Expert-System Scores for Complex Constructed-Response Quantitative Items: A Study of Convergent Validity

Randy Elliot Bennett
Educational Testing Service

Marc M. Sebrechts
Catholic University of America

Donald A. Rock
Educational Testing Service

This study investigated the convergent validity of expert-system scores for four mathematical constructed-response item formats. A five-factor model comprised of four constructed-response format factors and a Graduate Record Examination (GRE) General Test quantitative factor was posed. 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. *Index terms: automatic scoring, constructed response, expert system, free-response items, open-ended items.*

Large-scale testing programs have built their operations around the multiple-choice item, which provides an efficient and objective means of measuring cognitive abilities. The multiple-choice format has been criticized, however, because it putatively measures lower-level skills than less restricted formats, provides limited opportunities for partial credit, offers 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 these deficiencies of the multiple-choice format might be addressed by *complex* constructed-response items (Bennett, in press). A complex constructed-response item is an item for which scoring decisions cannot typically be made immediately and unambiguously using mechanical application of a limited set of explicit criteria; rather, scoring of this type of item requires expert judgment. Such items can be designed to reflect ''real-life'' tasks more accurately, to support partial-credit scoring, to facilitate instructional diagnosis through analysis of solution processes, and to 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 (e.g., Educational Testing Service, 1989c), the automated presentation of constructed-response questions becomes plausible. Moreover, advances in expert systems—computer programs that emulate the behavior of

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 15, No. 3, September 1991, pp. 227-239
© Copyright 1991 Applied Psychological Measurement Inc.
0146-6216/91/030227-13$1.90

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 assessed the convergent validity of expert-system scores for four complex constructed-response mathematical formats. The relations of the formats among themselves and to established measures were examined.

Method

Examinees

Examinees were drawn from a pool of more than 50,000 examinees taking a single form of the Graduate Record Examinations (GRE) General Test administered nationally in June 1989 (see Bennett, Sebrechts, & Rock, 1991, for full details of the sampling process). Up to the first 100 examinees available from each of eight regions were selected, with some individuals replaced to produce within-region samples composed of citizens and noncitizens in proportions similar to the General Test population. A final study sample of 249 was available for analysis.

Table 1 presents General Test scores and biographical information for the sample and the population taking the June 1989 administration. The sample differed somewhat from the population. The sample's General Test performance was significantly higher by .5, .4, and .4 standard deviations (SDs), for verbal, quantitative, and analytical, respectively; the most consequential of several statistically significant demographic differences was that the sample contained a greater proportion of nonwhites.

Table 1
Background Data for the Study Sample

Variable	June 1989 Population	Sample
N	50,548	249
General Test Performance		
Verbal		
Mean	476	534*
SD	122	130
Quantitative		
Mean	532	583*
SD	140	137
Analytical		
Mean	513	568*
SD	132	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 z test of sample value with total test population parameter.
[a]U.S. citizens only.

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 × 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 were divided among four formats such that each isomorphic item in a content class appeared in a different format. The four formats used were:

1. Open-ended: only the problem stem is presented and the examinee must provide a step-by-step solution.
2. Goal specification: the problem stem, a list of givens, and a list of unknowns is presented.
3. Equation setup: the problem stem and the equations to identify the unknowns are given.
4. Faulty solution: the stem is presented with an erroneous solution for the examinee 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 in Figure 1. (For a more detailed description of the item development process, see Sebrechts, Bennett, & Rock, 1991.)

A rubric and key for scoring the items were designed in consultation with Educational Testing Service (ETS) mathematics test development staff (for the rubric and key, see Sebrechts et al., 1991). 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.

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., inverting the values in a ratio), and computational mistakes.

