Appendix F: Bayesian Estimation in the Problem Solving in Technology-Rich Environments Study

Introduction

The Problem Solving in Technology-Rich Environments (TRE) study incorporates several design features that are not found in standard NAEP analysis. These features include

- an a priori hypothesized structure of the relationship among the set of latent proficiency variables,
- the potential to accommodate multivariate items (i.e., items that measure more than one latent proficiency), and
- inclusion of context effects; items sharing a context are related to each other more strongly than to other items.

All three of these features are beyond the scope of measurement models used in operational NAEP. Operational NAEP employs a univariate Item Response Theory (IRT) model that uses a simple structure, i.e., each item measures only one latent proficiency. Since the IRT model is univariate, there can be no structural relations among latent proficiencies, there can be no item that measures more than one proficiency, and there can be no context effect in addition to the latent proficiency.

This appendix outlines the cognitive models that were used in the TRE study. (The term, cognitive model, is used here to refer to the union of the student and evidence models described in chapter 2 of this report.) These are represented by directed graphs showing latent proficiency, observable, and context variables, with arrows showing direction of influence. Note that two scenarios, or separate computer tasks, were delivered. One was the Search scenario, in which students used a simulated web search to answer questions about scientific balloons. They conducted searches, gathered information, and then summarized results. The second scenario was Simulation. In this activity, students used a simulation tool to conduct a series of experiments in order to discover relationships among variables related to the physics of balloon behavior in the atmosphere.

This appendix also presents the Bayesian models used to analyze the data and estimate item parameters. These consist of the IRT model for items; the structural model for representing relationships among the latent proficiencies; the conditioning model, which describes the structured prior distribution of the latent problem-solving in TRE proficiency; and finally the population model for deriving estimates of population means, percents, and associated standard errors.

Finally, this appendix discusses the construction of a real-time inference engine for the Search scenario. Model parameters estimated from the Bayesian IRT analysis are imported as fixed quantities into an inference engine (ERGO 2001 by Noetic Systems, Inc.), enabling sensitivity testing of the model and scoring of student responses. Profiles of proficiencies can be selected to see what response probabilities of the observables will result. Also, a vector of observed responses can be selected, and the resulting proficiency scores can be estimated. The inference engine can also be used as a stand-alone application to get real-time estimates of proficiency as an examinee responds to the assessment. This aspect of the Bayesian inference engine demonstrates the feasibility of using a computer to assess and immediately provide proficiency estimates over the Web.

The Cognitive Models

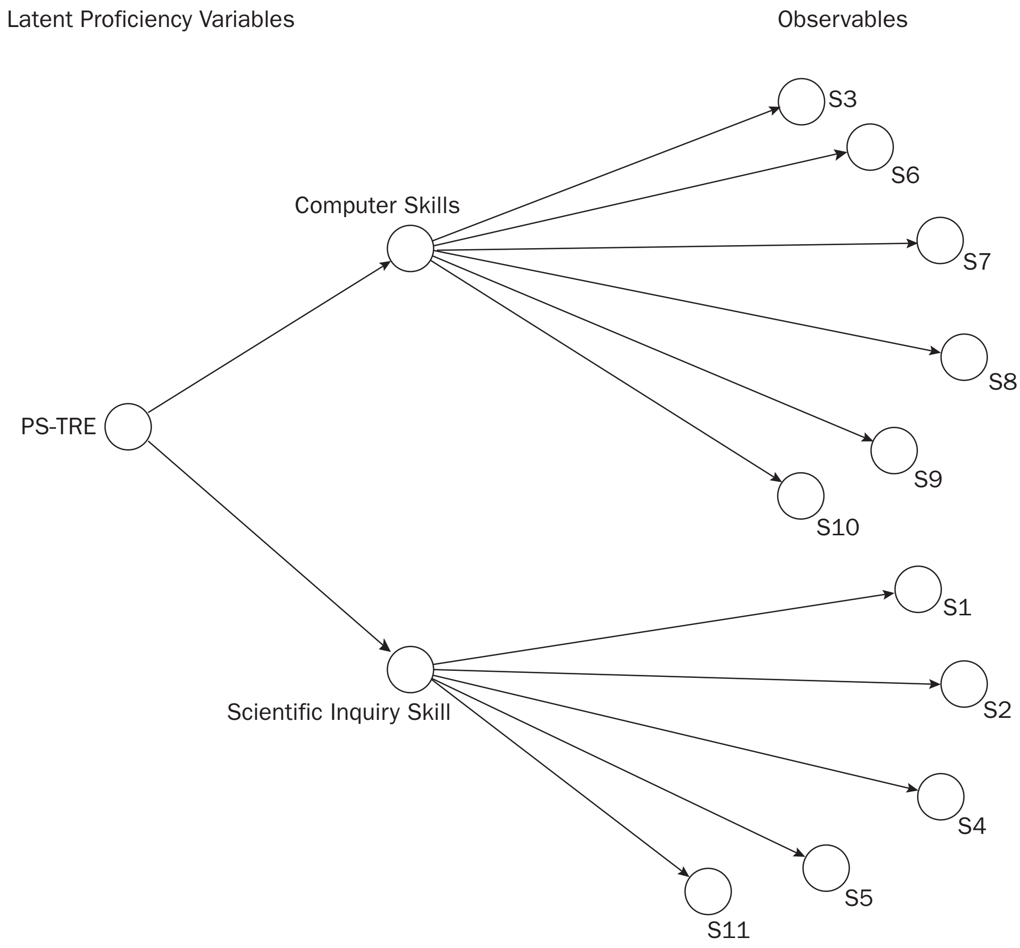
Two somewhat different cognitive models were fitted to the two TRE scenarios. First, consider the directed graph in figure F-1, which depicts the relationships among variables for the Search scenario. Two classes of variables are shown. To the left are latent proficiencies, and to the right are observables, representing observed scores on performance tasks.

This discussion of latent proficiencies follows customary usage in calling precursor variables “parents” and other latent variables “children” (to avoid use of causal language). In this model, the parent proficiency is problem solving in technology-rich environments (PS-TRE), which has computer skills and scientific inquiry skill as resultant or “child” proficiencies. Arrows between the latent skills indicate the direction of influence.[1]

[1] Note that scientific inquiry skill was originally proposed as having two component skills: scientific inquiry exploration skill and scientific inquiry synthesis skill. With the Search scenario, it was found that there were too few observables to reliably measure these constructs. As a result, they were combined into a single scientific inquiry proficiency in the final model.

To the right of figure F-1 are observables. These are summaries of observed behaviors that can be mapped onto several levels of partial credit (from two to four levels). The probability that a student will score at a specific level is a function of that student's latent skill. The nature of this function is defined by an IRT model. According to the model, computer skill contributes to a student's propensity to respond correctly to observables requiring computer-related abilities such as keyboarding, using menus correctly, and not needing to use the help function. Similarly, scientific inquiry skill contributes to a student's propensity to explore content and draw conclusions about scientific questions correctly.

Figure F-1. The TRE Search cognitive model, grade 8: 2003



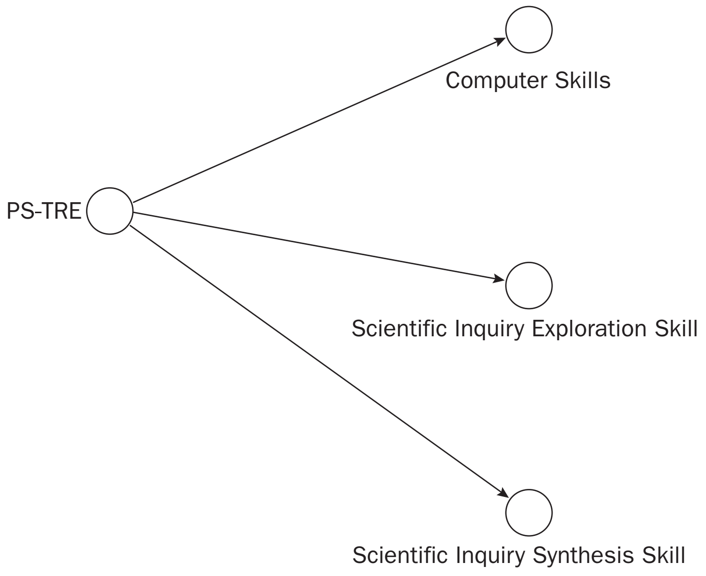
NOTE: PS-TRE = Problem solving in technology-rich environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure F-2 shows the directed graph depicting a structural (or student) model for the latent proficiencies in the Simulation scenario. In this model, PS-TRE is the parent of three other latent skills: computer skills, scientific inquiry exploration skill, and scientific inquiry synthesis skill. These latter three are proficiencies that contribute to the propensity to respond correctly to observables.[2]

Figure F-3 shows the cognitive model for the Simulation scenario. The variables on the left, PS-TRE, computer skills, scientific inquiry exploration skill, and scientific inquiry synthesis skill, are latent proficiencies. These are the direct precursors of observables, which are found in the middle of the diagram. Each observable measured (was the child of) just one latent proficiency. This simple structure was confirmed to fit the data best. On the far right of figure F-3 are three other latent variables, which define the effect of context.

The three context effects correspond to the three Simulation problems in the scenario. The context variables represent any knowledge, skill, or other factor that is specific to one Simulation task but not another. Students with a higher level of task-specific skill will tend to do better on all the items in the task. As a result, items sharing a common task tend to be more highly correlated than items in different tasks. The context effect can be thought of as controlling for a type of nuisance variation. With context effects in the model, conditional independence of observables, given a student's latent skills, holds. The assumption of conditional independence is a basic tenet of any explanatory model. This assumption also underlies all conventional IRT estimation.

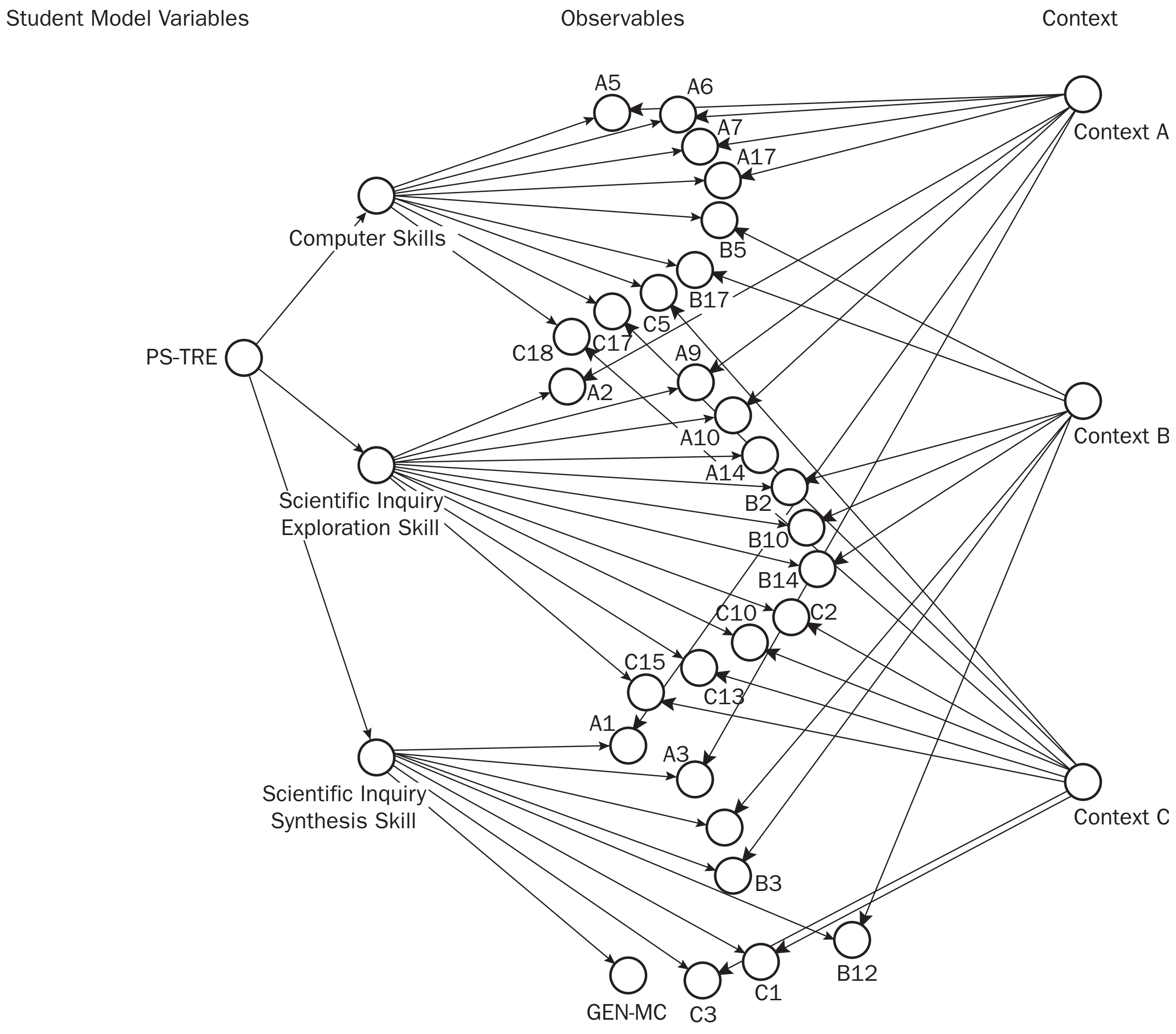
Figure F-2. Student model for TRE Simulation scenario, grade 8: 2003


Computer Skills
PS-TRE
Scientific Inquiry Exploration Skill
Scientific Inquiry Synthesis Skill

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

[2] Unlike the Search scenario, Simulation had a sufficient number of observables to reliably measure exploration and synthesis as separate skills. However, scientific inquiry skill was dropped as a precursor to the latter two proficiencies, because scientific inquiry skill was not reliably measured by its component skills.

Figure F-3. Cognitive model for TRE Simulation scenario, grade 8: 2003



NOTE: PS-TRE = Problem solving in technology-rich environments. GEN-MC = Synthesizing multiple-choice items.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

General Description of the Bayesian Model

The previous section gave an outline of the cognitive model behind this analysis. This section presents a detailed description of the models used to analyze data and estimate item parameters.

Item Response Model

In the TRE study, the item-category responses (i.e., the probability of responding correctly to a category of an observable) are modeled as dichotomous item responses. In the Simulation scenario, a student's behavior on an observable is influenced by a latent proficiency skill (student model variable) and a context effect. As a result, the item response is multivariate in form. In the present form, this is a compensatory model, with equal slopes for Θ_{ij}, the value for a student on the latent proficiency, and Φ_{im}, the value for a student on the latent context effect. This model is compensatory in that the two latent variables have an additive effect on item response. Other types of relationships (e.g., disjunctive) could have been modeled to represent different sorts of relationships between the latent variables (Almond et al. 2001).

For observables with a dichotomous response (i.e., that can either be correct or incorrect), the multivariate item response takes the form

$$p_{ij}(x_{ij}=1\,|\,a_j,b_j,\Theta_{ij},\Phi_{im}) = \frac{1}{1+\exp[-\mathrm{K}*a_j(\Theta_{ij}+\Phi_{im}-b_j)]} \qquad (1)$$

where

- K is a scaling constant,
- P_{ij} is the probability of student i correctly responding to item j,
- Θ_{ij} is the value of student i on the parent proficiency j,
- Φ_{im} is the value of student i on latent context effect m,
- a_j is the slope of the item response function for item j, and
- b_j is the difficulty of the item response function for item j.

The probability of responding incorrectly to the observable is the complement of success, 1-p_{ij}.

