RESEARCH
MEMORANDUM

July 2003
RM-03-03

Assessing Complex Problem-solving Performances

Randy Elliot Bennett

Frank Jenkins

Hilary Persky

Andy Weiss

ETS®
Educational
Testing Service

Research &
Development Division
Princeton, NJ 08541

Assessing Complex Problem-solving Performances

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

Educational Testing Service, Princeton, NJ

July 2003

Abstract

Computer-based simulations can give a more nuanced understanding of what students know and can do than traditional testing methods. These extended, integrated tasks, however, introduce particular problems, including producing an overwhelming amount of data, multidimensionality, and local dependence. In this paper, we describe an approach to understanding the data from complex performances based on evidence-centered design (Mislevy, Almond, & Lukas, in press), a methodology for devising assessments and for using the evidence observed in complex student performances to make inferences about proficiency. We use as an illustration the NAEP Problem-Solving in Technology-Rich Environments Study, which is being conducted to exemplify how nontraditional skills might be assessed in a sample-based national survey. The paper focuses on the inferential uses of ECD, especially how features are extracted from student performance, how these extractions are evaluated, and how the evaluations are accumulated to make evaluative judgments.

Key words: Performance assessment, computer-based testing, evidence-centered design (ECD), simulations

Acknowledgements

The TRE Study was conducted under cooperative agreement for the Office of Educational Research and Improvement, National Center for Education Statistics, Washington, DC 20208, CFDA#84.902F. The positions expressed herein do not necessarily represent those of OERI or of Educational Testing Service.

ETS staff members contributing to the TRE Study included Malcolm Bauer, Kevin Bentley, Doug Forer, Jeff Haberstroh, Cindy Hammell, Kathy Howell, Frank Jenkins, Holly Knott, Mary Lauko, Lou Mang, Chris O'Sullivan, Debbie Pisacreta, Peggy Redman, Linda Steinberg, Michael Wagner, and David Williamson. We are also indebted to the staff of Westat, Inc. for data collection. Finally, Holly Spurlock, Vonda Kiplinger, and Bob Evans gave invaluable advice as NCES project monitors.

Cognitive science research offers new understandings and methods that can help us describe what students know and can do in more meaningful and detailed ways than traditional methods can. This research emphasizes the use of extended performances that call upon examinees to employ multiple skills in an integrated fashion. Such problem-solving exercises, however, take lots of examinee time and that time investment cannot be well justified if the outcome is only a single score. Fortunately, extended exercises delivered on computer can provide an enormous amount of examinee information because every action can be recorded. But how would one make sense of all that data?

In addition to the *amount* of data, other factors make interpretation challenging. Extended performances are often *multidimensional*, meaning that multiple, intertwined skills are necessary to respond successfully. Further, response data that compose an extended performance are often *locally dependent*. That is, factors other than the skills of interest may influence responding on related aspects of a complex task. For example, if an assessment is composed of several extended exercises, an examinee's performance may be affected by differential familiarity with the context of the exercise or, on a concluding exercise, by diminished motivation. Commonly used measurement models, however, do not accommodate either local dependence or multidimensionality effectively. How do we deal with these challenges?

This paper describes an approach to understanding the data from complex problem-solving performances. The approach is described in the context of the NAEP Problem-solving in Technology-rich Environments (TRE) study. The TRE study uses simulation tasks to measure students' proficiency in scientific problem solving with technology. The study capitalizes on Evidence-centered Design, a methodology for devising assessments and for using the evidence observed in complex student performances to make inferences about proficiency (Mislevy, Almond, & Lukas, in press; Mislevy, Steinberg, Almond, Breyer, & Johnson, 2001). The paper focuses on the latter use of the methodology, in particular how features are extracted from student performance, how these extractions are evaluated, and how the evaluations are accumulated to make evaluative judgments.

Measurement Goals

The TRE study is one of several projects funded by the U.S. National Center for Education Statistics as groundwork for using new technology in the National Assessment of Educational Progress (NAEP). NAEP is a sample survey that reports to policy makers and the public on what U.S. school children know and can do. The program reports scores for groups but not for individuals. TRE provides an example of how to measure problem-solving with technology in the NAEP context.

