The Relationship of Expert-System Scored Constrained Free-Response Items to Multiple-Choice and Open-Ended Items

Randy Elliot Bennett, Donald A. Rock, and Henry I. Braun
Educational Testing Service

Douglas Frye and James C. Spohrer, Yale University

Elliot Soloway, University of Michigan

This study examined the relationship of an expert-system scored constrained free-response item (requiring the student to debug a faulty computer program) to two other item types: (1) multiple-choice and (2) free-response (requiring production of a program). Confirmatory factor analysis was used to test the fit of a three-factor model to these data and to compare the fit of the model to three alternatives. These models were fit using two random-half samples, one given a faulty program containing one bug and the other a program with three bugs. A single-factor model best fit the data for the sample taking the one-bug constrained free response and a two-factor model fit the data somewhat better for the second sample. In addition, the factor intercorrelations showed this item type to be highly related to both the free-response and multiple-choice measures. *Index terms: artificial intelligence, constructed-response items, expert-system scoring, free-response items, open-ended items.*

Over the better part of a century, the multiple-choice item has been the mainstay of standardized testing in the United States. The use of this format is justified by its objectivity and efficiency, and more recently by the development of such statistical models as item response theory (Lord, 1980) for its analysis.

Multiple-choice items have been criticized, however, because they often do not directly resemble criterion behaviors, they are of limited utility for instructional diagnosis, and they might not be capable of measuring certain cognitive processes or skills. To address these limitations, a heavier reliance on constructed response (e.g., essays, performance tasks) is often suggested. Constructed-response items can present tasks similar to those encountered in educational and work settings, they can offer information on problem-solving processes (Birenbaum & Tatsuoka, 1987), and they may measure somewhat different skills than multiple-choice formats (Ward, Frederiksen, & Carlson, 1980).

Although constructed-response formats offer attractive potential advantages, their main liabilities for major testing programs have been the subjectivity and high cost associated with scoring. For example, the College Board's Advanced Placement Program annually invests substantial resources to employ temporarily several hundred teachers who score hundreds of thousands of constructed responses. Although significant efforts are made to enhance objectivity (e.g., teachers are trained to score each question and two levels of re-reading occur for samples of papers), variation across readers is at times considerable (Braun, 1988). If a machine-scorable constructed-response item type could be developed, problems associated with scoring cost and reliability might be substantially reduced.

One example of progress toward developing such an item type is found in the domain of computer

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 14, No. 2, June 1990, pp. 151–162

0146-6216/90/020151-12$1.85

science (Braun, Bennett, Frye, & Soloway, in press). This item type presents the examinee with a specification describing a task to be performed by a computer program and a completed program that does not correctly perform that task. It is the examinee's assignment to correct the program by deleting and/or inserting the required code. The corrected program is then given to an expert system for scoring. In a recent study (Braun et al., in press), this experimental system was able to score 83% of the papers it encountered (it offered no analysis on the remaining papers), and agreed with a human rater at levels similar to those at which raters agree among themselves (product-moment correlations in the .80s).

The purpose of this study was to assess the relationship of this expert-system scored constrained free-response item type to multiple-choice and to free-response items contained on the College Board's Advanced Placement Computer Science (APCS) Examination. The magnitude of this relationship is central to evaluating the potential of this item type as an eventual replacement for more open-ended formats and as a supplement to multiple-choice questions.

Method

Examinees

Examinees were drawn from a prior study of the item type conducted with a sample of high school seniors taking the 1988 APCS examination (Braun et al., in press). Student selection procedures involved the following steps: (1) participation was solicited from all APCS teachers with class enrollments of 15 or more or who had participated in grading the 1987 APCS examination, (2) indications of interest were received from teachers at 70 of 112 solicited schools, (3) constrained free-response items were mailed to these teachers, (4) responses were received from 916 students in 59 schools, and (5) 1988 APCS scores were located in Educational Testing Service files for 737 of these students for whom responses were judged to be complete. Of these 737 completed records, the constrained free-response item type was able to be machine-scored for 614 students. For purposes of this study, this group was split into two samples, differentiated by having randomly been given variants of the faulty solution problem that contained either one or three bugs.

