DOCUMENT RESUME

ED 386 474 TM 024 028

AUTHOR Bennett, Randy Elliot; And Others
TITLE Fitting New Measurement Models to GRE General Test Constructed-Response Item Data. GRE Board Professional Report No. 89-11P.
INSTITUTION Educational Testing Service, Princeton, NJ. Graduate Record Examination Board Program.
SPONS AGENCY Graduate Record Examinations Board, Princeton, N.J.
REPORT NO ETS-RR-91-60
PUB DATE Dec 91
NOTE 59p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Algebra; College Entrance Examinations; *College Students; *Constructed Response; Error Patterns; *Goodness of Fit; Higher Education; *Item Response Theory; *Measurement Techniques; Models; Scoring; Test Construction; Word Problems (Mathematics)
IDENTIFIERS *Graduate Record Examinations; *Partial Credit Model

ABSTRACT

This exploratory study applied two new cognitively sensitive measurement models to constructed-response quantitative data. The models, intended to produce qualitative characteristics of examinee performance, were fitted to algebra word problem solutions produced by 285 examinees taking the Graduate Record Examinations (GRE) General Test. The two types of response data modeled, error diagnoses and partial-credit scores, were produced by an expert system. Error diagnosis, analyzed using K. Yamamoto's (1989) Hybrid model, detected a class of examinees who tended to miss important pieces of the problem solution, but made relatively few errors of other types. Comparisons with matched examinees whose response patterns were better captured by the unidimensional item response theory model suggested subtle differences in error frequency rather than sharp qualitative distinctions. In contrast with the error data, partial-credit scores modeled using D. A. Rock's Hierarchically Ordered Skills Test (Rock and J. Pollack, 1987) procedure did not fit well, in part owing to limitations of the task theory being tested. Implications for the development of refined task and error theories, improvements to expert-system scoring procedures, and response modeling are discussed. Three figures and 11 tables present analysis results. Four appendixes present items, the scoring rubric, and information about bugs. (Contains 36 references.) (Author/SLD)

GRE®
RESEARCH

Fitting New Measurement Models to GRE General Test Constructed-Response Item Data

Randy Elliot Bennett
Marc M. Sebrechts
and
Kentaro Yamamoto

December 1991

GRE Board Professional Report No. 89-11P
ETS Research Report 91-60

Educational Testing Service, Princeton, New Jersey

Fitting New Measurement Models to GRE General Test Constructed-Response Item Data

Randy Elliot Bennett
Marc M. Sebrechts
and
Kentaro Yamamoto

GRE Board Report No. 89-11P

December 1991

This report presents the findings of a research project funded by and carried out under the auspices of the Graduate Record Examinations Board.

Educational Testing Service, Princeton, N.J. 08541

The Graduate Record Examinations Board and Educational Testing Service are dedicated to the principle of equal opportunity, and their programs, services, and employment policies are guided by that principle.

Abstract

This exploratory study applied two new cognitively sensitive measurement models to constructed-response quantitative data. The models, intended to produce qualitative characterizations of examinee performance, were fitted to algebra word problem solutions produced by examinees taking the GRE General Test. Two types of response data were modeled--error diagnoses and partial-credit scores--both produced by an expert system. Error diagnoses, analyzed using Yamamoto's (1989a) Hybrid model, detected a class of examinees who tended to miss important pieces of the problem solution but made relatively few errors of other types. Group members were of low quantitative proficiency overall, though considerable variability was evident. Comparisons with matched examinees whose response patterns were better captured by the unidimensional IRT model suggested subtle differences in error frequency rather than sharp qualitative distinctions. In contrast with the error data, partial-credit scores modeled using Rock's (Rock & Pollack, 1987) HOST procedure did not fit well, in part owing to limitations of the task theory being tested. Implications for the development of refined task and error theories, improvements to expert-system scoring procedures, and response modeling are discussed.

Among other things, multiple-choice tests have been criticized for being comprised of seemingly artificial tasks that offer little instructionally relevant information (Fiske, 1990; Guthrie, 1984). A conception intended to move standardized testing programs toward greater task fidelity and instructional utility is "intelligent assessment" (Bennett, in press). This conception attempts to integrate research on complex constructed-response items, artificial intelligence, and cognitively driven measurement models. A complex constructed-response denotes a task whose solutions are composed of many elements, take a variety of correct forms, and, when erroneous, approximate accurate answers to different degrees; these tasks (e.g., writing the steps associated with solving an algebra problem) come closer to the ones typically encountered in academic and work settings. Artificial intelligence is invoked as a practical means of analyzing solutions, producing both partial-credit scores and qualitative analyses. Finally, the driving mechanisms underlying the tasks and their scoring are cognitively grounded measurement models that may dictate what the characteristics of items should be, which items from a large pool should be administered, how item responses should be combined to make more general inferences, and how uncertainty should be handled.

Considerable progress has been made toward realizing this integration. Progress has occurred primarily in applying methods derived from intelligent tutoring (Wenger, 1987) to analyzing solutions to constructed-response questions. These methods have been implemented in the context of two testing programs, the College Board's Advanced Placement Computer Science (APCS) examination and the Graduate Record Examinations (GRE) General Test. The methods take the form of several expert systems--that is, computer programs intended to emulate the analytic behaviors of a content expert. The systems, PROUST, MicroPROUST, and GIDE, are described in detail elsewhere (Johnson, 1986; Johnson & Soloway, 1985; Sebrechts, LaClaire, Schooler, & Soloway, 1986).

A variety of item formats have been explored. In computer science, these formats call for the examinee to write a short procedure or to correct a faulty one. Scoring accuracy has been moderate to high, with correlations between machine and human ratings ranging from the .70s to .90s, though typically some solutions cannot be scored (Bennett, Gong, Kershaw, Rock, Soloway, & Macalalad, 1990; Braun, Bennett, Frye, & Soloway, 1990).

For algebra word problems, four constructed-response formats have been probed: open-ended, which presents the examinee with only the problem stem; goal specification, including the problem stem, a list of givens, and a list of unknowns; equation setup, which gives the unknowns and the equations needed to derive them; and faulty solution, comprised of the problem stem and an incorrect solution for the examinee to correct. In contrast to computer science, all responses to these problems could be machine scored, in part because the hierarchical nature of algebra solutions permits the integrity of previous solution steps to be inferred from subsequent ones (Sebrechts, Bennett, & Rock, 1991). Agreement between machine and human content experts' scores was reasonably high, with a median correlation across 12 problems of .88.

Preliminary evidence relevant to the construct meaning of machine scores for several of the formats has also been gathered. The computer science faulty solution format appears to measure the same construct as the APCS test (Bennett, Rock, Braun, Frye, Spohrer, & Soloway, 1990), which is composed of both multiple-choice and open-ended programming tasks. In algebra, the constructed-response scores appear to measure a dimension highly related to--but somewhat different from--that underlying the quantitative section of the GRE General Test. The difference in the two dimensions may be due in part to the broader range of content and difficulty covered by the latter measure (Bennett, Sebrechts, & Rock, in press).

The effort devoted to developing and evaluating item formats and analytic programs has necessarily preceded application of measurement models to scoring and diagnosis, which has focused on item-level performance. Aggregating an examinee's responses across items, however, is necessary if diagnoses are to characterize examinee performance at a meaningful level (e.g., be indicative of stable errors or of particular skill deficiencies).

Measurement models for guiding such aggregations have only recently become available (e.g., Falmagne, 1989; Masters & Mislevy, in press; M. Wilson, 1989). Such models should be cognitively driven; that is, they should derive their structure in considerable part from domain characteristics. Further, a model should have a principled mechanism for handling uncertainty, or noise in the data, resulting from the sometimes irrelevant influences that help determine observed performance. By formalizing relationships among the domain, observed performance, and the characterizations to be inferred from that performance, measurement models should provide more efficient and psychologically meaningful statements than ad hoc approaches.

This study was undertaken to investigate the fit of two recently developed measurement models--Hybrid (Yamamoto, 1989a) and HOST (Hierarchically Ordered Skills Test) (Rock & Pollack, 1987)--to constructed-response data collected through the Graduate Record Examinations program and, secondarily, to explore theoretical notions about the cognitive structures associated with solving algebra word problems. In combination with complex constructed-response items and intelligent analysis methods, cognitively sensitive measurement models may eventually offer a foundation for powerful, interactive diagnostic assessment systems.

Method

Subjects

Subjects were participants in a series of studies concerned with automated scoring of constructed-response algebra word problems. The sample was composed of 285 volunteers drawn from a pool of more than 50,000 examinees taking a single form of the GRE General Test administered nationally in June 1989. (See Sebrechts, Bennett, & Rock, 1991, for details of sample selection.) The sample differed somewhat from the General Test population along several dimensions. For example, the sample's General Test performance was significantly, though not dramatically, higher (by .4, .3, and .3 standard deviations, for verbal, quantitative, and analytical, respectively), and the most notable of several statistically significant demographic differences was in a greater proportion of non-Whites (see Table 1).

Table 1

Background Data for Study Sample

Variable	June 1989 Population	Study Sample
N	50,548	285
General Test Performance		
Verbal Mean(SD)	476(122)	527(132)*
Quantitative Mean (SD)	532(140)	573(141)*
Analytical Mean (SD)	513(132)	558(129)*
Percentage female	55%	60%
Percentage non-White[a]	16%	24%*
Percentage U.S. citizens	79%	85%*
Undergraduate Major		
Business	4%	2%*
Education	14%	6%*
Engineering	13%	12%
Humanities/Arts	14%	21%*
Life Sciences	18%	18%
Physical Sciences	10%	9%
Social Sciences	18%	24%*
Other	9%	10%
Intended Graduate Major		
Business	2%	2%
Education	18%	12%*
Engineering	10%	9%
Humanities/Arts	8%	9%
Life Sciences	16%	15%
Physical Sciences	8%	9%
Social Sciences	13%	19%*
Other	11%	9%
Undecided	15%	18%

[a]U.S. citizens only.

*$p < .05$

Instruments

Constructed-response items were adapted from standard, five-option multiple-choice algebra word problems taken from disclosed forms of the General Test quantitative section administered between 1980 and 1988. Three prototype items--one from each of the rate x time, interest, and work sets--were selected and three "isomorphs" were written for each prototype. Isomorphs were intended to differ from the prototype in surface characteristics only, for example, in topic (filling a tank vs. sending characters to a printer, determining percent profit instead of simple interest), and linguistic form, but not in underlying processes used to reach a solution.

Each item in a set (i.e., the prototype and its three isomorphs) was cast into one of four formats, such that each isomorph appeared in a different format. The formats were open ended, goal specification, equation setup, and faulty solution (see Figure 1). The items presented in each format are contained in Appendix A.

Item responses were analyzed by GIDE, a batch-processing laboratory tool capable of numerically scoring and diagnostically analyzing solutions to selected mathematical problems (Sebrechts, LaClaire, Schooler, & Soloway 1986; Sebrechts, Schooler, LaClaire, & Soloway, 1987; Sebrechts & Schooler, 1987). For each problem, GIDE has a specification that identifies both the "given" information and the goals into which the problem has been decomposed, where a goal is one of several objectives to be achieved in reaching a solution. (Problem decompositions were derived through a previous cognitive analysis.) To be considered correct, a solution must satisfy each goal. GIDE attempts to discover how the student solution satisfies a particular goal by testing it against a series of alternative correct plans (i.e., stereotypical procedures or equations) drawn from its knowledge base. If no matching plan is found, GIDE attempts to discover the nature of the discrepancy by testing plans that incorporate errors commonly made in achieving that goal or bug rules that represent more general mistakes. When no plan, buggy or correct, can be matched, the goal is considered missing. This determination is revised if subsequent goals that depend for their own satisfaction on the putative missing goal are found to be achieved.

