DOCUMENT RESUME

ED 386 490		TM 024 059

AUTHOR	Bennett, Randy Elliot; Rock, Donald A.
TITLE	Generalizability, Validity, and Examinee Perceptions of a Computer-Delivered Formulating-Hypotheses Test. GRE Board Professional Report No. 90-02aP.
INSTITUTION	Educational Testing Service, Princeton, NJ. Graduate Record Examination Board Program.
SPONS AGENCY	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-93-46
PUB DATE	Nov 93
NOTE	73p.
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Tests/Evaluation Instruments (160)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	*Computer Assisted Testing; Difficulty Level; *Generalizability Theory; Graduate Students; Higher Education; *Hypothesis Testing; *Interrater Reliability; Judges; Scores; Scoring; *Test Items; Test Validity
IDENTIFIERS	Constraints; *Formulating Hypotheses Items; *Ideational Fluency

ABSTRACT

Formulating-Hypotheses (F-H) items present a situation and ask the examinee to generate as many explanations for it as possible. This study examined the generalizability, validity, and examinee perceptions of a computer-delivered version of the task. Eight F-H questions were administered to 192 graduate students. Half of the items restricted examinees to 7 words per explanation, and half allowed up to 15 words. Generalizability results showed high interrater agreement, with tests of between two and four items scored by one judge achieving coefficients in the .80s. As in studies of paper-and-pencil versions, validity analyses found that although F-H was highly reliable, it was only weakly related to Graduate Record Examinations (GRE) General Test scores, differing primarily in its strong relations to a measure of ideational fluency. Versions of F-H based on different response limits tapped different abilities, with items employing the 15-word constraint appearing to be more useful for graduate assessment. Although the overwhelming majority of examinees found the F-H interface easy to use, some did experience difficulty, suggesting the possibility that computer familiarity constitutes a source of irrelevant variance in F-H scores. Twelve tables and two figures illustrate the analysis. Five appendixes contain test directions, an F-H scoring ruburic, ideation fluency measure, accomplishments questionnaire, and an opinion questionnaire. (Contains 30 references.) (Author/SLD)

GRE

RESEARCH

Generalizability, Validity, and Examinee Perceptions of a Computer-Delivered Formulating-Hypotheses Test

Randy Elliot Bennett
and
Donald A. Rock

November 1993

GRE Board Professional Report No. 90-02aP
ETS Research Report 93-46

Educational Testing Service, Princeton, New Jersey

2

Generalizability, Validity, and

Examinee Perceptions of a Computer-Delivered

Formulating-Hypotheses Test

Randy Elliot Bennett
and
Donald A. Rock

GRE Board Report No. 90-02aP

November 1993

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

****

Researchers are encouraged to express freely their professional judgment. Therefore, points of view or opinions stated in Graduate Record Examinations Board Reports do not necessarily represent official Graduate Record Examinations Board position or policy.

****

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

Abstract

Formulating-Hypotheses (F-H) items present a situation and ask the examinee to generate as many explanations for it as possible. This study examined the generalizability, validity, and examinee perceptions of a computer-delivered version of the task. Eight F-H questions were administered to 192 graduate students. Half of the items restricted examinees to 7 words per explanation and half allowed up to 15 words. Generalizability results showed high interrater agreement, with tests of between two and four items scored by one judge achieving coefficients in the .80s. As in studies of paper-and-pencil versions, validity analyses found that although F-H was highly reliable, it was only weakly related to GRE General Test scores, differing from that test primarily in relating more strongly to a measure of ideational fluency. Versions of F-H based on different response limits tapped somewhat different abilities, with items employing the 15-word constraint appearing more useful for graduate assessment. These items added to conventional measures in explaining school performance and creative expression. Finally, although the overwhelming majority of examinees found the F-H interface easy to use, some experienced difficulty, suggesting the possibility that computer familiarity constitutes a source of irrelevant variance in F-H scores.

Acknowledgments

We are indebted to many people for their contributions to this study. Betty McGrail developed the F-H items and keys with the assistance of Jeff Wadkins and Robin Huffman. Ann Gallagher pilot tested the items and assisted in selecting the final set. Larry Stricker provided the Accomplishments Questionnaire. Harriet Trenholm developed the F-H interface (with help from Randy Kaplan), and Eric Thompson integrated it into the ETS OSA software. Jan Flaugher oversaw integration, testing, and revision of the interface, as well as development of the F-H tutorial (along with Holly Knott and Probal Tahbildar). Ann Gallagher recruited graduate departments, Lou Woodruff and Charlene Canny handled arrangements with the computer-based test centers, and Shelley Mendelowitz coordinated examinee scheduling. Pat Bakr, Shirley Epstein, Stephen McBrearty, Karyn Storti, Virginia Vandagriff, and Durelle Yarbro oversaw testing at the ETS centers. Mary Fowles conducted a session for scoring F-H responses, with the help of Kelli Boyles, Kathy Bradley, Ramon Lacayo, Della Collantone, and Betty McGrail. Nancy Robertson programmed the data analyses. Altamese Sherrell performed many important tasks in her role as project administrator and Bill Ward advised throughout. Finally, the GRE Board and ETS provided generous support.

Large-scale institutional testing programs are increasingly moving to computer delivery. For example, a computer-based GRE General Test has been introduced, as has an examination for prospective teachers, Praxis I: Academic Skills Assessments. The introduction of tests for nursing and architectural licensure also has been scheduled and a prototype computer-based SAT has been developed.

The move to computer-based tests is motivated by several factors. One factor is real-time scoring, which makes possible both dramatic reductions in test length (through adaptive testing) and instantaneous score reporting. A second factor is the belief that computer delivery will encourage development of new methods for measuring traditional constructs as well as the measurement of constructs not currently assessed. These new methods and measures will undoubtedly include a variety of item types from simple adaptations of multiple choice to sophisticated simulations. Some of these item types will be scorable by machine, some by machine with human assistance, and some--like the essay component of Praxis I--will need to call entirely on human analysis.

The current study concerns the potential of one experimental computer-based effort, Formulating Hypotheses (F-H), as a graduate admissions indicator. The F-H item type was created by Norman Frederiksen (1959) to measure "abilities of the sort required by the research scholar in trying to make sense of research findings" (p. 2). Developed under the sponsorship of the Office of Naval Research, Frederiksen's F-H problems were composed of a graph or table and a brief textual explanation describing a situation. Situations were chosen to avoid the need for specialized knowledge. Subjects were asked to generate possible explanations for the situation, each explanation taking the form of a short handwritten sentence or phrase.