Instruments

Constrained free-response item. The constrained free-response item was a more structured adaptation of an open-ended problem from the 1985 APCS examination. The open-ended version required the student to write a program that rotated the elements of an array. Eight constrained variants of this problem were developed as a means of increasing the breadth of the content domain studied. Each variant contained a program specification and a faulty solution to that specification; the student's task was to correct the solution by inserting and/or deleting lines of computer code. In six of the variants, the solution contained a single bug. In the remaining two variants, three bugs each were embedded, with care taken to select bugs whose results did not interact with the other bugs, thereby keeping the faulty program at a difficulty level appropriate for a novice.

Bugs were chosen to reflect three categories that have been found to capture most of the nonsyntactic errors produced by novices when writing programs (Spohrer, 1989). These categories were arrangement, completeness, and detail. An arrangement bug occurred when all of the parts of a program were present but not put together properly. A completeness bug existed when one component was missing. When a single part of a component (e.g., a variable or operator) was at fault and could be repaired by changing one word or operator, the bug fell into the detail category.

Two bugs were selected from each category, for a total of six different bugs (one for each single-bug variant). Each of the triple-bug variants contained one bug from each category. Examples of the items are presented in Braun et al. (in press).

Students' responses to these items were presented to the expert system MICROPROUST (Johnson & Soloway, 1985) as complete programs within which the student's correlations were embedded.

MICROPROUST scored the solutions by (1) breaking a problem down into a set of component goals, (2) comparing sections of the student's program to correct ways of achieving those goals, and where it could not find a match, (3) comparing those sections to common faulty implementations of the goals. On the basis of the faults detected, diagnostic comments were produced and numeric scores were assigned. Rater reliability was computed by correlating expert-system scores with those of a human grader. For the one-bug variants the correlation was .88 (n = 40), and r = .82 (n = 44) for the three-bug variants (Braun et al., in press).

The Advanced Placement Computer Science Examination. The APCS "A" Examination is intended to assess mastery of topics covered in the first semester of a college-level introductory course in computer science (College Board, 1988). The examination emphasizes programming methodology and procedural abstraction, but also includes some material on the study of algorithms, data structures, and data abstraction. The test comprised 35 multiple-choice and 3 free-response items. The free-response items, which are scored by human graders, require the student to write or design a program, subprogram, or data structure and, at times, to analyze the efficiency of certain operations involved in the solution. Examples of these items and of the multiple-choice questions can be found in College Board (1988).

Data Collection

Each student was asked to respond to one of the eight variants of the first problem, as well as to one of eight variants of a second probem. (Responses to the second problem were not included in this study because they were scored by a second expert system for which rater reliability was found to be suspect.) Variants were randomly assigned to students such that equal numbers of one- and three-bug versions were administered. Teachers were instructed to administer the problems in a single class period during the month prior to the APCS examination.

Though the faulty solution item type was envisioned for delivery by computer, problems were presented and responses collected in paper-and-pencil format. As a result, responses had to be converted to machine-readable form upon receipt. Because MICROPROUST will not analyze programs with syntax errors, each student program was run through an automatic parser to check its syntax. Those programs rejected by the parser were reviewed by one of the authors, who judged whether the error could be objectively corrected. If the reproducibility of a correction was considered questionable, the paper was eliminated from the study; otherwise, the correction was made. (The overwhelming majority of corrections made in this manner—approximately 80%—consisted of adding a semicolon as a delimiter at the end of a line of code.) Finally, each amended paper was again run through the parser to check that the error had been successfully removed.[1]

Data Analysis

The model. A three-factor model composed of multiple-choice, free-response, and faulty-solution factors was specified to test the relationship of the new item type to the two others. The hypothesized model consisted of factors marked by the three item types. For the first factor, these item types were parcels of APCS multiple-choice items balanced on difficulty. Three multiple-choice parcels (Multiple-Choice A, B, C) were constructed from every third item in each of four test specification content areas (programming methodology, features of languages, algorithms, and computer systems) and from a single item from each of two additional areas (data structures and applications). Items were then shifted among parcels (but within content categories) so that the mean difficulty values for each parcel were similar. Parcels were scored on a 12- or 13-point number-correct scale, based on the number of items in the parcel. The second

[1]The need to correct such errors was an artifact of the paper-and-pencil format. In a computer-delivered administration, syntax errors would be identified automatically—as they would in any programming language environment—and more than likely fixed by the student before the response was finalized.

factor was indicated by each of three APCS free-response problems (Free-Response A, B, C), with each free-response scored on a 10-point scale.

