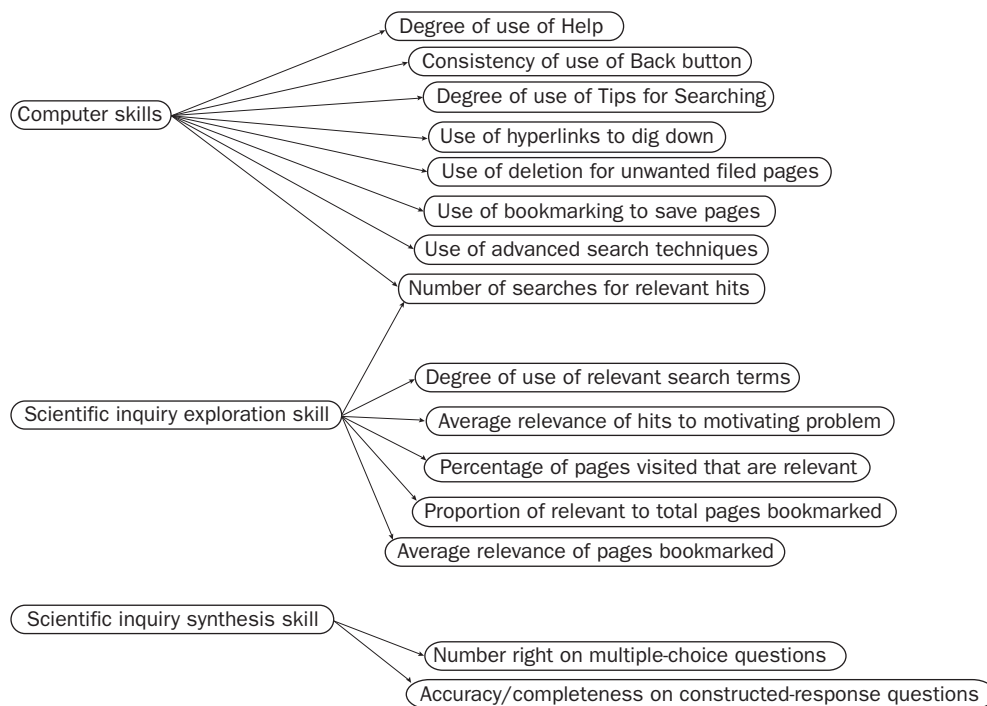
Figure 2-2 graphically depicts the evidence model for the Search scenario. The model is essentially a set of hypotheses about which observables are direct evidence of the proficiencies in the student model. In the center are the student-model proficiencies—computer skills, scientific inquiry exploration 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.<sup>9</sup> 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



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.

<sup>9</sup> 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.

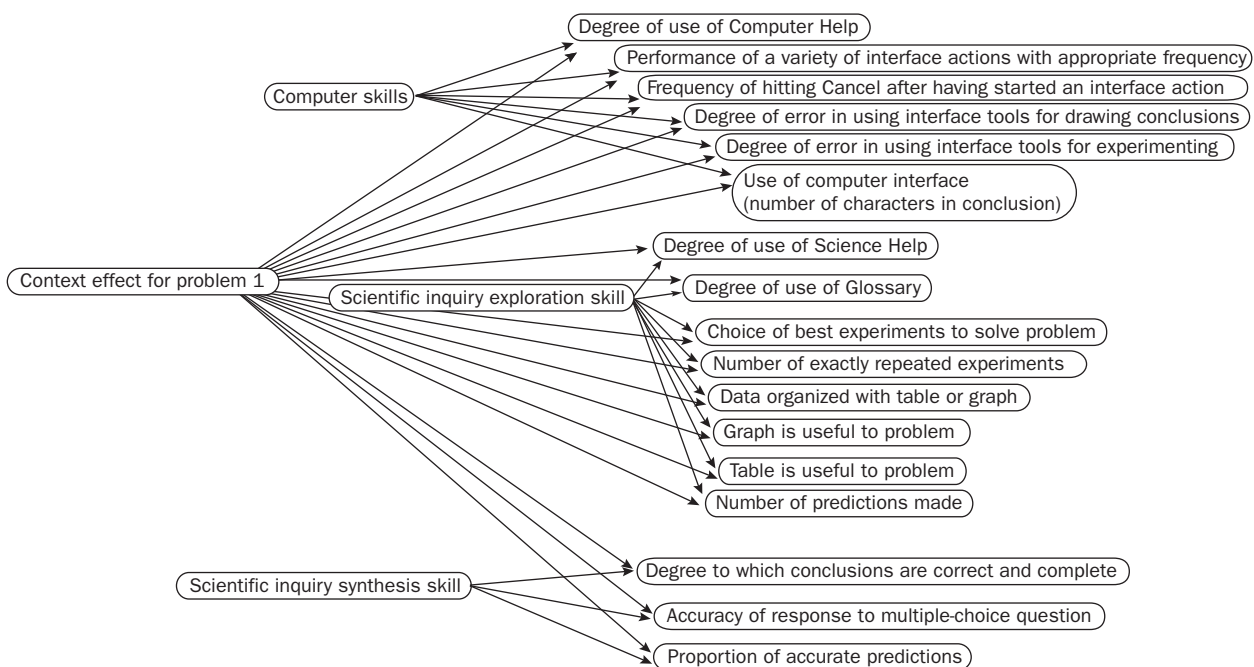
appropriate that table or graph was to the problem posed. Linked to scientific synthesis are the accuracy of answers to the constructed-response and multiple-choice questions that motivate the problem, and the proportion of accurate predictions. Some of these behaviors (such as how frequently Science 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