Studies of the F-H item type began to appear in the late 1960s and early 1970s (e.g., Klein, Frederiksen, & Evans, 1969; Frederiksen & Evans, 1974). The investigations most relevant to institutional testing were sponsored by the GRE Board and appeared a few years later.[1] For these studies, F-H items were written in a psychology context and subjects were undergraduates intending to pursue graduate work in that field. Construct validity analyses found that, although reliable, F-H scores correlated only minimally with GRE General Test performance, differing from the General Test and from an objectively scored version of the item type primarily in loading more highly on an ideational fluency factor (Frederiksen & Ward, 1978; Ward, Frederiksen, & Carlson, 1980). Also noteworthy was that the F-H items improved the prediction of some criteria; they were more effective than General Test scores in forecasting subsequent self-reports of certain professional accomplishments. The above results generally held when scoring was based on the quantity of the hypotheses posed by examinees. When based on response quality, F-H became less reliable and more highly related to the General Test. The relation with ideational fluency also was greatly diminished.

Work on F-H continued in the 1980s with studies of law school applicants and medical students (Carlson, 1985; Frederiksen, Ward, Case, Carlson, & Samph, 1981). Although the results were promising, the need to score responses manually limited the task's attractiveness for large-scale assessment. A standard, five-option multiple-choice version was created but found to measure the same reasoning ability as other General Test questions (Ward, Carlson, & Woisetschlaeger, 1983), reinforcing the value of the open-ended version.

Additional support for F-H came indirectly from studies of the requirements of graduate education, which confirmed the relevance of abilities similar to those tapped by the item type. For example, Powers and Enright (1987) asked graduate faculty from six disciplines to rate several dozen reasoning skills on the extent to which each differentiated marginal from successful students. Factor analysis reduced the ratings to five dimensions, including the ability to generate alternatives, which was rated as particularly important by professors in psychology and in education. Tucker (1985) asked cognitive psychologists, philosophers, and test developers to rank 19 analytical reasoning processes according to their importance for success in graduate programs. "Formulating alternative possibilities of conceptualization, classification, or explanation" was ranked first or second by each group.

Chances for realizing the promise of F-H increased considerably with the growing availability of personal computers and with concurrent advances in automatic approaches to natural language processing. Taking a new look at F-H, Carlson and Ward (1988) recommended that the item type be computer delivered and that an automated scoring system be developed.

In 1990, the GRE Board funded a project to move toward these ends. The current report is one of two emanating from this project and focuses on the operation of the computer-delivered F-H test. Score generalizability, score validity, and examinee perceptions of the test are investigated. The results of the automatic scoring analyses appear in a companion publication (Kaplan & Bennett, 1994).

Method

Subjects

Subjects were paid volunteers recruited from graduate departments at institutions proximal to 12 ETS computer-based test centers. Contacts were made primarily through education, psychology, English, chemistry, and biology departments. The test centers were Louisiana State University, the University of Houston, Miami-Dade Community College, Arizona State University (Tempe), Norfolk State University (Norfolk, VA), the University of Arizona (Tucson), and ETS field service offices in Atlanta, Austin, Emeryville (CA), Boston, Evanston, and Pasadena. Among the participation guidelines were that students had taken the GRE General Test during the 1990-91 academic year, were already enrolled in the first year of a graduate degree program (making follow-up studies easier), and were native English speakers (reducing confounding due to language differences). Of the students who indicated interest, 211 were tested. One-hundred ninety-two of these examinees provided usable data.

Table 1 shows the sample demographic data compared with the 1987-88 GRE General Test examinee population, the most recent year for which data were available. As expected, the sample diverged from the test population in noticeable ways. The sample scored considerably higher on the three General Test scales and had proportionally more females, U.S. citizens, individuals whose graduate objective was the Ph.D., and social science and humanities/arts majors. Engineering majors and majors classified as "other" were underrepresented.

It is also worth noting that because subjects were recruited from a limited number of graduate departments, the majors represented were generally more narrow than the table's broad categorizations might imply. So, for example, most social science majors were enrolled in psychology programs, humanities majors in English departments, and physical science majors in chemistry programs.

Finally, although recruitment guidelines stipulated that students be native speakers, a subsequent check of GRE General Test registration data revealed that 21 subjects had indicated a "best language of communication" other than English. (No indication of "best language" was given by 13 other examinees. On the basis of citizenship status, last name, and magnitude of the verbal-quantitative score discrepancy, 4 of these were presumed to belong to the "other language" group, 8 to be native speakers, and 1 remained unclassified.) Further examination suggested that although English was not their "best" language, it was a reasonably well-developed one. The group's GRE verbal scores (mean = 542, SD = 103) were considerably higher than those of the examinee population (mean = 486, SD = 122), and their performance on the study's main instrument, Formulating-Hypotheses (mean = 69, SD = 28), was very similar to that of subjects who indicated English was their primary language (mean = 71, SD =23). As a result, these examinees were included in the investigation.

Instruments

Formulating Hypotheses (F-H). The primary instrument was a computer-delivered Formulating-Hypotheses test (see Appendix A for test directions). The F-H items required no specific disciplinary knowledge but, rather, general knowledge about the world. Twenty-two items were written and pilot tested in paper-and-pencil form, each with a constraint on the length of the examinee's response of 7 or 15 words. This limitation was imposed to permit exploration of the effect of response constraint on the meaning of F-H scores and on the accuracy of automatic analysis, which, in earlier work, was higher for shorter responses (Kaplan, 1992).

From this pool, 10 items were chosen (8 for the test and 2 as samples), with the test items evenly distributed between the two constraint categories. Within constraint categories, items were selected to generate a broad distribution of scores and to vary situational contexts (roughly classified as humanities, science, social science), the presence of graphical information in the stimulus, and the phenomenon being described (gradual change, sudden change, presence or absence of some object or event).

Table 1
Demographic Data

Background Characteristic	Study Sample (n=192)[a]	1987-88 Examinee Population (n > 185,000)
General Test Performance		
Verbal mean (SD)	578 (108)	486 (122)
Quantitative mean (SD)	597 (116)	553 (139)
Analytical mean (SD)	612 (114)	529 (128)
Percentage Female	68%	53%
Percentage Non-White	19%	14%
Percentage U.S. Citizen	95%	81%
Percentage with Ph.D. Goal	72%	40%
Graduate Major		
Social Sciences	43%	18%
Humanities/Arts	18%	11%
Life Sciences	15%	18%
Education	14%	15%
Physical Sciences	7%	11%
Engineering	1%	12%
Business	0%	3%
Other	2%	12%

Note. Population data are from Examinee and Score Trends for the GRE General Test by D. M. Wah and D. S. Robinson. Copyright 1990 by Educational Testing Service. Percentage non-White is for U.S. citizens only. Graduate major percentages for population are based on those with decided majors only.