Expert system. Examinees' constructed responses were scored by GIDE (Sebrechts, LaClaire, Schooler, & Soloway, 1986; Sebrechts & Schooler, 1987), an expert system that was designed in earlier versions to detect examinee 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 examinee'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 et al., 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. These are high 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 30-minute blocks. Items are divided among real (i.e., practical problems) and pure arithmetic, algebra, and geometry. The items 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 infor-

Figure 1
Isomorphic Problems in Four Item Formats: a. Open-Ended; b. Goal Specification; c. Equation Setup; d. Faulty Solution (Print Size Is Reduced and Page Arrangement Modified for Publication)

a. 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: ______________

b. 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: ______________

c. 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: ______________

d. $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: ______________

mation 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 also administered in two 30-minute 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 (UGPA) (Educational Testing Service, 1989b). Finally, the median internal consistency reliabilities computed from multiple samples were .93 and .91 for quantitative and verbal, respectively.

Procedure

General Test scores and biographical 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 1-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 × time, and work. Items were presented in two orders (most to least constrained and the reverse) and administered 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 provided 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. (For the transcription rules, see Sebrechts et al., 1991.) 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 was computed. This analysis produced a median correlation of .96, with the lowest value at .87 and the 11 remaining correlations above .90. Also relevant is the scoring reliability analysis summarized above. 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 examinees' responses.

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 (GRE-Q), a reasonably well-established mathematical ability measure. This structural relation should be less than unity, however, not only because of format differences with the 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 (Bridgeman, in press; Traub, 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 needed 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-Q, 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 λs denote that a factor loading was to be estimated; a 0 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 more clearly reflect any format-related differences in covariance structure.

For the GRE-Q factor, each variable was a parcel of 20 items constructed by randomly sampling from each of the six test specification content areas: real arithmetic, pure arithmetic,

Table 2
Hypothesized Factor Model, Number of Items per Indicator, and Type of Problem for the Constructed-Response Variables [A = Five-Goal (Rate), B = Three-Goal (Interest), and C = Two-Goal (Work)]

		Factor				
Marker Variable	No. of Items	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

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. (Mean parcel difficulties in delta units calculated from GRE pretest data were 11.33, 11.65, and 11.92 with SDs of 2.21, 2.22, and 2.28, respectively.) Parcels were scored on a 21-point number-correct 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.

Summary statistics for the markers are presented in Table 3. As the table shows, 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 (Jöreskog & Sörbom, 1986) was used to estimate the sample product-moment correlation matrix for uncensored distributions (for this matrix, see Bennett, Sebrechts, & Rock, 1991). The maximum likelihood procedure from LISREL 7 (Jöreskog & Sörbom, 1988) was then employed to estimate the unknown factor loadings for the full sample collapsing across administration orders. (The preferred procedure would have been to estimate loadings for each administration order separately, but the small sample size precluded this.) 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. However, because the maximum likelihood procedure assumes multivariate normality, its estimates of parameter standard errors should, in this case, be cautiously interpreted.

Table 3
Summary Statistics for Marker Variables

Marker Variable	Scale	Mean	SD	Skewness
Quantitative A	0-20	13.7	4.2	-.5
Quantitative B	0-20	14.3	3.9	-.7
Quantitative C	0-20	13.4	4.0	-.4
Open-Ended A	0-15	12.7	3.3	-2.3
Open-Ended B	0-9	7.8	2.3	-2.3
Open-Ended C	0-6	5.0	2.0	-1.7
Goal Specification A	0-15	12.3	3.6	-2.2
Goal Specification B	0-9	8.6	1.2	-3.6
Goal Specification C	0-6	5.6	1.0	-3.6
Equation Setup A	0-15	12.6	3.4	-2.3
Equation Setup B	0-9	8.3	1.9	-3.4
Equation Setup C	0-6	5.3	1.4	-2.3
Faulty Solution A	0-15	11.9	4.2	-1.8
Faulty Solution B	0-9	5.9	2.6	-.5
Faulty Solution C	0-6	4.8	2.1	-1.5

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-Q and constructed-response factors intended to assess whether the constructed-response scores were collectively measuring a single attribute distinguishable from the quantitative section.

Assessing fit. 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:

1. χ^2/degrees of freedom (χ^2/df) ratio. This index is based upon the overall χ^2 goodness-of-fit test associated with the factor model. In samples of moderate size, 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).