The third factor was marked by the single indicator of the response to the "Rotate" problem. This problem was scored on a five-point scale for the sample taking the one-bug variants, and on a six-point scale for the group taking the three-bug versions. Differences in the scales emanated from the need to award points for correcting different numbers of seeded bugs and to deduct points for the expected introduction of different numbers of new bugs (e.g., students would be expected to introduce more new bugs in solving the three-bug variants than in the one-bug variants, because of the added complexity of the former items). Both scales were set to range from 0 to 2, with a score of 2 indicating a perfect solution. (Scale points for the one-bug problem were 0, .5, 1.0, 1.5, and 2.0.)

Table 1 depicts the hypothesized model. The asterisks indicate that a factor loading was to be estimated. Conversely a "0" denotes that the indicator variable was constrained to have a zero loading on that particular factor. To estimate the factor pattern from the data, the sample polychoric correlation matrix was computed using the program PRELIS (Jöreskog & Sörbom, 1986). The weighted least-squares factor estimation procedure from LISREL 7 (Jöreskog & Sörbom, 1988) was then used to estimate the unknown factor loadings (i.e., the asterisks) subject to the pattern of zero constraints and allowing the factors to be intercorrelated.

Parameter estimation. The factor pattern was estimated from the polychoric correlation matrix for each sample (Table 2) using the weighted least-squares procedure because the distributions for the marker variables were frequently non-normal. The weighted least-squares procedure provides for asymptotic standard errors and overall goodness-of-fit tests that do not assume normality. Further, the use of polychoric correlations tends to minimize the effects on factor-analytic results of differences in difficulty across marker variables.

To estimate accurately the relationship between factors, a reliability estimate for each factor must be available. For factors with multiple markers, this estimate is generated from within the factor model. However, because there was only one indicator of the constrained free-response factor, the reliability of this factor could not be estimated in this way. Hence an alternative estimate was needed.

To approximate the reliability of the faulty-solution item, the average reliability of the free-response items was used. This reliability estimate can be argued to be a lower bound for the faulty solution because the free-response estimate includes two sources of variation: topic (each problem poses a different task), and rater (each solution is graded by a different individual). The faulty solution is computer scored; thus there is no rater variance, leaving topic as the only source of variation. To compute the reliability estimate, the factor loadings for the model were estimated, the loading for each free response in the weighted least-

Table 1
Hypothesized Factor Model

		Factor		
Marker Variable	No. Items	Multiple-Choice	Free-Response	Constrained Free-Response
Multiple-Choice A	12	*	0	0
Multiple-Choice B	12	*	0	0
Multiple-Choice C	11	*	0	0
Free-Response A	1	0	*	0
Free-Response B	1	0	*	0
Free-Response C	1	0	*	0
Constrained Free-Response	1	0	0	*

*Loading estimated.

Table 2
Polychoric Correlation Matrices: Sample 1 (N = 314)
Below the Diagonal, Sample 2 (N = 300) Above the Diagonal

	1	2	3	4	5	6	7
1. Multiple-Choice A	--	.69	.73	.61	.62	.70	.42
2. Multiple-Choice B	.68	--	.70	.61	.64	.68	.46
3. Multiple-Choice C	.68	.67	--	.65	.64	.68	.46
4. Free-Response A	.55	.54	.52	--	.59	.59.	.47
5. Free-Response B	.65	.61	.59	.55	--	.65	.48
6. Free-Response C	.61	.62	.62	.54	.58	--	.41
7. Constrained Free-Response	.57	.49	.57	.54	.55	.55	--

squares solution was squared, and these squared loadings were averaged. The resulting reliabilities were .56 for Sample 1 and .62 for Sample 2. Finally, the solutions were rerun using these estimates for the reliability of the faulty solutions.