[a]The percentages for non-White, U.S. citizen, Ph.D. objective, and graduate major are based on n's of 176, 189, 180, and 159, respectively.

The F-H computer interface is illustrated in Figure 1. The top left-hand window shows an item, and directions for completing the task are given in the bottom left window. The examinee types a hypothesis, which appears in the lower right box. When the SAVE button is clicked with the mouse, the hypothesis is moved to the list in the upper right-hand window. To edit a hypothesis on the list, the examinee highlights it with the mouse and clicks on the EDIT button, moving the hypothesis back to the entry box where it can be changed.

Each F-H item was scored on a 0-15 scale, with one point awarded for each plausible, unduplicated hypothesis. This scheme was chosen based on earlier F-H research suggesting that the number of hypotheses made for more meaningful relations with criterion measures than did scoring response quality (Frederiksen & Ward, 1978).

To define the nature of creditable responses, rubrics were written for each of the eight items (see Appendix B for an example). These rubrics were developed after examining pilot test results and a subsample of responses from the main data collection. Each rubric listed several general categories--and, within these, several specific categories--into which correct responses might fall. In general, a response was considered creditable if it stated or implied a possible explanation that was readily apparent to the reader and did not duplicate another hypothesis generated by the student for that problem. Duplication was defined as more than one hypothesis falling into the same specific category, or one hypothesis in a general category and another in a corresponding specific category. Thus, the rubric attempted to discredit instances in which an examinee generated a series of hypotheses that were conceptually similar. Aside from duplication, a response was not to be considered creditable if it directly contradicted the situation, if no plausible explanation was readily apparent, or if it was based only on science fiction or the supernatural.

Ideational fluency marker. This paper-and-pencil measure was used to identify the extent to which the computer-delivered F-H test tapped ideational fluency, the facility to generate a number of ideas about a given topic within relatively broad constraints. The measure was composed of four items (see Appendix C). The first item was taken from the Topics Test of the Kit of Factor-Referenced Cognitive Tests (Ekstrom, French, & Harman, 1976) and required the examinee to generate ideas about a topic (e.g., a train journey). The second item came from the Verbal Edition of the Torrance Tests of Creative Thinking (Torrance, 1974). It asked the subject to pose questions about an object, in this case, a cardboard box. The third and fourth items were "pattern meaning" tasks (Wallach & Kogan, 1965). These items each presented an unfinished drawing and called for ideas about what the drawing might be if it were finished. For each ideational fluency item, the score was the number of responses given. Scores from the four items were then summed to give a total.

GRE General Test and background information questionnaire. The General Test is a multiple-choice examination designed to measure broad, developed abilities generally required for success in graduate work.

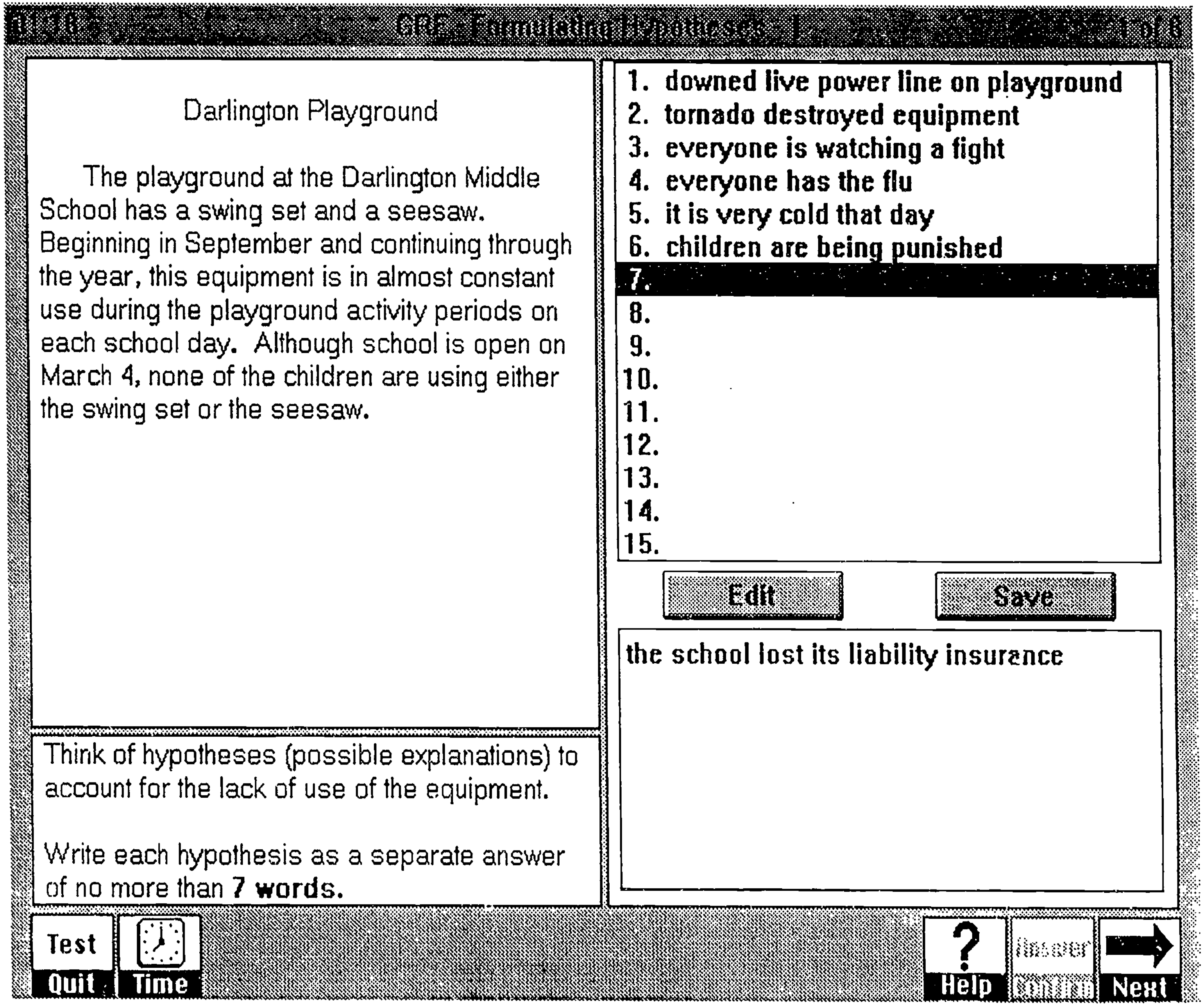


Figure 1

BEST COPY AVAILABLE

The test is composed of three sections. The verbal section (GRE-V) is intended to test the examinee's ability to reason with words in solving problems (Educational Testing Service, 1991). It contains 76 items falling into four categories (analogies, antonyms, sentence completion, and reading comprehension). The quantitative section (GRE-Q) is meant to measure basic mathematical skills, understanding of elementary mathematical concepts, and ability to reason quantitatively and solve problems in a quantitative setting. Items are divided among real (i.e., word 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 except basic mathematical knowledge), and data interpretation (based on information presented in tables or graphs). The analytical section (GRE-A) includes two item types. Analytical reasoning items, which compose the bulk of the section, evaluate the ability to understand a given structure of arbitrary relationships and to deduce new information from that structure. Logical reasoning questions test the ability to analyze and critique argumentation by understanding and assessing relationships among arguments or parts of arguments.