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

Scenario	<i>Is there a computer at home that you use?</i>	
	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

Scenario	<i>How often do you use a computer at school?</i>				
	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)
Scenario	<i>How often do you use a computer outside of school?</i>				
	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>I am more motivated to do schoolwork on a computer.</i>					
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>I have more fun learning on a computer.</i>					
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>I get more done when using a computer for schoolwork.</i>					
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

<i>Play computer games</i>				
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)
<i>Use a word processor</i>				
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)
<i>Make drawings/art on computer</i>				
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)
<i>Make tables, charts or graphs on computer</i>				
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)
<i>Look up information on a CD</i>				
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)
<i>Find information on the Internet</i>				
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)
<i>Use e-mail</i>				
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)
<i>Talk in chat groups</i>				
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

Scenario	<i>Which best describes the science course you are taking?</i>					
	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

<i>Design your own science experiment</i>				
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)
<i>Carry out science experiment</i>				
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)
<i>Write up results of science experiment</i>				
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)
<i>Talk to class about results of experiment</i>				
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)
<i>Collect data using computerized lab equipment</i>				
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)
<i>Download data from the Internet</i>				
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)
<i>Analyze data using computer</i>				
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)
<i>Use the Internet to exchange information with other students or scientists about experiments</i>				
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)
<i>Use computer simulations to perform experiments or explore science topics</i>				
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 it has a simple design and can hold a lot of weight. It also costs less to make a balloon rather than making a satellite. 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 satellite or space shuttle.
- Using balloons to do scientific experiments has several advantages which I will only name a few. The first advantage is that they allow the payloads that they are carrying to lift without vibrations or G-forces that a rocket would, and may damage the payload. Another advantage is that the balloons are quickly launched and they are quickly recovered 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 expensive. A human does not have to be in one and there is no risk of losing lives.
- Scientists use balloons for space and atmospheric experiments because they can offer capabilities 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 up almost anywhere 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 balloon 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 put into the balloon it could not support the weight of the payload mass that balloon barely lifted off of the ground.
- 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 experiment 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 pound 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 height 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 height of 22326 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 amount 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 to 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 amount 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

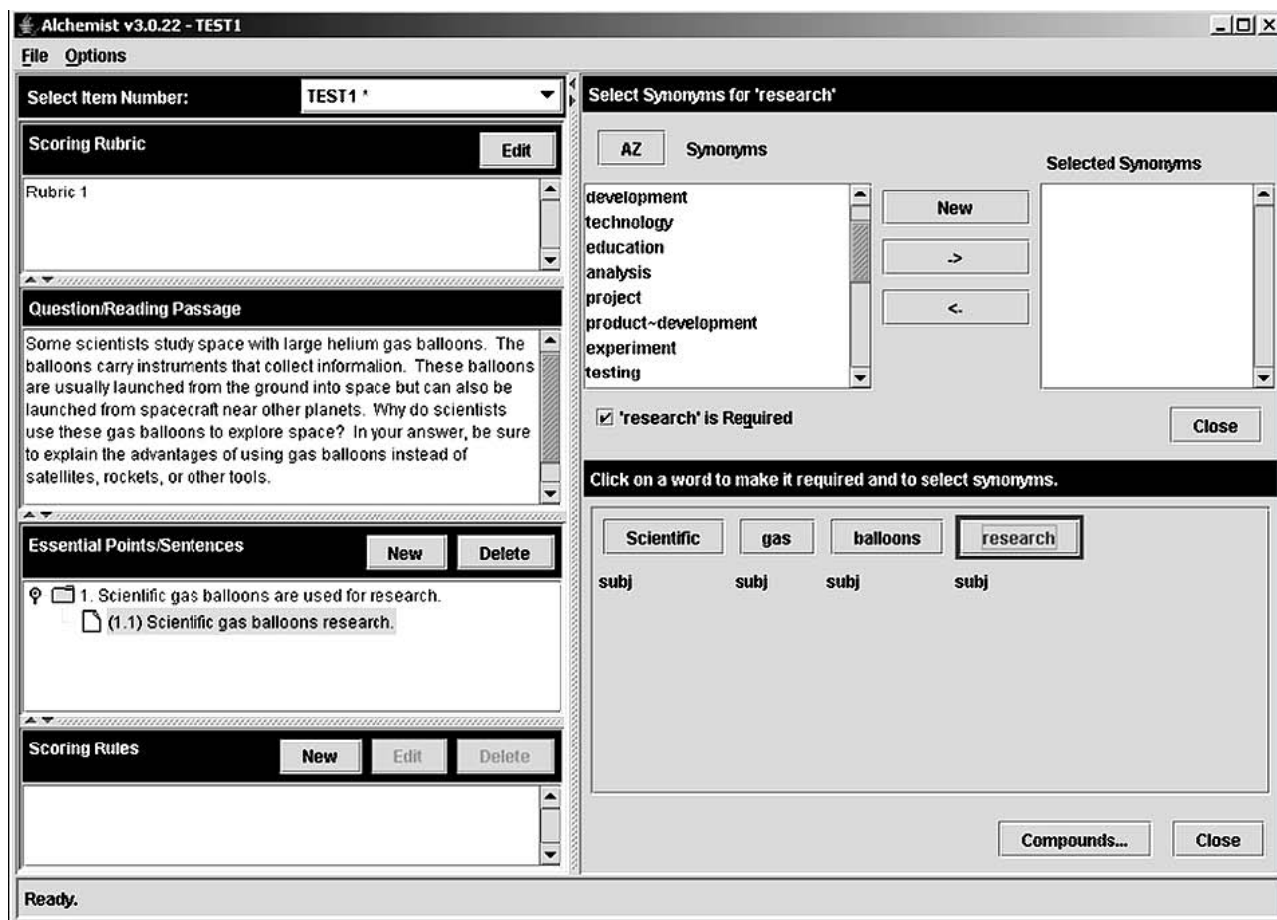
One measure of students' exploration skill for the Search scenario was the degree to which students used terms relevant to the motivating problem in their search queries.