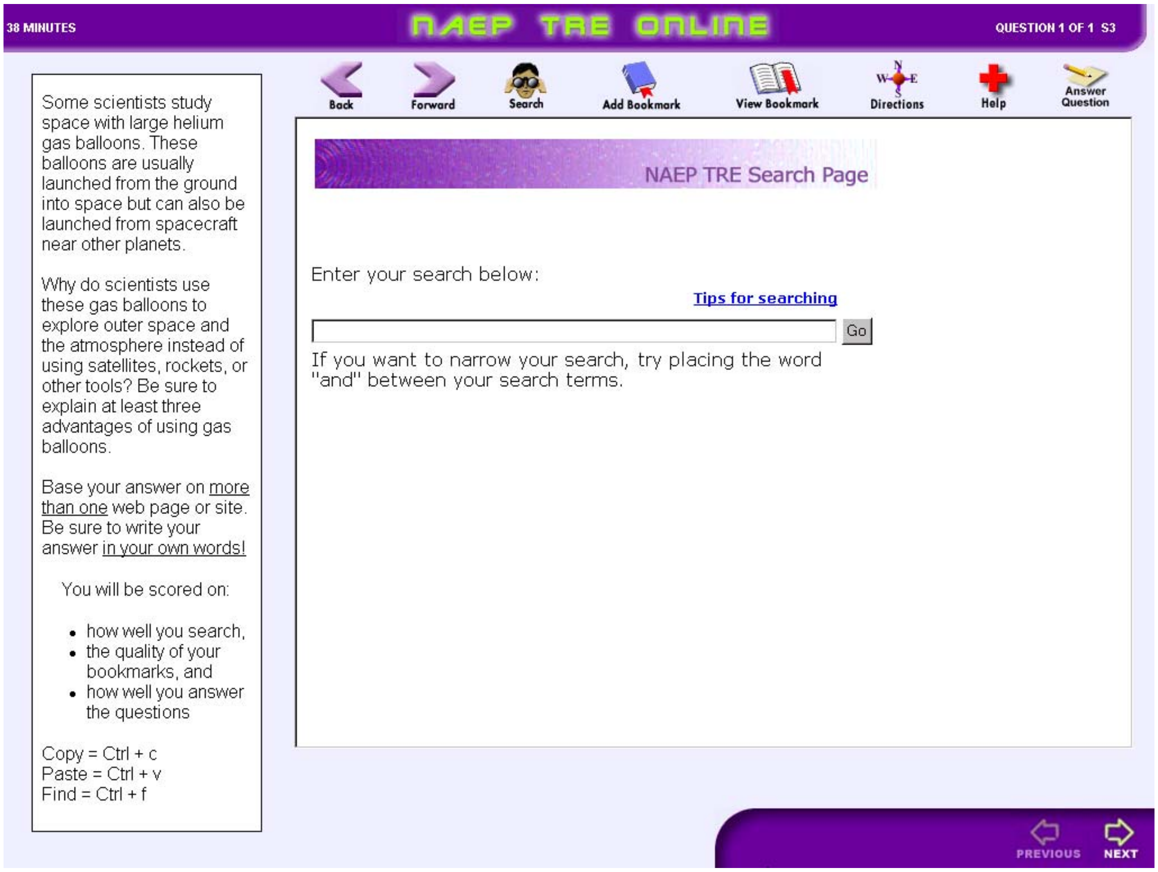
For purposes of the TRE project, problem-solving with technology is viewed as an integration of scientific inquiry and computer skill. By *scientific inquiry*, we mean being able to find information about a given topic, judge what information is relevant, plan and conduct experiments, monitor one’s efforts, organize and interpret results, and communicate a coherent interpretation. By *computer skill*, we mean (1) being able 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; and (2) being able to monitor one’s efforts.

The conception of scientific inquiry embodied in TRE is one of partial inquiry; full inquiry gives greater attention to question choice, explanation, and connections of those explanations with scientific knowledge than we are able to give (Olson & Loucks-Horsley, 2000, pp. 28-30). We chose partial inquiry 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 an example for NAEP that could be taken either in the direction of a content-based assessment like science or a more general problem-solving with technology assessment.

The project samples from a universe of content domains and technology environments. Problem-solving with technology can occur in such substantive domains as physics, biology, ecology, economics, and history. In each domain, a variety of technology tools can be applied, including word processors, search engines, spreadsheets, and e-mail, presentation, and graphics programs.

For TRE, we sampled from that universe so that the *same* content domain—the science associated with gas balloons—carries through different computer uses. Thus, we kept the domain constant and asked students to use various tools instead of asking them to use a single tool (e.g., a search engine) to answer questions from a variety of domains. We believe that the chosen approach is more consistent with the emphasis in the cognitive science literature on extended problem-solving because the student remains situated in the same substantive context throughout the assessment. In addition, giving salience to substance communicates our view that in real-world settings problem-solving with technology is driven by the problem and not by the technology: Successful problem solvers in a domain tend to look for the tool that's best suited to their problem, not a problem suited to the tool that happens to be closest at hand.

Two modules were created to measure the aforementioned skills, a search module and a simulation module (Bennett & Persky, 2002). The search module presents one open-ended question and several multiple-choice items about the science and uses of gas balloons. The student must search a simulated World Wide Web composed of some 5,000 pages to gather relevant information and fashion a response. Figure 1 shows a screen from the module. The motivating question, shown on the left, asks why scientists use such balloons—as opposed to satellites, rockets, and other mechanisms—to explore outer space. The area on the right functions like a standard Web browser. It allows the student to enter search queries, see results, bookmark pages, visit Web sites, and type a response.