GIDE assigns numeric scores based on a rubric and set of keys (see Appendix B). Full credit was awarded if all goals were achieved, suggesting the student was able to decompose the problem, correctly structure each goal, and compute its solution. Credit was deducted differentially depending on the errors detected for each goal. The largest deduction was made for missing goals because these absences suggest the student was unaware that addressing the goal was necessary to achieve a correct result. Less credit was deducted for structural bugs because such errors suggest both recognition of the goal's importance and a coherent, though incorrect, attempt to solve the goal. The smallest deduction was for computational errors, which may imply failures in basic calculation skills or procedural "slips" (Matz, 1982). Score scales for the items were based on the number of goals required for solution. Isomorphs developed from the work prototype contained two goals and were scored on a 0-6 scale. Problems based on the interest item were decomposed into three goals and scored on a 0-9 continuum. A 0-15 scale was employed for the rate items, which required solving five goals for a correct response.

Figure 1
Isomorphs in Four Item Formats

Open Ended

How many minutes will it take to fill a 2,000-cubic-centimeter tank if water flows in at the rate of 20 cubic centimeters per minute and is pumped out at the rate of 4 cubic centimeters per minute?

_____
_____
_____
_____

ANSWER:_____

Goal Specification

One of two outlets of a small business is losing $500 per month while the other is making a profit of $1750 per month. In how many months will the net profit of the small business be $35,000?

Givens
Profit from Outlet 1 = _____
Profit from Outlet 2 = _____
Target Net Profit = _____

Unknown
Net Monthly Profit = _____
= _____
Months to Reach Target Net Profit = _____
_____

ANSWER:_____

Equation Setup

A specialty chemical company has patented a chemical process that involves 2 reactions. Reaction 1 generates 24 grams of molecule B per minute and reaction 2 consumes 5 grams of molecule B per minute. If 4,560 grams of molecule B are desired as a product of this process, how many minutes must it continue?

Equations that Will Provide a Solution:

Net Amount of B Per Minute = Amt. Produced by Reaction 1 + Amt. Produced by Reaction 2
Time for Desired Amount of B = Desired Amount of B/Net Amount of B Per Minute

Your Solution:

_____
_____
_____
_____

ANSWER:_____

Faulty Solution

$3.50 in tolls is received each minute at an automated toll booth while the rate at a booth with an operator is $2.80 each minute. How many minutes elapse before the automated booth receives $14.00 more in tolls than does the person-operated booth?

Tolls per Minute = $3.50/min + $2.80/min
Tolls per Minute = $6.30/min
Time for $14 lead = $14/$6.30 per minute
Time for $14 lead = 2.22 minutes

Your Corrected Solution:

_____
_____
_____
_____

ANSWER:_____

Note. Print size is reduced and page arrangement modified for publication purposes.

Data Collection

Items were presented in paper-and-pencil format in individual and small group sessions. Examinees were asked to complete the problems at their own pace, though a one-hour period was suggested. Handwritten responses were typed to machine-readable form according to transcription rules (see Sebrechts, Bennett, & Rock, 1991) and the resulting 3,420 transcripts (12 items x 285 examinees) scored by GIDE. GIDE's scores for a subsample of examinees were then compared with those given by content experts to the original hand-written responses (Sebrechts, Bennett, & Rock, 1991). For the 12 items, correlations between GIDE and the mean of the humans' scores ranged from .74 to .97 with a median of .88.

Hybrid Model Description, Analyses, and Results

General Description

Yamamoto's Hybrid approach combines latent class models and item response theory (IRT) (Lord, 1980), with the former intended to capture such cognitive components as error tendencies, problem-solving strategies, mental models, and levels of operation.

Latent class models are built on the idea of a categorical or nominal latent variable (Lazarsfeld, 1960). There are two major assumptions in such models: (1) the classes (e.g., types of misconception, tendency toward a particular error) are mutually exclusive and together exhaustive--that is, each examinee belongs to one and only one class, and (2) responses are conditionally independent given the class of the responder. In applications similar to that reported here, a unique, idealized response pattern is associated with each latent class. Subjects who belong to the latent class should give responses similar to the idealized pattern. The imperfect fit of a subject's responses to that ideal pattern is characterized by a vector of conditional probabilities that suggests the likelihood with which the subject might be considered a member of one or another of the latent classes.

In practice, there will be individuals whose performance is not well captured by a limited set of classes. This eventuality may be owed to the existence of more classes than are represented in the model, or to responding in an inconsistent fashion, for example, making particular errors on some items but not on others where their occurrence would be expected. Performance that does not fit one of the hypothesized latent classes may be characterized more appropriately by the IRT model, which makes no strong assumptions about the qualitative understandings that examinees have. The Hybrid model accounts for these individuals by scaling them along a general dimension underlying performance on the problem set, while simultaneously providing diagnostic information for those examinees who fit a latent class.

There are three sets of parameters for the Hybrid model: (1) IRT parameters (item parameters for each item and an ability parameter for each examinee), (2) mixture proportions of IRT and latent classes for the population as a whole, and (3) a set of conditional probabilities for each of the latent classes. These parameters are estimated using the marginal maximum likelihood method (Bock & Aitkin, 1981; Mislevy, 1983).

At present, the fit of the Hybrid model cannot be precisely tested using statistical methods. For comparing the fit of two nested models--such as the IRT-only model versus IRT-with-latent-classes--the improvement of the log-likelihood ratio relative to the number of degrees of freedom expended can be examined using the chi-square statistic (if the number of subjects is large and the number of items is small). Additionally, the Akaike Information Criterion (AIC)--a parsimonious fit index--can be employed. When the competing models are not nested, a clear-cut statistical index is not available and greater weight must be placed on subjective judgments of the reasonableness of model parameters.

The performance of the Hybrid model has been assessed using data on electronic technicians' ability to interpret logic gate symbols (Gitomer & Yamamoto, 1991). Five latent classes were represented based on specific errors commonly made by technicians. The model's latent class portion was able to capture 25% of the response patterns, a respectable performance given the specificity of the error classes. In addition, for individuals picked up by the latent classes, the distinction among error classes given particular response patterns was quite sharp, making class assignments very clear. Finally, the probability of belonging to any latent class was unrelated to overall ability estimates, supporting the model's capacity to represent qualitative states.

Data Analyses

Hypothesized model. The cognitive structures modeled were major error classes evidenced by examinees in problem solving. Major error classes, rather than specific bug types, were modeled because research has suggested that the former may be stable whereas the latter generally are not (Payne & Squib, 1990; Tatsuoka, Birenbaum, & Arnold, 1989; VanLehn, 1982). The analyses were intended to explore whether groups of examinees could be distinguished based on these classes.

Major error classes were defined with reference to a general theory of problem solving propounded by Newell and Simon (1972), and applied to diagnosis by Johnson and Soloway (Johnson, 1986; Johnson & Soloway, 1985). The theory posits that problems can be decomposed into goals and each goal solved with a stereotypical method or plan. Errors are conceptualized as deviations from these goal-plan structures and can occur in failing to address a goal, in posing an incorrect plan, or in carrying out low-level operations as part of a plan.

This theoretical perspective suggests four major error classes within the algebra word problem domain. Mathematical errors involve a failure to execute a low-level operation (e.g., by inappropriately shifting a decimal, by incorrectly treating the remainder of a division as a decimal). Specific plan errors are inappropriate procedures for solving a goal linked to a particular problem class (e.g., confusing the rates for different trip segments). General plan errors suggest more universal failures to formulate procedures, with the same malformation having the potential to occur across problem contexts (e.g., dividing when multiplication is called for). Finally, missing goals, as noted, suggest the omission of a critical solution component. The specific errors composing each class were identified through a detailed cognitive analysis of GRE quantitative algebra problems (see Sebrechts,

Bennett, & Rock, 1991). Descriptions of the specific errors appear in Appendix C.[1]

Preliminary analyses. Preliminary analyses were conducted to describe bug occurrence and to suggest what latent classes might be represented in the data. The latter purpose was accomplished by estimating the relations within and among the four major bug categories across items, thereby indicating the extent to which examinees tended to repeat bugs from a category or to make errors from one category in conjunction with those of another. For each of the 12 items, each examinee was assigned four 0/1 scores, where 0 indicated the absence of a bug of a given class and 1 indicated the presence of one or more bugs from that class. This produced 44 scores per examinee (12 items x 4 bug categories minus the specific plan category for the two-goal work items, on which these errors were not observed). The 44 x 44 tetrachoric correlation matrix among these scores was then calculated and factor analyzed using Testfact (D. T. Wilson, Wood, & Gibbons, 1987). Loadings were computed via the marginal maximum likelihood estimation method with promax rotation. Omitted items were treated as indicating missing goals and not-reached items were ignored.

Model-fitting analyses. Using the HYBIL program (Yamamoto, 1989b), two models were fitted to the 44-score examinee vectors: (1) a two-parameter logistic IRT-only model in which examinees were arrayed along a scale of propensity to make errors and (2) a Hybrid model. Several points should be noted about these models. First, the IRT model employed two--rather than three--parameters because the probability of guessing correctly for constructed-response items is extremely low. Second, IRT models conventionally array individuals according to the probability of getting items correct. The convention was reversed here because the information of interest related to the error(s) committed. Finally, the assumption of conditional independence was violated in the limited instance in which all goals were missing. Such responses, by definition, could contain no other bugs.

The Hybrid model contained several levels of constraint on the latent class parameters. These were:

1. $\alpha_k = (1-\beta_k)$, where:

   $\alpha_k = p(x_i=1, T_{ki}=1 | \text{class } k)$;

   $\beta_k = p(x_i=0, T_{ki}=0 | \text{class } k)$;

   x_i is one of four 0/1 scores on each of 12 items where each 0/1 score indicates the presence or absence of a bug from one of the four major bug classes; and

   T_{ki} is an element from the idealized response pattern, T_k, of a particular latent class, k, that indicates the response that is expected given membership in that class.

2. $\alpha_k \neq \beta_k$, where α_k, β_k, x_i and T_{ki} are defined as above.

3. No constraints, where:

$\alpha_{ki} = p(x_i=1, T_{ki}=1 | \text{class } k)$;

$\beta_{ki} = p(x_i=0, T_{ki}=0 | \text{class } k)$; and

x_i and T_{ki} are defined as above.

The idealized response pattern indicated that, given membership in the latent class, an examinee should make at least one bug of that type on each item and no other categories of error. This pattern is, of course, unrealistic: few examinees will make an error from the same class on every item, and fewer still will consistently make that error in the absence of all others. More likely is that some examinees, even though committing errors from all four classes, will show relative tendencies toward one or another type. This probabilistic reality is reflected in the three constraint levels that allow different degrees of slippage from the idealized pattern, with the greatest slippage permitted in the unconstrained case.

Several indicators of model fit were evaluated. First, improvement in the fit of the Hybrid model over the separately estimated continuous IRT-only model was evaluated via the -2 log-likelihood index and the Akaike Information Criterion, for which the smaller the value, the better the fit. Because the log likelihood statistic is not chi-square distributed when the number of items and the sample size are small, this statistic needs to be interpreted cautiously. Second, the distribution of examinees across Hybrid latent and IRT classes was examined to determine the extent to which individuals were well represented in the latent classes; if only a minute portion of the sample is captured, the model will have limited diagnostic value. Third, for each item the conditional probabilities of making a particular error given membership in a latent class were inspected to see how severely they diverged from the idealized response pattern. Finally, the posterior probability distribution of the latent and IRT classes was computed. These probabilities were then compared for each latent class member to see how much better the observed pattern was described by the latent class than by the IRT model.

Scoring analyses. To understand better the meaning of bug information, relations with the General Test were investigated. These relations were of interest because the General Test's verbal and quantitative sections, in particular, are established reasoning measures with well-known psychometric characteristics. (At the same time, the limitations of this measure must be recognized, especially the potential of the test's multiple-choice format to constrain the type of problem solving assessed.)

