DOCUMENT RESUME

ED 392 814 TM 024 452

AUTHOR	Bennett, Randy Elliot; And Others
TITLE	The Convergent Validity of Expert System Scores for Complex Constructed-Response Quantitative Items. GRE Research. GRE Board Professional Report No. 88-07bP.
INSTITUTION	Educational Testing Service, Princeton, N.J.
SPONS AGENCY	Graduate Record Examinations Board, Princeton, N.J.
REPORT NO	ETS-RR-91-12
PUB DATE	Jun 91
NOTE	41p.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*College Students; *Constructed Response; Correlation; *Expert Systems; Goodness of Fit; Higher Educati'n; Mathematical Models; Mathematics Tests; Matrices; *Scores; Test Format; Test Items; *Test Validity
IDENTIFIERS	Confirmatory Factor Analysis; *Convergent Validation; Graduate Record Examinations; Scoring Rubrics

ABSTRACT

This study investigated the convergent validity of expert-system scores for four mathematical constructed-response item formats. A five-factor model was proposfd comprised of four constructed-response format factors and a Graduate Record Examinations (GRE) General Test quantitative factor. Subjects were drawn from examinees taking a single form of the GRE General Test in June 1989. The final study sample consisted of 249 test takers. Confirmatory factor analysis was used to test the fit of this model and to compare it with several alternatives. The five-factor model fit well, although a solution comprised of two highly correlated dimensions--GRE quantitative and constructed-response--represented the data almost as well. These results extend the meaning of the expert system's constructed-response scores by relating them to a well-established quantitative measure and by indicating that they signify the same underlying proficiency across item formats. Appendices present test item stems, the scoring rubrics, and an estimated correlation matrix. (Contains 1 figure, 8 tables, and 34 references.) (Author/SLD)

GRE®

RESEARCH

The Convergent Validity of Expert System Scores for Complex Constructed-Response Quantitative Items

Randy Elliot Bennett
Marc M. Sebrechts
and
Donald A. Rock

June 1991

GRE Board Professional Report No. 88-07bP
ETS Research Report 91-12

Educational Testing Service, Princeton, New Jersey

The Convergent Validity of Expert System Scores for Complex Constructed-Response Quantitative Items

Randy Elliot Bennett
Marc M. Sebrechts
and
Donald A. Rock

GRE Board Report No. 88-07bP

June 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

Abstract

This study investigated the convergent validity of expert-system scores for four mathematical constructed-response item formats. A five-factor model was posed comprised of four constructed-response format factors and a GRE General Test quantitative factor. Confirmatory factor analysis was used to test the fit of this model and to compare it with several alternatives. The five-factor model fit well, although a solution comprised of two highly correlated dimensions--GRE-quantitative and constructed-response--represented the data almost as well. These results extend the meaning of the expert system's constructed-response scores by relating them to a well-established quantitative measure and by indicating that they signify the same underlying proficiency across item formats.

The Convergent Validity of Expert System Scores for Complex Constructed-Response Quantitative Items

Large-scale testing programs like the Graduate Record Examinations (GRE) have built their operations around the multiple-choice item, which provides an efficient, objective cognitive measure. The multiple-choice format has, however, been criticized because it putatively measures lower-level skills than less restricted formats, provides limited opportunity for partial credit and little diagnostic information, does not faithfully reflect the tasks performed in academic settings, and encourages students to focus on learning decontextualized facts (Fiske, 1990; Frederiksen & Collins, 1989; Guthrie, 1984; Nickerson, 1989).

Some of the supposed deficiencies of the multiple-choice format might be addressed by complex constructed-response items (Bennett, in press). A complex constructed-response is one for which scoring decisions cannot typically be made immediately and unambiguously using mechanical application of a limited set of explicit criteria, but rather require expert judgment. Such items can be designed to reflect "real-life" tasks more accurately, support partial-credit scoring, facilitate instructional diagnosis through analysis of solution processes, and highlight behaviors considered important to success in academic settings.

The primary impediment to using these items in large-scale testing programs is scoring, which typically must be done at great expense by human judges over several days or weeks. With plans for introducing computerized administration in large programs ("ETS research plan," 1989), the automated presentation of constructed-response questions becomes plausible. Moreover, advances in expert systems--computer programs that emulate the behavior of a human content specialist--make immediate scoring of even relatively lengthy responses a possibility (Bennett, Gong, Kershaw, Rock, Soloway, & Macalalad, 1990; Braun, Bennett, Frye, & Soloway, 1990).

The meaning of scores produced by these systems can be evaluated from several perspectives including agreement with human judges and relations with established tests. This study assesses the convergent validity of expert-system scores for four complex constructed-response mathematical formats. The relations of the formats among themselves and to the GRE General Test's quantitative section are examined.

Method

Subjects

Subjects were drawn from a pool of more than 50,000 examinees taking a single form of the GRE General Test administered nationally in June 1989. Examinees living within approximately 30 miles of an ETS field office were identified and asked by letter to participate for a fee in research to develop instructionally relevant test items.[1] Expressions of interest were received from 1,236 of 3,244 individuals contacted. Respondents were removed from consideration if they were not on the original mailing list, if their citizenship could not be determined from their GRE registration data, or if they no longer lived near an ETS office. From the remaining group, up to the first 100 persons from each region were selected, with some individuals replaced to produce within-region samples composed of citizens and noncitizens in proportions similar to the General Test population. Attempts were made to schedule the resulting 684 people, of whom 285 participated. Twenty-four examinees were deleted because they did not provide complete response data; an additional 11 were used for testing and revising the expert scoring system, leaving a final study sample of 249.[2]

Table 1 presents General Test scores and Biographical Information Questionnaire data for the sample and population taking the June 1989 administration. As can be seen, the sample differed somewhat from the population. The sample's General Test performance was significantly higher (by .5, .4, and .4 standard deviations, for verbal, quantitative, and analytical, respectively), and the most consequential of several statistically significant demographic differences was in a greater proportion of nonwhites.