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



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.<sup>10</sup> 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.

<sup>10</sup> 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).<sup>11,12</sup> 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

**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.

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 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.<sup>13</sup> 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.

<sup>11</sup> A NAEP hands-on block is a section of experimental tasks and constructed-response test items administered to a student.

<sup>12</sup> The TRE observables may not be completely independent, so the internal consistency estimates for the TRE scales may be inflated.

<sup>13</sup> 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.<sup>14</sup>

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 <sup>1</sup>	.17	<b>.71</b>
Accuracy/completeness on constructed-response question	.39	<b>.70</b>
Degree of use of relevant search terms	.33	<b>.51</b>
Number right on final multiple-choice questions	.28	<b>.44</b>
Average relevance of hits to motivating problem	.20	<b>.34</b>
Use of hyperlinks to dig down	<b>.69</b>	.37
Consistency of use of Back button	<b>.65</b>	.36
Number of searches for relevant hits <sup>2</sup>	<b>.65</b>	.33
Use of bookmarking to save pages	<b>.60</b>	.45
Use of advanced search techniques	<b>.46</b>	.30
Use of deletion for unwanted filed pages	<b>.24</b>	.08

<sup>1</sup> 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.

<sup>2</sup> 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.

<sup>14</sup> 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 <sup>1</sup>	.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 <sup>2</sup>	.20*
Consistency of use of Back button	.19*
Use of deletion for unwanted filed pages	.03

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

<sup>1</sup>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.

<sup>2</sup>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.<sup>15</sup> 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

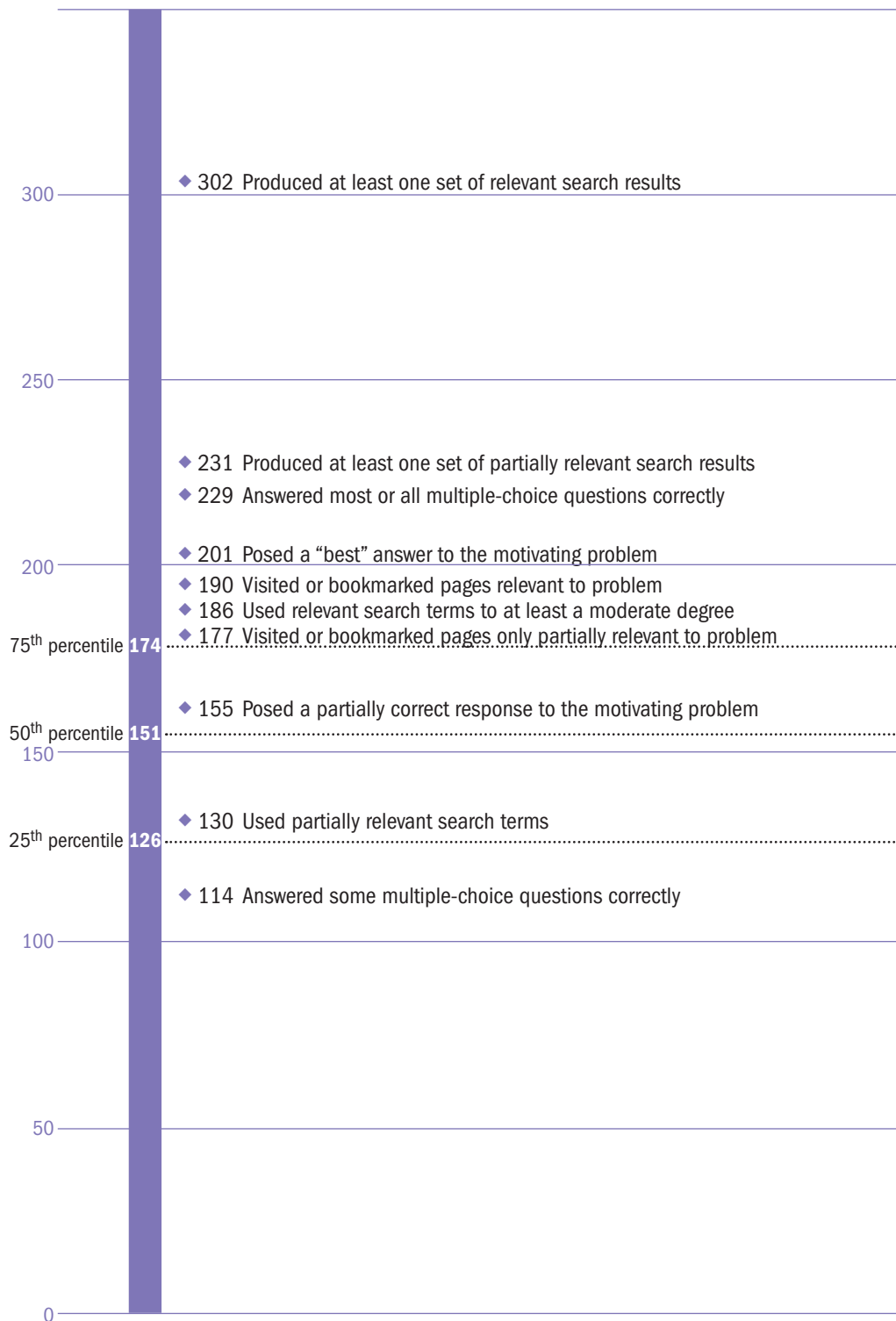
credit for responding to the motivating problem maps to a score of 201. Students with a score of 201 would have at least a 65 percent chance of submitting an answer achieving a top score of 3.