Relations with the General Test were estimated using the full examinee sample with both model-based and model-free methods. In the model-based method, a two-parameter logistic IRT-only model was fitted to the 44-score examinee vectors described above using marginal maximum likelihood estimation (Bock & Aitkin, 1981) as implemented in HYBIL (Yamamoto, 1989b). Model fit was evaluated via the chi squares associated with the estimated item parameters. The proportions of variance explained in GRE General Test scores by the IRT theta estimates generated from this model were then computed. To determine how these relations changed when bug information was omitted, the

0/1 scoring commonly employed with multiple-choice questions was simulated by fitting a two-parameter logistic IRT-only model. Scoring for this model was conventional, with each item graded to indicate the complete absence of bugs (1) or the presence of one or more errors (0). The fit of the model to these 12-score examinee vectors was evaluated and the proportions of variance explained in General Test scores calculated. For both model-based analyses, omitted items were treated as indicating missing goals, and not-reached items were ignored.

The model-free method used least-squares linear multiple regression. GRE General Test scores were regressed separately on the 44-score and 12-item examinee vectors. Twenty-five examinees with multiple not-reached items were excluded from this analysis.

Data analysis summary. The Hybrid approach was used to model four major error classes evidenced by examinees in problem solving: mathematical, specific plan, general plan, and missing goal. Preliminary analyses were conducted to describe bug occurrence and to suggest which of the potential latent classes deriving from the four error types might be represented in the data. Next, an IRT-only model and a Hybrid model were fit to the examinee responses, each of which consisted of 44 scores indicating the errors made on each item. Finally, bug classes were related to General Test scores to investigate the meaning of error information.

Results

Preliminary analyses. On average, examinees made relatively few errors. The mean number of bugs per examinee taken across all items was 11, or almost 1 per item, with a standard deviation of 8. (The median was 9 and the range extended from 0 to 35.) The distribution was essentially unimodal, with the majority of examinees making between 1 and 13 errors. Seven examinees got all items correct.

Table 2 shows the percentages of subjects evidencing at least one bug from a major category for a given item. Because not all examinees finished the test, the chances of bug occurrence vary across items. Consequently, differences in bug incidence need to be carefully interpreted. As the table indicates, the distribution of examinees across the four major categories was reasonably similar from one isomorph to the next. An obvious exception was the "active ingredient" problem, characterized by an unusually high proportion of specific plan errors, the overwhelming majority of which turned out to be "percent-as-decimal" bugs--treating .25% as .25. These errors occurred far less frequently for the other isomorphs, in which whole-number percentages were used (e.g., 5% rather than .25%). (See Appendix D for the incidence of specific bugs.)

In most instances, less than a third of examinees made a particular category of error on any given item. The exceptions were for the specific plan category on the "active ingredient" problem noted above and for math errors on the five-goal items. The latter mistakes were owed primarily to providing an answer that was very close to--but not precisely--correct (e.g., 6.30 for 6.33).

Table 2

Percentages of Examinees Evidencing Major Bug Types by Item

Two-Goal Items (Work)				
Bug Type	Faulty Solution--$3.50 Tolls (n=273)	Equation Setup--Chemical Co. (n=284)	Goal Specific.--Small Bus. (n=277)	Open-Ended--2000cc Tank (n=274)
Math	8%	13%	3%	6%
General plan	12%	19%	16%	12%
Specific plan	0%	0%	0%	0%
Missing goal	18%	8%	5%	17%

Three-Goal Items (Percent)				
Bug Type	Faulty Solution--Active Ingd. (n=271)	Equation Setup--Graphics (n=283)	Goal Specific.--Load Cement (n=275)	Open-Ended--Investment (n=272)
Math	4%	4%	4%	8%
General plan	19%	6%	9%	10%
Specific plan	62%	2%	5%	11%
Missing goal	17%	8%	2%	15%

Five-Goal Items (D=RT)				
Bug Type	Faulty Solution--DOT Crew (n=278)	Equation Setup--720 Pages (n=284)	Goal Specific.--2400g Tank (n=283)	Open-Ended--600-Mile (n=269)
Math	47%	42%	40%	44%
General plan	22%	24%	29%	16%
Specific plan	11%	24%	31%	15%
Missing goal	26%	16%	20%	19%

Note. Percentages are of the number of examinees responding to each item that is indicated in parentheses.

Table 3 shows the percentage of students evidencing specific bugs on one or more items. (Bugs were included if they were made one time or more by at least 10% of examinees.) The results are generally consistent with other findings on the stability of specific bugs (e.g., Payne & Squibb, 1990). Of the 21 errors listed, only 4 were made repeatedly by 10% of the sample; 14 were made consistently by at least 5%. Of the four most persistent bugs, two were math errors (decimal and unit precision mistakes made repeatedly by 29% and 27% of examinees, respectively), one was an unexplained-value general-plan error (made by 43% of examinees), and the last was a missing-goal bug (committed by 18% of students). The stability of even these bugs was relatively weak, however: less than six percent of the sample made any one of them on four or more questions and almost no one showed them on five or greater items.

The stability of, and relations among, major error classes were investigated via exploratory factor analysis of the 44 x 44 matrix of correlations among the four bug classes. This analysis produced two factors with eigenvalues of 7.9 and 2.6. (The remaining eigenvalues were less than 1.9 and gradually decreased.) On the first factor, missing-goal bugs generally had the highest loadings (mean = .54, SD = .19), and the loadings for the other bug categories were considerably smaller (mean = .19, SD = .26). Factor two had a reverse pattern: low loadings for missing goals (mean = -.03, SD = .23) and higher ones for the other bug categories (mean = .25, SD = .21). The two factors were correlated at .51.

Model-fitting analyses. Since the factor analysis suggested that missing-goal errors were somewhat independent of other bugs, a Hybrid model was fitted that captured examinees either in a single latent class (i.e., missing goals in the absence of other errors) or with the IRT model. Attempts to fit Hybrid models with more latent classes increased the number of captured examinees only marginally and are not reported.

Table 4 shows the -2 log-likelihood fit statistic for the IRT-only model and the Hybrid model under three levels of constraint. Each model was fit using the 44-score bug occurrence vector. As the table indicates, the Hybrid models provided significant improvements over the IRT-only model, with the unconstrained Hybrid fitting best. The Akaike Information Criterion (AIC) presented a similar picture, though the best fit was suggested for the moderately constrained model. (For the least to most constrained Hybrid models, the AICs were 9280, 9267, and 9283; for the IRT-only model, the value was 9324.) Because the fit statistics did not agree on which of these two Hybrid models was to be preferred, the unconstrained model was selected for further examination. This model captured a larger latent class and, through its free conditional probabilities, permitted exploring the association of error tendency with item content and format (the latter a central concern of the HOST analysis reported below).

Under the unconstrained model, 12% of the sample--32 examinees--were encompassed by the missing-goals class. Conditional probabilities for the four error classes given membership in the missing-goals latent class are presented in Table 5. As the table indicates, there is a modest tendency toward missing-goals errors: the mean conditional probability for this class was .33 compared with .20 for the three other error classes combined (.19 for

Table 3

Percentages of Examinees Evidencing Specific Bugs
on One or More Items (n = 285)

Bug Type	Percentages of Students Evidencing Bug		
	In Total	On Only One Item	On More than One Item
Math			
Decimal Shift	33%	24%	9%
Close Enough Tenths	61%	32%	29%
Close Enough Units	55%	28%	27%
General Plan			
Unexplained Value	16%	14%	3%
Unknown Value	72%	29%	43%
Add for Subtract	16%	13%	3%
Times for Divide	11%	11%	0%
Specific Plan			
Result per Unit=Amt/Rate	19%	15%	4%
Percent as Decimal	58%	52%	6%
Unexplained Distance	11%	9%	1%
Close Enough Minutes	22%	22%	0%
Time as Decimal	10%	10%	0%
Decimal Portion as Time	13%	9%	4%
Missing Goal			
Work Goal #2	26%	18%	8%
Work Goal #1	22%	16%	6%
Interest Goal #3	30%	22%	8%
D=RT goal #5	44%	26%	18%
D=RT goal #4	26%	17%	9%
D=RT goal #3	18%	12%	6%
D=RT goal #2	19%	14%	6%
D=RT goal #1	17%	12%	5%

Note. Bugs were included if they were made at least one time by at least 10% of examinees. Not-reached items are omitted from tabulations.

Table 4

Fit of the IRT-Only and Hybrid Models

Model	-2 log-likelihood		df		Difference		
Contrast	Model 1	Model 2	Model 1	Model 2	-2 log	df	p
H3 v. H2	9014	9085	133	91	71	42	<.01
H2 v. H1	9085	9103	91	90	18	1	<.01
H1 v. IRT	9103	9148	90	88	45	2	<.01

Note. H3 = Unconstrained Hybrid model; H2 = Hybrid model with $\alpha_k \neq \beta_k$; H1 = Hybrid model with $\alpha_k = (1-\beta_k)$. Model 1 is the less constrained model in each contrast.

Table 5

Conditional Probabilities of Making Four Different Error Types Given Membership in the Missing-goals Latent Class

Item Number & Bug Class	Item Type		Idealized Probability	Conditional Probability
	Format	Content		
1. Math	Open-Ended	5-goal	.00	.45
General			.00	.14
Specific			.00	.18
Missing			1.00	.33
2. Math	Open-Ended	3-goal	.00	.11
General			.00	.00
Specific			.00	.18
Missing			1.00	.24
3. Math	Open-Ended	2-goal	.00	.16
General			.00	.25
Missing			1.00	.34
4. Math	Goal Specification	3-goal	.00	.06
General			.00	.10
Specific			.00	.16
Missing			1.00	.09
5. Math	Goal Specification	2-goal	.00	.00
General			.00	.22
Missing			1.00	.24
6. Math	Goal Specification	5-goal	.00	.24
General			.00	.18
Specific			.00	.14
Missing			1.00	.53
7. Math	Equation Setup	2-goal	.00	.26
General			.00	.34
Missing			1.00	.20
8. Math	Equation Setup	5-goal	.00	.40
General			.00	.27
Specific			.00	.07
Missing			1.00	.40
9. Math	Equation Setup	3-goal	.00	.14
General			.00	.23
Specific			.00	.11
Missing			1.00	.28
10. Math	Faulty Solution	5-goal	.00	.18
General			.00	.21
Specific			.00	.10
Missing			1.00	.64
11. Math	Faulty Solution	2-goal	.00	.17
General			.00	.21
Missing			1.00	.22
12. Math	Faulty Solution	3-goal	.00	.09
General			.00	.44
Specific			.00	.66
Missing			1.00	.43

math, .22 for general, and .20 for specific). In several instances, however, the conditional probabilities diverged--sometimes markedly--from this tendency (items 1, 4, 7, 8, and 12). For item 12, perhaps the most seriously misfitting item, the probability of a specific-plan error was .66, compared with .43 for a missing goal, a reversal of the idealized model. This was the "active ingredient" item on which a large number of "percent-as-decimal" bugs occurred.

In Table 6, the conditional probabilities of missing-goal errors given membership in that class are shown by format and content. For five-goal problems, the probability of making a missing-goal error was consistently higher than for the other two content types, perhaps because the fourth goal (i.e., find the activity's duration) is frequently the terminal one in other, similar problems. No distinction among formats was evident.

The posterior probabilities of the latent and IRT classes were compared for each member of the missing-goals group to see how much better the latent class characterized the observed pattern than did the IRT model. The mean and median IRT probabilities for the group were .06 and .02, respectively; the corresponding latent class probabilities were .91 and .97. Twenty-seven of the 32 missing-goals group members had latent class probabilities greater than .85 and IRT probabilities less than .10.