Instruments

Constructed-response items. Three prototype items were selected from standard, five-option multiple-choice algebra word problems appearing on disclosed forms of the quantitative section of the General Test. One prototype each was drawn from the *rate x time*, *interest*, and *work* content classes. Three "isomorphs" for each prototype were written to produce a set of four items intended to differ in surface characteristics (e.g., topic, linguistic form), but not underlying structure (i.e., the operations for solving the problem). The resulting 12 items (see Appendix A) were divided among four formats such that each isomorphic item in a content class appeared in a different format. The formats were open-ended (only the problem stem is presented and the student must provide a step-by-step solution); goal specification (the problem stem, a list of givens, and a list of unknowns is presented); equation setup (the problem stem and the equations to identify the unknowns are given); and faulty solution (the stem is presented with an erroneous solution for the student to correct). The formats impose different degrees of response constraint (Bennett, Ward, Rock, & LaHart, 1990) and, consequently, would seem to present qualitatively different cognitive tasks. Examples of each format appear as Figure 1. (See Sebrechts, Bennett, & Rock, 1991, for a more detailed description of the item development process.)

Table 1

Background Data for Study Sample

	June 1989	
Variable	Population	Sample
N	50,548	249
General Test Performance		
Verbal Mean(SD)	476(122)	534(130)*
Quantitative Mean (SD)	532(140)	583(137)*
Analytical Mean (SD)	513(132)	568(127)*
Percentage Female	55%	60%
Percentage Non-White[a]	16%	27%*
Percentage U.S. Citizens	79%	84%
Undergraduate Major		
Business	4%	2%
Education	14%	5%*
Engineering	13%	13%
Humanities/Arts	14%	21%*
Life Sciences	18%	19%
Physical Sciences	10%	9%
Social Sciences	18%	23%*
Other	9%	8%
Intended Graduate Major		
Business	2%	2%
Education	18%	11%*
Engineering	10%	9%
Humanities/Arts	8%	8%
Life Sciences	16%	15%
Physical Sciences	8%	9%
Social Sciences	13%	19%*
Other	11%	8%
Undecided	15%	19%*

* $p < .05$, two-tailed $\underline{z}$-test of sample value with total test population parameter.

[a] U.S. citizens only.

Figure 1
Isomorphic problems 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.

A rubric and key for scoring the items were designed in consultation with ETS mathematics test development staff (see Appendix B). The rubric and key were based on decompositions of the solution process associated with each item, where a decomposition constituted the set of goals (i.e., intermediate and terminal objectives) for achieving a solution (e.g., for the first item in Figure 1, compute the net filling rate, compute the time to fill the tank). Decompositions were derived from a cognitive analysis of expert and novice responses to open-ended versions of the three prototype items (see Sebrechts, Bennett, & Rock, 1991).

Score scales were based on the number of goals required for solution. Because the number of goals differed across the three problem classes, questions were graded on different scales: 0-6 for the *work* items, 0-9 for *interest*, and 0-15 for *rate*. Points were deducted for missing goals (i.e., solution components), structural errors (e.g., dividing when a multiplication should have occurred), and computational mistakes.

Expert system. Students' constructed responses were scored by GIDE (Sebrechts, LaClaire, Schooler, & Soloway, 1986), an expert system that was designed in earlier versions to detect student errors in statistics questions and rewritten to analyze solutions to open-ended algebra word problems. For each problem, GIDE has access to a knowledge base of goals and plans (i.e., step-by-step procedures for solving a goal) derived from the cognitive analysis. GIDE scores solutions by (1) identifying in its knowledge base the set of goals for solving a problem, (2) comparing portions of the student's response to correct plans for achieving those goals, and (3) where a match is not found, comparing those portions with common faulty plans for solving the goals. On the basis of the faults detected, diagnostic comments are produced and numeric scores assigned based on the scoring rubric and key. A companion study investigated agreement between GIDE's scores and those of content experts (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, impressive levels of scoring reliability given the complexity of the item responses.

GRE General Test. The General Test is a multiple-choice examination designed to measure broad, developed abilities generally required for success in graduate work. The test is composed of three sections--quantitative, verbal, and analytical --two of which were used in this study.

The quantitative section is meant to measure basic mathematical skills, understanding of elementary mathematical concepts, and ability to reason quantitatively (Educational Testing Service, 1989a). The section's 60 questions are administered in two half-hour blocks. Items are divided among real (i.e., practical problems) and pure arithmetic, algebra, and geometry, and are presented in three formats: quantitative comparison (comparing the relative sizes of two quantities or discerning that not enough information is available), discrete quantitative (containing all the information needed to answer the item), and data interpretation (based on information presented in tables or graphs).

The verbal section is intended to test the examinee's ability to reason with words in solving problems. It is administered in two half-hour segments and contains 76 items falling into four categories (analogies, antonyms, sentence completion, and reading comprehension).