The psychometric characteristics of the General Test have been extensively studied. For example, factor analytic investigations have repeatedly supported the existence of distinguishable verbal and quantitative 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). (Studies have typically determined the analytical section to be more factorially complex, however.) Predictive validity analyses have found correlations with first-year grades averaged across 1,038 graduate departments to be .30 for verbal, .29 for quantitative, .28 for analytical, and .34 for a weighted composite of the three (Educational Testing Service, 1992). The median internal consistency reliabilities computed from test-analysis samples for four recent test forms were .91, .92, and .88 for verbal, quantitative, and analytical, respectively.

The background information questionnaire (BIQ) is part of the registration form for the General Test. This forced-choice, machine-scannable questionnaire contains demographic items, as well as some questions on such achievement-related indicators as college grades. Self-reported grades, like those found on the BIQ, have generally been found to accurately portray school-reported marks (Baird, 1976, p. 8). School-reported grades, in turn, are useful predictors of graduate performance. Undergraduate grade-point average (UGPA) is slightly more predictive of first-year graduate performance than the General Test; its correlation with grades taken across 1,038 departments was .37 (Educational Testing Service, 1992). Also, its independent contribution to prediction is substantial: When added to the General Test it increases the multiple correlation with first-year grades from .34 to .46.

<u>Activities and Accomplishments Questionnaire</u>. There is considerable evidence that (a) the best predictors of future, high-level

accomplishment in science, writing, music, art, and leadership are similar (usually lower level) achievements in prior years and (b) past accomplishments can be reliably documented through self-reports (Baird, 1976, pp. 35-36). Based on this work, a 52-item paper-and-pencil measure was adapted from L. Stricker (personal communication, October 10, 1991). This measure asked the examinee to indicate whether or not a given accomplishment had been achieved and, if it had, to provide documentary information on that achievement. (See Appendix D for the questionnaire.) One point was awarded for each accomplishment. Scores were computed for the total questionnaire and for six subscales: academic achievement (5 items), leadership (5 items), linguistic (composed of 12 ordinary speaking and ordinary writing questions), aesthetic expression (composed of 20 creative writing, art, music, and dramatics questions), science (5 items), and mechanical (5 items).

Opinion Questionnaire. This 10-item paper-and-pencil instrument was used to gather examinee impressions of the F-H item type and its computer delivery. All items were forced-choice except for the final question, which asked for additional comments. Appendix E contains the questionnaire.

Procedure

Subjects were assessed in computer-based test centers managed by ETS or its institutional affiliates. The centers were composed of a small number of individual stations at which examinees could work at their own pace. All sessions followed essentially the same format, involving a sign-in and orientation, the testing session, and sign-out. The orientation involved an interactive tutorial that instructed the examinee in how to use the computer to respond to F-H items. This was followed by the F-H test, the Ideational Fluency measure, the Accomplishments Questionnaire, and the Opinion Questionnaire. Subjects were informed repeatedly that they would not be paid unless they answered all questions on each instrument.

Two forms of the F-H test, differing only in item order, were administered to random halves of the group. Form A presented items with the 7-word response limitation first, whereas in Form B these items followed questions with the 15-word restriction. A limit of 80 minutes was imposed for each 8-item test form, with no restriction on the time devoted to any one item. Revisiting a question after moving on to the next item was not allowed.

After data were returned to ETS, all F-H responses were electronically checked for spelling errors and corrected, where appropriate. Finally, examinee records were matched with ETS files, and GRE scores and background information questionnaire data were extracted.

Data Analysis

Generalizability. This analysis was directed at determining the main sources of variation in F-H scores and the number of judges and items needed for acceptable levels of generalizability. For this analysis, a subsample of 30 examinees' F-H responses was randomly selected and given in hard copy form to four ETS test developers and one ETS consultant to score. Four of the readers had earned M.A. degrees and one a Ph.D.; three had majored in English literature, one in education, and one in physics. Before each item was scored, the rubric was introduced, sample responses were discussed, and several responses were graded for practice purposes. All five readers then independently scored all 30 responses. This process was repeated until all eight items had been evaluated.

A three-way repeated measures analysis of variance (with the between-group effect only for persons) was used to estimate the variance components of the following mixed model:

$$Y_{ijk} = \mu + R_j + Q_k + \pi_i + RQ_{jk} + \pi R_{ij} + \pi Q_{ik} + \pi RQ_{ijk}$$

where Y_{ijk} is the score assigned to the ith person by the jth rater for the kth item, R, the rater effect, Q, the question effect, and π the person effect, with all of the effects random facets presumed to be sampled from infinite universes of raters, questions, and persons, respectively. Analyses were conducted separately for the 7-word and 15-word items, in each case using the scores assigned by each of the five raters to 30 examinees' responses to each of four questions. Generalizability coefficients were generated as per Thorndike (1982, pp. 165-167) for different numbers of raters and items. In calculating these coefficients, the variance component for questions was dropped out under the assumption that examinees would receive tests equated for difficulty in any operational use of F-H, thereby eliminating difficulty differences as a source of error variance.

Score validity. Score validity analyses centered on determining if F-H items with different response limits measured the same dimension; if F-H and the General Test were measures of the same construct and, if not, how F-H was different; and if F-H contributed anything over conventional indicators in explaining accomplishments and school performance. Differences between F-H items were assessed through confirmatory factor analysis. This analysis tested the hypothesis that the two F-H item types measured somewhat different factors. Although the 7-word limit item should be easier to machine score, having to state a hypothesis in fewer words should place relatively greater demands on verbal facility. Thus, the more restricted item was expected to show a higher relation with GRE verbal and a lower relation to ideational fluency than the version with the 15-word limit.

For this analysis, a three-factor model was fitted to the sample correlation matrix using the EQS program (Bentler, 1989). This model was comprised of F-H 7-word, F-H 15-word, and ideational fluency factors, in which the factors were assumed to be correlated. Each of the factors was marked by four items constrained to load only on that factor. This zero constraint was imposed to make each factor as pure as possible. Consequently, the factor intercorrelations should more clearly reflect any differences in covariance structure.