As previously explained, the context effect represents the correlation among responses to observables having a common context. In the Simulation scenario, there are three problems of increasing complexity. Each problem forms a context. Any task-specific skills contribute to a latent context propensity in the student. This parameterization of the context effect follows the item cluster effect model of Scott and Ip (2002). In the Bayesian IRT model, the context effect has prior

$$\Phi_{im} \sim N(0,\tau_m)$$

for task m. The precision of the context is given a gamma prior:

$$\frac{1}{\tau_m} \sim Gamma(.01,.01).$$

Gelman and colleagues (1995) point out that a gamma distribution with parameter values approaching zero constitutes a noninformative prior. In this case, the sampled values would be very dispersed, approaching a uniform distribution.

For observables with polytomous responses, i.e., that can be responded to in two or more categories of partial credit, the item response is more complicated. The probability of responding to each category of partial credit, or higher, is modeled as a compensatory multivariate item response as above, but with an additional item-category parameter, d_{jk}, for item j and category k. Since the probability is for a given category, *or higher,* it is referred to here as p^{cum}. Such a formulation follows Samejima (1969).

$$p^{CUM}(x_{ij}=\mathrm{k}\,|\,a_j,b_j,\Theta_{ij},\Phi_{im}) = \frac{1}{1+\exp[-\mathrm{K}*a_j(\Theta_{ij}+\Phi_{im}-b_j-d_{jk})]} \qquad (2)$$

where $p^{\mathrm{cum}}_{\mathrm{i,j,k}}$ is the probability of responding in categories k, k+1, …Q, where Q is the highest category of partial credit.

Although these parameters will be estimated by Bayesian techniques using a Markov Chain Monte Carlo (MCMC) algorithm, constraints to assure identifiability of item-category parameters were employed. This was accomplished by stipulating that the item-category associated with the first category, d_{j0} , is zero, and setting $\sum_{m=1}^{M} d_{jm} = 0$. In practice, only a single item category parameter was estimated. For three-category items, d_{j1} had a positive prior, $N(1{,}1000)$, and $d_{j2} = - d_{j1}$. For four-category items, $d_{j1} \sim N(1{,}1000)$, $d_{j2} = 0$ and $d_{j3} = -d_{j1}$. The positive prior means that d_{j1} will likely be associated with more difficult levels of item response.

Since the response probabilities are cumulative in that they are the probability of responding in category k or higher, the item-category probabilities (except for the last one) must be calculated by subtraction:

$$p_{i,j,0} = 1 - p_{i,j,1}^{cum}$$

$$p_{i,j,1} = p_{i,j,1}^{cum} - p_{i,j,2}^{cum}$$

.

.

.

$$p_{i,j,Q} = p_{i,j,Q}^{cum}$$

Determining the Scale of the Latent Proficiencies

The scale of the latent proficiencies is indeterminate. This indeterminacy can be resolved in a Bayesian model either by specifying strong informative priors or by constraining the item parameters. The latter course was taken. The scale for each of the measured latent proficiencies was determined by setting the following constraints on the item parameters corresponding to the observables that measure that scale:

$$\sum_{j=1}^{J_p} b_{jp} = 0$$

where J_p is the number of items in proficiency p, and b_{jp} is the difficulty parameter for item j in proficiency p; and

$$\prod_{j=1}^{J_p} a_{jp} = 1 \quad ,$$

where a_{jp} is the slope associated with item j in proficiency p.

Structural Equation Model

There is a network of relations among the student model variables. These structural relations are modeled as simple linear regressions:

$$\Theta_i^{child} = B_0 + B_1 * \Theta_i^{parent} + e_i$$

with, (3)

$$VAR(e_i) = \sigma^2_{child\text{-}parent}$$

In the Simulation scenario, for example, these describe how PS-TRE influences computer skills, how PS-TRE influences scientific inquiry exploration skill, and how PS-TRE influences scientific inquiry synthesis skill.

Because of the complexity of the overall model, the structural equations were constrained to a limiting case with slopes fixed to 1.0. An informative prior was set for B_0 , at $N(0,1)$. Finally, $Var(e_i)$ was set to 1.0, as a way to control the overall variance of the proficiency estimates.

Structured Prior for the Summary Proficiency, Problem Solving in Technology-Rich Environments

With all NAEP assessments, the average number of items measuring each subproficiency for an examinee is small. Such sparseness of measurement can lead to biased estimates of group quantities. A way to remedy this problem is to use auxiliary information related to an examinee's ability in the estimation of group means and percents. This is accomplished by regressing latent proficiency scores on student background information. In operational NAEP, a Bayesian estimation procedure is employed in which item response information is combined with student background information to get posterior distributions of proficiency for each examinee (Mislevy 1991). In the present application, background information is introduced by defining a structured prior on the unmeasured summary proficiency, PS-TRE.

Auxiliary information is introduced by assuming that an examinee's prior ability is structured (i.e., derived from a regression of proficiency on background variables),

$$\text{PS–TRE}_{\text{i}} \sim N(\mathbf{\Gamma}'\mathbf{y}_{\text{i}}, \sigma^2), \tag{4}$$

where $\mathbf{y_i}$ is a vector of background variables for examinee i, Γ is a vector of regression effects, and σ^2 is a common variance for all examinees. In the present application, there are 10 categorical background variables that are recoded into 21 dummy variables. These variables consist of gender, race/ethnicity, whether the student had disabilities or was an English language learner, whether the scenario was administered to the student on a laptop computer, prior computer knowledge level, and socioeconomic status (SES), including parents' education level, number of reading-related materials in the home, whether the student was eligible for free/reduced-price school lunch, and whether the student was in the Title I program.

In order to control the contribution to proficiency variance made by the structured prior, two conditions were imposed. First, regression parameters were given informative priors with high precision,

$$\Gamma_{\text{p}} \sim N(1,1), \tag{5}$$

for regression weight p (p = 1 to 21). Next, the predictors, $\mathbf{y}_{\text{i}}$, were standardized and weighted by approximately $\frac{1}{\sqrt{21}}$ (the square root of the inverse of the number of predictors), so that the variance would not increase as the number of predictors increased. The R-squares of the conditioning models for the Search and Simulation scenarios were modest, between .34 and .41, but within the range of operational NAEP assessments.

In the present application, regression parameters, variance components, and the prior proficiency distribution of PS-TRE are estimated by using an MCMC algorithm, in which all model parameters are jointly estimated, conditional on the data. A general outline of the MCMC algorithm will be given in the next section.

General Description of MCMC Estimation Techniques

In operational NAEP procedures, item parameters are estimated using a marginal maximum likelihood approach (Muraki and Bock 1997). Multivariate proficiencies with a structured prior distribution are estimated in a conditioning phase in which item parameters in the first phase are introduced as fixed parameters (Mislevy 1991). In TRE, an MCMC algorithm to estimate all parameters simultaneously was employed. For item parameter estimates, the MCMC approach has been shown to produce point estimates and standard errors that are similar to those in operational NAEP estimates (Patz and Junker 1999). Further, if the scope is extended to include item parameters, conditioning parameters, and sampling variances, MCMC estimation produces results similar to those produced by operational NAEP techniques, when models are parallel (Johnson and Jenkins 2005). In the present research, MCMC estimation is applied to a model that is unlike an operational NAEP model in several key aspects (e.g., multivariate items and structured relationships among latent proficiencies). Also, unlike that in Johnson and Jenkins, the present model does not incorporate estimates of sampling variances. These are estimated by a separate jackknife procedure, which is an approach similar to that of Scott and Ip (2002).

A Markov chain is a sequence of random variables,

$$\psi^{1}, \psi^{2}, \ldots, \psi^{\mathrm{T}},$$

such that the probability of observing ψ^{t} is the transition probability,

$$p(\psi^{t} \mid \psi^{t-1}). \tag{6}$$

So ψ^{t} depends only on the previous state of the chain.

Under certain regularity conditions (Tierney 1994, section 3.1), the Markov chain converges to a stationary distribution (i.e., is invariant over time *t*). The general idea behind MCMC estimation is to set up a chain, which converges to a stationary distribution that equals the joint conditional distribution of model parameters, given data:

$$p(\psi \mid \mathrm{X}).$$

The procedure for deriving statistical estimates from a Markov chain is the following: Simulate a series of "burn in" observations from the chain until it is judged that the chain has converged to its stationary distribution,

$$\psi^{\text{-M}}, \psi^{\text{-(M-1)}}, \ldots, \psi^{0}.$$

The Gelman-Rubin diagnostic gives one test for convergence (Gelman and Rubin 1992). The M iterations till convergence are called "burn in iterations." For the burn-in phase, 5000 iterations were required. These were then tested for convergence.

After convergence, a series of T further observations are drawn from the joint distribution of the model parameters:

$$\psi^{1}, \psi^{2}, \ldots, \psi^{\mathrm{T}}.$$

Typically, between 5,000 and 10,000 samples of each parameter were drawn from the joint posterior.

Point estimates of model parameters are calculated from sample averages:

$$\hat{\psi}_{\mathrm{p}} = \frac{1}{\mathrm{T}} \sum_{t=1}^{\mathrm{T}} \psi_{\mathrm{p}}^{t}, \tag{7}$$

where T is the number of MCMC iterations.

This procedure would yield a point estimate of parameter p, such as an item difficulty or the proficiency score for examinee i. However, for more complex parameters, such as "percent above achievement-level cut-point K," estimates are averages of functions of parameters:

$$\begin{aligned} \hat{\Theta}_{\mathrm{p}} &= \sum_{t=1}^{\mathrm{T}} f(\psi_{\mathrm{p}}^{t}) \\ &= \frac{1}{\mathrm{T}} \sum_{t=1}^{\mathrm{T}} \frac{1}{\mathrm{N}} \sum_{i=1}^{\mathrm{N}} I(\Theta_{i}^{t}), \end{aligned} \tag{8}$$

where $I(\Theta_{i}^{t}),$ is an indicator of whether proficiency Θ for examinee i is at or above achievement-level cutscore K, and N is the sample size.

It is often difficult to simulate multivariate draws from the joint conditional distribution. A way to simplify the process is to take univariate draws from a distribution conditional on the data and all other model parameters. This has been shown to approximate draws from the joint posterior distribution (Geman and Geman 1984). By this approach, one draw of the parameters at iteration t, ψ^{t}, would consist of P univariate draws, each draw conditioned on the data and the rest of the parameters. If a set of parameters is symbolized by Ω, then the sequential set of draws for iteration t is described by:

$$\begin{aligned} &\psi_{1}^{t+1} \sim \pi(\psi_{1} \mid X, \psi_{1 \notin \Omega}^{t}) \\ &\psi_{2}^{t+1} \sim \pi(\psi_{2} \mid X, \psi_{2 \notin \Omega}^{t}) \\ &\vdots \\ &\psi_{\mathrm{P}}^{t+1} \sim \pi(\psi_{\mathrm{P}} \mid X, \psi_{\mathrm{P} \notin \Omega}^{t}) \end{aligned},$$

where $\pi(* \mid *)$ is the stationary distribution of a parameter, and $\psi_{p \notin \Omega}^{t}$ is the most current vector of parameters with parameter p excluded.

The MCMC simulating package BUGS (Spiegelhalter et al. 2004) was used to get Bayesian estimates of parameters. When posterior distributions can be explicitly defined, BUGS uses a Gibbs sampler. When posterior distributions of a particular parameter are not explicitly available, it uses two types of approximation for the univariate draw: Metropolis Hastings (Metropolis et al. 1953) and slice sampling (Neal 2003). In the present research, BUGS employed all three types of sampling.

Estimation of Population Parameters

Point estimates for most model parameters (e.g., item parameters and regression coefficients) were calculated from MCMC sample averages as described in equation 7. However, for estimates of mean proficiencies of student groups and their associated standard errors, approximation procedures from operational NAEP were employed.

Plausible Values of Latent Proficiencies

Plausible values consist of a set of M independent draws from each examinee's posterior proficiency distribution. With MCMC estimation, drawing plausible values consists of systematically selecting 5 values from the thousands of MCMC draws, taking care that each draw has a minimum of 50 draws between them. Equation 6 implies that each MCMC draw is dependent on the previous draw. As a result, the MCMC series of parameter draws are autocorrelated. Diagnostics indicated that it took about 25 to 50 draws for the autocorrelation to fall to zero. In practice, the 5 independent draws were separated by several hundred iterations. Following NAEP terminology, these 5 independent draws will be called plausible values (Allen, Carlson, and Zelenak 1999, section 12.3.3).

Calculating Student Group Means

The Bayesian model did not contain a model for the population. Such a model would have to include proficiency distributions corresponding to all primary sampling units and schools in the sampling frame. This would have been impractical for the present analysis. As a result, sampling weights are used to approximate population estimates.

The targets of reporting are student group means and standard errors. Student group means are calculated on each of the 5 plausible values and then averaged:

$$\hat{\mu}_{\mathrm{kG}} = \frac{1}{N_{\mathrm{G}}} \sum_{i \in \mathrm{G}} w_i PV_{\mathrm{k}i}, \tag{9}$$

where $\hat{\mu}_{\mathrm{kG}}$ is the estimated population mean of student group G, for the kth set of plausible values, w_i is a sampling weight for examinee I, N_G is the weighted size (sum of sample weights) of student group G, and PV_{ki} is the plausible value k for examinee i.

Point estimates are averages over plausible values (Allen, Carlson, and Zelenak 1999, section 12.4.1),

$$\hat{\mu}_{\mathrm{G}} = \frac{1}{\mathrm{M}} \sum_{k=1}^{\mathrm{M}} \hat{\mu}_{k\mathrm{G}}, \tag{10}$$

where M is the number of plausible values (which is 5 in this application).

Estimating Standard Errors

Measurement variance

Measurement variance is the variance across plausible values of the target statistic. The first step in the procedure is to calculate t_{im}, a sample statistic, based on the mth plausible value. It is equal to either a student group mean or a student group percent above achievement level. The variance over plausible values is:

$$U_{\mathrm{G}} = \frac{1}{\mathrm{M}-1} \sum_{m=1}^{\mathrm{M}} (t_{\mathrm{G}}^{m} - \bar{t}_{\mathrm{G}})^2, \tag{11}$$

where U_G is the measurement variance, $t_{\mathrm{G}}^{\mathrm{m}}$ is the value of the statistic over all examinees in group G for plausible value m, and $\bar{t}_{\mathrm{G}}$ is the mean value of the statistic averaged over plausible values.

Sampling variance

The procedure used to estimate sampling variance followed operational NAEP procedures. Typically, schools are grouped into 2P primary sampling units (PSUs). These are stratified into P pairs of PSUs, where the PSUs within a pair are similar on various SES measures. The procedure of the jackknife is to work through the P pairs one by one. Each time a PSU pair is selected, a single PSU is dropped from the pair, the data are suitably reweighted, and an estimated sample statistic (called a pseudoestimate), $t_{\mathrm{G}}^{\mathrm{p}}$, is calculated on the remaining sample. In the present case, this statistic is a group mean. This process is followed till a series of P sample statistics is estimated, $t_{\mathrm{G}}^{1}, t_{\mathrm{G}}^{2}, \ldots, t_{\mathrm{G}}^{\mathrm{P}}$. The sampling variance is calculated as

$$V_{\mathrm{G}} = \sum_{p=1}^{\mathrm{P}} (t_{\mathrm{G}}^{p} - \bar{t}_{\mathrm{G}})^2, \tag{12}$$

where $\bar{t}_{\mathrm{G}}$ is the average statistic over P pseudoestimates.

Note that the proper estimate of V_G is the average of the estimate calculated over the k set of plausible values. Practice in NAEP has shown that using an estimate based on one plausible value is sufficiently accurate.

Standard errors

The total variance of a sample statistic is a weighted combination of measurement and sampling variances (Mislevy 1991). As a result, the standard error for a sample statistic for group *G* is

$$SE_G = \sqrt{V_G + (1+\frac{1}{M})U_G}, \tag{13}$$

where M is the number of plausible values (Allen, Carlson, and Zelenak 1999, section 12.4.1).