The psychometric characteristics of the quantitative and verbal sections have been extensively studied. For example, factor analytic investigations have repeatedly supported the existence of distinguishable quantitative and verbal dimensions that are stable across population subgroups and related to demographic variables in predictable ways (Rock, Bennett, & Jirele, 1988; Rock, Werts, & Grandy, 1982; Stricker & Rock, 1987; Swinton & Powers, 1980). Predictive validity analyses have found correlations with first-year grades averaged across 606 graduate departments to be .28 for quantitative and .29 for verbal, only slightly lower than that for undergraduate grade-point-average (Educational Testing Service, 1989b). Finally, the median internal consistency reliabilities computed from multiple test-analysis samples were .93 and .91 for quantitative and verbal, respectively.

Procedure

General Test scores and Biographical Information Questionnaire data for all examinees were drawn from ETS files. Constructed-response items were presented in individual and small group sessions conducted at ETS field offices. Examinees were asked to complete the problems at their own pace, though a one-hour period was suggested.

To reduce the chances of examinees recognizing isomorphic relations, the three problems of a given format were presented together and examinees were asked to complete items in sequence without referring to earlier work. In addition, to limit recall each format was separated by "filler" questions--two General Test multiple-choice items taken from quantitative content areas other than interest, rate x time, and work. Finally, items were presented in two orders (most to least constrained and the reverse), given to random halves of the sample at each location. These orders permitted some degree of control over an order effect in which solutions to the more constrained items provide guidance in solving the less constrained ones.

Items were presented in paper-and-pencil format. Handwritten responses were transcribed to machine-readable form according to rules intended to place solution elements into linear order and to translate each line to a syntactically correct equation. (See Sebrechts, Bennett, & Rock, 1991, for the transcription rules.) To assure that responses were consistently transcribed, two coders independently typed a random sample of 14 examinees' responses to each of the 12 problems. Each set of responses was then scored by GIDE and the Pearson product-moment correlation between the two sets computed. This analysis produced a median correlation of .96, with the lowest value at .87 and the 11 remaining ones above .90. Also relevant is the previously cited scoring reliability analysis (Sebrechts, Bennett, & Rock, 1991). In this analysis, content experts graded examinees' original solutions, whereas GIDE graded the transcribed versions. The high correlations between GIDE and the experts suggests that the transcriptions generally captured the substance of the examinee productions.

Data Analysis

This investigation used confirmatory factor analysis (CFA). In contrast with exploratory methods, CFA is intended primarily for testing hypotheses about covariance structures. The proposed theoretical model hypothesized, first, that the factor(s) underlying GIDE's constructed-response scores would be substantially related to the dimension indicated by the GRE-quantitative section, a reasonably well-established mathematical ability measure. This structural relation should be less than unity, however, not only because of format differences with the General Test's multiple-choice quantitative section, but also because of that test's broader content coverage (arithmetic, algebra, and geometry vs. algebra only) and more stringent timing.

Second, the theoretical model posited that scores on the four constructed-response formats would measure related, but distinguishable, dimensions. The limited psychometric work undertaken in quantitative domains offers little evidence of format differences (Traub, in press; Bridgeman, in press). Work in cognitive psychology, however, does suggest an influence on problem solving. Newell and Simon (1972) conceptualize problem solving as a search for a path from given information to goals, where the path constitutes a solution method. The four constructed-response formats offer varying degrees of given information, thus differentially constraining the search and conceivably calling into play moderately different skills.

To illustrate, algebra word problems appear to cluster into families that share similar solution paths (Mayer, 1981); expertise in solving these problems is in part the ability to recognize a problem's class and retrieve a "template" representing the appropriate path (Hinsley, Hayes, & Simon, 1977). Success on open-ended formats might, therefore, depend upon the extent to which this process has been schematized and can be rapidly executed, as well as on the procedural knowledge to solve the specific equations derived from the template and the given information. In contrast, equation setup problems provide a template in the stimulus and, relative to open-ended questions, should call more on procedural skills.

In the present study, these hypotheses were tested by posing a five-factor model composed of GRE-quantitative, open-ended, goal-specification, equation-setup, and faulty-solution factors in which the factors were assumed to be correlated. Each of the five factors was marked by three variables (see Table 2). The lambdas indicate that a factor loading was to be estimated; a zero denotes that the indicator was constrained not to load on that factor. This zero constraint was imposed to make each factor as pure as possible with respect to item format. Consequently, the factor intercorrelations should reflect any format-related differences in covariance structure more clearly.

For the GRE-quantitative factor, each variable was a parcel of 20 quantitative section items constructed by randomly sampling from each of the six test specification content areas in turn: real arithmetic, pure arithmetic, real algebra, pure algebra, real geometry, and pure geometry. The resulting parcels were, consequently, parallel in content and difficulty, and therefore more apt to produce a single quantitative factor against which to compare the constructed-response formats.[3] Parcels were scored on a 21-point number-right scale. Each of the remaining factors was indicated by three constructed-response problems of the same format, with each problem scored on a 7-, 10-, or 16-point scale depending on its content class.

Means and standard deviations for each of the markers are presented in Table 3. As the table suggests, the distributions for the constructed-response indicators were often extremely curtailed, with many examinees scoring in the upper portions of the score range.