What characteristics distinguished the missing-goals group? Compared with the total sample, missing-goals examinees had lower partial-credit scores on the 12-item constructed-response test (a mean of 64 and standard deviation of 26 vs a mean of 97 and standard deviation of 22); had lower mean General Test scores (427 vs 527 for verbal, 409 vs 573 for quantitative, 427 vs 558 for analytical); were more frequently noncitizens (34% vs 15%), and, of those who were citizens, were more frequently non-White (54% vs 24%).

Although missing-goals examinees had relatively low mean quantitative scores, these individuals were not necessarily the least adept performers on the GRE quantitative section. The point-biserial correlation between group membership and GRE quantitative score was -.42, $t(283) = 7.79$, $p < .001$, suggesting a significant but moderate relationship. Seven of 32 members had quantitative scores exceeding the total sample's lower quartile (i.e., > 470). The mean score for the quartile was 378 with a standard deviation of 65; the mean for the missing-goals group was 409 (SD = 142).

As expected, latent class members made frequent missing-goal errors. These examinees made a median of four such errors with a range of 2-10. Item responses containing missing-goal errors were divided between those that were partially misspecified and those that were completely malformed. Partially misspecified responses were lacking at least one solution component (but not all), and were generally accompanied by other errors. The completely malformed responses varied from no response or only a restatement of given information, to partial solutions following incorrect paths. These completely malformed responses generated a missing-goal bug for each goal in the problem decomposition and, by definition, were accompanied by no other errors. (Figure 2 gives examples of each response type.) Examinees tended to make more partially misspecified responses (median = 3, range = 1-5) than completely malformed ones (median = 1, range = 0-8); only one student showed the reverse pattern.

Table 6

Conditional Probabilities of Making Missing-goal Errors Given Membership in the Missing-goals Latent Class by Item Format and Item Content

	Item Content			
Item Format	Five-Goal	Three-Goal	Two-Goal	Mean
Open Ended	.33	.24	.34	.30
Goal Specif.	.53	.09	.23	.28
Equation Setup	.40	.28	.20	.29
Faulty Solution	.64	.43	.23	.33
Mean	.48	.26	.25	

Figure 2
Responses Containing Missing Goals

(a) A partially misspecified response with no error other than a single missing goal.

Money in a certain investment fund earns an annual dividend of 5 percent of the original investment. In how many years will an initial investment of $750 earn total dividends equal to the original investment? (Open-ended format)

12 months = 5%
$750 original amount
750 x 5/100 = 37.50

(b) Two partially misspecified responses: one with a final missing goal and an unreduced expression bug; the other with a final missing goal, a close-enough tenths bug, and a decimal-as-time bug.

A graphics designer earns 2% of a $1500 yearly bonus for each shift of overtime she works. How many shifts of overtime must she work to earn the equivalent of the entire yearly bonus? (Equation-setup format)

X = 2% x 1500
X = 1500 x .02

On a 600-hundred mile motor trip, Bill averaged 45 miles per hour for the first 285 miles and 50 miles per hour for the remainder of the trip. If he started at 7:00 a.m., at what time did he finish the trip (to the nearest minute)? (Open-ended format)

600 miles - 285 = 315
285/45 = 6.33
315/50 = 6.33
12.66 = 13.06 min
13 hrs 6 min

(c) A completely malformed response with all goals missing.

A Department of Transportation road crew paves the 15 mile city portion of a 37.4 mile route at the rate of 1.8 miles per day and paves the rest of the route, which is outside the city, at a rate of 2.1 miles per day. If the Department of Transportation starts the project on day 11 of its work calendar, on what day of its work calendar will the project be completed? (Faulty solution format)

15 miles
37.4 = 1.8 day
37.4:1.8 day as X:2.1

Note. The formats in which problems were administered are indicated in parentheses after each item. Responses are from examinees with posterior probabilities of belonging to the missi g-goals class = 1.00.

Latent class members were also more likely than the overall sample not to complete the constructed-response test (though because the Hybrid modeling ignored not-reached items, nonresponse was not directly a factor in forming the missing-goals group). Some 31% of the group did not reach items, compared with about 11% of the total sample.

Table 7 compares the responses of several missing-goals examinees with IRT-modeled subjects matched on GRE quantitative and total constructed-response test scores. The matches were somewhat imprecise due to the need to satisfy both criteria simultaneously and to eliminate examinees who had multiple not-reached items. As expected from the latent class conditional probabilities, these examinees do tend to make more missing-goal errors in the relative absence of other mistakes. Also as expected, this tendency is modest, perhaps better characterized as one of degree rather than a sharp qualitative distinction.

Closer inspection of the responses of missing-goals examinees revealed a limited number of mistakes in transcribing examinee responses from written to machine-readable form and in GIDE's processing of the transcriptions. The transcription errors generally involved failing to record the minimal notations that constituted some responses (e.g., a partially executed plan for the first goal with no attempt to solve subsequent goals); this sometimes caused GIDE to interpret a response as a series of missing goals when, in fact, it should have been described as containing a plan error in conjunction with missing goals. The processing errors took several forms. In some cases, examinees represented clock-time results using decimals instead of colons (e.g., 7:37.8 instead of 7:37:48 or simply 7:38), which GIDE misinterpreted as a missing goal. In a second instance, the examinee formulated a solution component as a nested equation, causing the same erroneous analysis. Finally, GIDE misinterpreted an examinee's restatement of given information (24=B) as an approximation of the correct answer (240 minutes) with a decimal-shift error, thus failing to recognize an instance in which the goal was, in fact, missing.

Scoring analyses. For the two-parameter IRT-only model using the 44-score number-correct examinee vector, the fit of the estimated item parameters produced chi squares between .23 and 10.25 (8 df) for 43 of the 44 item scores; the chi square for the remaining score was 32.04. Proportions of variance accounted for by the regressions of GRE verbal, GRE quantitative, and GRE analytical on these scores were .20, .50, and .33, respectively, with all values significant at $p < .001$.

The two-parameter IRT-only model was also fitted to the 12-score number-correct examinee vector. The fit of the estimated item parameters produced chi squares from .11 to 4.09 (8 df). Proportions of variance accounted for by the regressions of GRE verbal, GRE quantitative, and GRE analytical on these scores were similar to those produced by the 44-score vectors (.18, .54, and .31, respectively, all significant at $p < .001$).

The multiple regressions were computed after deleting 25 examinees with multiple not-reached items, 8 of whom were members of the missing-goals group. With bug information, the proportion of variance accounted for in GRE verbal was .29, .62 for GRE quantitative, and .43 for GRE analytical. For the

Table 7

Response Patterns of Three Pairs of Matched Examinees
Captured by the Hybrid and IRT Models

	Contrast #1		Contrast #2		Contrast #3	
Variable	Missing Goals	IRT	Missing Goals	IRT	Missing Goals	IRT
GRE quantitative	340	340	460	480	610	580
Partial-credit score	55	64	95	94	84	85
Latent class prob.	.97	.00	.94	.00	.99	.01
Item						
1 600-mile	*	*M*	A			*M*,A
2 Investment	*M*	G				
3 2000cc Tank		*M*	*M*,G			
4 Load Cement				G		
5 Small Business		*M*				
6 2400g Tank	*M*,G	A,G	G,S	*M*	*M*	A
7 Chemical Company	*M*	A	G	A	A,G	
8 720 Pages	*M*	*M*,A,G,S	A	G,S	A	*M*,G,S
9 Graphics	*M*,G		*M*,S		A,S	*M*
10 DOT Crew	*M*,A	*M*,A,S		A	*M*,A,S	A
11 $3.50 Tolls	*M*	*M*,G				
12 Active Ingredient	G,S	*M*,G,S	G,S	S	*M*	A,S

Note. Latent class probability = probability that the examinee's response pattern belongs to the missing-goals latent class. Partial-credit score = the sum of the scores on the 12 constructed-response items (scale = 0 - 120). Error codes are *M* = missing goal, A = math, G = general plan, S = specific plan. Each error code indicates the presence of one or more errors of that type. * = not-reached item ignored by the Hybrid model.

number-correct scoring, the values were .18, .56, and .33, respectively. All values for these analyses were significant at $p < .001$.

Table 8 shows the median and range of correlations between the presence and absence of each of the four bug categories on an item and General Test score (as computed from the multiple regression analyses). As is evident, missing-goal errors consistently predicted General Test performance.

HOST Model Description, Analyses, and Results

General Description

The Hierarchically Ordered Skills Test (HOST) model (Rock & Pollack, 1987) can be viewed as a restricted case of the Hybrid approach. In the former model, groups of items represent levels of proficiency, with each succeeding level requiring the cognitive operations of the preceding one plus som thing additional. If the model fits, standing on the HOST scale denotes what operations the student is and is not able to perform.

Besides indicating level of proficiency, the HOST model has two useful properties. First, it provides a measure of individual fit that advises the user on the appropriateness of model-based interpretations. Because individuals often come to proficiency by different paths, the same hierarchy does not hold for everyone, thereby making HOST-based interpretations sometimes inapplicable (though describing the student's standing on a more general proficiency scale might still be justified). Second, the model provides estimates of the probabilities associated with being at particular skill levels. These probabilities have proven particularly useful for measuring the extent to which individuals change because the probabilities seem less sensitive than other metrics to the ceiling and floor effects that have perennially hampered attempts to assess individual growth (Rock & Pollack, 1987).

Rock has studied the fit of the HOST model to mathematics proficiency data from the 1980 sophomore High School and Beyond (HS&B) cohort and from item subsets extracted from the Scholastic Aptitude Test (SAT) (Gitomer & Rock, in press; Rock & Pollack, 1987). Because no statistical index exists, fit was evaluated primarily though the proportions of students whose response patterns were consistent with the hypothesized hierarchical ordering. In these studies, the overwhelming majority of students fit the model: 90% for the HS&B sample and 96%-98% for the SAT sample. Further, the model fit equally well for males and females, and for majority and minority students.

Data Analyses

Hypothesized model. For purposes of applying the HOST model, the four item formats--open ended, goal specification, equation setup, and faulty solution--were hypothesized to form a hierarchy based on the degree of constraint imposed on the response, where increased constraint was expected to aid problem solution (i.e., if one can solve a problem in the open-ended format, one should be able to solve its isomorph in the equation-setup format). Figure 3 delineates the cognitive operations suggested to underlie each level of this hierarchy and, consequently, the statements that might be made about individuals at each level. Proficiency at each level was measured

Table 8

Medians and Ranges of Correlations Between the Presence of Bug Classes on Individual Items and GRE General Test Score (n = 260)

	GRE General Test Score		
Bug Class	Verbal	Quantitative	Analytical
Math	.08 (.00-.16)	.19* (.02-.29)	.12 (.01-.26)
General	.15* (.02-.21)	.18* (.06-.27)	16* (.04-.20)
Specific	.12 (.03-.19)	.23* (.12-.34)	.17* (.13-.26)
Missing	.17* (.08-.27)	.27* (.14-.37)	.20* (.09-.35)

Note. Each cell is based on 12 correlations except for the Specific bug cells, which are based on 8 values.

*$p < .05$ (two-tailed test)

Figure 3

Proposed Hierarchical Arrangement of Item Formats

Level	Format	Operations
4	Open ended	Identify givens and unknowns. Create representation for problem based on knowns and unknowns. Map equations onto problem statement. Solve equations. Check solution, detect error(s), and recover.
3	Goal specification	Create representation for problem based on knowns and unknowns. Map equations onto problem statement. Solve equations. Check solution, detect error(s), and recover.
2	Equation setup	Map equations onto problem statement. Solve equations. Check solution, detect error(s), and recover.
1	Faulty solution	Check solution against problem statement, detect error(s), and recover.

by a three-item parcel (one item from each content set), with each item scored on a different partial-credit scale. Because items were constructed to be isomorphic to one another, difficulty across levels should be roughly similar with the exception of that introduced by question format.