Model fit. The fit of the three-factor model was assessed by examining its factor intercorrelations and goodness-of-fit indicators, and by comparing the model's fit 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 seven 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 APCS test and constrained free-response factors intended to assess whether the constrained responses were measuring attributes different from the test. These alternative models allowed the goodness-of-fit indices to be investigated as a function of factorial complexity, where changes in the indices suggested how much fit was lost by moving from more to less complex models.

Evaluating model fit was complicated by the fact that, in confirmatory factor analysis, universally accepted measures of fit do not exist (Marsh & Hocevar, 1985; Sobel & Bohrnstedt, 1985). Consequently, several goodness-of-fit indicators were used, particularly in comparing the three-factor model to the alternatives. These indicators were:

1. Tucker-Lewis index. The Tucker-Lewis (TL) index (Tucker & Lewis, 1973) represents the ratio of the variance associated with the model to the total variance, and may be interpreted as indicating how well a factor 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.
2. Root mean square residual. The root mean square residual (RMSR) is the average correlation among the markers that remains after the hypothesized model has been fitted (Jöreskog & Sörbom, 1988). The lower the RMSR, the better the fit.
3. Chi-square/degrees of freedom ratio. The χ^2/df ratio is based on the overall χ^2 goodness-of-fit test associated with each factor model. Ratios up to 5.0 indicate a reasonable fit (Marsh & Hocevar, 1985).
4. Goodness-of-fit index. Ranging from 0 to 1.00, the goodness-of-fit index (GFI) is a measure of the relative amount of variance and covariance jointly accounted for by the factor model (Jöreskog & Sörbom, 1988). The higher the magnitude of this index, the better the model fit.
5. Akaike information criterion. The Akaike information criterion (AIC) is an index of parsimony in which the best-fitting model is defined as having a small χ^2 with few unknowns (Loehlin, 1987). As scaled here, the AIC was always negative, with the best-fitting model having the index closest to 0.
6. Hierarchical chi-square test. Hierarchical χ^2 tests can be conducted to determine which of

two models that share a nested relationship has the better fit (Loehlin, 1987). The χ^2 for this test was the difference between the separate χ^2s of the two models. The number of degrees of freedom was computed analogously.

7. Standardized residuals. Standardized residuals can be used to judge fit and to locate the specific causes of a lack of fit. In general, residuals larger than 2.0 in magnitude suggest a problem with the model (Jöreskog & Sörbom, 1988).

Results

Table 3 presents APCS means and standard deviations for the two samples and for the population taking the 1988 APCS examination. (Scores in this and all other analyses are number-correct raw scores, as opposed to the formula scores used in the APCS program.) Also presented are the summary statistics for performance on the faulty solution items for the two samples. For each APCS score, a two-tailed z test was used to contrast each sample mean with the population mean, which was treated as a population parameter. Although the sample means proved to be significantly higher than the test population mean for most contrasts, the magnitude of these differences was marginal, ranging from .14 to .26 standard deviations. These marginal differences suggest that the samples were not dramatically different in computer science knowledge from the population taking the examination.

Table 4 presents the loadings for each variable as estimated from the three-factor model. In both samples, all loadings are significant ($p < .001$, t range 14.01 to 39.95). Loadings for the multiple-choice factor are generally slightly higher than those for the free-response factor, probably because the multiple-choice indicators were constructed so as to be parallel in content and difficulty. Hence these indicators share substantial variance. In contrast, each free-response indicator deals with a different topic, thereby reducing the common variance and, hence, the loading of each on the common factor.

The absolute fit of the three-factor model can be evaluated through inspection of several indices. The goodness-of-fit indices and standardized residuals suggest the extent to which the model is complex enough to account for the data. For Samples 1 and 2, the TL index was 1.00 and .99, respectively, which indicates that the three-factor model accounts for virtually all of the variance among the markers. The RMSRs of .02 for both samples present a similar picture. Also, inspection of the standardized residuals reveals that none were larger in magnitude than 2.0 in Sample 1 and only one of 28 was larger than 2.0 in Sample 2, but this is a finding expected on the basis of chance alone.