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).

<sup>15</sup> 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);

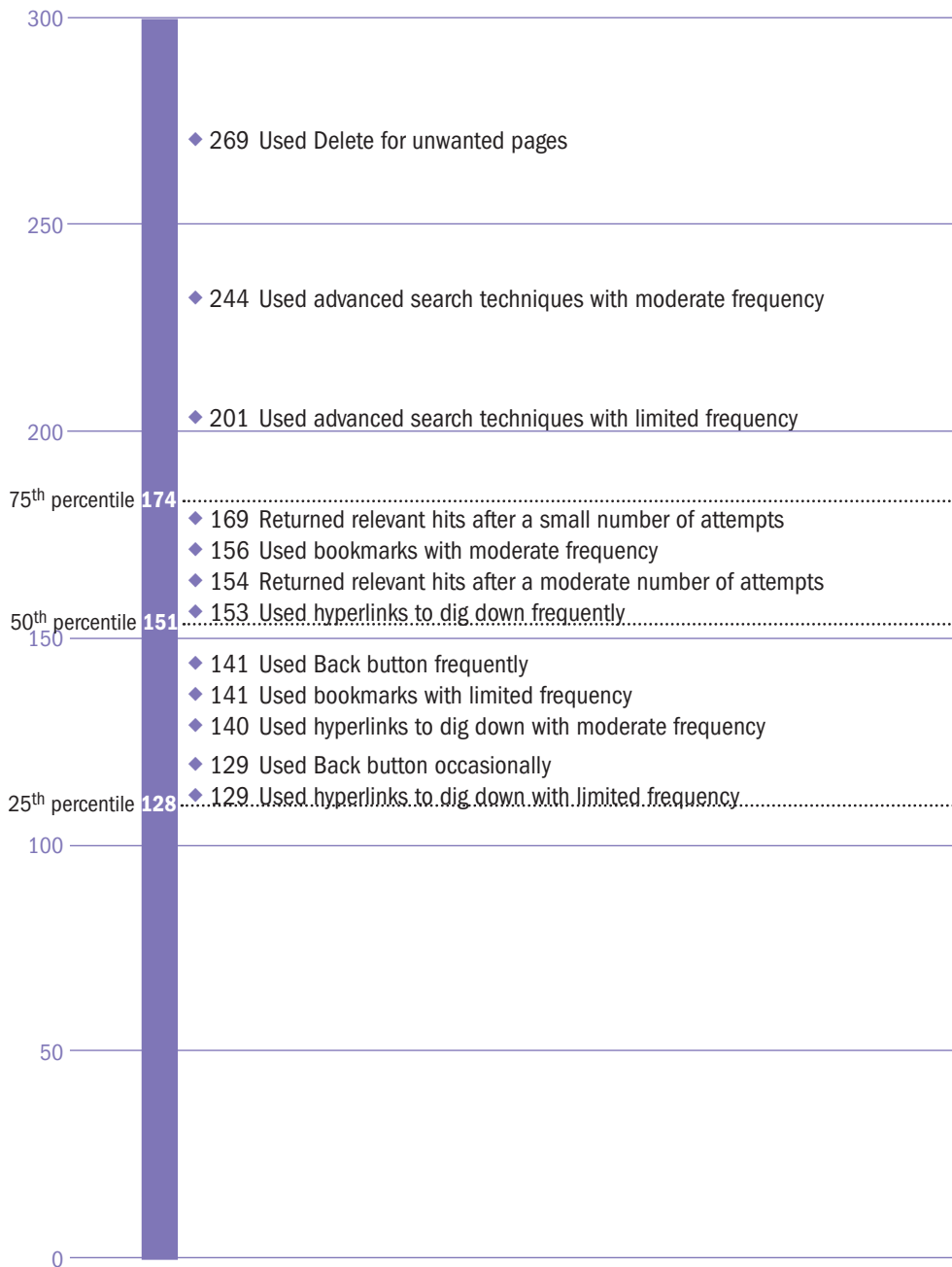
- using advanced search techniques with limited frequency (1–2 searches);
- using advanced search techniques with at least moderate frequency (3 or more searches); and
- using Delete to remove a page that had been bookmarked.