As Figure 3 indicates, the formats are suggested to form a Guttman-type simplex (Guttman, 1954). That is, items at level 2 call for the same operations as those at level 1 (i.e., check solution, detect errors, and recover), but also demand additional processes (i.e., map equations onto problem statement and solve equations). As a result, examinees who are proficient at level 2 should have a high probability of being proficient at level 1.

Because the HOST model is predicated on item format and the constructed-response items were administered in format sequence, not-reached items would be expected to introduce spurious effects. Consequently, examinees who did not reach two or more items were removed from the analysis. Twenty-five of the 285 examinees were excluded by this criterion; 19 of these did not reach 3 or more items (an entire HOST level).

Model-fitting analyses. To test the fit of the HOST model, a pass/fail score was generated for each proficiency level. Scores were aggregated across items in a level because previous attempts to fit Guttman-type scales to item data have almost always produced disappointing results, in part due to the low reliability of individual items. Variations on two pass/fail scoring schemes were used. The first method computed the level score by taking the sum of the item scores (from 0-30) and considering an examinee to have passed the parcel if the score equaled or exceeded the cut value. Cut values were 23, 27, and 30. Under the second method, each item score was calculated as a percentage of the maximum possible item score and the student was considered to pass the item if the score equaled or exceeded the cut point. Cut points under this method were 75%, 90%, and 100% of the scale maximum. A student was considered to have passed the parcel if any two items were passed.

The fit of the HOST model was assessed through several means. First, mean scores (taken across students and items) were computed for each level to see if the expected item ordering held. Second, a four-element vector was formed for each individual, with each 0/1 element indicating proficiency for a given level. The elements were ordered from level 1 through 4 so that if the formats constituted a perfect hierarchical scale, 0 should never precede 1. The percentage of scale reversals was computed separately for single (e.g., 0111), double (e.g., 1001), and triple (e.g., 0001) reversals.

Data analysis summary. The HOST procedure was used to model a hierarchy of cognitive operations composed of four proficiency levels. Each proficiency level was marked by three items of the same format, with formats varying in the degree of constraint imposed upon the response. Pass/fail scores were generated for each proficiency level. Mean differences among levels were compared and student proficiency vectors examined to determine if the hypothesized hierarchy was supported.

Results

Model-fitting analyses. Table 9 shows the mean parcel scores for the four proficiency levels, where a parcel score was the sum of the partial-credit scores for each constituent item. As the table indicates, the distributions were skewed, showing a marked ceiling effect. Additionally, three of the four means fell within an item of one another. The single outlying mean--for level 1 (faulty solution)--ran counter to the expected ordering, showing that level to be the most difficult instead of the least challenging.

Mean partial-credit scores for the individual items are depicted in Table 10. Items are arranged by level (format) within content set, the latter indicated by score range (0-6, 0-9, 0-15), permitting the difficulties of isomorphic items of different levels to be compared. Again, level 1 problems were always the most difficult; within content sets, however, the means were closely similar. The singular exception was for the "active ingredient" item, a level 1 question that differed in difficulty from the nearest item in its content group by a full standard deviation (where the standard deviation was the mean of the standard deviations of the content set).

Table 11 presents the percentages of students with different parcel pass/fail patterns under three cut scores for each of two scoring methods. Two points are noteworthy. First, regardless of method, only a small percentage of examinees unequivocally fit the hypothesized model. Although between 43% and 67% of response patterns were consistent with the model, the overwhelming majority of these patterns were either consistent passes (1111) or failures (0000), both of which may mask model misfit. As the cut score was adjusted to increase fit within either method, the percentage of consistent mixed patterns (i.e., 1000, 1100, 1110) remained almost constant (changing from 3% to 6% to 5% under the item-score method). At the same time, the combined percentage of perfect and failure patterns increased while the percentage of inconsistent patterns (i.e., single, double, and triple reversals) decreased. This covariation suggests that the total percentage of consistent patterns may significantly overestimate model fit. Second, some of the misfit can be traced to the level 1 (faulty solution) items. Under both scoring methods, the most frequent single reversals usually involved these problems (i.e., 0111, 0110, 0100).

Discussion

This exploratory study investigated the fit of two cognitively oriented measurement models to constructed-response item data and, secondarily, probed ideas about cognitive structure in solving algebra word problems. The Hybrid model captured a small percentage of examinees whose performance was not well characterized by the IRT model. This group tended to miss critical problem components while making relatively few errors of other types. No association with item format was evident, although this error tendency was more prevalent with five-goal problems, perhaps because these problems contained an easy-to-forget final step. Comparison of the responses of these individuals with those represented by the IRT model suggested that the groups were distinguished more by degree than by a sharp qualitative difference.[2] Still,

Table 9

Means and Standard Deviations for Parcel Scores at Each Proficiency Level (n =260)

Proficiency Level	Mean	Standard Deviation
4. Open ended	25.6	5.4
3. Goal specification	26.6	4.5
2. Equation setup	26.3	5.4
1. Faulty solution	22.7	6.8

Note. Scores are on a 0-30 scale.

Table 10

Means and Standard Deviations for Item Scores (n =260)

Item	Format & Level		Scale	Mean	Standard Deviation
2000cc Tank	Open	(4)	0-6	5.0	2.0
Small Business	Goal	(3)	0-6	5.6	1.0
Chemical Co.	Equation	(2)	0-6	5.3	1.5
$3.50 in Tolls	Faulty	(1)	0-6	4.9	2.1
Investment	Open	(4)	0-9	7.9	2.3
Load Cement	Goal	(3)	0-9	8.6	1.2
Graphics	Equation	(2)	0-9	8.4	1.9
Active Ingrd.	Faulty	(1)	0-9	5.9	2.6
600-Mile	Open	(4)	0-15	12.7	3.3
2400g Tank	Goal	(3)	0-15	12.3	3.6
720 Pages	Equation	(2)	0-15	12.6	3.4
DOT Crew	Faulty	(1)	0-15	11.9	4.2

Table 11

Numbers of Students with Different Patterns of Parcel Pass/Fail Scores Under Two Scoring Methods and Three Cut Scores (n = 260)

Parcel Pass/Fail Pattern	Total-Score Method			Item-Score Method		
	23	27	30	75%	90%	100%
Consistent						
0000	4%	16%	53%	2%	9%	11%
1000	0%	0%	1%	0%	1%	0%
1100	0%	2%	3%	1%	0%	0%
1110	7%	4%	2%	2%	5%	5%
1111	54%	23%	3%	62%	36%	27%
Total	65%	45%	60%	67%	51%	43%
Single reversal						
0111	17%	17%	3%	19%	18%	17%
1011	2%	2%	2%	2%	3%	2%
1101	1%	2%	0%	0%	3%	3%
0110	4%	10%	4%	2%	7%	8%
1010	0%	0%	1%	0%	0%	0%
0100	3%	4%	8%	1%	1%	2%
Total	27%	34%	19%	25%	33%	33%
Double reversal						
1001	0%	0%	1%	0%	0%	0%
0101	1%	6%	4%	1%	3%	6%
0011	2%	5%	2%	3%	6%	8%
0010	4%	6%	5%	5%	4%	6%
Total	7%	17%	11%	8%	13%	21%
Triple reversal						
0001	1%	3%	10%	0%	2%	3%
Total	1%	3%	10%	0%	2%	3%

Note. In the total-score method, the parcel score is computed by summing the item scores (0-30) and considering a examinee to have passed the parcel if the score equaled or exceeded the cut value. For the item-score method, each item score was calculated as a percentage of the maximum possible item score and the student was considered to pass the item if the score equaled or exceeded the cut point. A student was considered to have passed the parcel if any two items were passed.

a tendency to misspecify important solution components may comprise the more salient of a given examinee's quantitative skill difficulties, even if it co-exists with other less pronounced deficits. That this error tendency may be important was supported by its significant relation with General Test performance. Thus, if these results can be replicated, this information may have value for descriptive or remedial purposes, possibly for a larger percentage of individuals as the focus shifts to more representative segments of the General Test population.

Why weren't more error classes detected and why wasn't the detected class more distinct? A persistent issue in the error analysis literature is stability. Several investigators have found substantial inconsistency in individuals' math errors (Payne & Squibb, 1990; Tatsuoka, Birenbaum, & Arnold, 1989). VanLehn (1982), however, has argued that systematicness can be detected when bugs are viewed from the perspective of repair theory, which suggests that students make local fixes upon encountering an impasse in problem solving. The particular fix may vary from one time or problem to the next, but having the same genesis, the resulting bugs should be related. Although GIDE's analyses are based on a problem-solving theory that conceptualizes bugs as deviations from correct goal-plan structures, the specific bugs composing a general class do not necessarily emanate from the same underlying source. Consequently, a reorganization of error classes according to underlying generative mechanisms might permit greater consistency to be observed.

Other factors may have limited the discovery of multiple, distinct examinee classes. For one, examinees made relatively few errors--less than one per item on average--possibly too little to evidence much consistency. Second, the number of items was such that there was limited opportunity to observe repeated errors. Finally, transcription and processing mistakes may have introduced noise. The frequency and gravity of these errors would appear to be low given the extensive checks placed on transcription (see Sebrechts, Bennett, & Rock, 1991) and the high relations of GIDE's scores with both experts' grades and GRE quantitative performance (Bennett, Sebrechts, & Rock, in press; Sebrechts, Bennett, & Rock, 1991). Even subtle transcription and processing mistakes, however, can change the general bug category detected, thereby masking consistency and diluting distinctions among examinees.

The second model explored, HOST, unequivocally fit only a small percentage of examinee responses. The primary source of misfit appeared to be the faulty solution format, which was hypothesized to be the easiest but consistently proved to be the hardest item type. This result suggests that the hypothesized hierarchical structure needs to be modified to recognize this format's greater cognitive complexity. The other formats were not readily distinguishable--possibly because many individuals were able to perform successfully regardless and because others were not clear as to how the formats could be used. This result left open the possibility that, excepting faulty solutions, the proposed hierarchy might fit in a less skilled sample instructed in how the formats might differentially aid problem solving. Transcription and processing errors were less likely to be relevant here because of the positive outcomes of previous investigations using the same numeric score data.

What are the implications of this work for GRE Program research and development? As noted, this study is part of an integrated research program on intelligent assessment (Bennett, in press). This program is pursuing goals related to constructed-response testing, interactive systems incorporating artificial intelligence, and cognitively driven measurement models. With respect to constructed-response testing, the top priority is a more refined task theory to account for the functioning of the various formats, faulty solution in particular. Stronger task theory might, in addition, suggest classes of bugs that are more likely to be associated with one format than another. Finally, it might better clarify the role of missing goals and its meaning for characterizing a subgroup of examinees.

The beginnings of a task theory exist in the goal-plan structures used in this study, which were previously derived from an analysis of open-ended items (Sebrechts, Bennett, & Rock, 1991). Further progress might be achieved through protocol analysis as well as by experimental studies. The former approach should help in building a more finely grained process model for each format (including how examinees use goal-plan information given in the item stem). The latter method might involve testing particular hypotheses developed from that process model. For example, given a model in which providing correct goal-plan structures facilitates problem solving, the same items in different formats might be randomly assigned to examinees (avoiding the confounding sometimes introduced by isomorphs). A second design would require presenting difficult problems by computer in open-ended format with one group getting progressively more goal-plan information if the problem is not solved. (This design could be implemented using the data collection interface being developed under a related GRE-funded effort.)

How might the faulty solution format fit into such a task theory? It is conceivable that examinees work these relatively unfamiliar problems by first solving them as if they were open-ended and then comparing the solution to the given one to identify the discrepancies. Difficulty would increase because of this added step and because the given solution might represent a strategy different from the one followed by the examinee, making comparison more complex. Another possibility is that examinees use the faulty solution to help generate a response but are misled by the given information, and thus replicate parts of the erroneous solution. This supposition might be tested in the present data by looking for a higher incidence of these replicated bugs for the faulty solution format.