Creation of a Real-Time Inference Engine for the Search Scenario

As part of the demonstration of the feasibility of delivering an assessment that uses the full potential of the computer, a Bayesian inference engine for the Search scenario was developed. A Bayesian inference engine is a system of variables like those depicted in figures F-1 and F-3. It is assumed that beliefs about the system, i.e., the conditional probability of any variable given the values of any precursor (parent) variables, can be defined. These conditional probabilities may come from the judgments of experts or from parameters estimated from the Bayesian analysis of data (as is the case with the present research). The goal of using an inference engine is to be able to estimate the probability distribution of any variable in the system given the observed or hypothesized value of any other variables in the system. On one hand, there is interest in being able to score an examinee; that is, given that a certain pattern of responses on the observables is obtained, it is desirable to estimate the distribution of the latent variables. On the other hand, there might be interest, given a certain profile of scores on the latent variables, in gauging the sensitivity of the model by estimating the probability of responding correctly on the observables.

Estimating probabilities in an inference engine is not straightforward. This is because often some variables in a network are not conditionally independent. As a result, information about observed values of variables may be redundantly accounted for when updating the system. To avoid such overcounting of evidence, a Bayes net has to be transformed into a structure that can propagate information throughout the network without redundancy. To accomplish this, a directed graph (such as the ones in figures F-1 and F-3) and conditional probabilities are translated into a linear inference tree, or clique tree. For details, see Lauritzen and Spiegelhalter (1988) and Pearl (1988). To make calculations in such a system tractable, all variables have to be defined as categorical. A program package called ERGO (Noetic Systems, Inc. 2001) automatically accomplishes the task of compiling a Bayes net into a linear inference tree.

There were several steps in defining an inference engine from the results of the Bayesian MCMC analysis.

1. Point estimates for all model parameters had to be extracted from the MCMC estimation.
2. The estimated sample distributions of the latent proficiency variables had to be made discrete. This was done by partitioning the distribution into 15 equal-probability regions. The values associated with these were the inverse normal probability functions of the midpoints.
3. Conditional probability tables that represent the relationship between the variables had to be constructed. The structural relations between latent proficiencies are represented with a normal translation model (Almond forthcoming), where the discrete values of the child variable are a linear function of the parent variable. This representation reflects the structural regression estimated in the MCMC phase. For the observables, the conditional probabilities of each observable are a function of the parent latent proficiency. This procedure employs an IRT model using item parameters from the Bayesian estimation.
4. The conditional probability tables were then imported into the ERGO program and compiled into a linear inference tree.

With the inference engine, it was possible to input profiles of latent proficiencies and see what probabilities of response resulted for the observables. For example, if a high level of computer skills was stipulated, there should be a high probability of a high score on all of the computer observables.

The inference engine was confirmed with the MCMC algorithm. This was done in the following way. The data were augmented by a few dozen dummy cases which had profiles of latent proficiencies fixed. This data set, which included some 1,100 *real* cases, was input into a run of the MCMC estimation program.[3] Average response probabilities of the observables corresponding to the dummy cases were then estimated. In a parallel analysis, the same profiles of latent proficiencies were input into the

[3] The n of ~1,100 was the number of students responding to the TRE Search scenario. This sample size was based on the minimum assumed for scaling in main NAEP and for detecting mean differences among reporting groups of interest.

inference engine, and the resulting response probabilities for observables were noted. It was found that the response probabilities derived from the MCMC algorithm almost exactly matched with those derived from the inference engine.

The ultimate utility of such a Bayes net would be to score results immediately from a computer-delivered assessment. It could also be part of a tailored test, in which the interim proficiency estimates would be used as a basis for deciding how to branch the assessment to more or less challenging activities.

In the current research, the inference engine provided a proof of concept for an approach to Bayesian IRT estimation. In an assessment using an inference engine, the model to estimate parameters from data could involve continuous latent-proficiency variables. It has been demonstrated that parameters from such a model can be translated into a discrete system.

Appendix G: C-rater Rules for Scoring Students’ Search Queries

Terms are assigned to the following seven categories:

1. *Comparative terms:* better, advantages, disadvantages, prefer, more, over, worse
2. *Relevant terms:* weather, atmosphere, space, outer space, cost, helium, science, scientist, astronomer, astronomy, astrophysics, NASA, study, research, explore, learn, experiment
3. *Tool terms:* satellite, rocket, telescope, space shuttle
4. *Weak balloon terms:* balloon, air balloon, hot air balloon
5. *Good balloon terms:* gas balloon, helium balloon, helium gas balloon, weather balloon
6. *Special balloon terms:* scientific balloon, scientific gas balloon, scientific helium balloon, super pressure balloon, long duration balloon, zero pressure balloon
7. *Explore terms:* study, research, explore, learn, experiment

Scoring rules (numbers represent categories):

SCORE = 2

1. 1 & 3 & 4
2. 1 & 2 & 7
3. 1 & 3 & 7
4. 2 & 5
5. 3 & 5
6. 6
7. 2 & 3 & 4
8. 4 & 2 (at least two from 2)
9. 4 & 3 (at least two from 3)

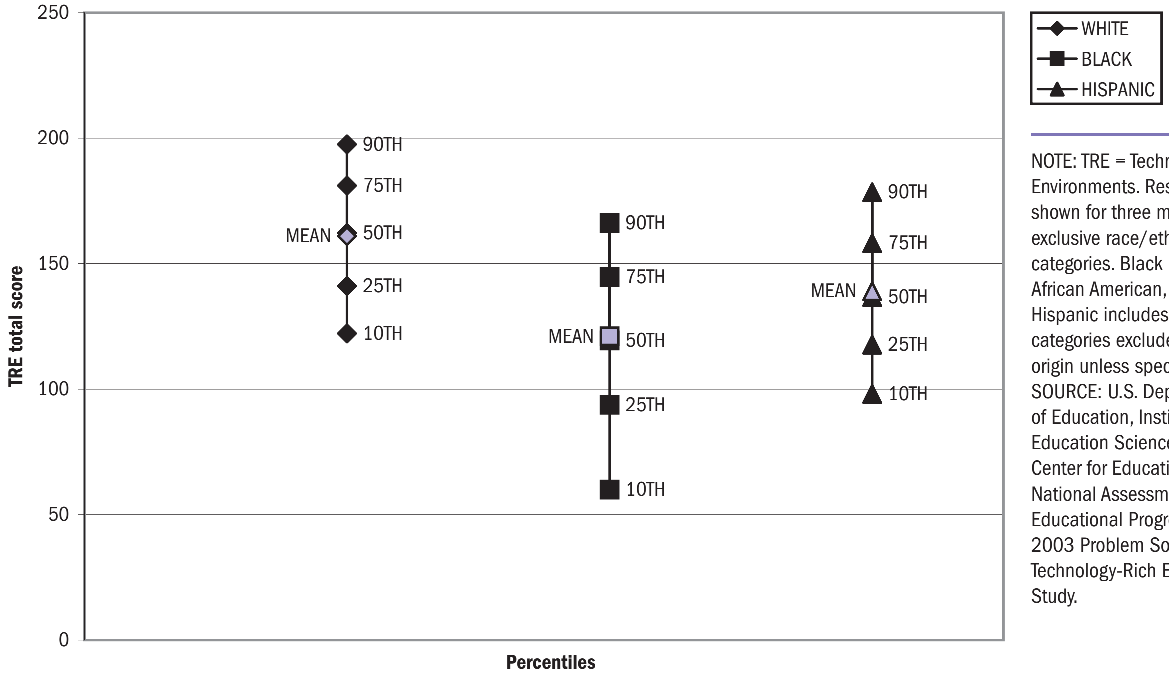
SCORE = 1

10. 2 & 3
11. 2 & 4
12. 3 & 4
13. 5

SCORE = 0 if no rules are met.

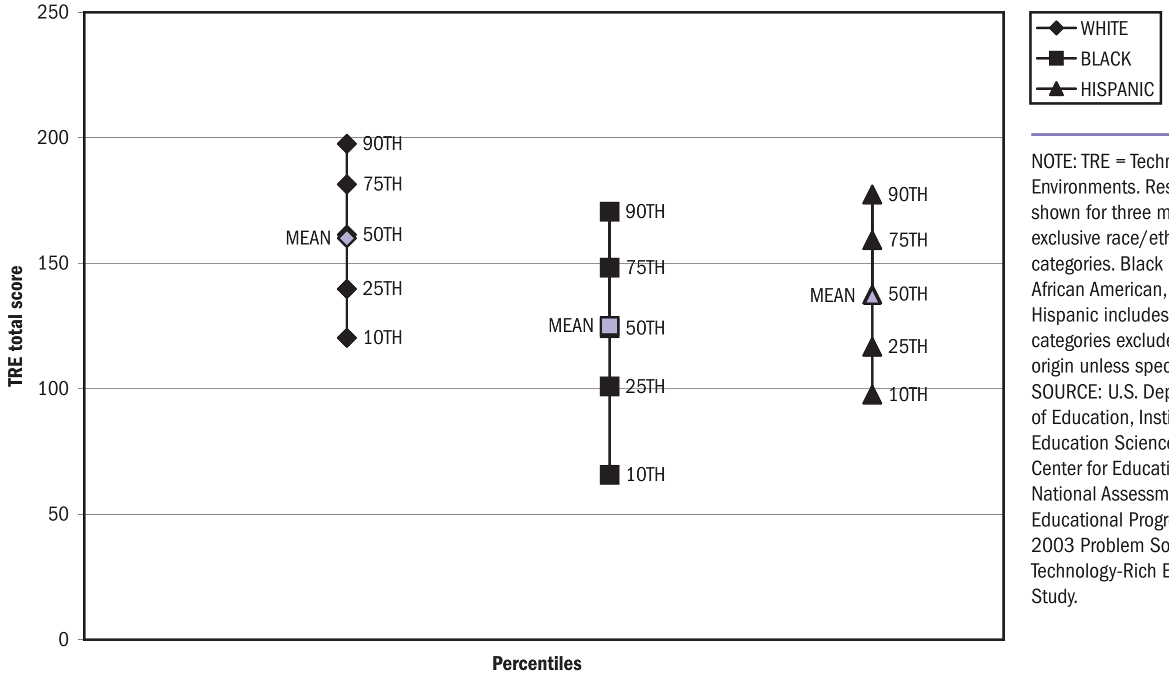
Appendix H: TRE Search and Simulation Scale Scores and Percentiles by Student Reporting Groups for Scales on Which Statistically Significant Group Differences Were Observed

Figure H-1. TRE Search total score distribution, by race/ethnicity, grade 8: 2003



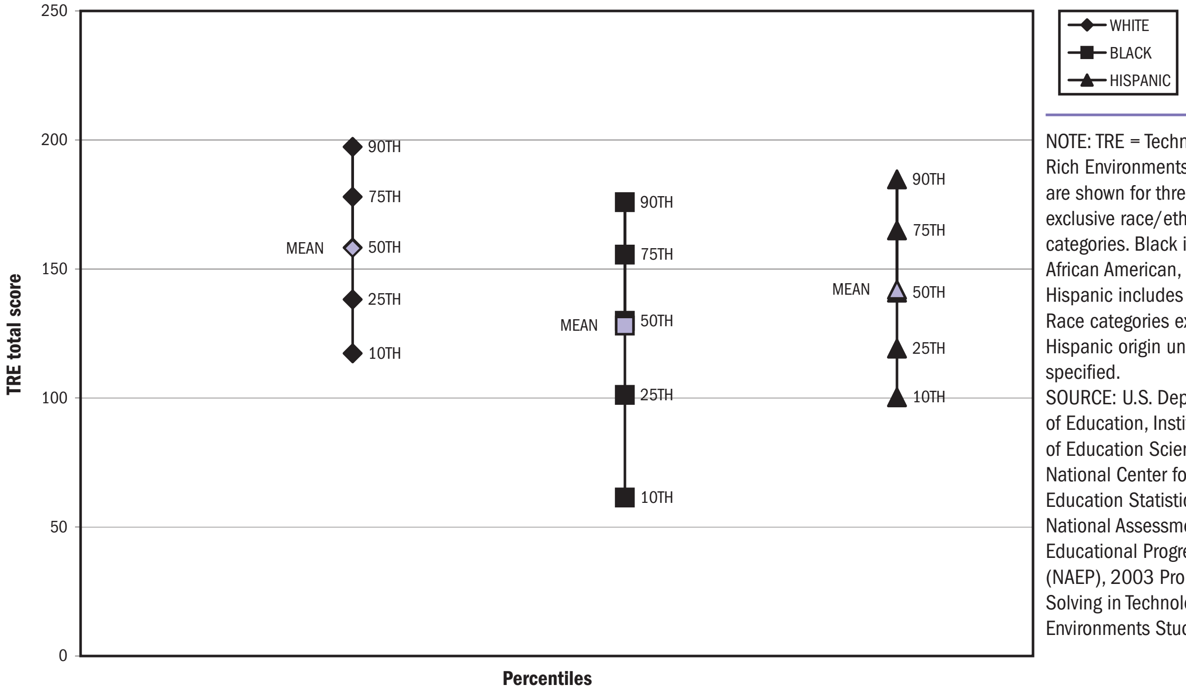
NOTE: TRE = Technology-Rich Environments. Results are shown for three mutually exclusive race/ethnicity categories. Black includes African American, and Hispanic includes Latino. Race categories exclude Hispanic origin unless specified.
SOURCE: U.S. Department of Education, Institute of Education Sciences,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

Figure H-2. TRE Search scientific inquiry skill score distribution, by race/ethnicity, grade 8: 2003



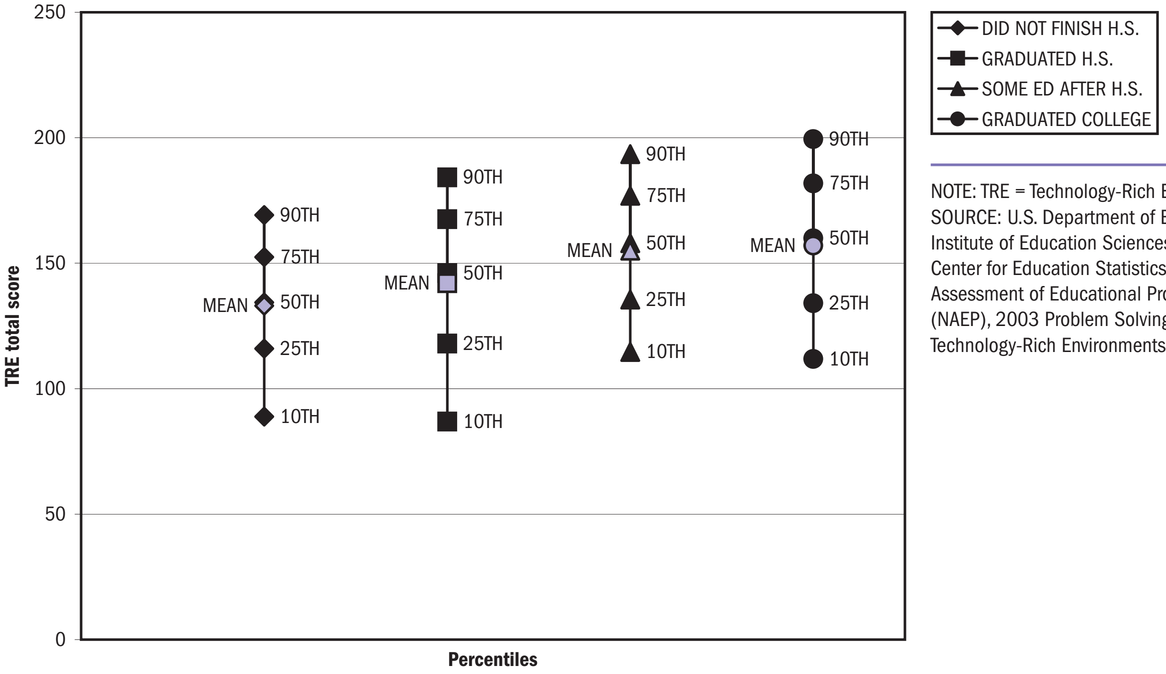
NOTE: TRE = Technology-Rich Environments. Results are shown for three mutually exclusive race/ethnicity categories. Black includes African American, and Hispanic includes Latino. Race categories exclude Hispanic origin unless specified.
SOURCE: U.S. Department of Education, Institute of Education Sciences,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