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 dif-

ferent 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

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

<sup>1</sup> No credit, partial credit, and full credit are the levels of correctness of response specific to each observable.

<sup>2</sup> "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

Observable	Low level of computer skills				Medium level of computer skills				High level of computer skills			
	No credit <sup>1</sup>	Low-partial credit	High-partial credit	Full credit	No credit <sup>1</sup>	Low-partial credit	High-partial credit	Full credit	No credit <sup>1</sup>	Low-partial credit	High-partial credit	Full credit
Use of hyperlinks to dig down	<b>.46</b>	.25	.17	.13	.09	.13	.23	<b>.55</b>	.01	.02	.06	<b>.91</b>

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

Observable	Low level of computer skills		Medium level of computer skills		High level of computer skills	
	No credit <sup>1</sup>	Full credit	No credit <sup>1</sup>	Full credit	No credit <sup>1</sup>	Full credit
Use of deletion for unwanted filed pages	<b>.96</b>	.04	<b>.91</b>	.09	<b>.81</b>	.19

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

<sup>2</sup> 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

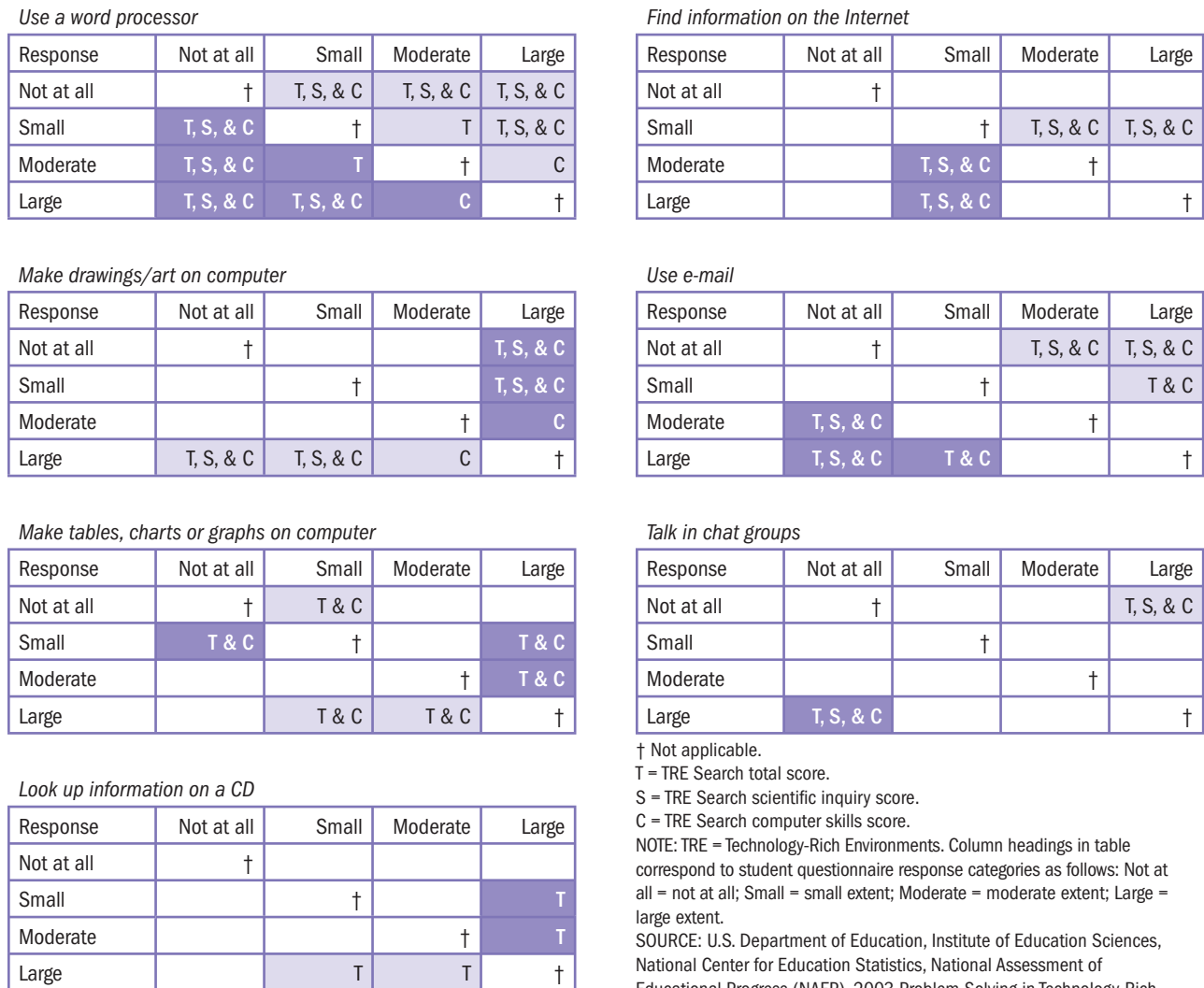
overall. For example, as figure 5-3 shows, students who reported more frequent use of a word processor (background question 2 in appendix D) scored better on average on all three TRE scales than those who reported not using a word processor at all. Other statistically significant differences in scores 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



† 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).<sup>16</sup>

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) ( $t, 43 = -3.63, p < .05$ ).