Regarding the research program's second goal, the development of interactive item delivery and analysis technology, it should be noted that such delivery eliminates the need for transcription and the errors inevitably associated with it. With this problem aside, emphasis should be on systematically studying the accuracy of GIDE's diagnoses and improving its analytical mechanism. This investigation might be conducted similarly to the prior study of scoring accuracy (Sebrechts, Bennett, & Rock, 1991), which compared GIDE's scores with those of human content experts. Experts might be asked to diagnostically analyze solutions from the existing data set. Expert diagnoses would be compared with the analyses already produced by GIDE and disagreements discussed to ensure that they represented erroneous machine processing. After the processing errors were corrected, the study would be replicated with data collected interactively from a new examinee sample taking both the existing items and an overlapping set now being developed.

The third goal, integrating cognitively driven measurement models, should follow the development of more refined task and bug theories. As stronger theories are posed, the HOST and Hybrid models might again be applied to test theoretical predictions. Other measurement models that should be considered include Mislevy's (in press) inference nets and Tatsuoka's (1983) rule space. These studies should be conducted with more items chosen to provide multiple opportunities for observing theoretically related or otherwise salient bugs (e.g., the "active ingredient" decimal-as-percent bug), as well as with more representative examinees samples. On this latter point, an argument might be made for oversampling students whose skills are somewhat lower than average as these individuals need diagnostic feedback and might be more likely to fall into latent classes. This strategy needs to be carefully considered as error consistency appears to be particularly low among unskilled students (Payne & Squibb, 1990; Tatsuoka, Birenbaum, & Arnold, 1989).

The three goals enunciated by this research program are intended to lead to components for enhancing existing assessment programs and building new products and services. These components might, for example, be incorporated in software for preparing students to take the General Test, in the General Test itself to provide more specific information to examinees, or in new program offerings intended to alert students early in their undergraduate careers to the fundamental skills they need to increase their chances for success in graduate education. As suggested, considerable progress has been achieved in developing innovative item formats, automatically scoring responses, and studying the meaning of the resulting partial-credit scores. In addition, substantial work has been invested in understanding how examinees correctly and incorrectly solve specific problems. This study has taken an initial step toward aggregating that information across items to produce more general qualitative characterizations. Such characterizations will likely play an important role as the GRE and other testing programs expand their foci to satisfy increasing needs for information service.

References

Bennett, R. E. (in press). Intelligent assessment: Toward an integration of constructed-response testing, artificial intelligence, and model-based measurement. In N. Frederiksen, R. J. Mislevy, and I. Bejar (Eds.), Test theory for a new generation of tests. Hillsdale, NJ: Lawrence Erlbaum Associates.

Bennett, R. E., Gong, B., Kershaw, R. C., Rock, D. A., Soloway, E., & Macalalad, A. (1990). Assessment of an expert system's ability to grade and diagnose automatically student's constructed responses to computer science problems. In R. O. Freedle (Ed.), Artificial intelligence and the future of testing (pp. 293-320). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bennett, R. E., Rock, D. A., Braun, H. I., Frye, D., Spohrer, J. C. & Soloway, E. (1990). The relationship of expert-system scored constrained free-response items to multiple-choice and open-ended items. Applied Psychological Measurement, 14, 151-162.

Bennett, R. E., Sebrechts, M. M., & Rock, D. A. (in press). Expert-system scores for complex constructed-response quantitative items: A study of convergent validity. Applied Psychological Measurement.

Birenbaum, M., & Tatsuoka, K. K. (1987). Open-ended versus multiple-choice response formats--It does make a difference for diagnostic purposes. Applied Psychological Measurement, 11, 385-395.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika, 46, 443-459.

Braun, H. I., Bennett, R. E., Frye, D., & Soloway, E. (1990). Scoring constructed responses using expert systems. Journal of Educational Measurement, 27, 93-108.

Falmagne, J. C. (1989). A latent trait theory via a stochastic learning theory for a knowledge space. Psychometrika, 54, 283-303.

Fiske, E. B. (1990, January 31). But is the child learning? Schools trying new tests. The New York Times, pp. 1, B6.

Gitomer, D. H., & Rock, D. A. (in press). Addressing process variables in test analysis. In N. Frederiksen, R. Mislevy, and I. Bejar (Eds.), Test theory for a new generation of tests. Hillsdale, NJ: Lawrence Erlbaum.

Gitomer, D. H., & Yamamoto, K. (1991). Performance modeling that integrates latent trait and class theory. Journal of Educational Measurement, 28, 173-189.

Guthrie, J. T. (1984). Testing higher level skills. Journal of Reading, 28, 188-190.

Guttman, L. (1954). A new approach to factor analysis: The radex. In P. F. Lazarsfeld (Ed.), Mathematical thinking in the social sciences. Glencoe, IL: Free Press.

Johnson, W. L. (1986). Intention-based diagnosis of novice programming errors. Los Altos, CA: Morgan Kaufmann Publishers.

Johnson, W. L., & Soloway, E. (1985). PROUST: An automatic debugger for Pascal programs. Byte, 10(4), 179-190.

Lazarsfeld, P. F. (1960). Latent structure analysis and test theory. In H. Gulliksen and S. Messick (Eds.), Psychological scaling: Theory and applications (pp. 83-86). New York: Wiley.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates.

Masters, G. N., & Mislevy, R. J. (in press). New views of student learning: Implications for educational measurement. In N. Frederiksen, R. J. Mislevy, and I. Bejar (Eds.), Test theory for a new generation of tests. Hillsdale, NJ: Lawrence Erlbaum Associates.

Mislevy, R. J. (1983). Item response models for grouped data. Journal of Educational Statistics, 8, 271-288.

Mislevy, R. J. (in press). A framework for studying differences between multiple choice and free response. In R. E. Bennett & W. C. Ward (Eds.), Choice vs. construction in cognitive measurement. Hillsdale, NJ: Lawrence Erlbaum Associates.

Matz, M. (1982). Towards a process model for high school algebra. In D. H. Sleeman and J. S. Brown (Eds.), Intelligent tutoring systems. London: Academic Press, Inc.

Newell, A., & Simon, H. A. (1972). Human problem solving. Englewood Cliffs, NJ: Prentice-Hall.

Payne, S. J., & Squibb, H. R. (1990). Algebra mal-rules and cognitive accounts of error. Cognitive Science, 14, 445-481.

Rock, D. A., & Pollack, J. (1987). Measuring gains--A new look at an old problem. Paper presented at the Educational Testing Service/Department of Defense Conference, San Diego, CA.

Sebrechts, M. M., Bennett, R. E., & Rock, D. A. (1991). Machine-scorable complex constructed-response quantitative items: Agreement between expert system and human raters' scores (RR-91-11). Princeton, NJ: Educational Testing Service.

Sebrechts, M. M., LaClaire, L., Schooler, L. J., & Soloway, E. (1986). Toward generalized intention-based diagnosis: GIDE. In R. C. Ryan (Ed.), Proceedings of the 7th National Educational Computing Conference (pp. 237-242). Eugene, OR: International Council on Computers in Education.

Sebrechts, M. M., Schooler, L. J., LaClaire, L., & Soloway, E. (1987). Computer-based interpretation of students' statistical errors: A preliminary empirical analysis of GIDE. Proceedings of the 8th National Educational Computing Conference (pp. 143-148). Eugene, OR: International Council on Computers in Education.

Sebrechts, M. M., & Schooler, L. J. (1987). Diagnosing errors in statistical problem solving: Associative problem recognition and plan-based error detection. Proceedings of the Ninth Annual Cognitive Science Meeting (pp. 691-703). Hillsdale, NJ: Lawrence Erlbaum Associates.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. Journal of Educational Measurement, 20, 345-354.

Tatsuoka, K. K., Birenbaum, M., & Arnold, J. (1989). On the stability of students' rules of operation for solving arithmetic problems. Journal of Educational Measurement, 26, 351-361.

VanLehn, K. (1982). Bugs are not enough: Empirical studies of bugs, impasses and repairs in procedural skills. The Journal of Mathematical Behavior, 3(2), 3-71.

Wenger, E. (1987). Artificial intelligence and tutoring systems. Los Altos, CA: Morgan Kaufmann Publishers.

Wilson, D. T., Wood, R., & Gibbons, R. (1987). Testfact: Test scoring, item statistics, and item factor analysis. Mooresville, IN: Scientific Software.

Wilson, M. (1989). Saltus: A psychometric model of discontinuity in cognitive development. Psychological Bulletin, 105, 276-289.

Yamamoto, K. (1989a). Hybrid model of IRT and latent class models (RR-89-41). Princeton, NJ: Educational Testing Service.

Yamamoto, K. (1989b). HYBIL: A computer program to estimate Hybrid model parameters [Computer program]. Princeton, NJ: Educational Testing Service.

Footnotes

1. The association between the four major error categories and the three scoring categories described earlier (computational, structural, missing goal) is indirect. Mathematical errors were considered to be computational and cost one point. General and specific plan errors may be computational (e.g., failing to reduce an answer) or they may more obviously affect the plan structure (e.g., dividing when multiplication was called for), in which case two points were deducted. Missing-goal bugs always cost three points.

2. In part, the uniqueness of this latent class can be attributed to the fact that for any given goal, a missing-goal bug precludes the existence of other errors. This dependency does not appear to be a strong one, however, as evidenced by the close similarity of the response patterns for matched examinees and by the fact that latent class examinees more often than not commit a missing goal in conjunction with a coherent attempt to solve other components of the same problem.

Appendix A

Items

Faulty Solution

A Department of Transportation road crew paves the 15 mile city portion of a 37.4 mile route at the rate of 1.3 miles per day and paves the rest of the route, which is outside the city, at a rate of 2.1 miles per day. If the Department of Transportation starts the project on day 11 of its work calendar, on what day of its work calendar will the project be completed?

The active ingredient is 0.25 percent of a 3-ounce dose of a certain cold remedy. What is the number of doses a patient must take before receiving the full 3 ounces of the active ingredient?

$3.50 in tolls is received each minute at an automated toll booth while the rate at a booth with an operator is $2.80 each minute. How many minutes elapse before the automated booth receives $14.00 more in tolls than does the person-operated booth?

Equation Setup

Of the 720 pages of printed output of a certain program, 305 pages are printed on a printer that prints 15 pages per minute and the rest are printed on a printer that prints at 50 pages per minute. If the printers run one after the other and printing starts at 10 minutes and 15 seconds after the hour, at what time to the nearest second after the hour will the printing be finished?

A graphics designer earns 2% of a $1500 yearly bonus for each shift of overtime she works. How many shifts of overtime must she work to earn the equivalent of the entire yearly bonus?

A specialty chemical company has patented a chemical process that involves 2 reactions. Reaction 1 generates 24 grams of molecule B per minute and reaction 2 consumes 5 grams of molecule B per minute. If 4,560 grams of molecule B are desired as a product of this process, how many minutes must it continue?

Goal Specification

800 gallons of a 2,400 gallon tank flow in at the rate of 75 gallons per hour through a clogged hose. After the hose is unclogged, the rest of the tank is filled at the rate of 250 gallons per hour. At what time to the nearest minute will the filling of the tank be finished if it starts at 5:30 a.m.?

On every $150 load of cement it delivers to a construction site, Acme Cement Company earns a 4 percent profit. How many loads must it deliver to the site to earn $150 in profit?

One of two outlets of a small business is losing $500 per month while the other is making a profit of $1750 per month. In how many months will the net profit of the small business be $35,000?

Open Ended

On a 600-hundred mile motor trip, Bill averaged 45 miles per hour for the first 285 miles and 50 miles per hour for the remainder of the trip. If he started at 7:00 a.m., at what time did he finish the trip (to the nearest minute)?

Money in a certain investment fund earns an annual dividend of 5 percent of the original investment. In how many years will an initial investment of $750 earn total dividends equal to the original investment?