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

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

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

You will be scored on:

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

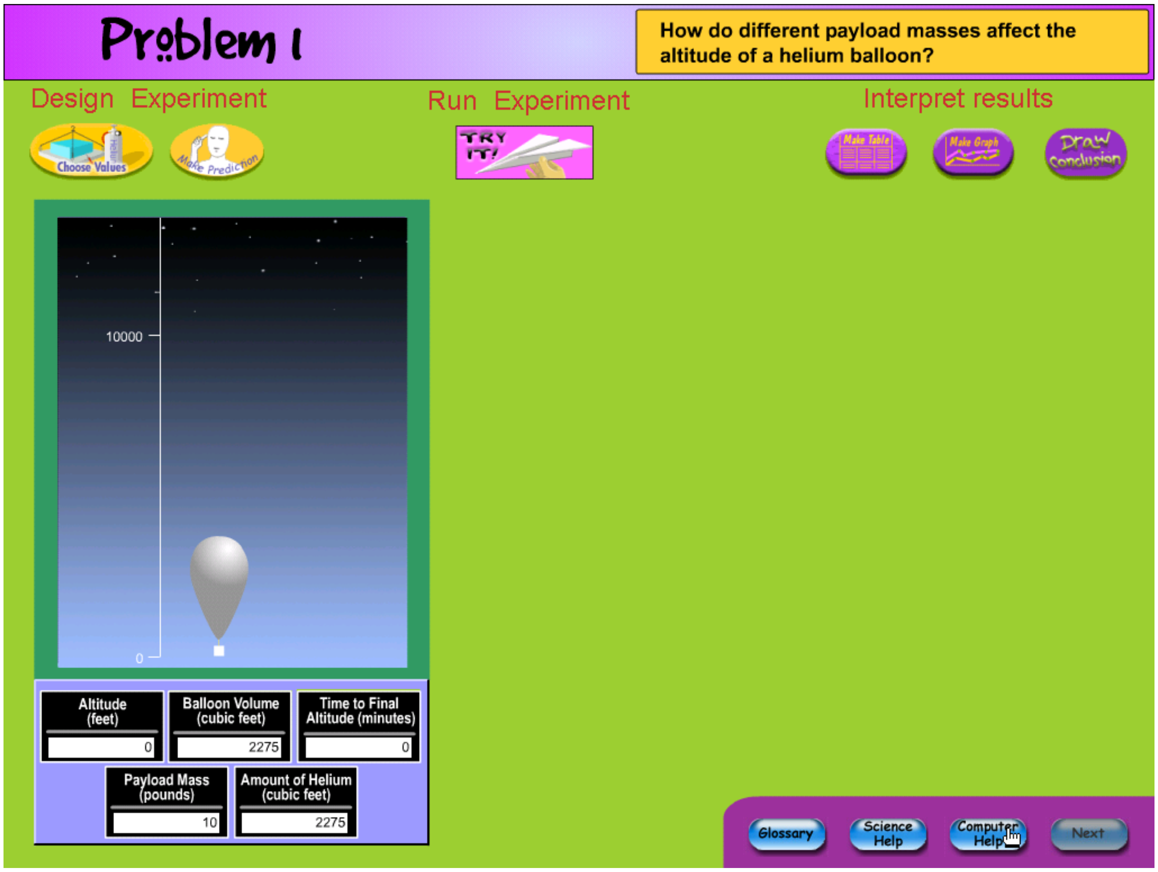
Copy = Ctrl + c
Paste = Ctrl + v
Find = Ctrl + f

Enter your search below:

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

***Figure 1.* A screen from the TRE search module.**

The simulation module presents a “what-if” tool that the student must use to discover an underlying scientific relationship for each of three problems. Figure 2 shows an example screen. The first problem is in the upper right-hand corner and involves determining the relationship between the payload mass carried by the balloon and its ultimate altitude.



Problem 1

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

Altitude (feet): 0

Balloon Volume (cubic feet): 2275

Time to Final Altitude (minutes): 0

Payload Mass (pounds): 10

Amount of Helium (cubic feet): 2275

Figure 2. **A screen from the TRE simulation module.**

The simulation tool is organized to facilitate a structured inquiry process built around designing experiments, running experiments, and interpreting results. To design an experiment, the student can choose values for an independent variable and make predictions. To interpret results, he or she can create a graph, make a table, or draw conclusions. The student can proceed through this process in any order, though some orders will be more productive than others and can conduct as many experiments as desired from a predetermined set.

Simulation results are presented in a flight box that depicts the behavior of the balloon under the selected conditions. Instruments at the bottom of the flight box also show results. These instruments dynamically give the balloon's altitude, volume, time to final altitude, payload mass, and the amount of helium put in it.

We attempt to minimize the impact of limited science knowledge by providing a glossary and problem-solving hints. We also attempt to minimize the impact of limited computer skill by providing an interface tutorial and computer help and by using common interface conventions like standard dialog boxes.

The TRE Interpretive Framework

In scoring responses to the search and simulation modules, we use both process and product information to make inferences about students' proficiency in problem-solving with technology. That is, we consider what students *do* in addition to what they answer. How do we do this scoring?

The approach to creating the TRE interpretive framework is based on evidence-centered design (Mislevy et al., in press; Mislevy et al., 2001). In this approach, we develop initial specifications for scoring and interpretation as part of assessment planning. These specifications take the form of *student* and *evidence* models. The student model constitutes our "theory" of how the components of proficiency are organized in the domain of problem solving in technology-rich environments. The evidence model describes how to connect student responses to the proficiencies about which we wish to know.

Graphically, the student model looks like Figure 3. Reading from left to right, the graph indicates that problem-solving in technology-rich environments is composed of computer skill and scientific-inquiry skill. Scientific-inquiry skill, in turn, is composed of two lower-level proficiencies—exploration and synthesis. These five proficiencies will be the reporting variables for the population and subpopulations in the TRE project.