<sup>16</sup> 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 6: The TRE Simulation Scenario Scales and Results

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).

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.<sup>17</sup>

**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.

<sup>17</sup> 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 inter-correlations; 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).<sup>18</sup> 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.

<sup>18</sup> 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 rel-

evant 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	<b>.69</b>
Accuracy of response to final multiple-choice question	.22	.26	<b>.31</b>
Graph is useful to problem	.45	<b>.60</b>	.52
Choice of best experiments to solve problem	.35	<b>.53</b>	.40
Table is useful to problem	.41	<b>.50</b>	.44
Degree of use of Glossary	-.17	<b>-.17</b>	-.19
Use of computer interface (number of characters in conclusion)	<b>.72</b>	.49	.54
Degree of error in using interface tools for drawing conclusions	<b>-.32</b>	-.25	-.28
Degree of error in using interface tools for experimenting	<b>-.28</b>	-.24	-.27
Degree of use of Computer Help	<b>-.26</b>	-.22	-.24
Simulation problem 2			
Degree to which conclusions are correct and complete	.59	.61	<b>.72</b>
Accuracy of response to final multiple-choice question	.31	.31	<b>.37</b>
Proportion of accurate predictions	.22	.22	<b>.25</b>
Choice of best experiments to solve problem	.45	<b>.64</b>	.52
Table is useful to problem	.41	<b>.52</b>	.44
Graph is useful to problem	.40	<b>.49</b>	.44
Use of computer interface (number of characters in conclusion)	<b>.78</b>	.52	.55
Degree of error in using interface tools for drawing conclusions	<b>-.27</b>	-.21	-.23
Simulation problem 3			
Degree to which conclusions are correct and complete	.52	.52	<b>.67</b>
Accuracy of response to final multiple-choice question	.36	.36	<b>.43</b>
Proportion of experiments controlled for one variable	.51	<b>.74</b>	.56
Choice of best experiments to solve problem	.44	<b>.56</b>	.46
Graph is useful to problem	.32	<b>.42</b>	.35
Table is useful to problem	.14	<b>.21</b>	.20
Use of computer interface (number of characters in conclusion)	<b>.76</b>	.53	.59
Use of computer interface (use of various interface functions, e.g., making tables and graphs)	<b>.42</b>	.54	.42
Degree of error in using interface tools for drawing conclusions	<b>-.21</b>	-.19	-.20
Conclusion			
Degree of correctness of responses to multiple-choice items	.47	.48	<b>.58</b>

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 Simu-

lation 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.

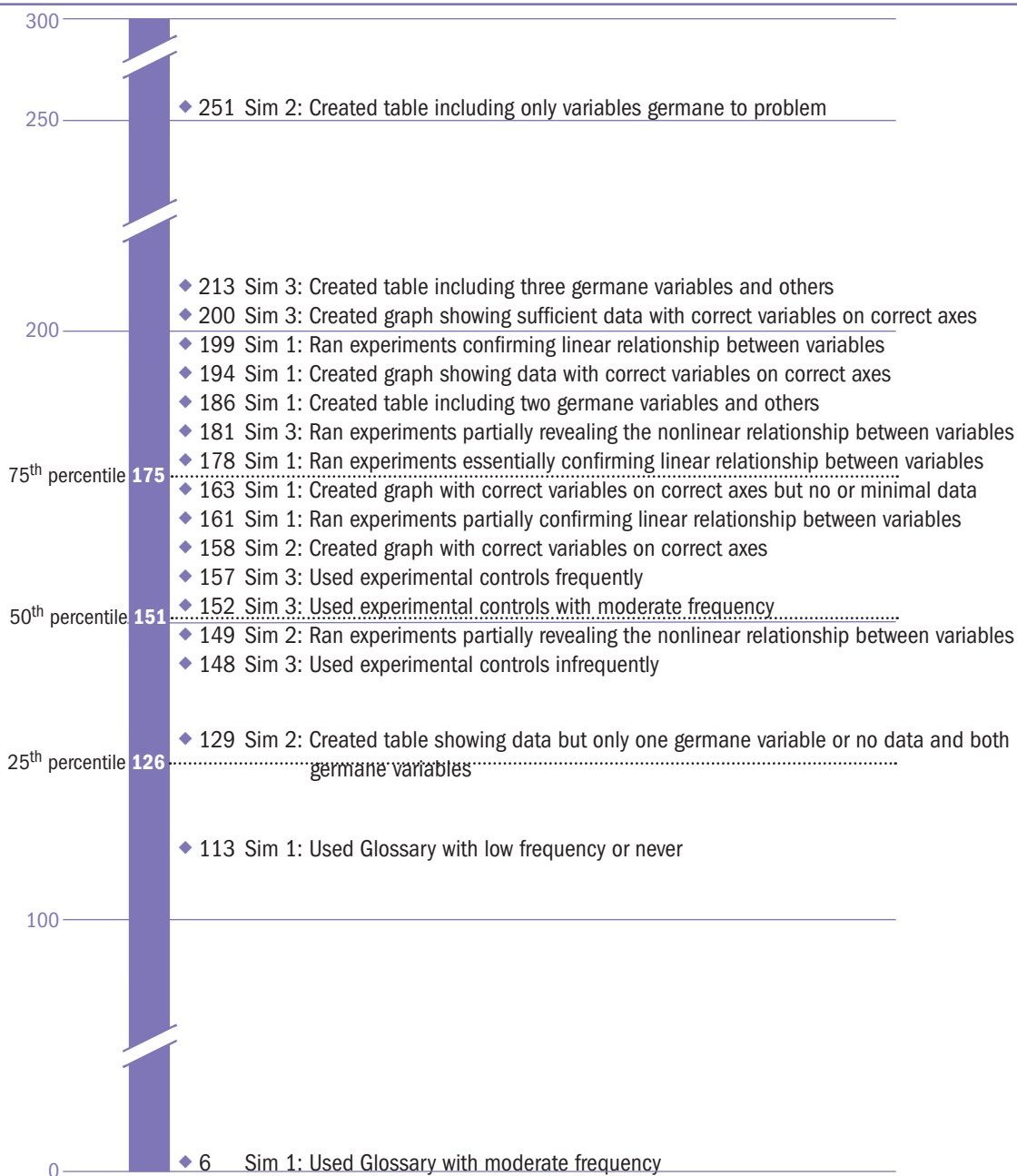
Figure 6-1 shows an item map for the scientific exploration scale. For mapping purposes, each observable has been transformed into one or more dichotomous variables, where the number of such variables is one less than the number of levels of correctness for the observable. Thus, each location on the 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.