2. Tucker-Lewis index (TLI). The TLI (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.
3. Aikake 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, smaller values indicate better fit.
4. Root-mean-square residual (RMSR). This measure indicates the average discrepancy between the elements in the sample and hypothesized covariance matrices (Jöreskog & Sörbom, 1988). When the sample correlation matrix is analyzed, the RMSR can be interpreted as the average correlation among the markers that is left over after the hypothesized model has been fitted. Lower values of RMSR indicate better fit.
5. Goodness-of-fit index (GFI). Ranging from 0 to 1.00, the GFI (Jöreskog & Sörbom, 1988) is a measure of the relative amount of variance and covariance jointly accounted for by the factor model. For GFI, higher values indicate better model fit.
6. Standardized residuals (SRs). SRs—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 lack of fit. Ideally, the residuals should be symmetric and centered around 0 (Bentler, 1989). Large residuals (>2.58) may suggest a possible problem with the model or reflect nonlinearity in the data (Jöreskog & Sörbom, 1988).
7. Hierarchical χ^2 test. These tests can be conducted to determine which of two models that share a nested relationship has the better fit (Loehlin, 1987). The χ^2 is the difference between the separate χ^2s of the two models. The number of *df* is computed analogously.

Relations with other variables. To explore the meaning of the preferred model, its factor solution was extended onto several external variates. The variates were General Test verbal score, UGPA, and number of mathematics courses taken. General Test scores were available for all examinees. The UGPA and course data, both taken from the biographical 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 loadings of the external variables 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 nonzero loadings were significant at $p < .001$ (t ranged from 6.18 to 18.93). The goodness-of-fit results were consistently acceptable: $\chi^2/df = 1.57$, TLI = .90, GFI = .94, RMSR = .04, and median SR = 0. The only indication of a potential misfit, or perhaps simply a reflection of the nonnormality of the data, was several SRs larger than the recommended 2.58 criterion. 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

Table 4
Factor 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

GRE-Q 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, χ^2/df changed by .02 (from 1.57 to 1.59); in contrast, the loss in fit by moving to the single-factor model was an additional 1.27. The distributions of the SRs 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 χ^2 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 and uniquenesses for the two-factor model. All non-zero loadings 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 in-

Table 5
Factor Intercorrelations for the Five-Factor Solution (N = 249): All Correlations Were Significantly Different From 0 at $p < .001$ (t Range = 13.44 to 21.04)

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

Table 6
Fit Index Values for Comparisons of Hypothesized and Alternative Factor Models (N = 249)

	Fit Index					
Factor Model	*df*	χ^2/df	TLI	RMSR	GFI	AIC
Five-Factor	80	1.57	.90	.04	.94	−34.61
Two-Factor	89	1.59	.90	.05	.93	−36.39
One-Factor	90	2.86	.82	.06	.85	77.73
Null	105	15.98	—	.36	.32	1,467.65

dicators were computed to explore differences in the meaning of the two factors. Coefficients for the GRE-Q factor 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 were statistically significant at $p < .05$.

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-Q and four constructed-response factors, fit the data well. However, a more parsimonious alternative comprised of two highly related dimensions—GRE-Q 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 related to the GRE-Q 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 constructed-response scores derived from expert systems. A companion study demonstrated that GIDE and human content experts agreed highly in grading a common set of examinee responses (Sebrechts et al., 1991). The present study extended this finding by suggesting a high structural relation with GRE-Q, a well-established mathematical ability measure. Further evidence is provided 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 et al., 1990). In combination, these findings supply promising evidence for interpreting constructed-response scores derived from expert systems as academic proficiency indicators.