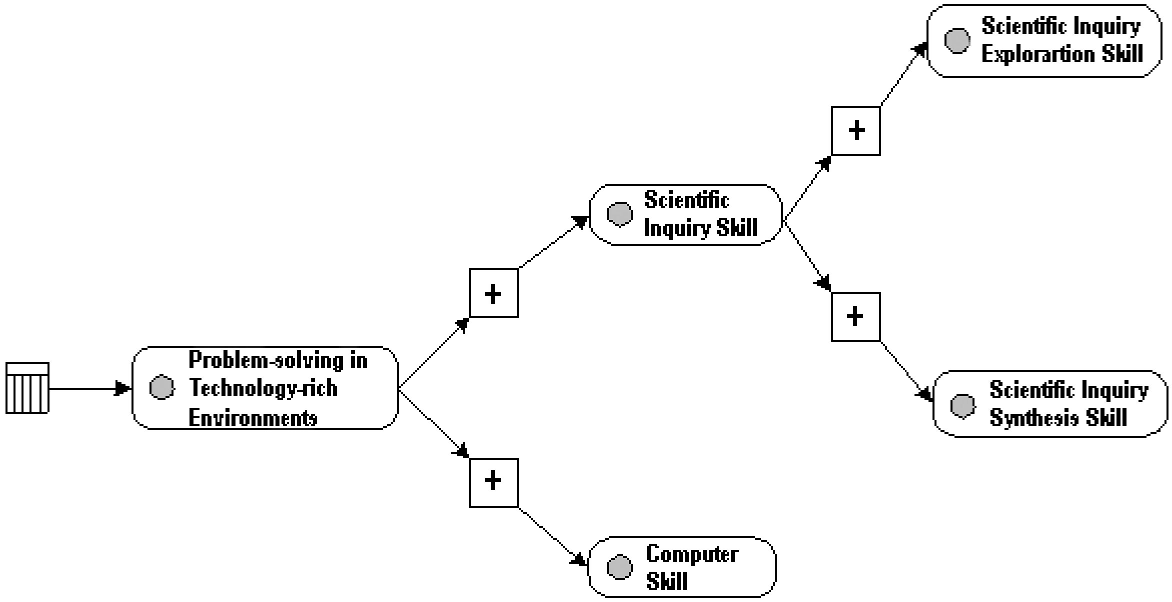

Scientific Inquiry Explorartion Skill
+
Scientific Inquiry Skill
+
+
Problem-solving in Technology-rich Environments
Scientific Inquiry Synthesis Skill
+
Computer Skill

Figure 3. **The TRE student model.**

Standing on each student-model variable is expressed in terms of a proficiency level. Any number of levels may be assigned, but for our purposes three or four might be sufficient (e.g., low, medium, high). Our uncertainty regarding a student's level takes the form of a distribution indicating the probability that he or she is at each level (e.g., the distribution for a given student might be: *low* = 10% probability, *medium* = 10%, *high* = 80%, which would indicate with considerable certainty that the student's level was *high*).

When a student takes a TRE module, each action is connected to one or more variables in the student model. We use a three-step evidence-modeling process to make these connections. The three steps are feature extraction, feature evaluation, and evidence accumulation.

For each TRE module, all student actions are logged in a transaction record. *Feature extraction* involves pulling out (or computing) particular observations from the record (e.g., in TRE simulation, one observation is the specific experiments the student chose to run). These observations may include both process and product variables. Table 1 shows an extraction from the first minute of the record for simulation problem 1. The extraction shows what the student did, the time at which each event occurred, and the value associated with any given action. Reading through the record, we can see that in designing the experiment the student pressed the

Choose Values button and selected a payload mass of 90 for the balloon to carry. He or she next pressed *Try It* to launch the balloon and was given a message suggesting that a prediction be made first. The student pressed the *Make Prediction* button and chose option C. He or she then pressed *Try It* again. 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 helium on the horizontal axis.