The fit of the three-factor model was assessed by examining its factor loadings, goodness-of-fit indicators, and factor intercorrelations, and by comparing it with three alternatives: a two-factor model composed of F-H and ideational fluency factors, a single-factor model, and a null model in which each marker was constrained to load only on its own factor. Several fit indices were used, each sensitive to different departures: the chi-square/degrees-of-freedom ratio; the nonnormed fit index (Bentler & Bonnett, 1980) (an indicator of the proportion of reliable variance accounted for by the factor model); the Akaike information criterion (a parsimony index); and the average off-diagonal absolute standardized residual (the average residual correlation among the markers after the model is fitted). Improvements in fit also were statistically tested with a hierarchical chi-square test (Loehlin, 1987).

Finally, several outside variables were extended onto the preferred factor solution, including General Test scores, self-reported undergraduate grade-point average (UGPA), Accomplishments total score, gender, and total time spent answering F-H items.[2] Extension loadings were computed by introducing the external variables into the factor model, allowing them to load only on their own factor(s), and comparing the model parameter estimates with the original runs to assure that the factor solution was not materially affected. In the case of the General Test scores, internal consistency reliabilities based on the examinee population were included in the model, fully correcting those extensions for attenuation.

To determine if F-H and the General Test were measures of the same construct, the observed correlations of F-H and the General Test with other variables were examined. Differences were evaluated via a two-tailed t-test for the difference between correlations derived from the same sample, as per McNemar (1962, p. 140).

The incremental validity of F-H scores was assessed through least-squares linear multiple regression. For the first analysis, self-reported UGPA served as an indicator of school achievement and was regressed on the three General Test scores (verbal, quantitative, and analytical entered as a set) and F-H. For the second analysis, the Accomplishments scores were the outcome criteria and UGPA was returned to its traditional role as a predictor. Here, the Accomplishments total score and each of the subscales were regressed, in turn, on General Test scores, UGPA, and F-H. For all analyses, F-H was entered into the model last on the premise that it must demonstrate value over established measures to justify the added costs that would be associated with its use in admissions.

Examinee perceptions. The proportion of examinees responding to each of the forced choices was computed. For the "additional comments" item, responses were grouped into categories and the categories with the greatest relative frequencies were identified. Finally, some perception items were extended onto the F-H and ideational fluency factors to elucidate the meaning of F-H scores.

Results

Summary Statistics

Table 2 presents summary statistics. Of note is that scores for several of the Accomplishments subscales--which are intended to measure unusual achievements--are substantially skewed.

Generalizability

Results of the generalizability analysis for F-H 7-word and 15-word items are presented in Table 3. In each case, the variance components assume a one-item F-H test scored by one judge. The only appreciable unwanted source of variance was the persons x question interaction, which is often large for complex constructed-response tasks (Bennett, 1993, p. 9), and indicates that some examinees do well on some items but poorly on others. At the same time, the variance components associated with judges were uniformly small, indicating high scoring agreement across raters.

Figure 2 shows the generalizability levels that would be expected from an F-H test scored by one judge given different numbers of items and from a one-item test scored by different numbers of judges. F-H 7-word items (represented by the darkened points) showed marginally higher generalizability than the 15-word items (depicted by the lighter points). To achieve generalizability in the .80s would require a test composed of two or three F-H 7-word items (taking 20-30 minutes to administer) or three to four 15-word items (taking 30-40 minutes) scored by one judge.

In Table 4 are the mean number of plausible, unduplicated hypotheses credited by the judges compared with the raw number produced by the examinees for each item. Interestingly, the judges credited most responses that the examinees offered, disallowing less than one hypothesis per item on average (out of nine or so offered). Further, the two indices were almost perfectly correlated (median r = .98). These results are understandable in that the F-H item type is intended to generate a large set of creditable responses. In an experimental setting in which examinees intend to present their skills accurately, wrong responses should be relatively rare. Once moved to an operational, high-stakes environment, however, test-taking tricks will come into play and more stringent scoring rules will probably be required.

Validity

Because the number of credited responses was almost perfectly correlated with the raw number of responses generated, the latter index was used in the validity analyses.

Differences between F-H 7-word and F-H 15-word items. Loadings for the three-factor model (given in the correlational metric) ranged from .83 to .90 for the F-H 7-word items, .78 to .92 for the 15-word

Table 2
Summary Statistics (n = 192)

Variable	Scale	Mean	SD	Skewness
F-H Total	0-120	71	24	.37
F-H 7-Word	0-60	37	13	.22
#1	0-15	9.5	3.7	.20
#2	0-15	9.3	3.6	.18
#3	0-15	9.2	3.4	.19
#4	0-15	9.0	3.6	.25
F-H 15-Word	0-60	34	12	.53
#5	0-15	7.5	3.4	.80
#6	0-15	8.3	3.4	.48
#7	0-15	9.0	3.3	.33
#8	0-15	8.9	3.6	.20
F-H Total Time Spent	0-80	64	15	-.92
GRE verbal	200-800	578	108	-.08
GRE quantitative	200-800	597	116	-.25
GRE analytical	200-800	612	114	-.50
Accomplishments	0-52	6.2	3.4	.81
Academic	0-5	1.7	1.2	.26
Leadership	0-5	1.0	1.0	.69
Linguistic	0-12	1.2	1.5	1.56
Aesthetic Exp.	0-20	1.0	1.5	1.60
Science	0-5	1.1	1.2	1.04
Mechanical	0-5	0.2	0.5	4.32
Ideational Fluency	0-86	55	18	.22
#1	0-36	19.6	9.0	.50
#2	0-20	14.0	4.6	-.16
#3	0-15	10.8	3.6	-.24
#4	0-15	10.7	3.4	-.18
UGPA	1-7 (D-A)	5.6 (A-)	1.0	-.22

Table 3
Estimated Variance Components for 1 Judge Scoring a Test Containing 1 F-H Item (n = 30)

Source of Variability	Sum of Squares	Mean Square	F	p	Estimated Variance Component	Percent of Total Variance
		F-H 7-Word				
Persons	5409.9	186.5			8.65	72%
Judges	32.0	8.0			.05	0%
Questions	59.8	19.9			.03	0%
Persons x Judges	74.8	0.6	1.15	.169	.02	0%
Persons x Questions	1176.2	13.4	23.91	<.001	2.57	22%
Questions x Judges	26.9	2.2	4.00	<.001	.06	0%
Persons x Questions x Judges	195.1	.6			.56	5%
		F-H 15-Word				
Persons	4626.5	159.5			7.22	63%
Judges	49.6	12.4			.07	1%
Questions	282.3	94.1			.51	4%
Persons x Judges	132.0	1.1	1.64	<.001	.11	1%
Persons x Questions	1278.6	14.7	21.25	<.001	2.80	24%
Questions x Judges	40.9	3.4	4.93	<.001	.09	1%
Persons x Questions x Judges	240.7	.7			.69	6%