Factor intercorrelations suggest whether a simpler model might account for the data. Table 5

Table 3
Mean and Standard Deviation of APCS Number-Correct Score and Faulty-Solution Scores for Samples 1 and 2 and the APCS Population

Score	Score Range	Sample 1 (N = 314) Mean	SD	Sample 2 (N = 300) Mean	SD	Population (N = 10,719) Mean	SD
APCS							
35-Item Objective	0-35	17.3**	6.4	17.8***	6.6	16.1	6.5
Free-Response 1	0-9	4.6*	3.7	4.8***	3.7	4.1	3.5
Free-Response 2	0-9	5.8**	2.8	5.9***	2.8	5.3	2.9
Free-Response 3	0-9	1.8	2.7	1.9	2.8	1.6	2.7
Rotate							
1-Bug Variant	0-2	1.2	.9				
3-Bug Variant	0-2			.9	.7		

*$p < .05$, **$p < .01$, ***$p < .001$.

Table 4
Factor Loadings for the Three-Factor Model

Sample and Marker Variable	Factor: Multiple-Choice	Factor: Free-Response	Factor: Constrained Free-Response
Sample 1 (N = 314)			
Multiple-Choice A	.84	.00	.00
Multiple-Choice B	.81	.00	.00
Multiple-Choice C	.81	.00	.00
Free-Response A	.00	.69	.00
Free-Response B	.00	.77	.00
Free-Response C	.00	.77	.00
Constrained Free-Response	.00	.00	.75
Sample 2 (N = 300)			
Multiple-Choice A	.84	.00	.00
Multiple-Choice B	.83	.00	.00
Multiple-Choice C	.86	.00	.00
Free-Response A	.00	.75	.00
Free-Response B	.00	.77	.00
Free-Response C	.00	.82	.00
Constrained Free-Response	.00	.00	.79

Note. All loadings were significant at the .001 level (t range for Sample 1 = 14.01 to 35.50; t range for Sample 2 = 15.16 to 39.95).

gives the factor intercorrelations for the three-factor model. Each correlation was tested for significant differences from 1.00 with a t test using the standard errors of estimate generated by the factor model. For Sample 1 (which took the one-bug variants), none of the disattenuated correlations was significantly different from 1.00; this called into question the need for a three-factor model. For Sample 2 (which took the three-bug variants), the correlations between the constrained free-response factor and the other factors were significantly less than 1.00, although those between the free-response and multiple-choice factors were not, which suggests the need for a simpler model here as well.

Table 5
Factor Intercorrelations for the Three-Factor Solution for Sample 1 (N = 314, Upper Triangle) and Sample 2 (N = 300, Lower Triangle)

Factor	Multiple-Choice	Free-Response	Constrained Free-Response
Multiple-Choice	--	.97	.89
Free-Response	.98		.98
Constrained Free-Response	.68*	.74**	

*Different from 1.00 at $p < .001$, $t = -5.02$, $df = 297$.
**Different from 1.00 at $p < .001$, $t = -3.58$, $df = 297$.
Note. All correlations were significantly different from 0 at $p < .001$, t range = 10.14 to 35.14, df range = 297 to 311).

The fit of the three-factor model in relation to several more parsimonious alternatives is presented in Table 6. For Sample 1, negligible losses in fit occurred for most indices in moving from the three- to the single-factor solutions. The changes are substantial, though, once the null model is reached. For example, the RMSR remains the same from the three-factor to the single-factor models, but it increases by .49 from the single-factor to the null models. In contrast to the other indices, the AIC—a measure of parsimony—shows marginal improvements in fit through the single-factor solution.

For Sample 2, the pattern is similar. The largest losses are associated with the move from the single-factor to the null models, and most indices show only trivial changes from the three- to the one-factor solutions. A hint of a slightly better fit for the two- over the one-factor model is given, however, by the AIC, which is at its lowest for the two-factor solution.

Table 7 presents hierarchical χ^2 tests for the competing models. In Sample 1, the only instance of a significant improvement in fit is for the single-factor versus null model contrast. In Sample 2, this contrast is also statistically significant, but so is the improvement in fit of the two-factor over the single-factor model.

Relative fit also can be assessed by examining the distributions of the standardized residuals (not shown). For Sample 1, the residuals changed marginally from the three-factor to the single-factor solutions, but became dramatically larger when the null model was reached. For Sample 2, a comparable pattern was observed.