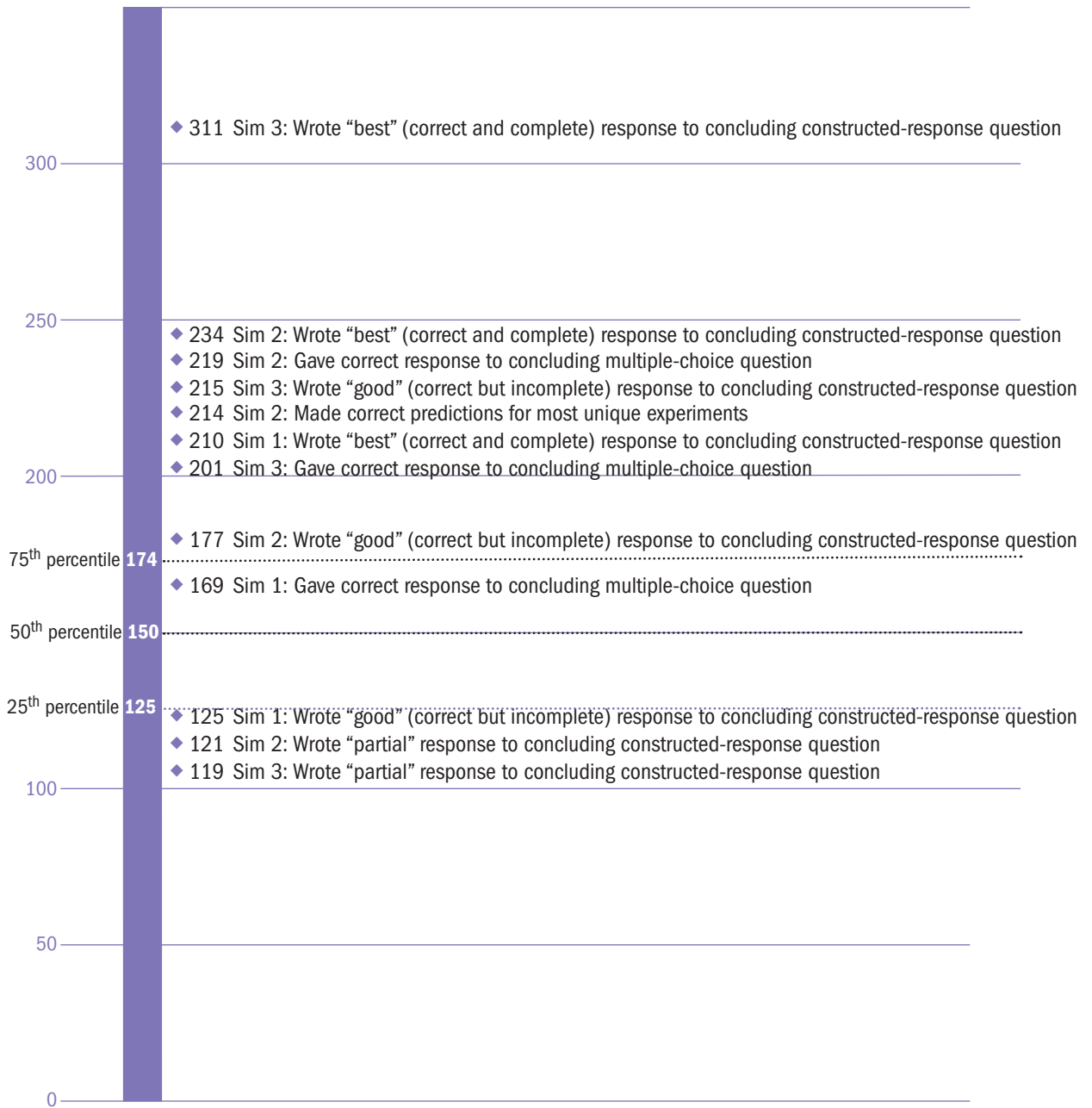
- creating a table for Simulation problem 2 that includes only the dependent and independent variables germane to the problem.