Figure 2

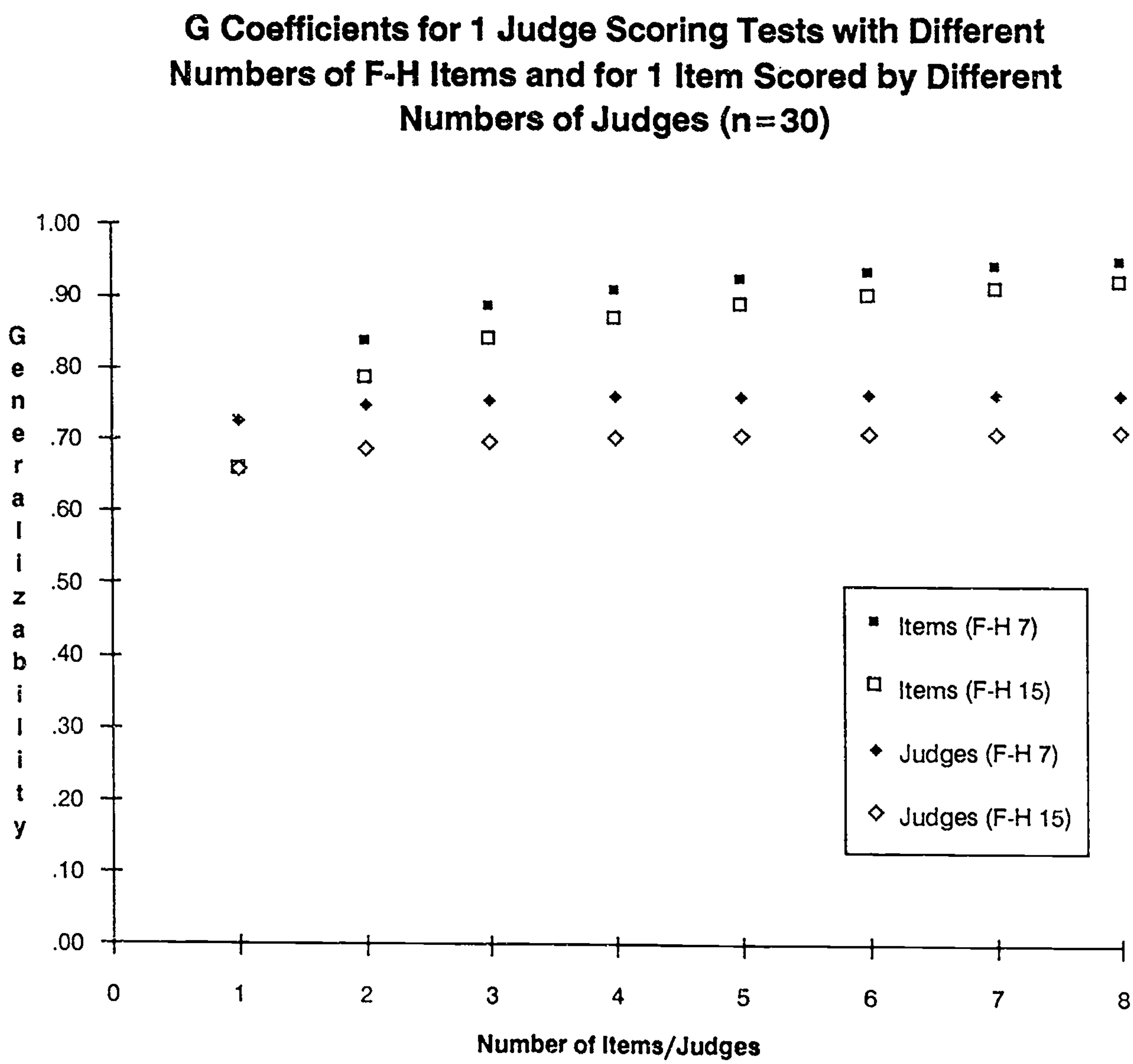

G Coefficients for 1 Judge Scoring Tests with Different Numbers of F-H Items and for 1 Item Scored by Different Numbers of Judges (n=30)
Generalizability
1.00
.90
.80
.70
.60
.50
.40
.30
.20
.10
.00
0
1
2
3
4
5
6
7
8
Number of Items/Judges
Items (F-H 7)
Items (F-H 15)
Judges (F-H 7)
Judges (F-H 15)

Table 4
Number of Credited versus Number of Offered Hypotheses ($\underline{n}$=30)

Item	Mean # of Hypotheses	Mean # Correct	Difference	$\underline{r}$ # of Hypotheses with # Correct
F-H 7-Word				
1	9.7	8.8	1.0	.97
2	9.7	9.2	.5	.99
3	9.1	8.5	.6	.99
4	9.0	8.4	.6	.99
F-H 15-Word				
5	7.6	6.5	1.1	.94
6	8.8	7.9	.9	.96
7	8.9	8.1	.8	.96
8	8.8	8.2	.7	.99

items, and .75 to .85 for the ideational fluency markers; all were significant at $p < .001$ (t-range = 11.7 to 16.6). The goodness-of-fit results were consistently acceptable: a chi-square/degrees-of-freedom ratio of 2.13, nonnormed fit index of .96, and an average off-diagonal absolute standardized residual of .036. The correlation between the two F-H factors was .90, which, while quite high, may not be sufficient to consider the item types equivalent. (An approximate 99% confidence interval for this correlation extends from .83 to .96.) The F-H 7-word and 15-word factors correlated with the ideational fluency dimension at .66 and .71 respectively, levels similar to that typically found between the factors underlying GRE verbal and GRE quantitative (Rock, Bennett, & Jirele, 1988; Rock, Werts, & Grandy, 1982).