The suggestion of a reasonable fit for the single-factor model in Sample 1 and possibly the two-factor model in Sample 2 can be further evaluated by inspecting the intercorrelations from the two-factor model. For Sample 1, the disattenuated correlation of .93 was not significantly different from 1.00 ($p > .05$, $t = -.87$, $df = 311$), which is too high to support a two-factor solution; for Sample 2, it was .71, significantly less than 1.00 ($p < .001$, $t = -4.94$, $df = 297$), and it is therefore more consistent with a two-factor model.

Table 8 shows the loadings for the two-factor solution. Again, all loadings are significant ($p < .001$; t range = 14.01 to 40.27). As for the three-factor solution, the loadings for the multiple-choice markers are slightly higher than those for the free responses. The probable explanation is similar: The multiple-choice markers share more variance because they are parallel, and as a result, they play a bigger role in defining the common factor than do the free-response indicators.

Discussion

Results suggested that the three item types formed a single factor in one sample, but that a two-factor model with faulty solutions defining a separate factor might better account for the data in the second

Table 6
Fit Indices for Hypothesized and Alternative Factor Models

Sample and Factor Model	χ^2/df	T-L Index	RMSR	GFI	Akaike Information Criterion
Sample 1 (N = 314)					
Three-factor	.32	1.00	.02	1.00	-17.92
Two-factor	.48	.99	.02	1.00	-17.39
One-factor	.50	.99	.02	1.00	-16.72
Null	72.47	--	.51	.42	-767.96
Sample 2 (N = 300)					
Three-factor	.48	.99	.02	1.00	-18.89
Two-factor	.51	.99	.02	1.00	-17.54
One-factor	1.30	.98	.03	.99	-22.72
Null	80.59	--	.52	.38	-853.14

Table 7
Hierarchical Chi-Square Tests of Competing Factor Models

Model Contrast	χ^2 Model 1	χ^2 Model 2	*df* Model 1	*df* Model 2	χ^2 Difference	*df* Difference	*p*
Sample 1 (N = 314)							
3- vs. 2-factor	3.83	6.78	12	14	2.95	2	NS
2- vs. 1-factor	6.78	7.43	14	15	0.65	1	NS
1-factor vs. Null	7.43	1521.91	15	21	1514.48	6	<.01
Sample 2 (N = 300)							
3- vs. 2-factor	5.77	7.08	12	14	1.31	2	NS
2- vs. 1-factor	7.08	19.44	14	15	12.36	1	<.01
1-factor vs. Null	19.44	1692.28	15	21	1672.84	6	<.01

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

sample. What might explain the differences in fit between the two samples? One potential explanation is that the timing guidelines under which the items were administered allotted less time per bug to those taking the three-bug problems. This differential might have created a power-versus-speed situation in which the major source of individual differences among students taking the one-bug variants was programming skill, whereas speed of processing might also have come into play for those taking the three-bug variants. The effects of speededness on cognitive test performance are well known. In the present case, however, it is not known if the dissimilarity in time allotted per bug was enough

Table 8
Factor Loadings for the Two-Factor Model

Marker Variable	Factor APCS	Factor Constrained Free-Response
Sample 1 (*N* = 314)		
Multiple-Choice A	.84	.00
Multiple-Choice B	.81	.00
Multiple-Choice C	.81	.00
Free-Response A	.68	.00
Free-Response B	.76	.00
Free-Response C	.76	.00
Constrained Free-Response	.00	.75
Sample 2 (*N* = 300)		
Multiple-Choice A	.84	.00
Multiple-Choice B	.82	.00
Multiple-Choice C	.85	.00
Free-Response A	.75	.00
Free-Response B	.77	.00
Free-Response C	.82	.00
Constrained Free-Response	.00	.79

Note. All loadings were significant at $p < .001$ level (t range for Sample 1 = 14.01 to 35.91; t range for Sample 2 = 15.16 to 40.27).

to cause individual differences in speed of processing to appear in one and not the other sample.

In addition to the variation in factor structure across samples, the faulty-solutions factor was virtually indistinct from the free-response factor in one sample and highly related to it in the other. This result suggests that the premise for the constrained free-response format is plausible: to combine in a single item type the surface characteristics and cognitive demands of free response with the machine-scorable efficiency of multiple choice. That faulty solutions might be reliably machine-scored is supported by a companion investigation that found that most student responses could be analyzed and that scores were generally similar to those awarded by a human grader (Braun et al., in press).