Because the distributions were curtailed, the PRELIS program (Joreskog & Sorbom, 1986) was used to estimate the sample product-moment correlation matrix for uncensored distributions (see Appendix C). The maximum likelihood procedure from LISREL VII (Joreskog & Sorbom, 1988) was then employed to estimate the unknown factor loadings.[4] This procedure was used instead of a distribution-free procedure because the latter methods are not yet well understood, and, consequently, there are no clear criteria for determining when they are to be preferred (Joreskog & Sorbom, 1988). However, because the maximum likelihood procedure assumes multivariate normality, its estimates of parameter standard errors should, in this case, be cautiously interpreted.

Table 2

Hypothesized Factor Model

	Factor				
Marker Variable	GRE-Q	Open Ended	Goal Spec.	Equation Setup	Faulty Solution
Quantitative-A (20)	λ	0	0	0	0
Quantitative-B (20)	λ	0	0	0	0
Quantitative-C (20)	λ	0	0	0	0
Open ended-A (1)	0	λ	0	0	0
Open ended-B (1)	0	λ	0	0	0
Open ended-C (1)	0	λ	0	0	0
Goal specification-A (1)	0	0	λ	0	0
Goal specification-B (1)	0	0	λ	0	0
Goal specification-C (1)	0	0	λ	0	0
Equation setup-A (1)	0	0	0	λ	0
Equation setup-B (1)	0	0	0	λ	0
Equation setup-C (1)	0	0	0	λ	0
Faulty solution-A (1)	0	0	0	0	λ
Faulty solution-B (1)	0	0	0	0	λ
Faulty solution-C (1)	0	0	0	0	λ

Note. The number of items per indicator is in parentheses. For the constructed-response variables, A, B, and C indicate five-goal (rate), three-goal (interest), and two-goal (work) problems, respectively.

Table 3

Means and Standard Deviations for Marker Variables

Marker Variable	Scale	Mean	Standard Deviation
Quantitative-A	0-20	13.7	4.2
Quantitative-B	0-20	14.3	3.9
Quantitative-C	0-20	13.4	4.0
Open ended-A	0-15	12.7	3.3
Open ended-B	0-9	7.8	2.3
Open ended-C	0-6	5.0	2.0
Goal specification-A	0-15	12.3	3.6
Goal specification-B	0-9	8.6	1.2
Goal specification-C	0-6	5.6	1.0
Equation setup-A	0-15	12.6	3.4
Equation setup-B	0-9	8.3	1.9
Equation setup-C	0-6	5.3	1.4
Faulty solution-A	0-15	11.9	4.2
Faulty solution-B	0-9	5.9	2.6
Faulty solution-C	0-6	4.8	2.1

The fit of the five-factor model was assessed by examining its loadings, goodness-of-fit indicators, and factor intercorrelations, and by comparing it to several reasonable alternatives. The alternative models were (1) a null model in which no common factors were presumed to underlie the data (i.e., each of the 15 markers was allowed to load only on its own factor), (2) a general model in which all variables loaded on a single factor, and (3) a two-factor solution composed of GRE-quantitative and constructed-response factors intended to assess whether the constructed-response scores were collectively measuring a single attribute distinguishable from the quantitative section.

Because hypothesized models are best regarded as imperfect representations of reality, assessing fit essentially involves judging, on the basis of both statistical and substantive criteria, how well a given model approximates the observed data (Cudeck & Browne, 1983; Marsh & Hocevar, 1985). It is generally advised that this judgment be made using several measures, as indicators are sensitive to different aspects of fit and, in many cases, are differentially affected by sample size (Marsh, Balla, & McDonald, 1988). The following indicators were used:

Chi-square/degrees of freedom ratio. This index is based upon the overall chi-square goodness-of-fit test associated with the factor model. In moderately sized samples, ratios of 2.0 or lower are commonly taken as evidence of good fit, though some investigators have suggested accepting values up to 5.0 (Marsh & Hocevar, 1985).

Tucker-Lewis (T-L) index. The T-L index (Tucker & Lewis, 1973) represents the ratio of the variance associated with the model to the total reliable variance, and may be interpreted as indicating how well a model with a given number of common factors represents the covariances among the markers. A low coefficient indicates that the relations among the markers are more complex than can be represented by that number of common factors.

Akaike information criterion (AIC). The AIC is an index of parsimony that takes into account both the statistical goodness-of-fit and the number of parameters that have to be estimated to achieve that fit (Bentler, 1989). For the AIC, the smaller the value the better the fit.

Root mean square residual (RMR). This measure indicates the average discrepancy between the elements in the sample and hypothesized covariance matrices (Joreskog & Sorbom, 1988). When the sample correlation matrix is analyzed, the RMR can be interpreted as the average correlation among the markers that is left over after the hypothesized model has been fitted. The lower the RMR, the better the fit.

Goodness-of-fit index. Ranging from 0 to 1.00, the Goodness-of-fit index (GFI) (Joreskog & Sorbom, 1988) is a measure of the relative amount of variance and covariance jointly accounted for by the factor model. The higher the index, the better the model fit.

Standardized residuals. Standardized residuals--the normalized discrepancies between each element in the sample and hypothesized matrices--can be used both to judge overall fit and to locate the specific causes of a lack of fit. Ideally, the residuals should be symmetric and centered around zero (Bentler, 1989). Large residuals (> 2.58 in magnitude) may suggest a possible problem with the model or reflect nonlinearity in the data (Joreskog & Sorbom, 1988).