How many minutes will it take to fill a 2,000-cubic-centimeter tank if water flows in at the rate of 20 cubic centimeters per minute and is pumped out at the rate of 4 cubic centimeters per minute?

Appendix B

Scoring Rubric

GRE Quantitative Constructed-Response Scoring Rubric

1. If the student provides two or more solutions, consider only the best one. In general, do not deduct credit if the student explicitly corrects errors.

2. Consider all available information including that in the "Calculations Space."

3. If only the final answer is present and it is correct, give full credit because there is no process on which to make any other decision. In all other cases, the total score for the problem is the sum of the scores for each goal.

4. Each goal is worth 3 points. Deduct points as follows:

a. Deduct 3 points if the goal is missing and is not implicitly satisfied. A goal is considered missing when there is no reasonable attempt to solve for it. A goal is considered to be implicitly satisfied if it can be inferred from other parts of the solution.

b. Deduct 2 points if the goal is present but contains an uncorrected structural error (e.g., inverting the dividend and the divisor, confusing operators). For a goal to be considered present but structurally incorrect, it must be clearly evident that the student is making an attempt--however misguided--to solve the goal (thereby showing awareness that solving for that goal is a step in the problem's solution process). The minimal evidence needed to indicate such an attempt is the presence of a reasonable expression bound to a label that can be unambiguously associated with that goal.

c. Deduct 1 point for each computational error within a present goal. Count as computational errors miscalculations (including those beyond the required level of precision), transcription errors (values incorrectly copied from one part of the problem to another), errors in copying a given from the problem statement, conversion errors (unless otherwise indicated), and, for the last goal only, failing to reduce the final answer to a single value. Only deduct for the same computational error once. For all computational errors, carry through the result to subsequent goals, giving full credit to those subsequent goals if they are structurally and computationally correct given their incorrect input.

d. Deduct 1 point for failing to carry the result of a goal to the required level of precision (i.e., two decimal places or the precision required by the individual problem, whichever is greater).

e. Deduct 0 points if the goal is present and correct. A goal should be considered to be present and correct if (1) the result and the method are correct, (2) the result is correct and the method is not identifiably faulty, or (3) the method is correct and the result is incorrect only because the inputs to the goal appropriately came from a previous goal that incorrectly computed those inputs.

In making the above deductions, try to distinguish between errors that can be explained by a single fault and those that are composites of two or more

faults. The following example could be conceived as a single error in which the student has mistakenly converted a decimal representation to time. This would constitute a single error for which 1 point would be deducted.

Time1 = 10.67
Time1 = 11 hr 7 min

In contrast, the following production could be interpreted as two separable errors, one in failing to round 10.66 to 10.67 (the result of 800/75), and the second in confusing decimal and time representations. For this goal, one point would be deducted for each of these computational mistakes.

Time1 = 800/75
Time1 = 11 hr 6 min

5. Unless the final answer (the value on the ANSWER line) is redundant with the culminating value in the student's solution, treat this final answer as part of the solution proper. That is, in many student solutions the ANSWER line value is <u>not</u> redundant but instead represents the result of the student's last goal. Such values should be included in scoring that goal.

6. Treat as equivalent the various operational notations (e.g., *, x, (), ·); mixed numbers and improper fractions (e.g., $8^1/_3$ and $^{25}/_3$); numbers with and without units (400 and 400 doses); and percentages, decimals, and fraction equivalents (e.g., $^1/_4$%, .25%, .0025, and $^1/_{400}$).

7. Treat as correct a goal that is satisfied except for the presence of a unit conversion if that conversion is made in a subsequent goal. In the example below, treat equivalently the conversion of hours to hours and minutes whether it occurs in goal #5, goal #4, or in goals #1 and #2.

Problem: On a 600-hundred mile motor trip, Bill averaged 45 miles per hour for the first 285 miles and 50 miles per hour for the remainder of the trip. If he started at 7:00 a.m., at what time did he finish the trip (to the nearest minute)?

a. Time1 = 285 miles / 45 miles per hour
   Time1 = 6.33 hours (6.33 hours = 6 hours and 20 minutes)
b. Distance2 = 600 miles - 285 miles
   Distance2 = 315 miles
c. Time2 = 315 miles / 50 mile per hour
   Time2 = 6.3 hours (6.3 hours = 6 hours and 18 minutes)
d. Total time = 6.33 hours + 6.3 hours
   Total time = 6 hours 20 min + 6 hours 18 min
   Total time = 12 hours 38 min
e. Finish time = 7:00 am + 12 hours 38 min (7:00 am + 12.63 hrs = 7:38 pm)
   Finish time = 7:38 pm

8. In some cases, the scoring key for a problem presents two alternative goal decompositions. Score the examinee response according to the decomposition that best characterizes the response. Be sure to use the same maximum scores and the same point deduction rules <u>regardless</u> of the decomposition being used to score the response. Under this rule, partially correct solutions that

follow more efficient decompositions will generally receive more points than similar quality solutions following less efficient decompositions.

9. The minimum score for a goal is 0 as is the minimum total score for a solution.

Appendix C

Bug Definitions

Canonical Solutions

Below are standard-form solutions. Following these canonical solutions are uniquely named and numbered bugs, their definitions, and examples. Examples are shown as deviations from the canonical solutions. The solution from which the example deviates is indicated by D=RT for distance problems, % for percent problems, and WORK for work problems. In most cases only the relevant modified lines are given. These lines are in most cases numbered and matched to the canonical version.

D=RT (Five-Goal Problems)

1. Time1 = 285 miles / 45 miles per hour
2. Time1 = 6.33 hours
3. Distance2 = 600 miles - 285 miles
4. Distance2 = 315 miles
5. Time2 = 315 miles / 50 miles per hour
6. Time2 = 6.3 hours
7. Total time = 6.33 hours + 6.3 hours
8. Total time = 6 hours 20 min + 6 hours 18 min
9. Total time = 12 hours 38 min
10. Finish Time = 7:00 am + 12 hours 38 min
11. Finish Time = 7:38 pm

Percent (%) (Three-Goal Problems)

Solution A:

1. 5% = .05
2. Annual Dividend = .05 * $750
3. Annual Dividend = $37.50
4. Investment Time = $750 / $37.50 per year
5. Investment Time = 20 years

Solution B:

1. 5% dividend per year * X years = 100% dividend
2. X years = 100% dividend / 5 % dividend per year
3. X = 20 years

Work (Two-Goal Problems)

1. Net filling rate = 20 cc per minute - 4 cc per minute
2. Net filling rate = 16 cc per minute
3. Filling Time = 2000 cc / 16 cc per minute
4. Filling Time = 125 minutes

Math Bugs

Remainder as Decimal (103): The remainder of a division is treated as a decimal. For example, 10/3 = 3.1

D=RT 2. Time1 = 6.15 hours

Close Enough (105): The value is not exact, but is accepted as being within a reasonable margin of error. This is used to catch potential spurious deviations, which are not precision errors (e.g. 8796 for 8795).

D=RT 4. Distance2 = 312 miles

Close-Enough Units (106): Obtained and expected values match up to the units place, but fail to match in the tenths place.

D=RT 2. Time1 = 6 hours

Close-Enough Tenths (107): Obtained and expected values match up to the tenths place, but fail to match in the hundredths place. Close-enough bugs are tested in order. First, close-enough-tenths is tested, then close-enough-units.

D=RT 2. Time1 = 6.30 hours

Decimal Shift (108): The obtained value has a shifted decimal with respect to the expected value.

D=RT 2. Time1 = 63.3 hours

General Plan Bugs

No Reduction (201): An expression is not sufficiently reduced. This bug is reported only if the nonreduced value is not resolved later in the solution.

D=RT
1. Time1 = 285 miles / 45 miles per hour
2. Time1 = 6.33 hours
3. Distance2 = 600 miles - 285 miles
4. Distance2 = 315 miles
5. Time2 = 315 miles / 50 miles per hour
7. Total time = 6.33 hours + 6.3 hours
8. Total time = 6 hours 20 min + 6 hours 18 min

No Final Reduction (202): The "final" answer to the problem is not reduced. It is like the no-reduction bug, but applies to the final goal. The difference is largely technical, indicating the way in which the bugs are treated during processing. A no-reduction bug can be subsequently resolved. A no-final-reduction bug cannot.

D=RT 1. Time1 = 285 miles / 45 miles per hour
2. Time1 = 6.33 hours
3. Distance2 = 600 miles - 285 miles
4. Distance2 = 315 miles
5. Time2 = 315 miles / 50 miles per hour
6. Time2 = 6.3 hours
7. Total time = 6.33 hours + 6.3 hours
8. Total time = 6 hours 20 min + 6 hours 18 min
9. Total time = 12 hours 38 min
10. Finish Time = 7:00 am + 12 hours 38 min

Times for Divide (203): The student uses multiplication where division is required.

D=RT 1. Time1 = 285 miles * 45 miles per hour

Divide for Times (204): The student uses division where multiplication is required.

%A 2. Annual Dividend = .05 / $750

Add for Subtract (205): Addition is used where subtraction is required.

D=RT 3. Distance2 = 600 miles + 285 miles

Unknown Value (206): The student uses a value in his or her solution that fits the required structure, but that cannot otherwise be accounted for by math or more specific plan errors.

D=RT 1. Time1 = 285 miles * 45 miles per hour
2. Time1 = 5

Out-of-Plan Value (207): A value used in a subsequent goal fits into the overall plan structure, but it is unclear what the source of the value is.

D=RT 7. Total time = 5.6 hours + 5.3 hours
8. Total time = 5 hours 36 min + 5 hours 18 min
9. Total time = 10 hours 54 min
10. Finish Time = 7:00 am + 10 hours 54 min
11. Finish Time = 5:54 pm

Unexplained Value (208): A plan has a single unexplained value within the correct structure. This bug is used in INTEREST and WORK problems. D=RT uses more specific bugs, 401 and 402.

WORK 3. Filling time = 2000 cc / 64 cc per minute

Common Symbol (209): This bug is triggered whenever the label is used for identification. It indicates that the combination of other plans and bugs could not adequately explain the value. At the same time, the value assigned to the bug must be within a reasonable range of the expected value, in this case within 1% of the larger of the obtained and expected values.

D=RT 3. absent
4. Distance2 = 312 miles

Specific Plan Bugs

D=RT Time Bugs

Remainder as Time (301): The remainder in a division is treated as a time unit.

D=RT 1. Time1 = 285 miles / 45 miles per hour
2. Time1 = 6 hours 15 min

Decimal as Time (302): The decimal portion of a division is treated as a time unit.

D=RT 1. Time1 = 285 miles / 45 miles per hour
2. Time1 = 6 hours 33 min

Decimal-Portion as Time (303): The same as decimal-as-time (302) except that it does not assume access to the current match frame.

D=RT 600 m - 285 m = 315 m
285 m / 45 min = 6.33 hr
315 / 50 min = 6.3 hr
6.33 * 45 = 28485
6.33 + 6.30 = 12.63
6.33 + 6.30 = 13.03 hrs
7 + 13.03 = 8:03 p.m.
ANSWER = 8:03 pm.

Time as Decimal (304): A time unit is treated as a decimal.

D=RT 1. Time1 = 285 miles / 45 miles per hour
2. Time1 = 6 hours 20 min
Time1 = 6.20 hours

AM/PM Shift (305): The student changes the time from am to pm or pm to am. This bug is sometimes triggered by student failure to indicate either am or pm, in which case am is assumed.

D=RT 11. Finish Time = 7:38 am

Close Enough Minutes (306): The observed and expected values are not exactly the same, but are within 1 minute of each other.

D=RT 2. Time1 = 6.34 hours

Other D=RT Bugs

Unexplained Rate (401): The value for the rate in a structurally matched plan is incorrect.

D=RT 1. Time1 = 285 miles / 55 miles per hour

Unexplained Distance (402): The value for the distance in a structurally matched plan is incorrect.