Figure H-3. TRE Search computer skills score distribution, by race/ethnicity, grade 8: 2003



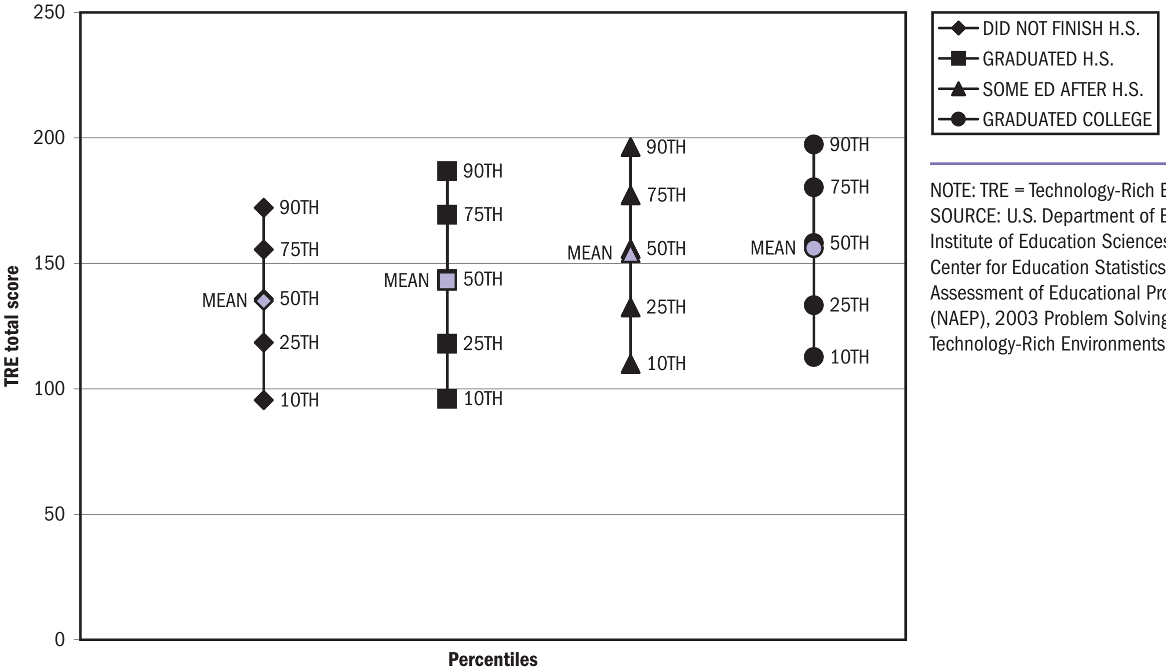
NOTE: TRE = Technology-Rich Environments. Results are shown for three mutually exclusive race/ethnicity categories. Black includes African American, and Hispanic includes Latino. Race categories exclude Hispanic origin unless specified.
SOURCE: U.S. Department of Education, Institute of Education Sciences,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

Figure H-4. TRE Search total score distribution, by student-reported parents' highest education level, grade 8: 2003



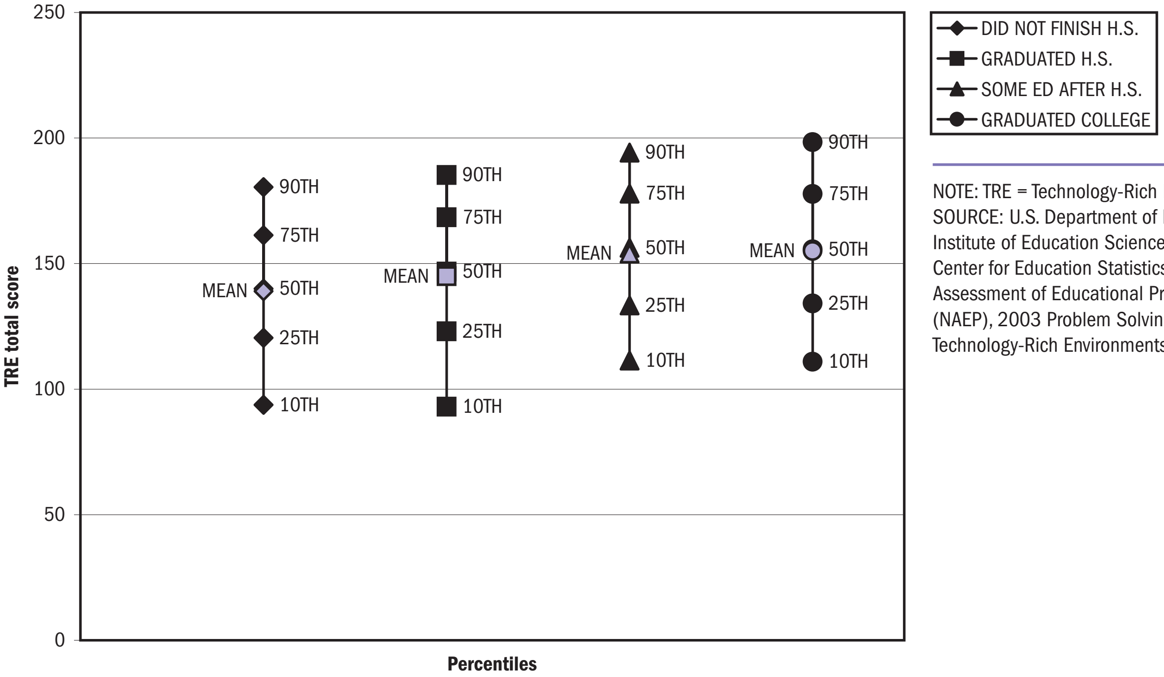
NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure H-5. TRE Search scientific inquiry skill score distribution, by student-reported parents' highest education level, grade 8: 2003



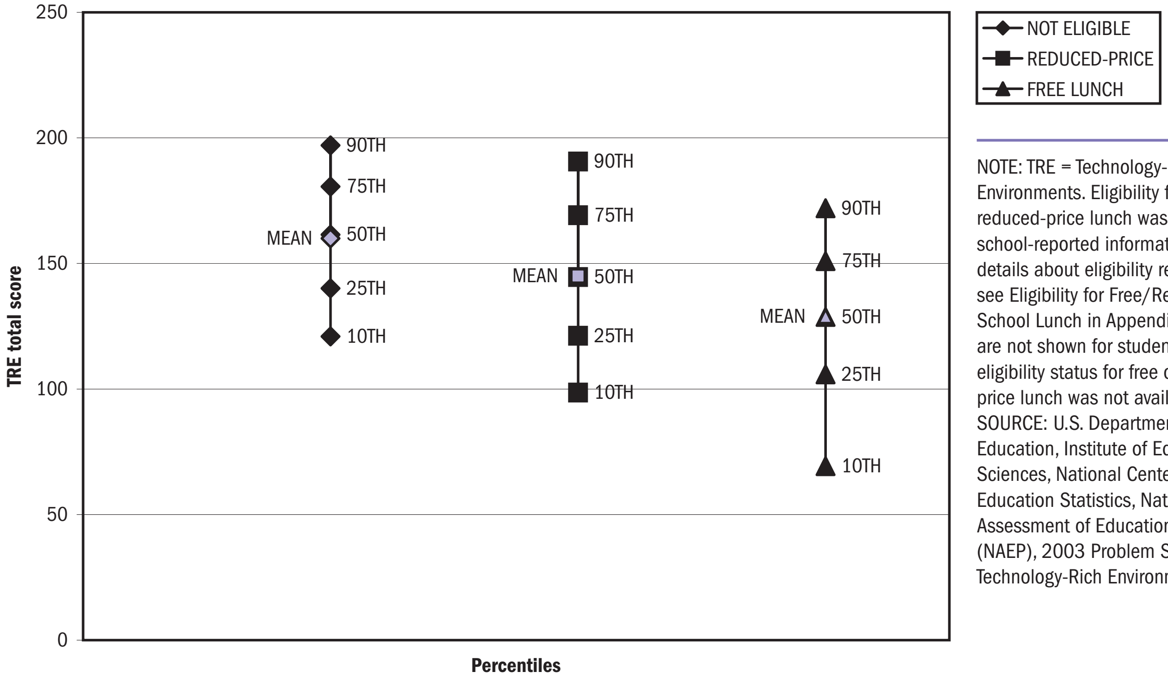
NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure H-6. TRE Search computer skills score distribution, by student-reported parents' highest education level, grade 8: 2003



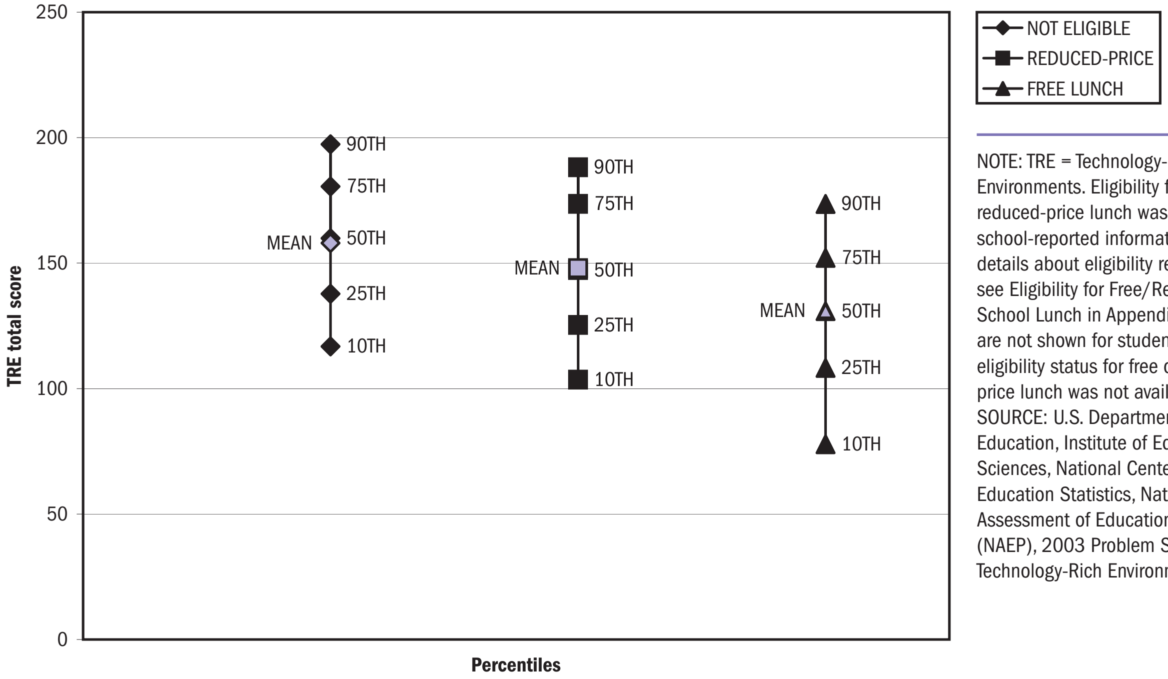
NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure H-7. TRE Search total score distribution, by eligibility for free or reduced-price school lunch, grade 8: 2003



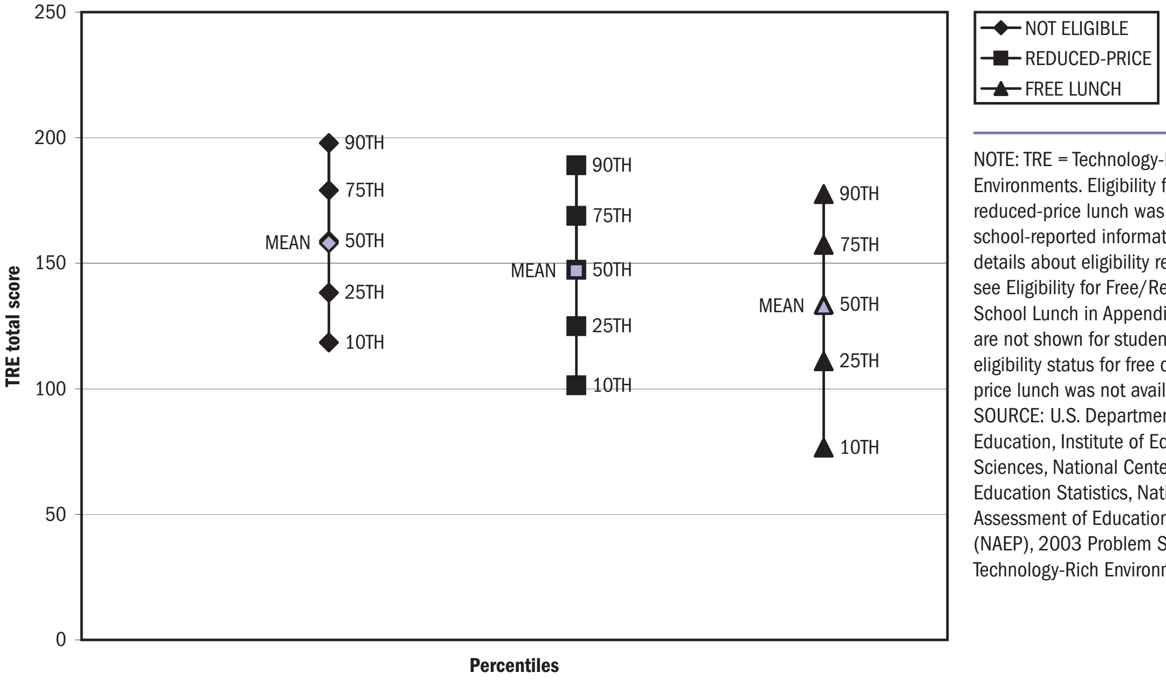
NOTE: TRE = Technology-Rich Environments. Eligibility for free or reduced-price lunch was based on school-reported information. For details about eligibility requirements, see Eligibility for Free/Reduced-Price School Lunch in Appendix K. Results are not shown for students whose eligibility status for free or reduced-price lunch was not available.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure H-8. TRE Search scientific inquiry skill score distribution, by eligibility for free or reduced-price school lunch, grade 8: 2003