Compared with the alternative models, the three-factor model did reasonably well (see Table 5). Notable losses in fit occurred as the models became less complex. The superiority of the three-factor model was confirmed by the hierarchical chi-square test, which showed a significant increment for this model over the two-factor solution (chi-square difference = 55.6, df difference = 2, $p < .01$).[3]

Extension loadings for several outside variables on the three-factor solution are given in Table 6. Most loadings were very similar across the F-H 7-word versus F-H 15-word factors. The exceptions were the loading for UGPA, on which F-H 15 looked more like ideational fluency than it did F-H 7, and the loading for total time spent on F-H (i.e., the sum of times for all eight items), which was more highly associated with the 7-word dimension than with the 15-word factor. Because the extension was for <u>total</u> time, its interpretation is unclear. However, the observed correlations between the F-H total scores and time spent on each test also suggest that performance and time may be more highly related for the 7-word than for the 15-word items ($r = .58$ for F-H 7 word questions and .49 for the 15-word items). This might be interpreted to mean that stating hypotheses within the 7-word limitation is more difficult and therefore more time-consuming. The predicted slightly lower relation of the F-H 7-word factor with ideational fluency supports this interpretation. The associated hypothesis, that the F-H 7-word factor would be more highly related to verbal ability, was not supported, however.

<u>Differences with the General Test</u>. Table 7 presents the observed correlations of F-H and the General Test with the criterion variables. First, note that although both the F-H and General Test scores were highly reliable, the correlations among these scores were quite low, ranging from .16 (F-H 7 with GRE-A) to .31 (F-H 15 with GRE-Q). Interestingly, the differentially higher relation with GRE quantitative ($t = 2.29$ for F-H 15 and 2.10 for F-H 7, $p < .05$) both replicates an earlier finding by Frederiksen and Ward (1978) and is found in the current data for the Ideational Fluency measure with the same General Test scales ($t = 1.99$, $p < .05$). From a theoretical perspective, why F-H and ideational fluency should relate more to GRE quantitative than GRE analytical is unclear.

Table 5
Confirmatory Factor Analysis Fit Results (n=192)

	Fit Index			
Model	NNFI	AIC	Chi/df	AODASR
3-Factor	.96	6.70	2.13	.036
2-Factor	.92	58.35	3.10	.045
1-Factor	.81	215.66	5.99	.062
Null	--	1694.24	27.67	--

Note. NNFI = nonnormed fit index, AIC = Akaike information criterion, Chi/df = Chi-square/degrees-of-freedom ratio, AODASR = Average off-diagonal absolute standardized residual. The three-factor model was comprised of F-H 7, F-H 15, and ideational fluency dimensions; the two-factor model was made up of F-H and ideational fluency factors.

Table 6
Extensions of Outside Variables on the
Three-Factor Solution ($\underline{n}$ = 192)

	Factor		
Outside Variable	F-H 7-word	F-H 15-word	Ideational Fluency
GRE verbal	.28	.28	.15
GRE quantitative	.30	.32	.22
GRE analytical	.19	.20	.09
Accomplishments	.18	.20	.24
UGPA	.21	.26	.27
Gender	.00	.01	.08
Total time spent on F-H	.57	.46	.28

Note. Extension loadings are in the correlational metric. GRE General Test extensions are fully corrected for attenuation.

Table 7

Observed Correlations of F-H and General Test Scores with Criterion Variables (n =192)

	F-H 7	F-H 15	G-V	G-Q	G-A	Acc	Acad	Lead	Ling	AE	Sci	Mech	Id Fl	UGPA
F-H 7	*.92*	.82*	.25*	.27*	.16*	.17*	.06	.04	.07	.18*	.04	.08	.61*	.20*
F-H 15		*.90*	.27*	.31*	.19*	.21*	.06	-.04	.11	.20*	.12	.11	.65*	.26*
GRE-V			*.91*	.59*	.62*	.10	.07	-.10	.09	.22*	-.06	-.06	.16*	.26*
GRE-Q				*.92*	.72*	.10	.16*	.01	-.08	.02	.18*	.00	.20*	.37*
GRE-A					*.88*	.05	.10	-.06	-.10	.07	.14*	-.07	.09	.29*
Accmp						*.54*	.47	.41	.68	.55	.42	.25	.23*	.34*
Acad							*.33*	.06	.14	-.02	.14	-.03	.17*	.57*
Lead								*.19*	.22*	-.05	.07	.07	-.04	-.01
Ling									*.54*	.33*	-.06	.10	.08	.12
AE										*.58*	.01	.00	.21*	.12
Sci											*.51*	.14*	.13	.11
Mech												*.47*	.06	-.05
Id. Fl.													*.79*	.27*
UGPA														----

Note. Internal consistency reliability estimates are on the main diagonal. Estimates for the Accomplishments scores were computed via KR-21. All other estimates are coefficient alpha. General Test estimates are taken from Educational Testing Service (1992).

* = $p < .05$

A second notable observation from Table 7 is that F-H was more strongly related to the Ideational Fluency measure than the General Test was. Differences were significant for all six comparisons (2 F-H correlations with ideational fluency x 3 General Test correlations with ideational fluency). For example, the correlation of the F-H 7-word items with the Ideational Fluency marker was .61 (the lower of the two values for F-H); the correlation for GRE-Q (the General Test scale most strongly related to the marker) was .20. The difference between these two correlations was significant at $p < .001$ ($t = 6.01$, $df = 189$).[4]

Differences between F-H and the General Test vis-a-vis relations to the Accomplishments measures were much less dramatic, in part because the correlations were uniformly low to begin with. The low correlations most likely derive from the limited reliability of the Accomplishments measures and the fact that such achievements are both relatively rare and difficult to predict. In addition to magnitude, the intercorrelations were imprecisely estimated at this sample size, making any true differences that much harder to detect. Of the 42 comparisons between the F-H and General Test correlations (for each of 7 Accomplishments scores, 2 correlations for F-H x 3 for the General Test), two would be expected to be significant by chance alone. Five significant differences were observed, all for the F-H 15-word scores. These scores were more positively correlated with (a) the Aesthetic Expression scale than was GRE-Q (.20 vs. .02, $t = 2.11$, $df = 189$, $p < .05$), (b) the Science scale than was GRE-V (.12 vs. -.06, $t = 2.05$, $df = 189$, $p < .05$), (c) the Mechanical scale than was GRE-A (.11 vs. -.07, $t = 1.98$, $df = 189$, $p < .05$), and (d) the Linguistic scale than was GRE-A (.11 vs. -.10, $t = 2.28$, $df = 189$, $p < .05$) or GRE-Q (.11 vs. -.08, $t = 2.23$, $df = 189$, $p < .05$).

With respect to UGPA, there were six comparisons (2 correlations for F-H x 3 for the General Test). Of these, only the one between GRE quantitative, which was related to UGPA at .37, and the F-H 7-word score, which was correlated with UGPA at .20, showed a significant difference ($t = -2.08$, $df = 189$, $p < .05$).

In sum, F-H scores are clearly distinct from the General Test. Both F-H item types were highly reliable but weakly correlated with this measure, and both item types were more highly related to the Ideational Fluency marker. Also, the F-H 15-word score was different from the individual General Test scales in relating to Accomplishments, generally showing more positive correlations. Finally, the F-H 7-word score was less related to UGPA than was the General Test.