Although the faulty-solutions factor was highly related to free response, the former was also highly related to the multiple-choice factor (though more so in Sample 1 than Sample 2). The melding of faulty solutions with both item types is seemingly due to the exceptionally close relationship observed between the multiple-choice and free-response factors themselves. This latter result would also appear to be a stable one because correlational analyses of student performance on other forms of the APCS examination with different samples have produced the same finding (Bleistein, Maneckshana, & McLean, 1988; Mazzeo & Bleistein, 1986; Mazzeo & Flesher, 1985). Similar relationships between multiple-choice and constructed-response formats have been reported in other content areas such as mathematical reasoning (Traub & Fisher, 1977) and verbal reasoning (Ward, 1982), though such a result is not universal (e.g., Ackerman & Smith, 1988; Ward et al., 1980).

In the present case, several mechanisms might explain the high relationship between the free-response and multiple-choice factors. First, some portion of the relationship likely results from general ability. Because the factors are defined by academic tasks, it is reasonable to expect each factor to be related to general ability, and in turn, to be positively correlated with each other.

Second, in some situations free-response and multiple-choice items may measure the same specific processes. Traub and Fisher (1977) made such an argument for mathematical reasoning when they suggested that the examinee must construct a solution regardless of the item format, although in the multiple-choice case the resulting answer is used as a basis for choosing among the response options. (Locating a constructed answer among the options, though, is still no guarantee that the answer is correct.)

In the APCS context, this argument would appear to have some merit. A cursory analysis of multiple-choice item content, for example, suggests that many of these items cannot be correctly answered with any consistency and efficiency by strategies other than construction. (These items call for such things as choosing the correct data structure, counting loop executions, and finding bugs.) For this reason, the processes used would arguably be identical or highly similar to those employed in writing a program or design.

Third, some part of the observed relationship is plausibly due to close relationships among specific processes or between processes and knowledges that are developed together. It is likely, for example, that the distinct processes sometimes invoked in responding to multiple-choice versus free-response items are correlated by virtue of being subcomponents of the same higher-level process (Sternberg, 1980). Alternatively, it is plausible that some of the knowledges tested by the multiple-choice items are taught along with programming skill or occur incidentally as a result of it (e.g., knowing the characteristics of a programming-language compiler).

Further research might help resolve much of this conjecture. In particular, cognitive analyses of the tasks posed by the APCS multiple-choice and free-response items, and by the faulty solutions, might better elucidate the degree to which these item types measure different processes. Analyses of the relations of these item types to external criteria, such as subsequent computer science course grades, would also be informative. Such cognitive and empirical criterion-related analyses might even identify how single- and multiple-bug faulty-solutions tasks differ. In addition, studies of the functioning of the faulty-solutions item type in other domains (e.g., algebra word problems) should help identify whether

and how this format might be used in assessing skills other than programming. Finally, development of a prototype intelligent assessment system might be explored (Bennett, in press). In such a system, multiple-choice items would be presented first. The information from these items would then be used to determine whether to present constructed-response items (i.e., faulty solutions and/or free-responses) to a given student and also to help the expert system interpret the student's answers. This combination of student screening and interpretive assistance might allow the level of successful analyses of constructed responses to approach 100%.

Several limitations of the present study should be noted. First, only a single instance of the constrained free-response item type was used within each sample. Even though multiple variants were employed, using only a single problem limits greatly the generalizability of results to faulty solutions as a class of constrained free-response, as well as to other classes of constrained free-response (e.g., completion items). Further, using a single exemplar prevented a reliability estimate for the item type from being generated by the factor model, and this forced the estimate to be approximated with the reliability of the free-response items. Although this is a reasonable approximation, it is upon this approximation that the intercorrelations between the constrained free-response and other factors are based. If, for example, this approximation is too low, the corrected intercorrelations may be too high. Future studies should include multiple instances to increase the likelihood of yielding accurate estimates and to enhance the generalizability of results.