Table 1

A Portion of the Student Transaction Record From TRE Simulation Problem 1

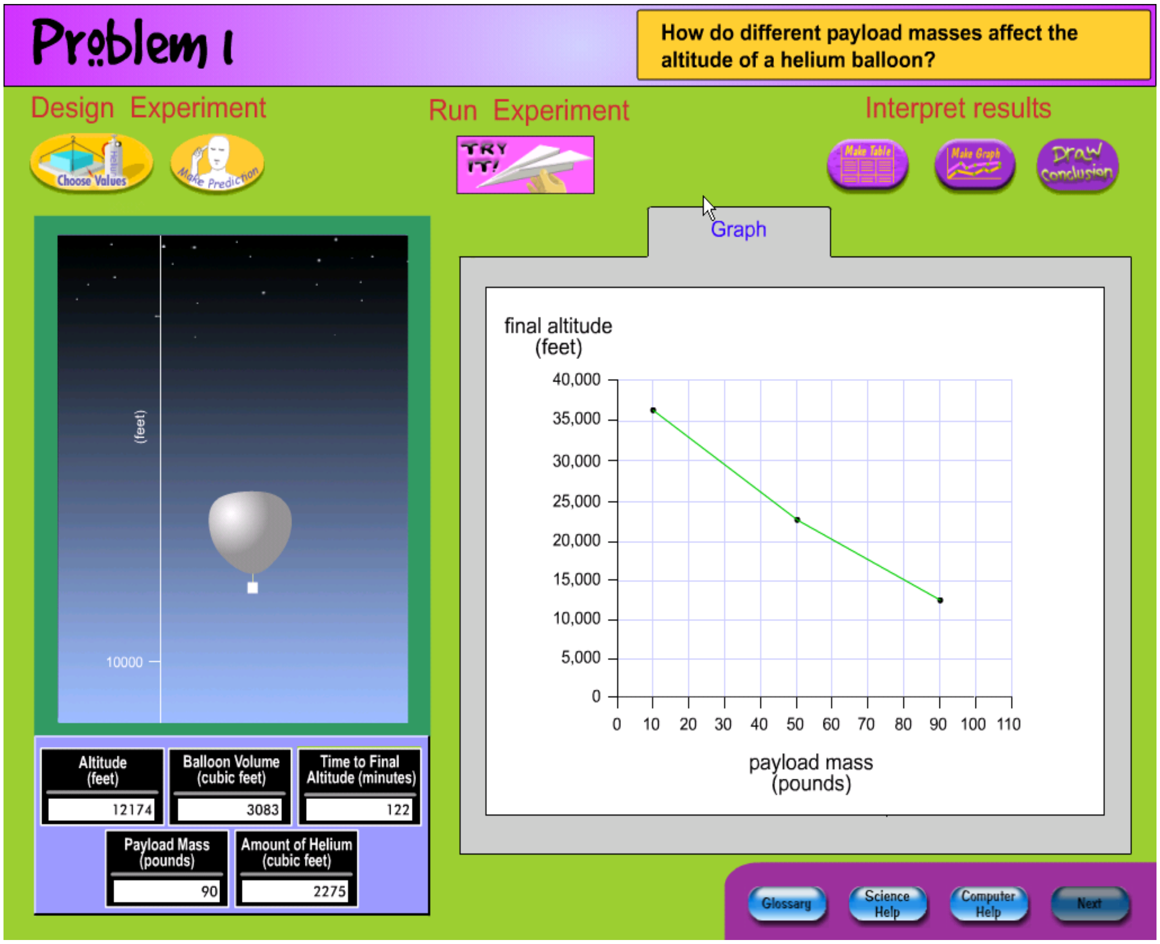
Time	Action	Value
137	Begin problem 1	
150	Choose values	90
155	Select mass	
157	Try it	
158	Prediction alert	
159	Make prediction	
161	Prediction option	C
162	Try it	
180	Make table	
182	Selected table variables	Payload mass
185	Make graph	
188	Vertical axis	Altitude
190	Horizontal axis	Helium

The second step in connecting observations to the student model is *feature evaluation*. Once desired elements are extracted, each extraction needs to be judged as to its correctness. Feature evaluation involves assigning scores to these observations. These scoring assignments may be done by machine or by human judges. In either case, the assignments are executed in keeping with evaluation rules. Below is a draft of such a rule, one that will be refined as we learn more about how the simulation functions. This rule is for determining if the *best* experiments to solve simulation problem 1 were run.

- 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 above 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.

Thus, 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 be a sufficient basis for the student to make a trustworthy inference. The claim behind the rule, then, is that the set of experiments in Figure 4 belongs in the top score-category because of the spread and number of values. This set is better than the set in Figure 5, for example, which receives the lowest score because it has only two closely spaced points.

The ordering described in the evaluation rule is, of course, only a claim. Formulating such a rule involves an iterative process in which logical challenges to the rule are posed and, if a challenge has merit, the rule is refined. At the end of that process, we have a *provisional* rule, one that we use *as is* until further challenge—primarily in the form of empirical data—suggests otherwise. Those data may come from asking students to think aloud during problem-solving or from a comparison of the responses of known-proficient and known-novice performers.