Hierarchical chi-square test. Hierarchical chi-square tests can be conducted to determine which of two models sharing a nested relationship has the better fit (Loehlin, 1987). The chi-square is the difference between the separate chi-squares of the two models. The number of degrees of freedom is computed analogously.

To explore the meaning of the preferred model, its factor solution was extended onto several external variates. The variates were General Test verbal score, undergraduate grade-point average (UGPA), and number of mathematics courses taken. General Test scores were present for all examinees. The UGPA and course data, taken from the Biographical Information Questionnaire, were available for 239 and 215 individuals, respectively. Missing values were estimated by the maximum likelihood method using the EM algorithm (Little & Rubin, 1987).

To generate maximum likelihood structure coefficients representing the correlation between each factor and each external variable, the external variables were simultaneously introduced into the model and allowed to freely load on each factor. Model parameter estimates with and without the external variates were then compared to assure that adding the variables had no material effect on the model. Finally, the structure coefficients were computed from the external variables' loadings and the factor intercorrelations.

Results

The absolute fit of the five-factor model was evaluated by inspecting its loadings and fit indicators. Factor loadings, expressed in the correlational metric, are presented in Table 4; all were significant at $p < .001$ (t-range = 6.18 to 18.93). The goodness-of-fit results were consistently acceptable: a chi square/degrees of freedom ratio of 1.57, Tucker-Lewis index of .90, GFI of .94, RMR of .04, and median standardized residual of zero. The only indication of a potential misfit, or perhaps simply a reflection of the nonnormality of the data, were several standardized residuals larger than the recommended 2.58 standard.

Table 4

Loadings for the Five-Factor Model (N=249)

	Factor				
Marker Variable	GRE-Q	Open Ended	Goal Spec.	Equation Setup	Faulty Solution
Quantitative-A	.92	.00	.00	.00	.00
Quantitative-B	.93	.00	.00	.00	.00
Quantitative-C	.87	.00	,00	.00	.00
Open ended-A	.00	.61	.00	.00	.00
Open ended-B	.00	.41	.00	.00	.00
Open ended-C	.00	.54	.00	.00	.00
Goal specification-A	.00	.00	.70	.00	.00
Goal specification-B	.00	.00	.45	.00	.00
Goal specification-C	.00	.00	.41	.00	.00
Equation setup-A	.00	.00	.00	.77	.00
Equation setup-B	.00	.00	.00	.66	.00
Equation setup-C	.00	.00	.00	.54	.00
Faulty solution-A	.00	.00	.00	.00	.70
Faulty solution-B	.00	,00	.00	.00	.65
Faulty solution-C	.00	.00	.00	.00	.42

Note. All loadings are significant at $p < .001$ (t-range = 6.18 to 18.93).

These residuals, however, suggested no consistent, substantively meaningful pattern.

Table 5 gives intercorrelations for the five-factor model. The relations among the constructed-response factors ranged from .89 to .98; the correlations between these factors and GRE-quantitative ran from .73 to .87. The magnitude of these relations suggests that a more parsimonious solution might account for the data almost as well.

The fit of the five-factor model in relation to the alternatives is presented in Table 6. Minimal losses occurred for most indices between the five- and two-factor models but increased from the two- to the single-factor solutions. For example, from the five- to the two-factor solutions, the chi-square/degrees of freedom ratio changed by .02, going from 1.57 to 1.59; in contrast, the loss in fit by moving to the single-factor model was an additional 1.27 points. The distributions of the standardized residuals displayed a similar pattern, with the number of large residuals (i.e., > 2.58) increasing considerably upon reaching the one-factor model.

Table 7 presents hierarchical chi-square tests for the competing models. As the table shows, the five-factor model did not lead to a significant improvement over the less complex solutions. The two-factor model, however, did fit significantly better than the single-factor solution.

Table 8 shows the loadings for the two-factor model, all of which were significant ($p < .001$, t-range = 6.32 to 18.91). The correlation between the factors was .83 (t = 29.31).

Structure coefficients representing the correlations between each factor and several external indicators were computed to explore differences in the meaning of the two factors. Coefficients for GRE-quantitative were .40 with GRE-verbal, .27 with UGPA, and .32 with the number of mathematics courses taken. The comparable coefficients for the constructed-response factor were .46, .25, and .13, respectively. All but the last coefficient was statistically significant.

Discussion

This study assessed the convergent validity of expert-system scores for mathematical items cast in four constructed-response formats. The hypothesized five-factor model, consisting of GRE-quantitative and four constructed-response factors, fit the data well. However, a more parsimonious alternative comprised of two highly related dimensions--GRE-quantitative and constructed-response--represented the data with almost no loss in fit. The structure coefficients between each factor and three external variates were comparable except for the relation with the number of mathematics courses taken. This variable might have been more

Table 5

Factor Intercorrelations:

Five-Factor Solution (N=249)

	Factor				
Factor	GRE-Q	Open Ended	Goal Spec.	Equation Setup	Faulty Solution
GRE-Q	---				
Open-ended	.87	---			
Goal specification	.79	.89	---		
Equation setup	.73	.91	.98	---	
Faulty solution	.87	.97	.94	.95	---

Note. All correlations were significantly different from 0 at $p <$ .001 (t-range = 13.44 to 21.04).