A second implication extends to the meaning of scores from the different constructed-response

Table 7
Hierarchical χ^2 Tests of Competing Factor Models: Model 1 is the More Complex of the Two Models in a Given Contrast

	χ^2		*df*				
Model Contrast	Model 1	Model 2	Model 1	Model 2	χ^2 Diff	*df* Diff	*p*
Five- vs. Two-Factor	125.4	141.6	80	89	16.2	9	NS
Two- vs. One-Factor	141.6	257.7	89	90	116.1	1	<.01
One-Factor vs. Null	257.7	1,677.7	90	105	1,419.9	15	<.01

Table 8
Factor Loadings and Uniquenesses for the Two-Factor Model (N = 249)

	Factor Loadings		
Marker Variable	GRE-Q	Constructed-Response	Uniqueness
Quantitative A	.92	.00	.16
Quantitative B	.93	.00	.14
Quantitive C	.87	.00	.25
Open-Ended A	.00	.59	.66
Open-Ended B	.00	.42	.82
Open-Ended C	.00	.53	.73
Goal Specification A	.00	.67	.55
Goal Specification B	.00	.45	.79
Goal Specification C	.00	.41	.84
Equation Setup A	.00	.73	.47
Equation Setup B	.00	.62	.62
Equation Setup C	.00	.54	.71
Faulty Soultion A	.00	.71	.49
Faulty Solution B	.00	.64	.59
Faulty Solution C	.00	.43	.82

formats. In contrast with expectation, 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 that diverse question formats may 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 (e.g., by one process causing another, or by contiguous learning) and thus not readily distinguishable through factor analysis. Yet there may be some purposes (e.g., instructional) for which these processing distinctions might be important to pursue.

In considering this study's results, several limitations should be noted. First, alternative approaches should be used in attempts to replicate these conclusions. In particular, exploratory factor analysis might be employed to determine if other substantively meaningful solutions fit the data.

A second 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 that results should be viewed as preliminary, pending replication. Similarly, the use of volunteers raises questions of motivation. In this instance, the sample's high levels of performance suggest that most individuals took the constructed-response measures seriously.

The divergences of the sample from the General Test population 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 for the constructed-response items. (Equated deltas from GRE program files were 13.0, 13.5, and 13.8.) One reasonable explanation lies in the test timing differences already noted. Regardless of the cause, the curtailed distributions introduce interpretational problems that limit the generalizability of results.

The combination of expert scoring systems and constructed-response tasks investigated here presents exciting new "intelligent" assessment possibilities that transcend graduate admissions testing (Bennett, in press). Among these more general 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. Building such systems will take considerable development (especially of knowledge bases) and research. 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 tools.

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: Erlbaum.

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: Erlbaum .

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., Sebrechts, M. M., & Rock, D. A. (1991). *The convergent validity of expert system scores for complex constructed-response quantitative items* (Research Rep. No. RR-91-12). Princeton NJ: Educational Testing Service.

Bennett, R. E., Ward, W. C., Rock, D. A., & LaHart, C. (1990). *Toward a framework for constructed-response items* (Research Rep. No. RR-90-7). Princeton NJ: Educational Testing Service.

Bentler, P. M. (1989). *EQS structural equations program manual.* Los Angeles CA: BMDP Statistical Software, Inc.

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.

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

Fiske, E. B. (1990, January 31). But is the child learning? Schools trying new tests. *The New York Times,* pp. A1, 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, (Eds.), *Cognitive processes in comprehension* (pp. 89-106). Hillsdale NJ: Erlbaum.

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

Jöreskog, K., & Sörbom, 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: Wiley.

Loehlin, J. C. (1987). *Latent variable models. Hills-*

dale NJ: Erlbaum.

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.

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

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.* (Research Rep. No. 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* (Research Rep. No. 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. (1987). Diagnosing errors in statistical problem solving: Associative problem recognition and plan-based error detection. In E. Hunt (Ed.), *Proceedings of the Ninth Annual Cognitive Science Meeting* (pp. 691–703). Hillsdale NJ: Erlbaum.

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. (in press). On the equivalence of the traits assessed by multiple-choice and constructed-response tests. In R. E. Bennett & W. C. Ward, (Eds.), *Construction versus choice in cognitive measurement.* Hillsdale NJ: Erlbaum.

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.

Author's Address

Send requests for reprints or further information to Randy Elliot Bennett, Educational Testing Service, Princeton NJ 08541, U.S.A.