Problem 1

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

Altitude (feet): 12174

Balloon Volume (cubic feet): 3083

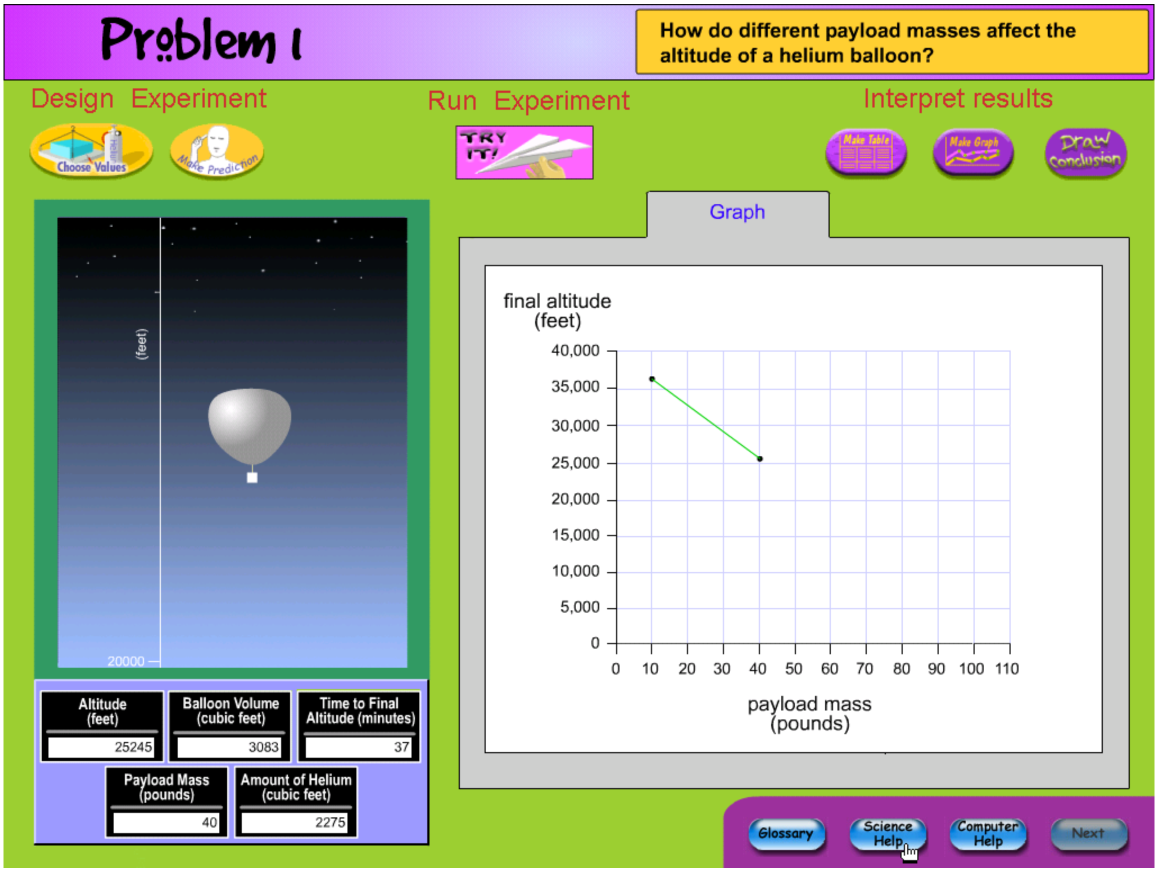
Time to Final Altitude (minutes): 122

Payload Mass (pounds): 90

Amount of Helium (cubic feet): 2275

***Figure 4.* A set of data points qualifying for the highest "Ran Best Experiments" score category.**

It is useful to note that the goal is to not to generate a perfect rule but rather a rule that works in the overwhelming majority of cases. No rule will accurately capture the behavior of all novice or all proficient 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 don't. As long as these positive and negative misclassifications are not too frequent and are not systematic (i.e., do not tend to occur in the same direction), they can be handled effectively through mechanisms that quantify uncertainty in examinee proficiency estimates, as described below.



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

Altitude (feet): 25245

Balloon Volume (cubic feet): 3083

Time to Final Altitude (minutes): 37

Payload Mass (pounds): 40

Amount of Helium (cubic feet): 2275

Figure 5. **A set of data points qualifying for the lowest "Ran Best Experiments" score category.**

The third step in connecting observations to the student model is *evidence accumulation*. Feature evaluations (like test items) need to be combined into summary scores that support the inferences we want to make from 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, we are doing this aggregation using Bayesian inference networks (Mislevy, Almond, Yan, & Steinberg, 2000). These networks offer a formal statistical framework for reasoning about interdependent variables in the presence of uncertainty. Bayesian networks are well suited to integrated tasks like TRE because these networks allow the various skills that compose performance to be modeled individually, along with the complex interrelationships that may exist among them.

Figure 6 graphically depicts the evidence model for the first simulation problem. There are similar evidence models for problem 2, problem 3, and for search.