D=RT 3. Distance2 = 500 miles - 285 miles

Rate2 for Rate1 (404): The second rate is used in place of the first rate.

D=RT 1. Time1 = 285 miles / 50 miles per hour

Rate1 for Rate2 (405): The first rate is used in place of the second rate.

D=RT 5. Time2 = 315 miles / 45 miles per hour

Whole for Part Distance (406): The total distance is used to determine one of the partial times instead of the partial distance.

D=RT 5. Time2 = 600 miles / 50 miles per hour

Part for Whole Time (407): The ending time uses only one of the elapsed times instead of the total elapsed time.

D=RT 10. Finish Time = 7:00 am + 6 hours 20 min

Unknown Time1, Unknown Time2, Unknown Total Time (411, 412, 413): These three "Unknown" bugs are part of a single plan that finds a global solution structure in the absence of reasonable constituent values. If only a single bug occurs or the numbers are reasonable deviations, the solution will usually be matched by a series of individual bugs or by the use of implicit matching. The following example incorporates the three "unknown" bugs (time1, time2, and total time).

D=RT 1. Time1 = 200 miles / 40 miles per hour
2. Time1 = 5 hours
5. Time2 = 300 miles / 60 miles per hour
6. Time2 = 5 hours
7. Total time = 5 hours + 5 hours
9. Total time = 10 hours
10. Finish Time = 7:00 am + 10 hours
11. Finish Time = 5:00 pm

Unweighted Average Rate (414): An average rate is computed improperly since it is not weighted according to the different times.

D=RT Average Rate = (15 pages per min + 50 pages per min)/2
Average Rate = 32.5 pages per min
Average Time = 720 pages / 32.5 pages per min
Average Time = 22.15 min

Percent Bugs

Percent as Decimal (501): The percent value is treated as a decimal (e.g. 5% is treated as 5.0).

%A 1. absent
2. Annual Dividend = 5 * $750
3. Annual Dividend = $3750
4. Investment Time = $750 / $3750 per year
5. Investment Time = 0.20 years

Decimal as Percent (502): A decimal value is treated as a percent (e.g., .05 as if it were .05% or .0005).

%A 2. Annual Dividend = 0.05 * $750
3. Annual Dividend = $0.3750

Mixed Percent and Decimal (503): The student mixes decimal and percent values.

%B 1. 0.05 dividend per year * X years = 100% dividend
2. X years = 100% dividend / 0.05 dividend per year
3. X = 2000 years

Result per Unit-Amount/Rate (504): The dividend is calculated as a division of investment by rate instead of a multiplication of investment times the rate. This bug also applies to calculating the profit per load, amount earned per shift, and active ingredient per dose.

%A 2. Annual Dividend = $750 / 0.05

Work Bugs

Subtract Order (603): The items are subtracted in the wrong order (i.e., volume out minus volume in).

WORK 1. Net filling rate = 4 cc per minute - 20 cc per min

Missing Goals

These bugs indicate that the stated goal is missing. No goal can have any other associated bug if it is "missing."

D-RT Bugs

Missing First Goal (911): Time1 is missing. The goal takes the canonical form, Time1 = Distance1 / Rate1.

D-RT
3. Distance2 = 600 miles - 285 miles
4. Distance2 = 315 miles
5. Time2 = 315 miles / 50 miles per hour
6. Time2 = 6.3 hours
10. Finish Time = 7:00 am + 12 hours 38 min
11. Finish Time = 7:38 pm

Missing Second Goal (912): Distance2 is missing. The goal takes the canonical form, Distance2 = Total Distance - Distance1.

D-RT
1. Time1 = 285 miles / 45 miles per hour
2. Time1 = 5.7 hours
5. Time2 = 2.85 miles / 50 miles per hour
6. Time2 = 6.3 hours
7. Total time = 5.7 hours + 6.3 hours
8. Total time = 5 hours 42 min + 6 hours 18 min
9. Total time = 12 hours
10. Finish Time = 7:00 am + 12 hours
11. Finish Time = 7:00 pm

Missing Third Goal (913): Time2 is missing. The goal takes the canonical form, Time2 = Distance2 / Rate2.

D-RT
1. Time1 = 285 miles / 45 miles per hour
2. Time1 = 6.33 hours
3. Distance2 = 600 miles - 285 miles
4. Distance2 = 315 miles
10. Finish Time = 7:00 am + 6 hours 20 min
11. Finish Time = 1:20 pm

Missing Fourth Goal (914): Total Time is missing. The goal takes the canonical form, Total Time = Time1 + Time2.

D-RT
1. Time1 = 285 miles / 45 miles per hour
2. Time1 = 6.33 hours
3. Distance2 = 600 miles - 285 miles
4. Distance2 = 315 miles
5. Time2 = 315 miles / 50 miles per hour
6. Time2 = 6.3 hours
10. Finish Time = 7:00 am + 6 hours 20 min
11. Finish Time = 1:20 pm

Missing Fifth Goal (915): Finish Time is missing. The goal takes the canonical form, Finish Time = Start Time + Total Time.

```
D=RT  1. Time1 = 285 miles / 45 miles per hour
      2. Time1 = 6.33 hours
      3. Distance2 = 600 miles - 285 miles
      4. Distance2 = 315 miles
      5. Time2 = 315 miles / 50 miles per hour
      6. Time2 = 6.3 hours
      7. Total time = 6.33 hours + 6.3 hours
      8. Total time = 6 hours 20 min + 6 hours 18 min
      9. Total time = 12 hours 38 min
```

Percent Bugs

Missing First Goal (921): The Percent Conversion is missing. The goal takes the canonical form, Decimal = .01 * percent

Missing Second Goal (922): Annual Dividend is missing. The goal takes the canonical form, Annual Dividend = Rate Per Year * Investment.

```
%A    4. Investment Time = $750 / 5
      5. Investment Time = 150 years
```

Missing Third Goal (923): Investment Time is missing. The goal takes the canonical form, Investment Time = Investment / Dividends Per Year.

```
%A    1. 5% = .05
      2. Annual Dividend = .05 * $750
      3. Annual Dividend = $37.50
```

Work Bugs

Missing First Goal (931): Net Filling Rate is missing. The goal takes the canonical form, Net Filling Rate = Rate In - Rate Out.

```
WORK  3. Filling time = 2000 cc / 20 cc per minute
      4. Filling time = 100 minutes
```

Missing Second Goal (932): Filling Time is missing. The goal takes the canonical form, Fill Time= Volume / Net Rate.

```
WORK  1. Net filling rate = 20 cc per minute - 4 cc per minute
      2. Net filling rate = 16 cc per minute
```

Appendix D

Bug Incidence by Item

Number of Examinees Making One of More Instances of a Specific Bug on an Item

	Two-Goal Items				Three-Goal Items				Five-Goal Items			
Bug	$3.50 Tolls (273)	Chem. Comp. (284)	Small Bus. (277)	2000 cc. (274)	Active Ingrd. (271)	Graph (283)	Load Cement (275)	Invst (272)	DOT Crew (278)	720 Pages (284)	2400g Tank (283)	600-Mile (269)
Math												
103	0	0	0	0	0	0	0	0	0	2	2	3
105	0	6	0	3	0	0	0	0	0	2	0	0
106	4	6	2	0	1	0	5	3	44	67	74	53
107	1	5	1	0	0	1	2	5	111	62	45	74
108	18	20	6	13	9	11	5	14	13	6	11	1
Gen.												
201	3	2	1	1	2	2	2	1	2	2	3	2
202	4	0	3	0	1	0	2	0	0	0	7	1
203	0	0	0	0	26	0	3	1	0	0	0	0
204					0	0	0	0				
205	23	26	1	4					0			
206	5	23	36	9	28	14	22	25	55	69	71	40
207	0	0	0	0	0	0	0	0	2	1	0	1
208	13	12	6	21	2	2	0	1	0	0	0	0
209	0	0	2	0	0	0	0	0	1	0	3	0
Spec												
301	0	0		0						2	0	3
302	0	0		0						6	2	3
303	0	0		0						19	16	15
304	0	0		0						28	0	0
305											5	6
306	0	0		0						2	59	1
401									0	4	1	2
402									9	14	4	7
404									0	0	0	1
405									0	0	0	0
406									12	2	9	3
407									2	1	0	3
411									0	1	0	0
412									0	1	0	0
413									11	7	7	0
414									0	0	0	0
501					161	5	9	10				
502					0	0	0	0				
503					5	0	0	0				
504					30	4	10	25				
603	0	0	0	0								
Miss												
911									22	18	24	6
912									31	19	21	6
913									26	20	20	12
914									34	29	29	23
915									58	40	55	47
921					7	12	1	10				
922					7	12	1	11				
923					47	21	5	39				
931	30	14	6	37								
932	44	17	12	29								

Note. Bug code definitions are listed in Appendix C. An empty cell indicates that the bug could not occur on that problem. The number of examinees responding to each item is given in parentheses.

Specific Bugs as Percentages of the Number of Examinees Responding to an Item

	Two-Goal Items				Three-Goal Items				Five-Goal Items			
Bug	$3.50 Tolls (273)	Chem. Comp. (284)	Small Bus. (277)	2000 cc. (274)	Active Ingrd. (271)	Graph (283)	Load Cement (275)	Invst (272)	DOT Crew (278)	720 Pages (284)	2400g Tank (283)	600-Mile (269)
Math												
103	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	1%	1%
105	0%	2%	0%	1%	0%	0%	0%	0%	0%	1%	0%	0%
106	1%	2%	1%	0%	0%	0%	2%	1%	16%	24%	26%	20%
107	0%	2%	0%	0%	0%	0%	1%	2%	40%	22%	16%	28%
108	7%	7%	2%	5%	3%	4%	2%	5%	5%	2%	4%	0%
Gen.												
201	1%	1%	0%	0%	1%	1%	1%	0%	1%	1%	1%	1%
202	1%	0%	1%	0%	0%	0%	1%	0%	0%	0%	2%	0%
203	0%	0%	0%	0%	10%	0%	1%	0%	0%	0%	0%	0%
204					0%	0%	0%	0%				
205	8%	9%	0%	1%					0%			
206	2%	8%	13%	3%	10%	5%	8%	9%	20%	24%	25%	15%
207	0%	0%	0%	0%	0%	0%	0%	0%	1%	0%	0%	0%
208	5%	4%	2%	8%	1%	1%	0%	0%	0%	0%	0%	0%
209	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%	1%	0%
Spec												
301	0%	0%		0%						1%	0%	1%
302	0%	0%		0%						2%	1%	1%
303	0%	0%		0%						7%	6%	6%
304	0%	0%		0%						10%	0%	0%
305											2%	2%
306	0%	0%		0%						1%	21%	0%
401									0%	1%	0%	1%
402									3%	5%	1%	3%
404									0%	0%	0%	0%
405									0%	0%	0%	0%
406									4%	1%	3%	1%
407									1%	0%	0%	1%
411									0%	0%	0%	0%
412									0%	0%	0%	0%
413									4%	2%	2%	0%
414									0%	0%	0%	0%
501					59%	2%	3%	4%				
502					0%	0%	0%	0%				
503					2%	0%	0%	0%				
504					11%	1%	4%	9%				
603	0%	0%	0%	0%								
Miss												
911									8%	6%	8%	2%
912									11%	7%	7%	2%
913									9%	7%	7%	4%
914									12%	10%	10%	9%
915									21%	14%	19%	17%
921					3%	4%	0%	4%				
922					3%	4%	0%	4%				
923					17%	7%	2%	14%				
931	11%	5%	2%	14%								
932	16%	6%	4%	11%								

Note. Bug code definitions are listed in Appendix C. An empty cell indicates that the bug could not occur on that problem. The number of examinees responding to each item is given in parentheses.

54020-07473 • Y121M.7 • 209075