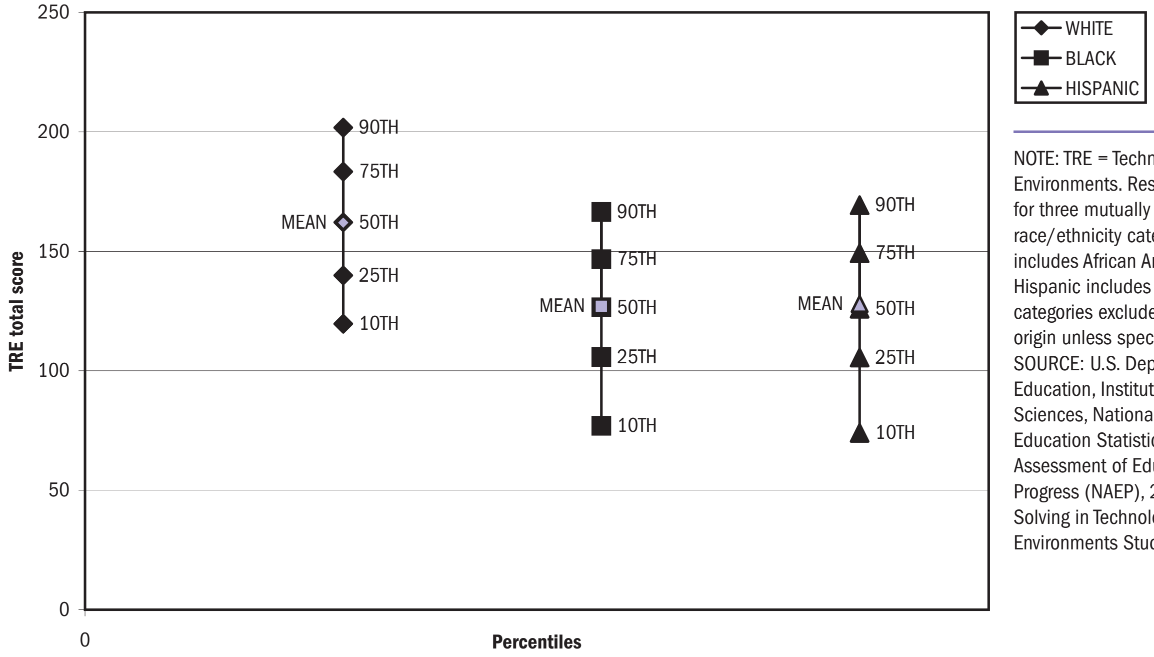
NOTE: TRE = Technology-Rich Environments. Eligibility for free or reduced-price lunch was based on school-reported information. For details about eligibility requirements, see Eligibility for Free/Reduced-Price School Lunch in Appendix K. Results are not shown for students whose eligibility status for free or reduced-price lunch was not available.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure H-9. TRE Search computer skills score distribution, by eligibility for free or reduced-price school lunch, grade 8: 2003



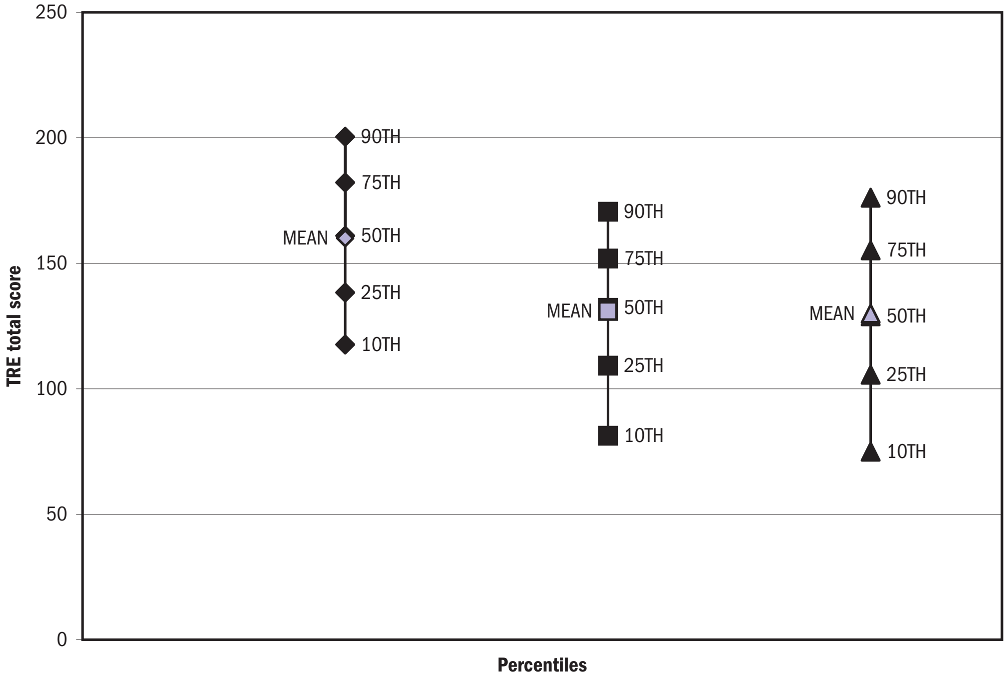
NOTE: TRE = Technology-Rich Environments. Eligibility for free or reduced-price lunch was based on school-reported information. For details about eligibility requirements, see Eligibility for Free/Reduced-Price School Lunch in Appendix K. Results are not shown for students whose eligibility status for free or reduced-price lunch was not available.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure H-10. TRE Simulation total score distribution, by race/ethnicity, grade 8: 2003



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

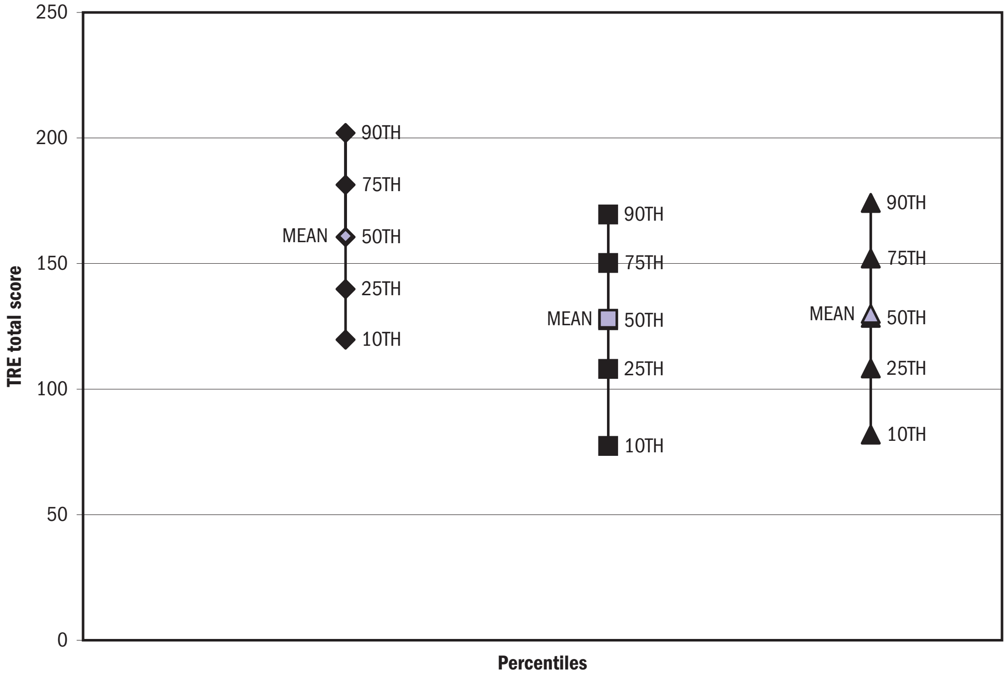
Figure H-11. TRE Simulation scientific exploration skill score distribution, by race/ethnicity, grade 8: 2003



NOTE: TRE = Technology-Rich Environments. Results are shown for three mutually exclusive race/ethnicity categories. Black includes African American, and Hispanic includes Latino. Race categories exclude Hispanic origin unless specified.

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

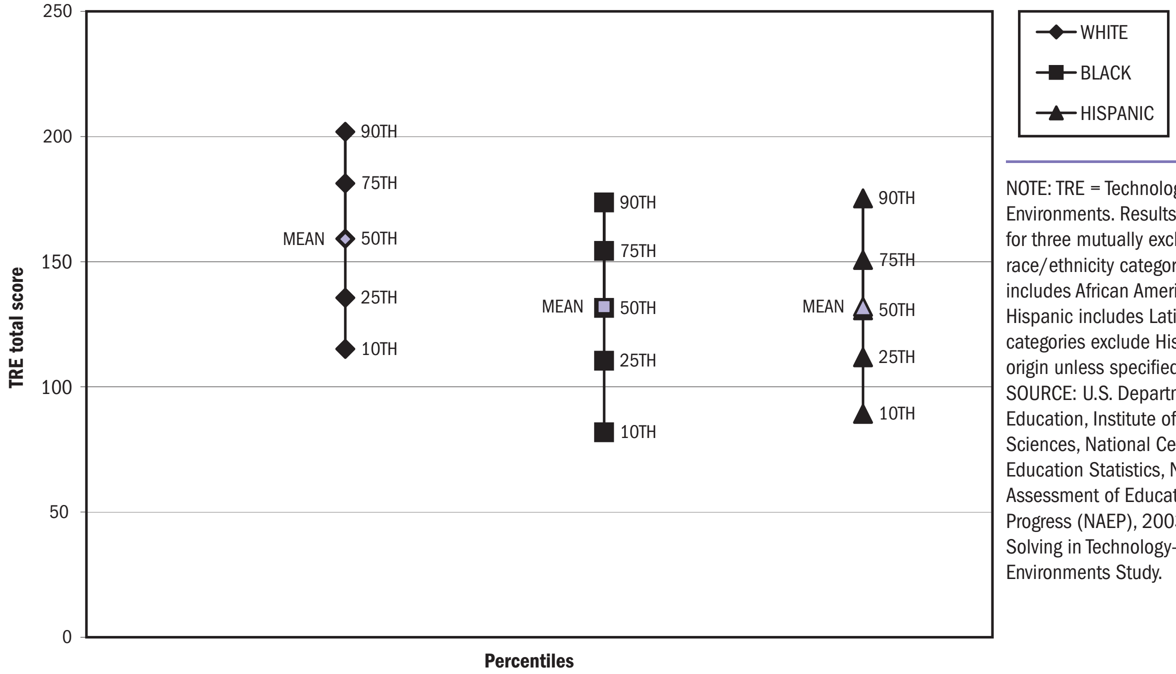
Figure H-12. TRE Simulation scientific synthesis score distribution, by race/ethnicity, grade 8: 2003



NOTE: TRE = Technology-Rich Environments. Results are shown for three mutually exclusive race/ethnicity categories. Black includes African American, and Hispanic includes Latino. Race categories exclude Hispanic origin unless specified.

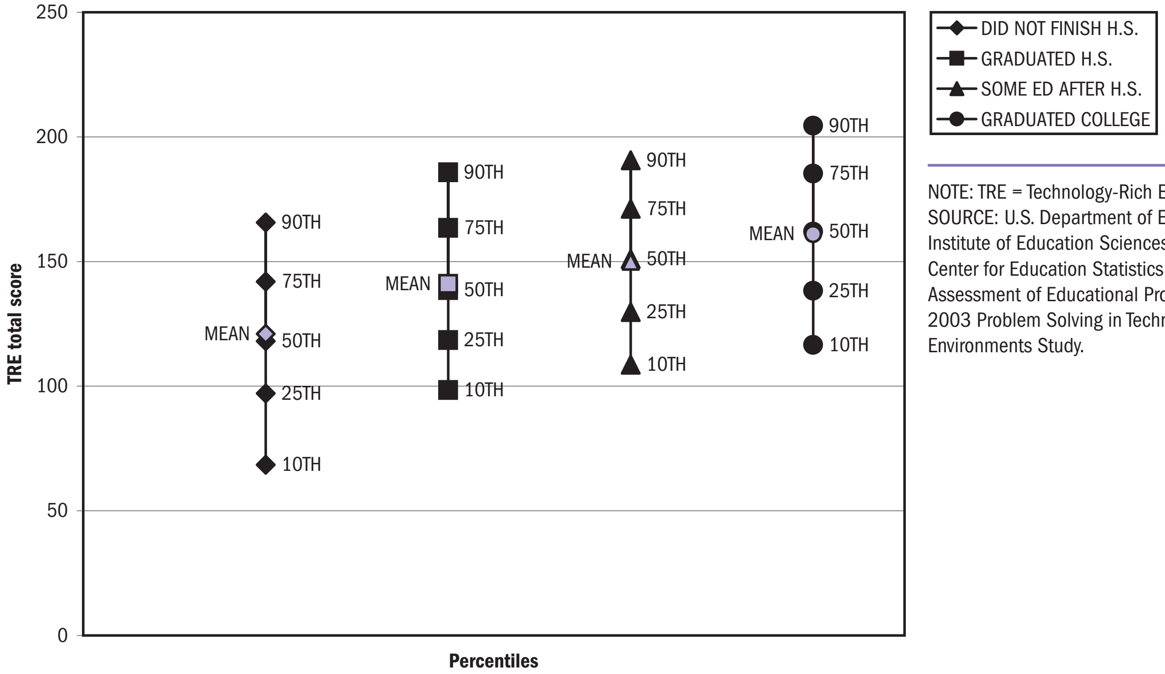
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure H-13. TRE Simulation computer skills score distribution, by race/ethnicity, grade 8: 2003



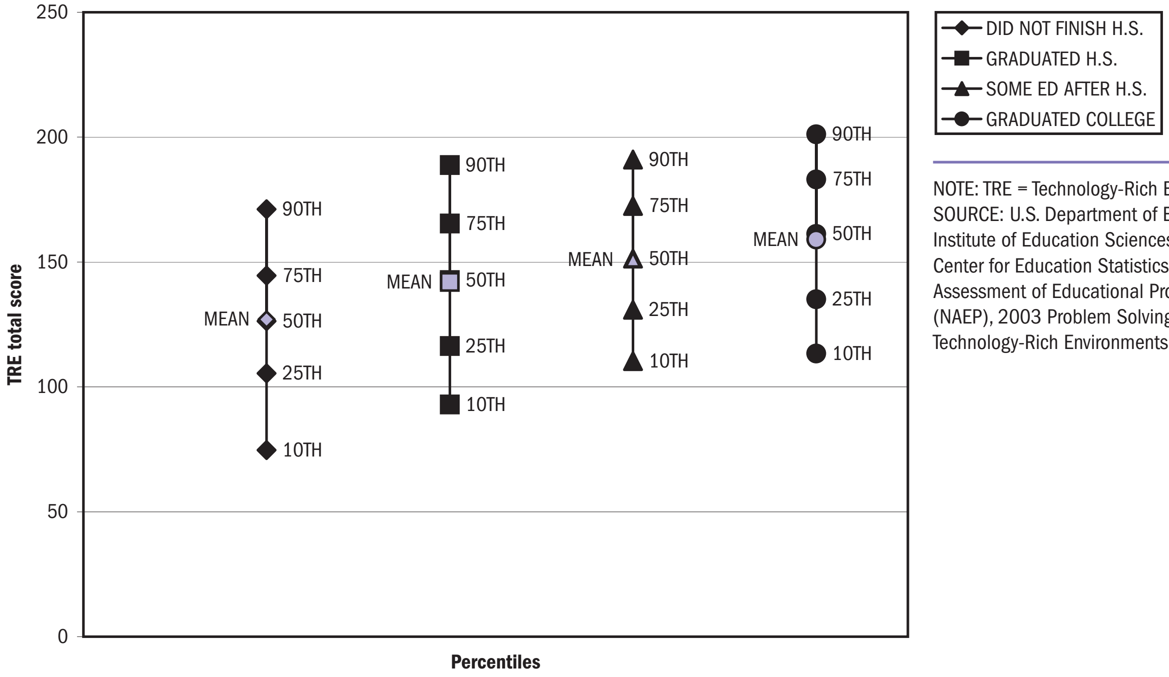
NOTE: TRE = Technology-Rich Environments. Results are shown for three mutually exclusive race/ethnicity categories. Black includes African American, and Hispanic includes Latino. Race categories exclude Hispanic origin unless specified.
SOURCE: U.S. Department of Education, Institute of Education Sciences,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

Figure H-14. TRE Simulation total score distribution, by student-reported parents' highest education level, grade 8: 2003