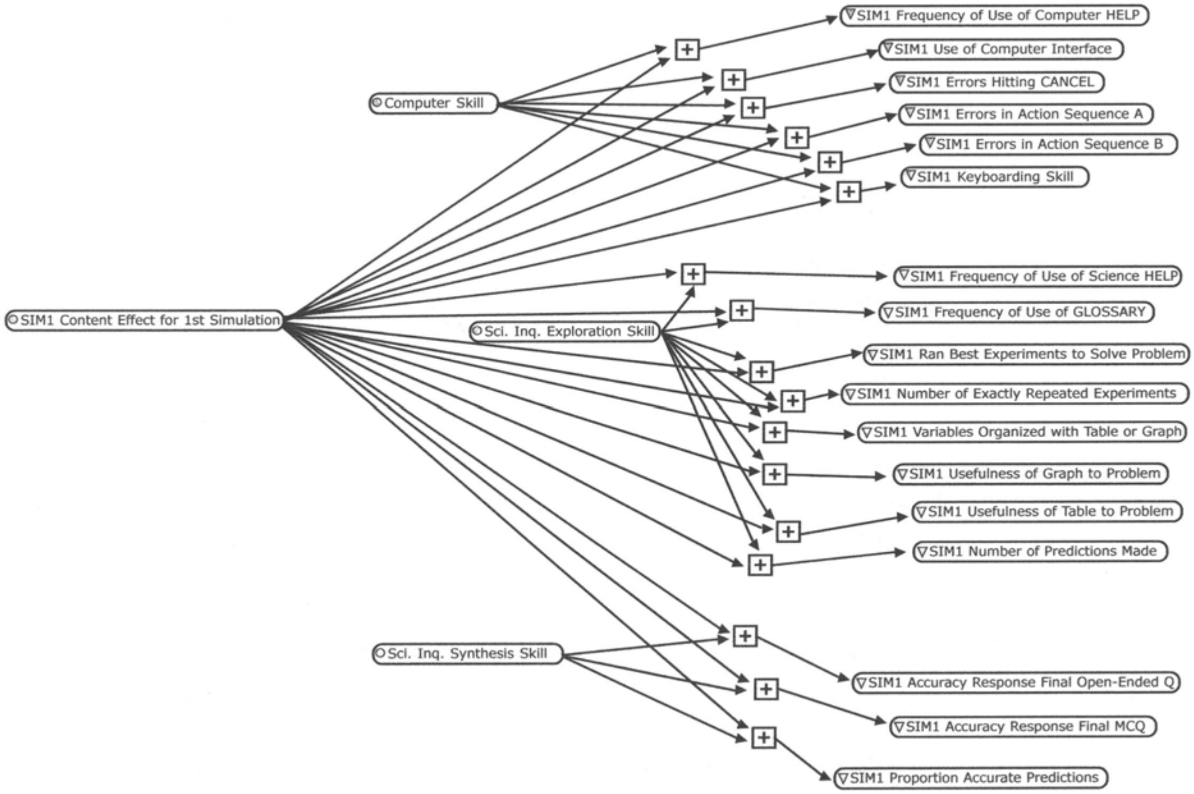

SIM1 Frequency of Use of Computer HELP
SIM1 Use of Computer Interface
SIM1 Errors Hitting CANCEL
SIM1 Errors in Action Sequence A
SIM1 Errors in Action Sequence B
SIM1 Keyboarding Skill
Computer Skill
SIM1 Content Effect for 1st Simulation
Sci. Inq. Exploration Skill
SIM1 Frequency of Use of Science HELP
SIM1 Frequency of Use of GLOSSARY
SIM1 Ran Best Experiments to Solve Problem
SIM1 Number of Exactly Repeated Experiments
SIM1 Variables Organized with Table or Graph
SIM1 Usefulness of Graph to Problem
SIM1 Usefulness of Table to Problem
SIM1 Number of Predictions Made
Sci. Inq. Synthesis Skill
SIM1 Accuracy Response Final Open-Ended Q
SIM1 Accuracy Response Final MCQ
SIM1 Proportion Accurate Predictions

***Figure 6.* A graphical depiction of the evidence model for TRE simulation problem 1.**

At the far left of the figure is a variable representing the context effect, that is, some local dependency among responses unrelated to the skills of interest. Extended problem-solving tasks are particularly susceptible to such effects, which may emanate from chance familiarity with a particular topic, personal interests, fatigue, or other sources. 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 isn't) interested in the passage topic. Conventional measurement models do not handle such dependency effectively. In the Bayes net framework, however, this dependency can be explicitly modeled for each extended exercise.

In the center of Figure 6 are those student-model variables that connect directly to the observables. Those variables are computer skill, scientific exploration, and scientific synthesis. Connected to computer skill are such observables as how frequently computer help is consulted, how extensively the various components of the simulation-tool interface are used, how often actions are cancelled, and whether actions are performed out of sequence. For scientific exploration, we include how frequently science help and the glossary are consulted, whether the best experiments were run, whether experiments were repeated, whether a table or graph was used, and how appropriate that table or graph was to the problem posed. Finally, for scientific synthesis, we look at the accuracy of answers to the open-ended and multiple-choice questions that motivate the problem and at the proportion of accurate predictions.

Note that some of these behaviors, like how frequently science help is consulted or the same experiment repeated, are expected to be negatively related to scientific inquiry. Others, like making a relevant graph, should be positively related. Also note that, until we have more data, these relations are nothing more than hypotheses, in principle similar to the ones an item writer makes in constructing an item and designating one option the key and the others distracters. For the item writer, as for us, that hypothesis will be confirmed, or refuted, through data.

How does such a model serve to facilitate inferences about proficiency? The reasoning in such a model runs from left to right, as indicated by the arrows in Figure 6. That is, the likelihood of observing a particular feature evaluation (e.g., *high* on running the best experiments) depends on a particular configuration of the student model. So, for *running the best experiments*, the likelihood of getting the top level will be greater for students who are highly proficient in scientific exploration than for those lower on that skill.

When a student takes a TRE module, we reason in the opposite direction: The feature evaluation for each observable is used to update our belief about standing on the student-model variable to which the behavior is connected. Thus, observing that a student *ran the best experiments* for simulation 1 would increase our belief in that student is being proficient in exploration skill. This increase would then propagate to other student-model variables linked to exploration, like scientific inquiry and problem-solving in technology-rich environments. This updating of the student model is carried out until all feature evaluations are incorporated from all three simulation problems (and potentially from the search module too). We can then generate a profile that gives the posterior probability distribution for each student-model variable for individuals, subgroups, and the test population. These probabilities quantify the uncertainty we have surrounding our proficiency estimates.

It is useful to note that standing on the student model variables constitutes a multidimensional picture of functioning that could not be generated through conventionally employed measurement models. Typically, multiple skills are modeled by creating separate measurement scales, each of which is indicated by a unique set of items. In the Bayes net framework, we 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.