Second, the effects of item format could not be strictly tested because content was not held constant across formats. That is, different problems were presented in the three formats, and in some cases the content measured was noticeably different (e.g., some multiple-choice questions dealt with knowledge incidental to the programming skill measured by the free-response items). Even with these content differences, however, the formats were highly intercorrelated.

Third, all measures were not given at the same point in time. The APCS multiple-choice and free-response problems were administered on the same day and the faulty solutions were administered up to a month previously, but the exact date within this period differed among the participating schools. It is possible that some relevant learning might have occurred between the two administrations; however, as both the one- and three-bug variants were administered within each school, additional earning (or other variables related to time between administrations) does not seem to be a plausible explanation for the observed differences in factor structure.

Finally, even though the faulty-solutions and free-response tasks involved construction, they were still somewhat removed from classroom debugging and programming behaviors. In the classroom, both behaviors are performed interactively, not in the paper-and-pencil mode employed in this study. Whether interactive environments that allowed examinees to execute the programs they were writing or debugging would still produce factor structures like those found here is an unresolved question.

These results have several implications for the APCS examination. If the results can be replicated with faulty solutions covering a wider range of programming skill, an argument might be made for eventually including the one-bug variant in future computer-delivered editions of the test. Substituting several faulty solutions for a free-response question would apparently not change the essential construct measured by the test and might possibly reduce scoring costs over the long term. This cost reduction is by no means assured; substantial effort is required to develop the knowledge base needed to score responses to each faulty solution, and it is not yet clear how much a problem can be changed before major modifications in the knowledge base need to be made. With respect to the three-bug faulty solution, a better understanding of the role of time limits and of any potential differences in cognitive requirements is required before use of this version can be seriously considered.

Lastly, even though multiple-choice and free-response items appear to measure the same essential APCS construct, there are good reasons to maintain—and, perhaps, increase—the role of constructed-response items. The most compelling reason

is that the ability to successfully complete free-response items—that is, to program—is central to the APCS curriculum. Including free-response items emphasizes to teachers and students the need to focus on developing this skill. Also, the multiple-choice format is viewed by many testing critics as measuring and encouraging the development of irrelevant skills. Because of their perceived relevance, the inclusion of constructed-response items should help respond to these concerns, thereby increasing the credibility of the measures taken.

References

Ackerman, T. A., & Smith, P. L. (1988). A comparison of the information provided by essay, multiple-choice, and free-response writing tests. *Applied Psychological Measurement, 12,* 117–128.

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

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

Bleistein, C., Maneckshana, B., & McLean, D. (1988). *Test analysis: College Board Advanced Placement Examination Computer Science 3JBP* (SR-88-63). Princeton NJ: Educational Testing Service.

Braun, H. I. (1988). Understanding scoring reliability: Experiments in calibrating essay readers. *Journal of Educational Statistics, 13,* 1–18.

Braun, H. I., Bennett, R. E., Frye, D., & Soloway, E. (in press). Scoring constructed-responses using expert systems. *Journal of Educational Measurement.*

College Board. (1988). *Advanced Placement course description: Computer science.* New York: Author.

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

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: SPSS, Inc.

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

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

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.

Mazzeo, J., & Bleistein, C. (1986). *Test analysis: College Board Advanced Placement Examination Computer Science 3IBP* (SR-86-105). Princeton NJ: Educational Testing Service.

Mazzeo, J., & Flesher, R. (1985). *Test analysis: College Board Advanced Placement Examination Computer Science 3HBP* (SR-85-180). Princeton NJ: Educational Testing Service.

Sobel, M. E., & Bohrnstedt, G. W. (1985). Use of null models in evaluating the fit of covariance structure models. In N. B. Tuma (Ed.), *Sociological methodology* (pp. 152–178). San Francisco CA: Jossey-Bass.

Spohrer, J. C. (1989). *MARCEL: A generate-test-and-debug (GTD) impasse/repair model of student programmers* (CSD/RR No. 687). New Haven CT: Yale University, Department of Computer Science.

Sternberg, R. J. (1980). Factor theories of intelligence are all right almost. *Educational Researcher, 9,* 6–18.

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.

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

Ward, W. C., Frederiksen, N., & Carlson, S. B. (1980). Construct validity of free-response and machine-scorable forms of a test. *Journal of Educational Measurement, 17,* 11–29.

Author's Address

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