Table 6

Comparison of Hypothesized and Alternative Factor Models (N=249)

	Fit Index				
Factor Model	Chi-square/ df ratio	T-L Index	RMR	GFI	Akaike Information Criterion
Five-factor	1.57	.90	.04	.94	-34.61
Two-factor	1.59	.90	.05	.93	-36.39
One-factor	2.86	.82	.06	.85	77.73
Null	15.98	---	.36	.32	1467.65

Table 7

Hierarchical Chi-Square Tests of Competing Factor Models

	Chi-Square		df		Chi-		
	Model	Model	Model	Model	Square	df	
Model Contrast	#1	#2	#1	#2	Diff	Diff	p
5- vs. 2-factor	125.4	141.6	80	89	16.2	9	NS
2- vs. 1-factor	141.6	257.7	89	90	116.1	1	<.01
1-factor vs. Null	257.7	1677.7	90	105	1419.9	15	<.01

Note. Model #1 is the more complex of the two models in a given contrast.

Table 8

Loadings for the Two-Factor Model (N=249)

Marker Variable	GRE-Q	Constructed-Response
Quantitative-A	.92	.00
Quantitative-B	.93	.00
Quantitative-C	.87	.00
Open ended-A	.00	.59
Open ended-B	.00	.42
Open ended-C	.00	.53
Goal specification-A	.00	.67
Goal specification-B	.00	.45
Goal specification-C	.00	.41
Equation setup-A	.00	.73
Equation setup-B	.00	.62
Equation setup-C	.00	.54
Faulty solution-A	.00	.71
Faulty solution-B	.00	.64
Faulty solution-C	.00	.43

Note. All loadings are significant at $p < .001$ (t-range = 6.32 to 18.91).

related to the GRE-quantitative factor because of the greater range of difficulty and content that characterized that factor's markers.

One major implication of these results is for the meaning of expert-systems' constructed-response scores. A companion study showed GIDE and human content experts to agree highly in grading a common set of examinee responses (Sebrechts, Bennett, & Rock, 1991). The current study extends this finding by suggesting a high structural relation with GRE-quantitative, a well-established mathematical ability measure. Further backing is lent by results from computer science, where another expert system, MicroPROUST, also produced scores reasonably consonant with raters' judgments and with an established achievement test (Bennett, Gong, Kershaw, Rock, Soloway, & Macalalad, 1990; Bennett, Rock, Braun, Frye, Spohrer, & Soloway, 1990; Braun, Bennett, Frye, & Soloway, 1990). In combination, these findings supply promising evidence for interpreting expert-systems' constructed-response scores as academic proficiency indicators.

Although the current study found a high structural relation between GIDE's scores and GRE-quantitative, the relation was not strong enough to justify interpreting the two indicators as measures of a single common factor. This distinction might reflect some fundamental difference between the constructed-response scores and GRE-quantitative (e.g., owed to dissimilarities in item format). Alternatively, the distinction may abate as the range of items scorable by GIDE broadens and time limits are made more comparable to those employed in the General Test.

The nature of the relationship between constructed-response scores and GRE-quantitative is worth pursuing. If the constructed-response scores do measure a proficiency that is both different from GRE-quantitative and important for success in graduate education, these scores might be used to supplement GRE-quantitative in the admissions process. If, on the other hand, the two measures indicate a single common factor, complex constructed-response "performance" tasks might be introduced into the General Test without changing the fundamental nature of the quantitative construct measured. This addition might have educational value by communicating to examinees the importance of practicing production tasks, as well as enhance the test's face validity. Once the General Test is delivered in computerized adaptive form, some of the time saved might be devoted to administering a small corpus of these items, perhaps as an institutional option. Also worth considering might be development of diagnostic tests intended to describe the errors individuals make in GRE-quantitative problem-solving. Taken early enough in an undergraduate's career, such information might be useful in strengthening certain problem-solving skills and thereby increase opportunity, especially for educationally disadvantaged groups.

A second implication of these results is for the meaning of scores from the different constructed-response formats. In contrast with expectation, GIDE's scores appeared to measure the same underlying mathematical proficiency regardless of format (i.e., the ordering of examinees on one format closely duplicated the orderings on the others). This finding is consistent with those psychometric studies that suggest diverse question formats sometimes tap a common dimension (Bennett, Rock, & Wang, 1991; Bridgeman, in press; Traub & Fisher, 1977; van den Bergh, 1990; Ward, 1982). For assessing level of general quantitative proficiency, it would seem that GIDE's scores can be combined across question formats.

Although the constructed-response formats measured the same general proficiency, the *specific* cognitive processes required by the item types may not be equivalent. As noted, some formats seem more oriented to procedural processes, for example, than to locating an appropriate problem representation. These processes could be highly intercorrelated in some populations--by one process causing another, by contiguous learning, etc.--and thus not readily distinguishable through factor analysis. Still, there may be some purposes (e.g., instructional) for which these distinctions might be important to pursue.

In the current data, such differential processing might have been reflected in the format means; that is, the more constrained items should have been consistently easier than the less constrained ones (which offered fewer clues to a problem solution). However, with the exception of faulty solutions, mean scores on the formats varied little--probably due to ceiling effects. Faulty solution items, which might have provided the most specific clues to the correct problem representation by including a complete--though subtly flawed--solution, were consistently the hardest questions, suggesting that working from a wrong solution may add, rather than reduce, cognitive complexity.