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.

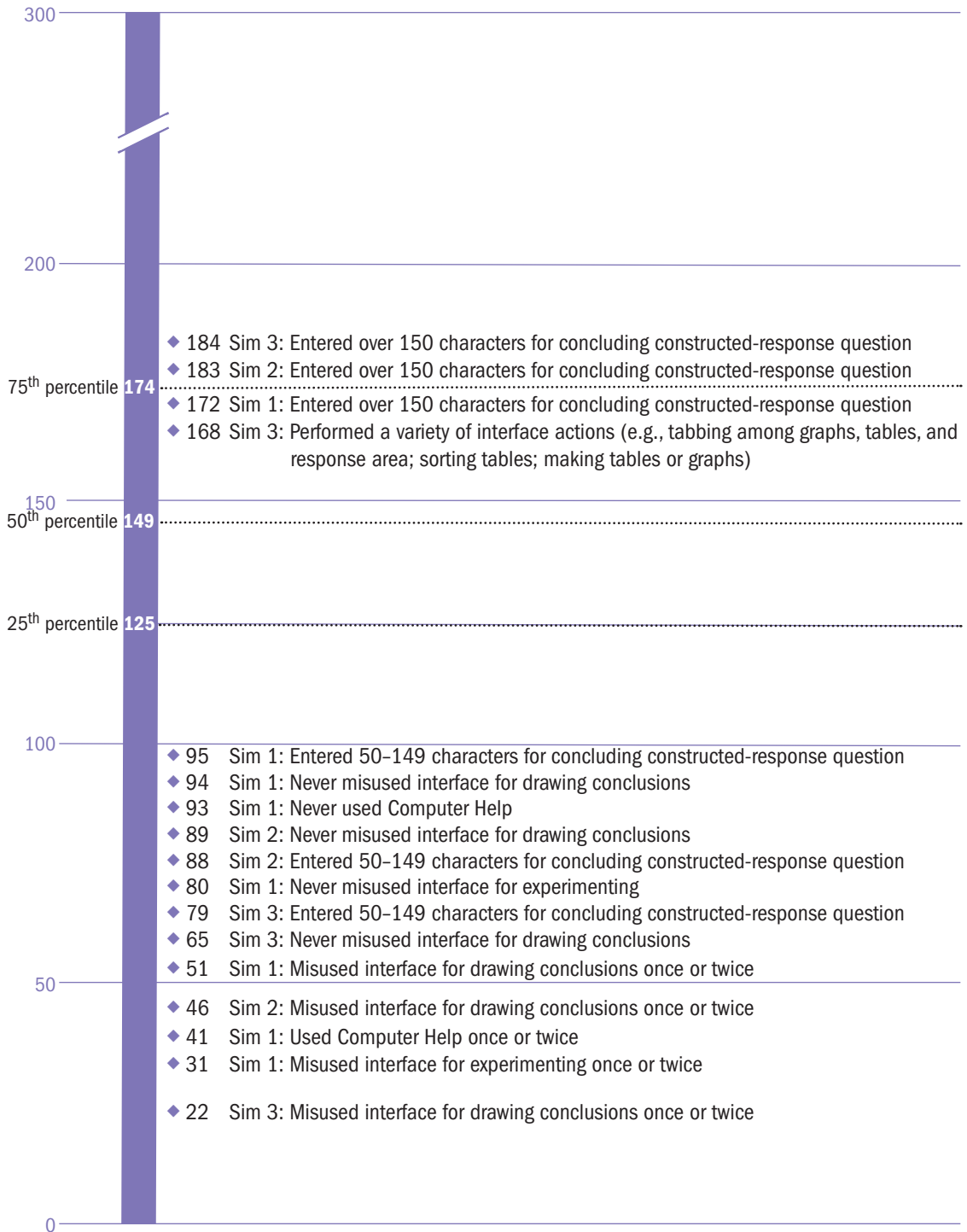
Figure 6-3 shows the locations of the levels of correctness for the observables on the computer skills scale. From the lowest scale point to the highest, the ordering is as follows:

- 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);<sup>19</sup>
- 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.

<sup>19</sup> 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.<sup>20</sup>

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.

<sup>20</sup> 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

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

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

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

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

<sup>2</sup> 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

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

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

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

<sup>1</sup> 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

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

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

<sup>1</sup> No credit, partial credit, and full credit are the levels of correctness of response specific to each observable.

<sup>2</sup> 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).<sup>21</sup>

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

<i>Play computer games</i>				
Response	Not at all	Small	Moderate	Large
Not at all	†		E	E
Small		†		
Moderate	E		†	
Large	E			†

<i>Use a word processor</i>				
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	†

<i>Make tables, charts, or graphs on computer</i>				
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				†

<i>Find information on the Internet</i>				
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.

<sup>21</sup> 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.