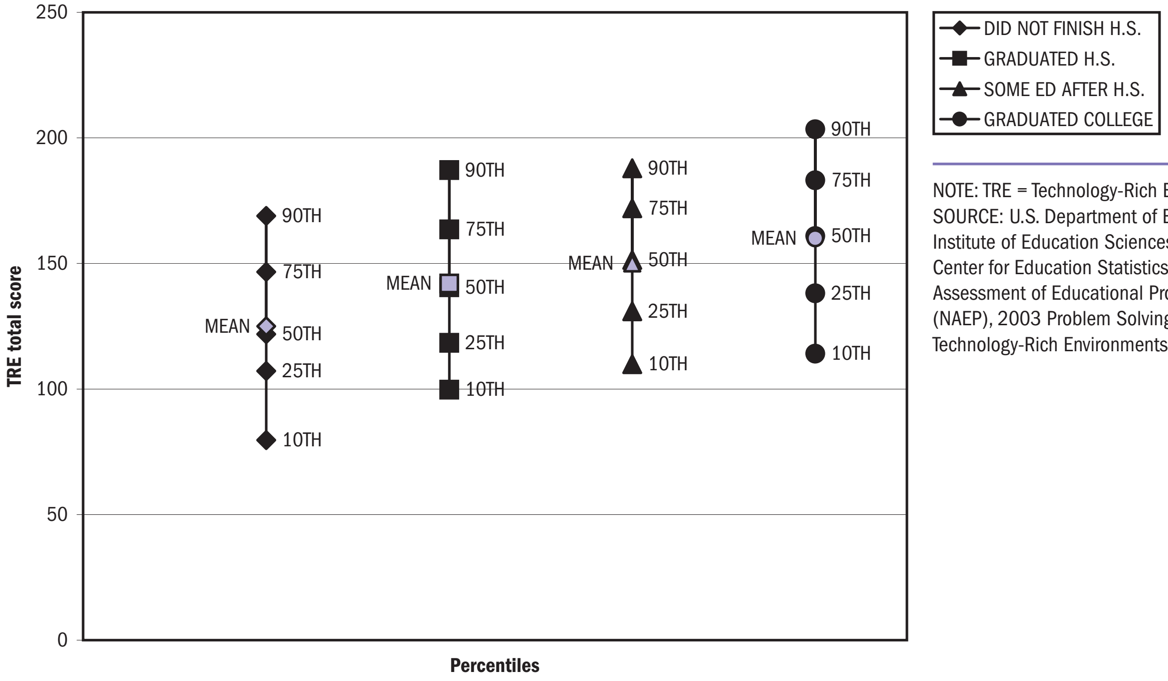
NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure H-15. TRE Simulation scientific exploration skill score distribution, by student-reported parents' highest education level, grade 8: 2003



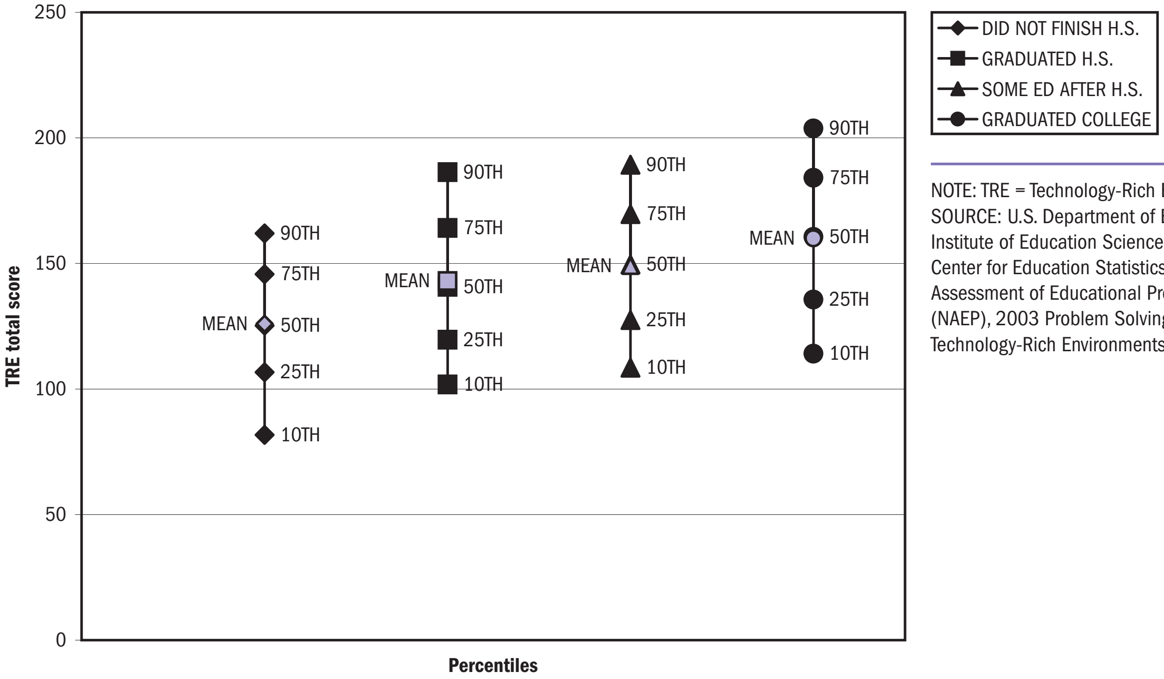
NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure H-16. TRE Simulation scientific synthesis score distribution, by student-reported parents' highest education level, grade 8: 2003



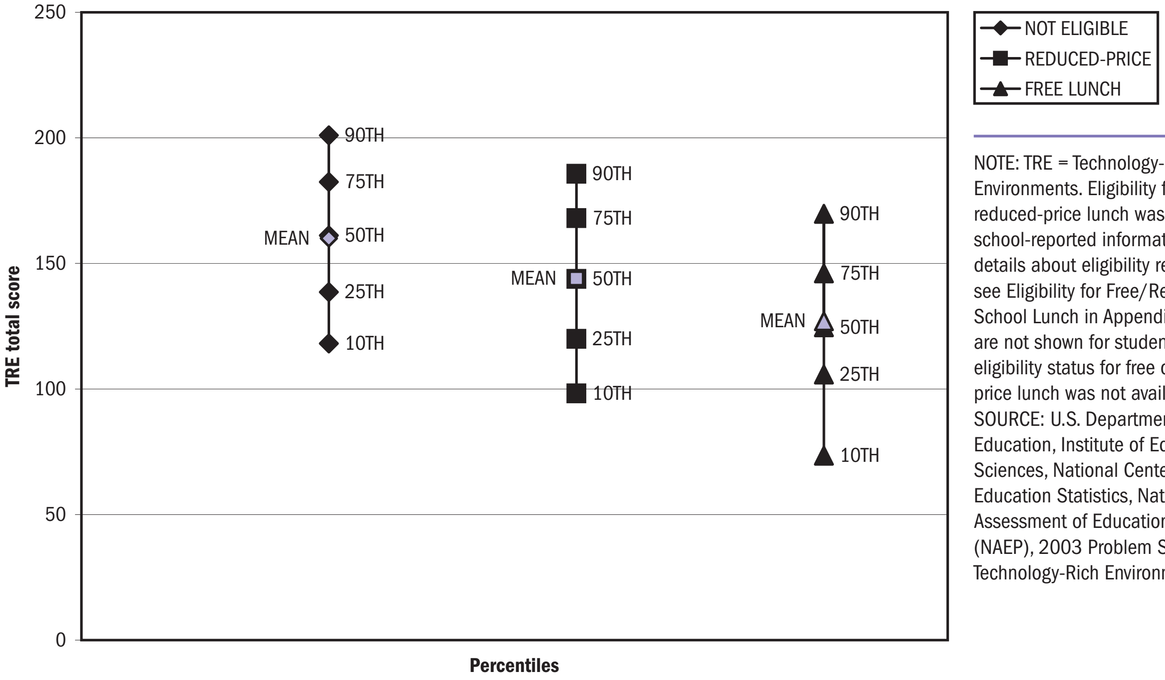
NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure H-17. TRE Simulation computer skills score distribution, by student-reported parents' highest education level, grade 8: 2003.



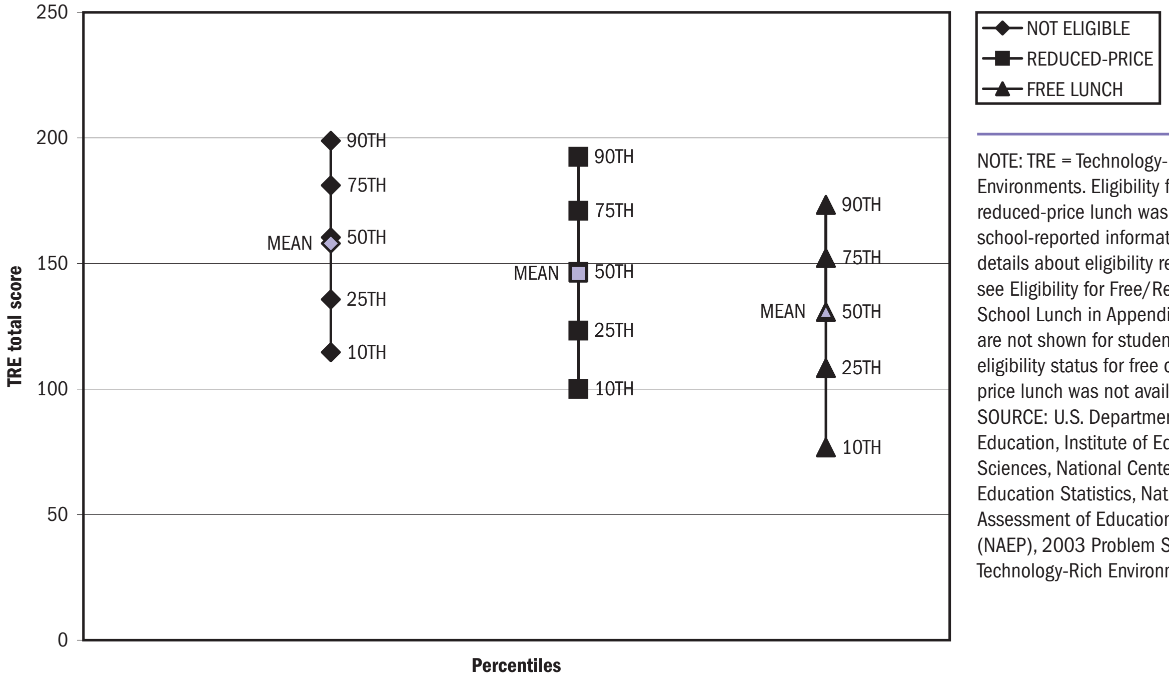
NOTE: TRE = Technology-Rich Environments.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Figure H-18. TRE Simulation total score distribution, by eligibility for free or reduced-price school lunch, grade 8: 2003



NOTE: TRE = Technology-Rich Environments. Eligibility for free or reduced-price lunch was based on school-reported information. For details about eligibility requirements, see Eligibility for Free/Reduced-Price School Lunch in Appendix K. Results are not shown for students whose eligibility status for free or reduced-price lunch was not available.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

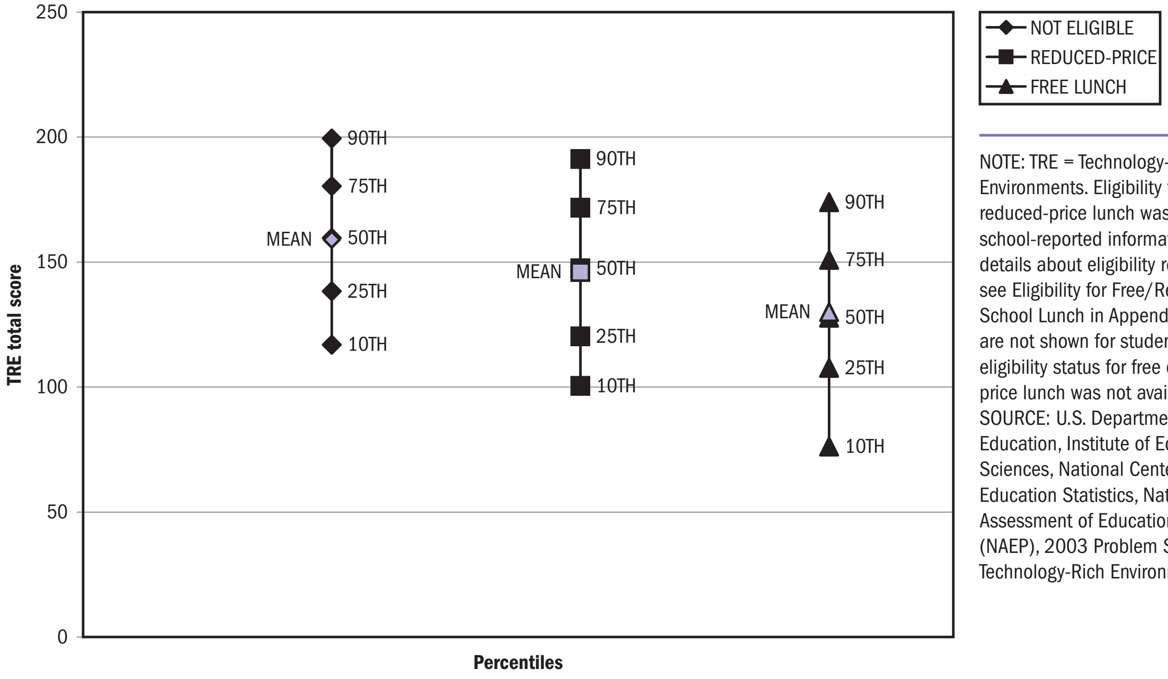
Figure H-19. TRE Simulation scientific exploration skill score distribution, by eligibility for free or reduced-price school lunch, grade 8: 2003



NOTE: TRE = Technology-Rich Environments. Eligibility for free or reduced-price lunch was based on school-reported information. For details about eligibility requirements, see Eligibility for Free/Reduced-Price School Lunch in Appendix K. Results are not shown for students whose eligibility status for free or reduced-price lunch was not available.

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

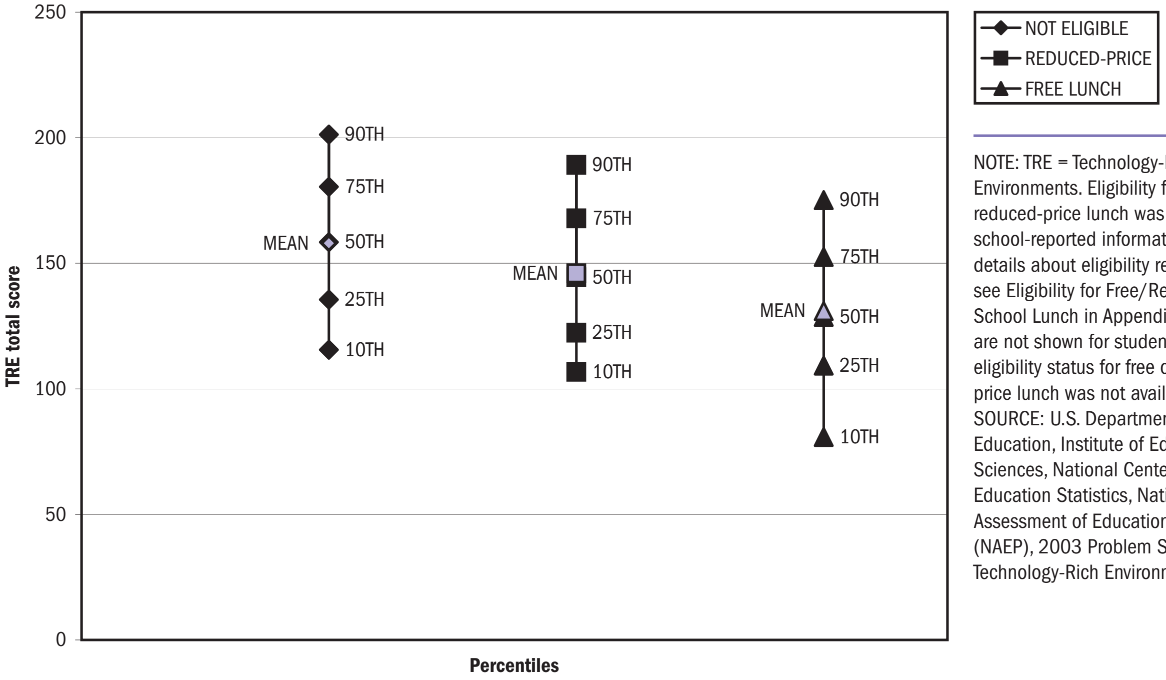
Figure H-20. TRE Simulation scientific synthesis score distribution, by eligibility for free or reduced-price school lunch, grade 8: 2003



NOTE: TRE = Technology-Rich Environments. Eligibility for free or reduced-price lunch was based on school-reported information. For details about eligibility requirements, see Eligibility for Free/Reduced-Price School Lunch in Appendix K. Results are not shown for students whose eligibility status for free or reduced-price lunch was not available.