In considering this study's results, several limitations should be noted. An important limitation relates to the sample, which was composed of a relatively small group of volunteers who differed somewhat from the June 1989 General Test population. Small samples always suggest results be viewed as preliminary, pending replication. Similarly, the use of volunteers raises questions of motivation. In this instance, the sample's high performance suggests that most individuals took the constructed-response measures seriously.

The divergences of the sample from the General Test population would seem only partially responsible for the curtailed score distributions observed. Although the sample was somewhat more mathematically able than the population, the moderate difficulties for the multiple-choice versions of the three prototype items appear inconsistent with the degree of skewness that occurred.[5] One reasonable explanation lies in the test timing differences already noted. Regardless of the cause, the curtailed distributions introduce interpretational problems which limit the generalizability of results.

What are the implications of this investigation for GRE program research and development? There is little doubt that future testing environments will be computer-based. The GRE program is among several large-scale operations that have begun work on interactive assessment systems ("ETS research plan," 1989). To permit meaningful generalizations to this delivery environment, it is important to move from the paper-and-pencil mode used in this study to an interactive data collection component capable of presenting constructed-response items and capturing examinee solutions directly. Studying the meaning of scores generated in this environment will eliminate the questions that inevitably arise when research data are collected in one mode and operational tests administered in another. Additionally, computer delivery should permit timing each item individually, increasing the chances that each item will be attempted and diminishing the interpretational problems associated with not-reached items.

A second implication is for broadening the pool of constructed-response items that can be scored by GIDE. A greater range of content and difficulty is needed to produce better appraisals of examinee ability and more precise estimates of the relations between constructed-response scores and GRE-quantitative. Developing the knowledge bases needed to support increased content coverage is labor-intensive. However, once this infrastructure is built, it can be used for multiple purposes: to score any item written to a given specification, to develop multiple-choice questions with cognitively meaningful distractors, and to underlie systems for training problem-solving expertise.

With an interactive data collection program and a broader corpus of items, the cognitive processes invoked in responding to different item formats and the formats' utility for instructional diagnosis could be more productively investigated. Particularly valuable insight might be obtained through cognitive analysis. From verbal protocol studies, a detailed model could be built of the processes employed in responding to the various constructed-response formats (as well as to multiple-choice items). Predictions made by the model regarding the processes invoked could be empirically tested by looking at the relations between item performance and precisely targeted process markers.

The combination of expert scoring systems and constructed-response tasks presents exciting new "intelligent" assessment possibilities (Bennett, in press). Among these possibilities are interactive classroom systems that diagnostically analyze constructed answers (and perhaps help remediate problem-solving errors), as well as batch-processing programs for grading open responses from large-scale testing operations. Validating the scores and diagnostic characterizations these devices produce will be an ongoing process providing both evidence for the meaning of the characterizations as well as information for improved assessment systems.

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., Rock, D. A., & Wang, M. (1991). Equivalence of free-response and multiple-choice items. Journal of Educational Measurement, 28, 77-92.

Bennett, R. E., Ward, W. C., Rock, D. A., & LaHart, C. (1990). Toward a framework for constructed-response items (ETS RR-90-7). Princeton, NJ: Educational Testing Service.

Bentler, P. M. (1989). EQS Structural equations program manual. Los Angeles: BMDP Statistical Software.

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

Bridgeman, B. (in press). A comparison of open-ended and multiple-choice question formats for the quantitative section of the Graduate Record Examination. Princeton, NJ: Educational Testing Service.

Cudeck, R., & Browne, M. W. (1983). Cross validation of covariance structures. Multivariate Behavioral Research, 18, 147-167.

Educational Testing Service. (1989a). GRE information bulletin. Princeton, NJ: Author.

Educational Testing Service. (1989b). 1989-90 GRE guide to the use of the Graduate Record Examinations Program. Princeton, NJ: Author.

ETS research plan designed to create a new generation of Graduate Record Examinations. (1989, February 23). ETS Examiner, 18(26), 1.

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

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. Educational Researcher, 18(9), 27-32.

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

Hinsley, D. A., Hayes, J. R., & Simon, H. A. (1977). From words to equations: Meaning and representation in algebra word problems. In P. Carpenter & M. A. Just, (Ed.), Cognitive Processes in Comprehension (pp. 89-106). Hillsdale, NJ: Lawrence Erlbaum Associates.

Joreskog, K., & Sorbom, D. (1986). PRELIS: A program for multivariate data screening and data summarization. Mooresville, IN: Scientific Software, Inc.

Joreskog, K., & Sorbom, D. (1988). LISREL 7: A guide to the program and applications. Chicago, IL: SPSS Inc.

Little, R. J. A., & Rubin, D. B. (1987). Statistical analysis with missing data. New York: John Wiley & Sons.

Loehlin, J. C. (1987). Latent variable models. Hillsdale, NJ: Lawrence Erlbaum Associates.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. Psychological Bulletin, 103, 391-410.

Marsh, H. W., & Hocevar, D. (1985). Application of confirmatory factor analysis to the study of self-concept: First and higher order factor models and their invariance across groups. Psychological Bulletin, 97, 562-582.

Mayer, R. E. (1981). Frequency norms and structural analysis of algebra story problems into families, categories, and templates. Instructional Science, 10, 135-175.

Nickerson, R. S. (1989). New directions in educational assessment. Educational Researcher, 18(9), 3-7.