As noted, the student model constitutes a "theory"—or more correctly, a structural proposal—for how the components of proficiency are organized in the domain. The evidence model is a set of *hypotheses* about what behaviors indicate proficiency. Obviously, those claims should be tested empirically and the results used to refine the models as appropriate. For TRE, among the questions that need to be answered are whether the evidence model functions as expected. To explore this question, we can observe the model's functioning when response patterns for hypothetical students are entered. For example, if we enter the responses we believe should be characteristic of the prototypic novice—creating a graph with the wrong variables, writing a wrong conclusion—does the evidence model generate the student-model profile we would expect? We should also observe the behavior of the model when values for specific student-model variables are fixed. That is, are the response probabilities for the observables sensible, given known settings for the student model? Additionally, we should observe behavior of the model with real student data to insure that the probability distributions we set for each observable are consistent with performance in the target population.

Besides asking if our evidence model functions as expected, we need to ask whether its composition and structure are empirically defensible. For one, does it incorporate the correct variables? To test this question, we can use quantitative and qualitative methods to identify behaviors that distinguish known-proficient from known-novice performers. The object of this exercise is to determine if those whom we know to be practiced at scientific inquiry are differentiated from those who are not on the basis of, for example, the spread and number of experiments they run, as our provisional evaluation rule claims they should be. With respect to model structure, are the empirical relationships among evidence- and student-model variables consistent with our expectations? For instance, assuming limited prior knowledge, one would expect students scoring high on synthesis to also score high on exploration. Those who gave answers that showed they understood the underlying relationships should have also conducted appropriate experiments, used tables and graphs, included the relevant variables in those displays, and so on.

These evaluations can lead to different kinds of evidence-model revision. We may, for example, modify an evaluation rule by changing the number of score levels into which we attempt to classify behavior or by amending the criteria for how the levels are defined. More fundamentally, we may eliminate an evidence-model variable entirely if it appears to duplicate another variable or if it fails to provide evidence that differentiates novices from proficient performers.

Applications to Operational Use

The TRE modules represent a limited example of how to measure complex problem-solving in technology environments. The modules are limited in that they cover a single substantive domain and utilize a few technology tools. In addition, they provide no estimate of substantive skill independent of technological skill or independent of general reasoning ability. A more complete assessment would entail some additional problem domains (e.g., in biology, ecology, history) and additional tools (e.g., e-mail and presentation programs). Also, to better separate technology skills from substantive ones, it might be useful to add a paper measure of substantive skills that represents as closely as possible the competencies contained in the online assessment and a measure of technology skill that requires no substantive proficiency. In a sample survey such as NAEP, each student would take some subset of these measures, which would allow estimating the population's standing on the student-model variables.

We believe that the approach to understanding student performance illustrated in this paper is applicable to assessment purposes other than national monitoring. The approach is at its most powerful when the assessment calls for responses composed of multiple related parts, as in a simulation or other complex performance task. In such cases, this approach can help organize test design and the framework for interpreting examinee performance so that the purpose of assessment—be it summative or formative—is more likely to be satisfied. The approach is also applicable to assessing both group and individual performance, although for the assessment of individuals a longer test would likely be necessary to accommodate the range of problem contexts required for score generalizability.

Conclusion

The TRE project illustrates how to make sense of the complex data examinees provide in response to extended performance tasks. Educators favor such tasks for assessment because they align more closely with instructional goals than the elemental questions that compose many large-scale tests. Cognitive science research also supports the use of such tasks because they bring to bear multiple processes that must be integrated and monitored for successful performance.

The TRE tasks ask the student to answer questions through electronic information search and "what-if" experimentation. The assessment includes tutorials on tool use, employs common interface conventions, and offers help facilities to decrease the chances that limited computer skill will prevent the student from showing inquiry skill (and vice versa). As students work with the modules, both process and product behaviors are tracked.

We represent our thinking about the structure of proficiency in the problem-solving with technology domain as a graphical student model. We connect behaviors to proficiencies using a formal statistical framework that provides a principled means of handling multidimensionality, local dependence, and uncertainty. With these cognitive and statistical modeling methods, we can report population and subgroup performance on complex tasks in terms of a structural description of proficiency. These descriptions may have more relevance to instructional policy and practice than purely content-based depictions because content-based depictions may present much too simple a view of what it means to be proficient in a domain.

References

Bennett, R. E., & Persky, H. (2002). Problem solving in technology-rich environments. In Qualifications and Curriculum Authority (Ed.), *Assessing gifted and talented children* (pp. 19-33). London, England: Qualifications and Curriculum Authority.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (in press). *A brief introduction to Evidence-Centered Design*. (CSE Technical Report). Los Angeles, CA: UCLA CRESST.

Mislevy, R. J., Almond, R. G., Yan, D., & Steinberg, L. S. (2000). *Bayes nets in educational assessment: Where do the numbers come from?* (CSE Technical Report 518). Retrieved November 14, 2002, from the UCLA CRESST Web site: http://cresst.org/products/reports_set.htm

Mislevy, R. J., Steinberg, L. S., Almond, R. G., Breyer, F. J., & Johnson, L. (2001). *Making sense of data from complex assessments* (CSE Technical Report 538). Retrieved April 19, 2002, from the UCLA CRESST Web site: http://www.cse.ucla.edu/CRESST/Reports/RML%20TR%20538.pdf

Olson, A., & Loucks-Horsley, S. (Eds.). (2000). *Inquiry and the National Science Education Standards: A guide for teaching and learning*. Washington, DC: National Academy Press.