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

Figure H-21. TRE Simulation computer skills score distribution, by eligibility for free or reduced-price school lunch, grade 8: 2003



NOTE: TRE = Technology-Rich Environments. Eligibility for free or reduced-price lunch was based on school-reported information. For details about eligibility requirements, see Eligibility for Free/Reduced-Price School Lunch in Appendix K. Results are not shown for students whose eligibility status for free or reduced-price lunch was not available.

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

Appendix I: Summary Statistics for Prior Knowledge Measures and Mean Scale Scores for Background-Question Response Options[1]

Table I-1. Unweighted summary statistics for Search scenario prior knowledge measures, grade 8: 2003

Statistic	Prior computer knowledge	Prior science knowledge
Number of students	1,059	1,062
Mean score	5.6	5.0
Standard deviation	2.1	1.8
Scale range	0–10	0–10
Coefficient alpha reliability	.58	.39

NOTE: Students' scores for a particular prior knowledge measure were deleted from this analysis if they did not answer all 10 questions in a scale.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Table I-2. Unweighted summary statistics for Simulation scenario prior knowledge measures, grade 8: 2003

Statistic	Prior computer knowledge	Prior science knowledge
Number of students	960	986
Mean score	5.5	5.3
Standard deviation	2.0	2.4
Scale range	0–10	0–10
Coefficient alpha reliability	.51	.67

NOTE: Students' scores for a particular prior knowledge measure were deleted from this analysis if they did not answer all 10 questions in a scale.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

[1] The items composing the Prior Computer Knowledge measure were the same for the Search and Simulation scenarios. For the Prior Science Knowledge measure, different items were used for each scenario.

Table I-3. Data for figure 5-3, mean scale scores, by extent of specific computer use and scale for Search scenario, grade 8: 2003

	Use a word processor			
Scale	Not at all	Small extent	Moderate extent	Large extent
Search total score	130 (4.1)	145 (2.8)	153 (2.2)	157 (2.5)
Search scientific inquiry score	132 (3.9)	145 (3.4)	153 (2.7)	156 (2.3)
Search computer skills score	133 (3.9)	146 (2.8)	151 (2.4)	159 (2.6)
	Make drawings/art on computer			
Scale	Not at all	Small extent	Moderate extent	Large extent
Search total score	151 (3.3)	152 (2.2)	149 (2.7)	138 (3.7)
Search scientific inquiry score	151 (3.2)	152 (2.2)	149 (3.3)	137 (4.0)
Search computer skills score	151 (2.4)	151 (2.3)	151 (2.5)	139 (4.2)
	Make tables, charts, or graphs on computer			
Scale	Not at all	Small extent	Moderate extent	Large extent
Search total score	145 (2.8)	155 (2.1)	150 (3.4)	134 (5.8)
Search scientific inquiry score	146 (2.9)	154 (2.7)	149 (2.8)	136 (5.8)
Search computer skills score	145 (2.8)	154 (1.8)	151 (3.7)	137 (5.6)
	Look up information on a CD			
Scale	Not at all	Small extent	Moderate extent	Large extent
Search total score	148 (2.8)	154 (2.7)	152 (2.7)	141 (3.4)
Search scientific inquiry score	149 (3.0)	154 (3.2)	151 (3.1)	143 (3.0)
Search computer skills score	148 (3.2)	153 (2.5)	152 (2.5)	144 (3.1)
	Find information on the Internet			
Scale	Not at all	Small extent	Moderate extent	Large extent
Search total score	‡	136 (3.8)	149 (2.7)	154 (2.2)
Search scientific inquiry score	‡	137 (4.4)	150 (3.4)	153 (2.3)
Search computer skills score	‡	134 (4.0)	149 (2.6)	154 (2.5)
	Use e-mail			
Scale	Not at all	Small extent	Moderate extent	Large extent
Search total score	138 (3.1)	146 (3.1)	151 (3.8)	156 (2.2)
Search scientific inquiry score	139 (3.8)	147 (3.7)	152 (2.6)	155 (2.2)
Search computer skills score	141 (3.3)	145 (3.0)	151 (2.7)	155 (2.1)
	Talk in chat groups			
Scale	Not at all	Small extent	Moderate extent	Large extent
Search total score	142 (2.6)	147 (3.6)	149 (3.3)	157 (2.3)
Search scientific inquiry score	143 (3.4)	147 (2.9)	149 (2.5)	156 (2.6)
Search computer skills score	143 (2.8)	147 (3.4)	149 (3.3)	157 (2.0)

‡ Reporting standards not met. Sample size was insufficient to permit a reliable estimate.

NOTE: The range of scores for each scale is 0–300. Standard errors of the estimated scores appear in parentheses.

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

Table I-4. Data for figure 5-4, mean scale scores, by frequency of computer use and scale for Search scenario, grade 8: 2003

	How often do you use a computer outside of school?				
Scale	Daily	2–3 times per week	Once a week	Once every few weeks	Never or hardly ever
Search total score	158 (2.4)	146 (2.2)	147 (3.6)	130 (5.8)	126 (5.1)
Search scientific inquiry score	157 (2.3)	147 (2.0)	147 (3.7)	131 (6.1)	129 (4.5)
Search computer skills score	157 (2.1)	148 (2.4)	147 (3.8)	129 (4.7)	131 (3.1)

NOTE: The range of scores for each scale is 0–300. Standard errors of the estimated scores appear in parentheses.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Table I-5. Data for figure 5-5, mean scale scores, by students indicating there is a computer at home that they use and scale for Search scenario, grade 8: 2003

	Is there a computer at home that you use?	
Scale	Yes	No
Search total score	153 (1.9)	125 (3.4)
Search scientific inquiry score	152 (1.9)	129 (3.3)
Search computer skills score	152 (1.9)	131 (3.5)

NOTE: The range of scores for each scale is 0–300. Standard errors of the estimated scores appear in parentheses.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Table I-6. Data for figure 5-6, mean scale scores, by frequency of school science activity and scale for Search scenario, grade 8: 2003

	Use the Internet to exchange information with other students or scientists about experiments			
Scale	Not taking science	Once a month or more	Sometimes, but less than once a month	Never
Search total score	‡	146 (3.3)	145 (3.5)	154 (2.2)
Search scientific inquiry score	‡	145 (3.4)	144 (3.3)	154 (1.8)
Search computer skills score	‡	147 (2.7)	147 (3.1)	153 (2.0)

‡ Reporting standards not met. Sample size was insufficient to permit a reliable estimate.
NOTE: The range of scores for each scale is 0–300. Standard errors of the estimated scores appear in parentheses.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Table I-7. Data for figure 6-4, mean scale scores, by extent of specific computer use and scale for Simulation scenario, grade 8: 2003

	Play computer games			
Scale	Not at all	Small extent	Moderate extent	Large extent
Simulation total score	140 (5.8)	149 (3.1)	153 (2.6)	152 (3.2)
Simulation scientific exploration score	137 (4.9)	149 (2.7)	153 (2.4)	154 (3.7)
Simulation scientific synthesis score	141 (4.8)	148 (3.4)	153 (2.2)	151 (3.3)
Simulation computer skills score	143 (6.0)	150 (3.7)	152 (3.7)	148 (4.0)

	Use a word processor			
Scale	Not at all	Small extent	Moderate extent	Large extent
Simulation total score	121 (4.3)	140 (3.6)	153 (2.6)	163 (2.7)
Simulation scientific exploration score	125 (5.3)	141 (4.0)	153 (2.3)	161 (2.3)
Simulation scientific synthesis score	124 (4.2)	141 (3.8)	153 (2.5)	161 (2.0)
Simulation computer skills score	123 (4.4)	138 (4.5)	152 (3.2)	165 (4.4)

	Make tables, charts, or graphs on computer			
Scale	Not at all	Small extent	Moderate extent	Large extent
Simulation total score	136 (2.9)	157 (2.4)	154 (3.7)	148 (5.3)
Simulation scientific exploration score	138 (3.2)	156 (2.1)	153 (3.4)	147 (5.4)
Simulation scientific synthesis score	136 (3.5)	156 (2.2)	154 (3.1)	149 (5.9)
Simulation computer skills score	135 (3.7)	156 (3.2)	155 (4.9)	151 (6.5)

	Find information on the Internet			
Scale	Not at all	Small extent	Moderate extent	Large extent
Simulation total score	‡	133 (4.4)	147 (3.5)	156 (2.5)
Simulation scientific exploration score	‡	137 (3.7)	147 (3.3)	155 (2.2)
Simulation scientific synthesis score	‡	136 (4.5)	147 (3.1)	155 (2.2)
Simulation computer skills score	‡	131 (4.4)	148 (4.0)	156 (3.6)

‡ Reporting standards not met. Sample size was insufficient to permit a reliable estimate.

NOTE: The range of scores for each scale is 0–300. Standard errors of the estimated scores appear in parentheses.

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

Table I-8. Data for figure 6-5, mean scale scores, by frequency of computer use and scale for Simulation scenario, grade 8: 2003

	How often do you use a computer outside of school?				
Scale	Daily	2–3 times per week	Once a week	Once every few weeks	Never or hardly ever
Simulation total score	160 (2.1)	147 (2.9)	134 (7.1)	130 (6.1)	118 (3.0)
Simulation scientific exploration score	159 (2.3)	148 (2.5)	136 (7.3)	134 (5.1)	119 (5.3)
Simulation scientific synthesis score	159 (2.0)	148 (2.4)	136 (9.1)	135 (5.9)	119 (2.7)
Simulation computer skills score	159 (3.4)	147 (3.6)	135 (7.3)	134 (6.7)	121 (3.7)

NOTE: The range of scores for each scale is 0–300. Standard errors of the estimated scores appear in parentheses.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Table I-9. Data for figure 6-6, mean scale scores, by students indicating there is a computer at home that they use and scale for Simulation scenario, grade 8: 2003

	Is there a computer at home that you use?	
Scale	Yes	No
Simulation total score	154 (2.1)	123 (4.4)
Simulation scientific exploration score	154 (1.8)	125 (5.2)
Simulation scientific synthesis score	154 (2.0)	125 (4.4)
Simulation computer skills score	153 (3.3)	128 (4.7)

NOTE: The range of scores for each scale is 0–300. Standard errors of the estimated scores appear in parentheses.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Appendix J: Performance on Problem Solving in Technology-Rich Environments (TRE) Observables

Table J-1. Weighted percentage of students achieving each level of correctness on each Search scenario scientific inquiry observable in order of first appearance on item map (figure 5-1), grade 8: 2003

Observable and level of correctness	Weighted percent
Correctly answering most, if not all (three or four), of the four multiple-choice items that require web searching.	18
Correctly answering some (one or two) of the four multiple-choice items that require web searching.	64
Correctly answering none of the four multiple-choice items that require web searching.	18
Using search terms that, on average, match those of proficient searchers to at least a moderate degree.	33
Using search terms that, on average, match those of proficient searchers only to a limited degree.	46
Using search terms that, on average, did not match those of proficient searchers.	21
Constructing a response that gives a reasonably complete answer to the motivating Search problem (i.e., three or more advantages of using gas balloons).	15
Constructing a response that only partially answers the motivating Search problem (i.e., giving only one or two advantages of using gas balloons).	35
Constructing a response that fails to answer the motivating Search problem (i.e., giving no advantages of using gas balloons).	43
Did not construct a response.	7
Bookmarking or visiting pages that are, on average, relevant to the question posed.	14
Bookmarking or visiting pages that are, on average, partially relevant to the question posed.	12
Bookmarking or visiting pages that are, on average, irrelevant to the question posed.	36
Did not bookmark, did not visit pages, did not search, or produced otherwise unscorable response for this observable.	38
Producing at least one set of search results with hits that are, on average, relevant to the question posed (i.e., have relevance scores averaging between 3 and 4 on a four-point scale, where a score of 4 denotes the most relevant hits).	1
Producing at least one set of search results with hits that are, on average, partially relevant to the question posed (i.e., have relevance scores averaging between 2 and 3 on a four-point scale, where a score of 4 denotes the most relevant hits).	11
Producing search results with hits that are, on average, irrelevant to the question posed (i.e., have relevance scores below 2 on a four-point scale, where a score of 4 denotes the most relevant hits).	83
Did not run any searches.	5

NOTE: Detail may not sum to totals because of rounding. Evaluation levels for certain observables were collapsed during analysis; hence, not all the levels for these observables shown in this table appear in the item map.

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

Table J-2. Weighted percentage of students achieving each level of correctness on each Search scenario computer skills observable in order of first appearance on item map (figure 5-2), grade 8: 2003

Observable and level of correctness	Weighted percent
Using the Back button frequently (at least five times) to navigate among web pages or from web pages to the search page.	69
Using the Back button occasionally (three or four times) to navigate among web pages or from web pages to the search page.	10
Using the Back button rarely (two times or less) to navigate among web pages or from web pages to the search page.	21
Using hyperlinks frequently (at least 5 times) to explore web pages linked to the page currently being viewed.	55
Using hyperlinks with moderate frequency (3 to 4 times) to explore web pages linked to the page currently being viewed.	11
Using hyperlinks with limited frequency (1 to 2 times) to explore web pages linked to the page currently being viewed.	15
Did not use hyperlinks to explore web pages linked to the page currently being viewed.	20
Using bookmarks with at least moderate frequency (two or more times).	58
Using bookmarks with limited frequency (one time).	13
Did not use bookmarks.	29
Returning relevant results after only a small number of attempts (1–3).	37
Returning relevant results after a moderate number of attempts (4–6).	24
Returning relevant results after many attempts (more than 6) or does not return relevant results at all.	34
Did not attempt any searches.	5
Using advanced search techniques with at least moderate frequency (3 or more searches).	8
Using advanced search techniques with limited frequency (1–2 searches).	24
Did not use advanced search techniques.	68
Using Delete with at least moderate frequency (2 or more times) to remove a page that had been bookmarked.	3
Using Delete with limited frequency (1 time) to remove a page that had been bookmarked.	8
Did not to use Delete to remove a page that had been bookmarked.	89

NOTE: Detail may not sum to totals because of rounding. Evaluation levels for certain observables were collapsed during analysis; hence, not all the levels for these observables shown in this table appear in the item map.