Rock, D. A., Bennett, R. E., & Jirele, T. (1988). The factor structure of the Graduate Record Examinations General Test in handicapped and nonhandicapped groups. Journal of Applied Psychology, 73, 383-392.

Rock, D. A., Werts, C., & Grandy, J. (1982). Construct validity of the GRE Aptitude Test across populations: An empirical confirmatory study (ETS RR 81-57). Princeton, NJ: Educational Testing Service.

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.

Stricker, L. J., & Rock, D. A. (1987). Factor structure of the GRE General Test in young and middle adulthood. Developmental Psychology, 23, 526-536.

Swinton, S. S., & Powers, D. E. (1980). A factor analytic study of the restructured GRE Aptitude Test (GREB Report No. 77-6P). Princeton, NJ: Educational Testing Service.

Traub, R. E., & Fisher, C. W. (1977). On the equivalence of constructed-response and multiple-choice tests. Applied Psychological Measurement, 1, 355-369.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. Psychometrika, 38, 1-10.

van den Bergh, H. (1990). On the construct validity of multiple-choice items for reading comprehension. Applied Psychological Measurement, 14, 1-12

Ward, W. C. (1982). A comparison of free-response and multiple-choice forms of verbal aptitude tests. Applied Psychological Measurement, 6, 1-11.

Footnotes

1. The locations were Atlanta, GA; Austin, TX; Brookline, MA; Emeryville, CA; Evanston, IL; Pasadena, CA; Princeton, NJ; and Washington, DC.

2. The twenty-four examinees were removed because they did not reach two or more of the 12 constructed-response items. Even a few not-reached items introduced spurious effects, since the test was sequenced by format and presented in only two orders. Failing to reach the last three items--which occurred for 18 of the 24 examinees--meant omitting an entire format. Items were considered to be not-reached only if the answer space was blank and no subsequent item was attempted; an item was considered attempted if any mark appeared suggesting that the question was considered.

3. Mean parcel difficulties calculated from GRE program pretest data were 11.33, 11.65, and 11.92 with standard deviations of 2.21, 2.22, and 2.28, respectively.

4. The preferred procedure would have been to estimate loadings for each administration order separately, but the small sample size precluded this. Distributions for the two orders were, however, similar: total test means were 101.8 (SD =17.8) and 99.9 (SD =19.1), $\underline{t} = .83$ (df = 247, $\underline{p} > .1$). Comparisons of the means for each item presented in one versus the other order produced $\underline{t}$-values of .12, .69, and 1.75 for the three open-ended items ($\underline{df} = 247$, $\underline{p} > .05$), .60, -.08, and .99 for the goal-specification items ($\underline{df} = 247$, $\underline{p} > .1$), -1.86, .07, and 1.07 for equation-setup ($\underline{df} = 247$, $\underline{p} > .05$), and .52, .11, and 1.46 for the faulty solutions ($\underline{df} = 247$, $\underline{p} > .1$).

5. Equated deltas from GRE program files were 13.0, 13.5, and 13.8.

Appendix A
Item Stems

Work Prototype (Two-goal problems)

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? (OE)

Isomorphs

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? (GS)

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? (ES)

$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? (FS)

Interest Prototype (Three-goal problems)

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? (OE)

Isomorphs

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? (GS)

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? (ES)

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? (FS)

Rate x Time Prototype (Five-goal problems)

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)? (OE)

Isomorphs

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.? (GS)

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? (ES)

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? (FS)

Note. The format in which the item was presented is indicated in parentheses after each item. OE = open ended, GS = goal specification, ES = equation setup, FS = faulty solution.

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 *in*correct, 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.

Timel = 10.67
Timel = 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.

Timel = 800/75
Timel = 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. Time 1 = 285 miles / 45 miles per hour
Time 1 = 6.33 hours (6.33 hours = 6 hours and 20 minutes)
b. Distance 2 = 600 miles - 285 miles
Distance 2 = 315 miles
c. Time 2 = 315 miles / 50 mile per hour
Time 2 = 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. End time = 7:00 am + 12 hours 38 min (7:00 am + 12.63 hrs = 7:38 pm)
End 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

Estimated Correlation Matrix

Marker Variable	Marker Variable													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Quantitative-A														
2. Quantitative-B	.84													
3. Quantitative-C	.80	.80												
4. Open ended-A	.48	.54	.46											
5. Open ended-B	.30	.30	.27	.20										
6. Open ended-C	.44	.43	.40	.32	.30									
7. Goal specification-A	.52	.50	.45	.38	.24	.28								
8. Goal specification-B	.33	.37	.26	.18	.31	.28	.30							
9. Goal specification-C	.33	.32	.31	.24	.24	.21	.30	.23						
10. Equation setup-A	.52	.52	.50	.44	.26	.31	.56	.29	.28					
11. Equation setup-B	.38	.45	.34	.35	.30	.37	.47	.31	.28	.54				
12. Equation setup-C	.40	.39	.37	.33	.30	.25	.34	.25	.14	.41	.26			
13. Faulty solution-A	.50	.54	.45	.35	.25	.35	.51	.36	.26	.58	.46	.43		
14. Faulty solution-B	.61	.57	.52	.37	.31	.37	.37	.26	.21	.41	.32	.36	.45	
15. Faulty solution-C	.32	.35	.31	.32	.22	.28	.24	.22	.21	.22	.22	.34	.30	.28

54020 06556 • Y81M 7 • 209138