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

Table J-3. Weighted percentage of students achieving each level of correctness on each Simulation scenario scientific exploration observable in order of first appearance on item map (figure 6-1), grade 8: 2003

Observable and level of correctness	Weighted percent
Using the glossary of science terms in Simulation problem 1 with low frequency or never.	80
Using the glossary of science terms in Simulation problem 1 with moderate frequency.	17
Using the glossary of science terms in Simulation problem 1 with high frequency.	2
Did not produce a scorable response for this observable.	1
Creating a table for Simulation problem 2 that includes only the dependent and independent variables germane to the problem.	9
Creating a table for Simulation problem 2 that includes both of the variables germane to solving the problem along with other variables.	19
Creating a table for Simulation problem 2 that either includes one of the variables germane to solving the problem along with experimental data, or both germane variables without data.	17
Creating a table for Simulation problem 2 that does not include either of the variables germane to solving the problem, or includes one germane variable without experimental data.	13
Did not create a table for Simulation problem 2.	42
Controlling for one variable in at least 66 percent of the experiments run for Simulation problem 3.	46
Controlling for one variable in 40 to 65 percent of the experiments run for Simulation problem 3.	9
Controlling for one variable in less than 40 percent of the experiments run for Simulation problem 3.	3
Running an insufficient number of experiments for controlled experimentation to be evaluated for Simulation problem 3.	40
Did not produce scorable response for this observable.	1
Running a set of experiments sufficient in number, range, and distribution to confirm that the relationship between altitude and amount of helium takes the form of a step function for Simulation problem 2.	#
Running a set of experiments sufficient in number, range, and distribution to confirm that the relationship between altitude and amount of helium is nonlinear for Simulation problem 2.	51
Running a set of experiments that suggests that the relationship between altitude and amount of helium takes the form of a two-piece linear one for Simulation problem 2.	9
Running a set of experiments that suggests that the relationship between altitude and amount of helium is linear for Simulation problem 2.	40
Running a set of experiments sufficient in number, range, and distribution to reveal the linear relationship between altitude and mass for Simulation problem 1.	24
Running experiments sufficient in number and range but not in distribution to confirm the linear relationship between mass and altitude for Simulation problem 1.	24
Running experiments either sufficient in number or in range to confirm the linear relationship between altitude and mass for Simulation problem 1.	10
Running experiments insufficient in number, range, or distribution to confirm the linear relationship between altitude and mass for Simulation problem 1.	42
Did not produce scorable response for this observable.	1
Creating a graph for Simulation problem 2 with the correct variables on the correct axes, with experimental data.	22
Creating a graph for Simulation problem 2 with the correct variables on the correct axes, with minimal experimental data or without data.	13
Creating a graph for Simulation problem 2 with only one or neither of the correct variables on the correct axes.	22
Did not create a graph for Simulation problem 2.	42

See notes at end of table.

Table J-3. Weighted percentage of students achieving each level of correctness on each Simulation scenario scientific exploration observable in order of first appearance on item map (figure 6-1), grade 8: 2003—Continued

Observable and level of correctness	Weighted percent
Creating a graph for Simulation problem 1 with the correct variables on the correct axes that shows at least two data points.	19
Creating a graph for Simulation problem 1 with the correct variables on the correct axes but that shows no experimental data or only one data point.	16
Creating a graph for Simulation problem 1 with only one or neither of the correct variables on the correct axes.	27
Did not create a graph for Simulation problem 1.	38
Running experiments for at least two values of mass and, for at least one of those values, conducting a set of experiments with amounts of helium sufficient in number and in range to confirm that the relationship between altitude and volume takes the form of a step function for Simulation problem 3.	9
Running experiments for at least one value of mass and conducting a set of experiments with amounts of helium sufficient in number and in range to confirm that the relationship between altitude and volume is nonlinear for Simulation problem 3.	4
Running experiments for at least one value of mass and conducting a set of experiments with amounts of helium that suggest that the relationship between altitude and volume takes the form of a two-piece linear function for Simulation problem 3.	15
Running experiments for at least one value of mass and conducting a set of experiments that suggest that the relationship between altitude and volume takes the form of a linear function for Simulation problem 3.	71
Creating a table for Simulation problem 1 that includes only the dependent and independent variables most germane to the problem.	8
Creating a table for Simulation problem 1 that includes the dependent and independent variables most germane to the problem as well as other variables.	18
Creating a table for Simulation problem 1 that includes the dependent OR independent variable most germane to the problem along with experimental data, OR that includes the dependent and independent variables most germane to the problem as well as other variables, but no data.	16
Creating a table for Simulation problem 1 that includes neither the dependent nor independent variable most germane to the problem, OR that includes either the dependent OR the independent variable most germane to the problem but no experimental data.	20
Did not create a table for Simulation problem 1.	37
Creating a graph for Simulation problem 3 with the correct variables on the correct axes that shows data for at least four experiments (two experiments for each of at least two values of mass).	20
Creating a graph for Simulation problem 3 with the correct variables on the correct axes that shows data for at least one experiment for each of two masses.	3
Creating a graph for Simulation problem 3 with the correct variables on the correct axes that shows data for one or no experiments.	27
Creating a graph for Simulation problem 3 that does not have the correct variables on the correct axes.	#
Did not create a graph for Simulation problem 3.	50
Creating a table for Simulation problem 3 that includes only the three variables most germane to the problem.	4
Creating a table for Simulation problem 3 that includes the three variables most germane to the problem along with other variables.	26
Creating a table for Simulation problem 3 that includes the three variables most germane to the problem along with other variables but no experimental data, OR any two of the most germane variables with data.	26
Creating a table for Simulation problem 3 that includes only one of the three variables most germane to the problem with experimental data, OR any two of the most germane variables without data.	#
Did not create a table for Simulation problem 3.	44

# The estimate rounds to zero.

NOTE: Detail may not sum to totals because of rounding. Evaluation levels for certain observables were collapsed during analysis; hence, not all the levels for these observables shown in this table appear in the item map.

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

Table J-4. Weighted percentage of students achieving each level of correctness on each Simulation scenario scientific synthesis observable in order of first appearance on item map (figure 6-2), grade 8: 2003

Observable and level of correctness	Weighted percent
Offering correct and complete ("best") responses to the constructed-response question that concludes Simulation problem 3 that explain how the relationship between amount of helium and balloon altitude for more than one payload mass takes the form of a series of step functions (e.g., "Once the balloon has enough helium to rise into the air, the balloon will rise to a maximum height and go no higher no matter how much helium is added.").	2
Offering correct but incomplete ("good") responses to the constructed-response question that concludes Simulation problem 3 by explaining either the top or the bottom of the step function (e.g., "Once in the air, the balloon will reach a maximum altitude no matter how much helium is added, and the maximum altitude the balloon can reach decreases as payload mass increases.").	7
Offering partially correct responses that can be derived from Simulation problems 1 or 2 to the concluding question for Simulation problem 3 (e.g., "Below a certain amount of helium the balloon cannot get off the ground.").	43
Offering wholly inaccurate responses to the concluding question for Simulation problem 3.	45
Did not produce scorable response for this observable.	4
Offering correct and complete ("best") responses to the constructed-response question that concludes Simulation problem 2 that explain how the relationship between amount of helium and balloon altitude for a payload mass of 100 lb. takes the form of a step function (e.g., "Once the balloon has enough helium to rise into the air, the balloon will rise to a maximum height and go no higher matter how much helium is added.").	13
Offering correct but incomplete ("good") responses referring either to the top or the bottom of the step function to the concluding question for Simulation problem 2 (e.g., "Once in the air, the balloon will reach a maximum altitude no matter how much helium is added.").	18
Offering partially correct responses that express a linear relationship between altitude and amount of helium to the concluding question for problem 2 (e.g., "More helium inside the balloon will make the balloon go higher.").	33
Offering wholly inaccurate responses to the concluding question for Simulation problem 2.	34
Did not produce scorable response for this observable.	2
Offering correct and complete ("best") responses to the constructed-response question that concludes Simulation problem 1 with specific references to experiments (e.g., "As the payload mass increases, the balloon's altitude decreases. For example, when I put 90 lb. of payload on the balloon, it only went to 10,000 feet. But when I put 50 lb. of payload mass on the balloon, it went to 22,326, and when I put 10 lb., it went to 36,211 feet.")	23
Offering correct but incomplete ("partial") responses that express the linear relationship between mass and altitude to the concluding question for Simulation problem 1 (e.g., "As the payload mass increases, the balloon's altitude decreases") with no specific references to experiments.	44
Offering wholly inaccurate response to the concluding question for Simulation problem 1.	31
Did not produce scorable response for this observable.	2
Correctly answering the multiple-choice question about the relationship between variables concluding Simulation problem 1.	59
Incorrectly answering the multiple-choice question about the relationship between variables concluding Simulation problem 1.	41
Correctly answering the multiple-choice question about the relationship among variables concluding Simulation problem 3.	31
Incorrectly answering the multiple-choice question about the relationship among variables concluding Simulation problem 3.	68
Did not produce scorable response for this observable.	1

See notes at end of table.

Table J-4. Weighted percentage of students achieving each level of correctness on each Simulation scenario scientific synthesis observable in order of first appearance on item map (figure 6-2), grade 8: 2003—Continued

Observable and level of correctness	Weighted percent
Making correct predictions for more than one half of unique experiments run for Simulation problem 2.	9
Making correct predictions for one half to one third of unique experiments run for Simulation problem 2.	6
Making correct predictions for less than one third of unique experiments run for Simulation problem 2.	6
Did not make predictions for Simulation problem 2.	79
Correctly answering the multiple-choice question about the relationship between variables concluding Simulation problem 2.	23
Incorrectly answering the multiple-choice question about the relationship between variables concluding Simulation problem 2.	77

NOTE: Detail may not sum to totals because of rounding. Evaluation levels for certain observables were collapsed during analysis; hence, not all the levels for these observables shown in this table appear in the item map.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Problem Solving in Technology-Rich Environments Study.

Table J-5. Weighted percentage of students achieving each level of correctness on each Simulation scenario computer skills observable in order of first appearance on item map (figure 6-3), grade 8: 2003

Observable and level of correctness	Weighted percent
Never using the interface tools in the wrong order for drawing conclusions in Simulation problem 3 (e.g., clicking on the Draw Conclusions button without having run any experiments).	93
Using the interface tools in the wrong order for drawing conclusions once or twice in Simulation problem 3 (e.g., clicking on the Draw Conclusions button without having run any experiments).	6
Using the interface tools in the wrong order for drawing conclusions at least 3 times in Simulation problem 3 (e.g., clicking on the Draw Conclusions button without having run any experiments).	#
Never using the interface tools in the wrong order for experimenting in Simulation problem 1 (e.g., clicking on the Make Predictions button without having chosen any values with which to experiment).	79
Using the interface tools in the wrong order for experimenting once or twice in Simulation problem 1 (e.g., clicking on the Make Predictions button without having chosen any values with which to experiment).	20
Using the interface tools in the wrong order for experimenting at least 3 times in Simulation problem 1 (e.g., clicking on the Make Predictions button without having chosen any values with which to experiment).	1
Did not produce scorable response for this observable.	1
Never using Computer Help in Simulation problem 1.	81
Using Computer Help once or twice in Simulation problem 1.	17
Using Computer Help at least 3 times in Simulation problem 1.	1
Did not produce scorable response for this observable.	1
Never using the interface tools in the wrong order for drawing conclusions in Simulation problem 2 (e.g., clicking on the Draw Conclusions button without having run any experiments).	90
Using the interface tools in the wrong order for drawing conclusions once or twice in Simulation problem 2 (e.g., clicking on the Draw Conclusions button without having run any experiments).	9
Using the interface tools in the wrong order for drawing conclusions at least 3 times in Simulation problem 2 (e.g., clicking on the Draw Conclusions button without having run any experiments).	1
Never using the interface tools in the wrong order for drawing conclusions in Simulation problem 1 (e.g., clicking on the Draw Conclusions button without having run any experiments).	75
Using the interface tools in the wrong order for drawing conclusions once or twice in Simulation problem 1 (e.g., clicking on the Draw Conclusions button without having run any experiments).	23
Using the interface tools in the wrong order for drawing conclusions at least 3 times in Simulation problem 1 (e.g., clicking on the Draw Conclusions button without having run any experiments).	1
Did not produce scorable response for this observable.	1
Key-entering a response of over 150 characters to the constructed-response question concluding Simulation problem 3.	51
Key-entering a response of 50 to 149 characters to the constructed-response question concluding Simulation problem 3.	37
Key-entering a response of less than 50 characters to the constructed-response question concluding Simulation problem 3.	11
Did not produce scorable response for this category.	1

See notes at end of table.

Table J-5. Weighted percentage of students achieving each level of correctness on each Simulation scenario computer skills observable in order of first appearance on item map (figure 6-3), grade 8: 2003—Continued

Observable and level of correctness	Weighted percent
Key-entering a response of over 150 characters to the constructed-response question concluding Simulation problem 2.	47
Key-entering a response of 50 to 149 characters to the constructed-response question concluding Simulation problem 2.	39
Key-entering a response of less than 50 characters to the constructed-response question concluding Simulation problem 2.	13
Did not produce scorable response for this observable.	#
Key-entering a response of over 150 characters to the constructed-response question concluding Simulation problem 1.	51
Key-entering a response of 50 to 149 characters to the constructed-response question concluding Simulation problem 1.	38
Key-entering a response of less than 50 characters to the constructed-response question concluding Simulation problem 1.	10
Did not produce scorable response for this observable.	1
Performing a variety of interface actions (e.g., tabbing among graphs, tables, and the response area; sorting tables) in Simulation problem 3.	47
Performing some interface actions (e.g., tabbing among graphs, tables, and the response area; sorting tables) in Simulation problem 3.	28
Performing few interface actions (e.g., tabbing among graphs, tables, and the response area; sorting tables) in Simulation problem 3.	25

# The estimate rounds to zero.

NOTE: Detail may not sum to totals because of rounding. Evaluation levels for certain observables were collapsed during analysis; hence, not all the levels for these observables shown in this table appear in the item map.

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

Appendix K: Understanding NAEP Reporting Groups

NAEP results are provided for groups of students defined by shared characteristics—gender, race/ethnicity, parental education, and eligibility for free/reduced-price school lunch. Based on participation rate criteria, results are reported for subpopulations only when sufficient numbers of students and adequate school representation are present. The minimum requirement is at least 62 students in a particular subgroup from at least five primary sampling units (PSUs).[1] However, the data for all students, regardless of whether their subgroup was reported separately, were included in computing overall results. Definitions of the subpopulations are presented below.

Gender

Results are reported separately for male students and female students.

Race/Ethnicity

In all NAEP assessments, data about student race/ethnicity is collected from two sources: school records and student self-reports. Prior to 2002, NAEP used students' self-reported race as the primary race/ethnicity reporting variable. As of 2002, the race/ethnicity variable presented in NAEP reports is based on the race reported by the school. When school-recorded information is missing, student-reported data are used to determine race/ethnicity. The mutually exclusive racial/ethnic categories are White, Black, Hispanic, Asian/Pacific Islander, American Indian (including Alaska Native), and Other. Information based on student self-reported race/ethnicity is available on the NAEP Data Explorer (http://nces.ed.gov/nationsreportcard/nde/).

Parental Education

Eighth-graders were asked the following two questions, the responses to which were combined to derive the parental education variable.

How far in school did your mother go?

A. She did not finish high school.

B. She graduated from high school.

C. She had some education after high school.

D. She graduated from college.

E. I don't know.

Students were also asked

How far in school did your father go?

A. He did not finish high school.

B. He graduated from high school.

C. He had some education after high school.

D. He graduated from college.

E. I don't know.

The information was combined into one parental education reporting variable in the following way: If a student indicated the extent of education for only one parent, that level was included in the data. If a student indicated the extent of education for both parents, the higher of the two levels was included in the data. If a student responded "I don't know" for both parents, or responded "I don't know" for one parent and did not respond for the other, the parental education level was classified as "I don't know." If the student did not respond for either parent, the student was recorded as having provided no response.

Eligibility for Free/Reduced-Price School Lunch

As part of the Department of Agriculture's National School Lunch Program, schools can receive cash subsidies and donated commodities in turn for offering free or reduced-price lunches to eligible children. Based on available school records, students were classified as either currently eligible for free/reduced-price school lunch or not eligible. Eligibility for the program is determined by students' family income in relation to the federally established poverty level. Free lunch qualification is set at 130 percent of the poverty level, and reduced-price lunch qualification is set at between 130 and 185 percent of the poverty level. Additional information on eligibility may be found at the Department of Agriculture website (http://www.fns.usda.gov/cnd/lunch/). The classification applies only to the school year when the TRE scenarios were administered (i.e., the 2002–2003 school year) and is not based on eligibility in previous years. If school records were not available, the student's information was recorded as "Unavailable." If the school did not participate in the program, all students in that school were classified as "Unavailable."

[1] A PSU is a selected geographic region (a county, group of counties, or metropolitan statistical area).