Incremental validity. Table 8 gives the results of regressing UGPA on General Test and F-H 15-word scores; Table 9 shows the effect of regressing Accomplishments on the General Test, UGPA, and F-H in turn. F-H added significantly to prediction for two of the eight outcome variables: UGPA and the Aesthetic Expression subscore. In both cases, the percentage of variance added was quite small (2%-3%). However, the amount of variance accounted for by the other independent variables was also relatively small, such that the increment explained by F-H was proportionally more substantial. Thus, F-H yielded a 13% increase in the proportion of variance explained for UGPA and a 26% increase for Aesthetic Expression.[5] In cerms of the relative importance of each

Table 8
Multiple Regression of UGPA on General Test and F-H 15-word Scores (*n*=192)

Independent Variable	R	R^2	Increment in R^2	Incremental *F*	*p*	Standardized Regression Weight
1. GRE-V						.03
GRE-Q						.27**
GRE-A	.37	.14	.14	10.15	.00	.04
2. F-H (15)	.40	.16	.02	4.85	.03	.16*

Note. General Test scores were entered as a set. Due to rounding, changes in R^2 may not equal the difference between the R^2 values. Regression weights are for the full model.

* = $p < .05$.

** = $p < .01$.

Table 9
Multiple Regression of Accomplishments on General Test, UGPA, and F-H 15-word Scores (n=192)

Independent Variable	R	R^2	Increment in R^2	Incremental F	p	Standardized Regression Weight
			Accomplishments Total Score			
1. GRE-V						.05
GRE-Q						-.04
GRE-A	.13	.02	.02	1.01	.39	-.07
2. UGPA	.35	.12	.10	22.01	.00	.33**
3. F-H (15)	.37	.14	.02	3.41	.07	.14
			Academic Subscore			
1. GRE-V						-.05
GRE-Q						.01
GRE-A	.16	.02	.02	1.60	.19	-.04
2. UGPA	.57	.33	.30	84.99	.00	.61**
3. F-H (15)	.58	.33	.00	1.35	.25	-.07
			Leadership Subscore			
1. GRE-V						-.13
GRE-Q						.16
GRE-A	.14	.02	.02	1.28	.28	-.09
2. UGPA	.14	.02	.00	.01	.92	.00
3. F-H (15)	.15	.02	.00	.24	.62	-.04
			Linguistic Subscore			
1. GRE-V						.24*
GRE-Q						-.17
GRE-A	.22	.05	.05	3.19	.02	-.18
2. UGPA	.27	.07	.02	4.43	.04	.14
3. F-H (15)	.28	.08	.01	1.70	.19	.10
			Aesthetic Expression Subscore			
1. GRE-V						.28**
GRE-Q						-.23*
GRE-A	.25	.06	.06	4.28	.01	.00
2. UGPA	.27	.07	.01	2.10	.15	.08
3. F-H (15)	.32	.10	.03	5.48	.02	.18*
			Science Subscore			
1. GRE-V						-.31**
GRE-Q						.19
GRE-A	.28	.08	.08	5.52	.00	.16
2. UGPA	.29	.08	.00	.73	.39	.05
3. F-H (15)	.31	.09	.01	1.75	.18	.10
			Mechanical Subscore			
1. GRE-V						-.08
GRE-Q						.12
GRE-A	.12	.01	.01	.85	.47	-.11
2. UGPA	.13	.02	.00	.44	.51	-.07
3. F-H (15)	.18	.03	.02	3.08	.08	.14

Note. General Test scores were entered as a set. Due to rounding, changes in R^2 may not equal the difference between the R^2 values for any two steps. Regression weights are for the full model.
* = $p < .05$.
** = $p < .01$.

variable in the full model, the standardized regression weights suggest that F-H carried over half the weight of the "best" GRE scale (GRE-Q) in explaining UGPA and 62% of GRE verbal's power toward predicting Aesthetic Expression.[6]

Incremental validity results for the F-H 7-word items are given in Tables 10 and 11. These items did not add significantly to prediction for any outcome variable (although for Aesthetic Expression, F-H 7 barely missed the $p < .05$ criterion).

To determine whether the major incremental validity findings were associated with certain sample characteristics, three additional analyses were run. The first analysis eliminated examinees whose motivation to take the F-H test was questionable because they spent relatively little time on it. Sixteen examinees (8% of the sample) were eliminated whose total time on the 8-item test was under 40 minutes (i.e., less than 5 minutes per question on average). All regressions were then recomputed. Results were essentially the same as in the full sample, with the minor exception of changing GRE-A to a significant contributor in the full model for the Linguistic subscore.[7]

The second analysis tested the effect of including 25 examinees in the sample whose best language was other than English. For this purpose, the correlations of the regression model variables with a binary language-group indicator (English/other) were examined and then each regression was recomputed with the indicator stepped in just before the F-H scores. The language indicator did not correlate significantly with any predictor or outcome variable, including GRE-V ($r = .13$, $p > .05$), and did not add significantly to prediction. The only notable effect of controlling for best language was to change the status of some borderline predictors, including the F-H 7-word score, which became a significant incremental predictor of Aesthetic Expression, and the 15-word score, which became a significant predictor of Accomplishments.

In the last analysis, the four Accomplishments subscores with skewness indices greater than 1.00 (Linguistic, Aesthetic Expression, Science, Mechanical) were transformed to a logarithmic scale and the regressions recomputed. This transformation had no effect on the status of the F-H scores, changing only GRE quantitative in one instance and UGPA in another, both from marginal nonsignificant to significant contributors to the full model.

Examinee Perceptions

Examinees' opinions of the F-H item type and its delivery are given in Table 12. In general, examinees thought the item's difficulty level and timing were about right (92% and 74%, respectively). In addition, they preferred F-H to the kinds of multiple-choice items found on GRE-A (51% to 28%) and thought that F-H was a fairer indicator than GRE-A questions of the ability to succeed in graduate school (63% to 16%). With respect to computer delivery, examinees indicated they would rather take a computer-based than paper-and-pencil test (44% to 32%). The overwhelming majority appeared to be comfortable and familiar with

Table 10
Multiple Regression of UGPA on General Test and F-H 7-word Scores (n=192)

Independent Variable	R	R^2	Increment in R^2	Incremental F	p	Standardized Regression Weight
1. GRE-V						.04
GRE-Q						.29**
GRE-A	.37	.14	.14	10.15	.00	.04
2. F-H (7)	.39	.15	.01	2.25	.14	.11

Note. General Test scores were entered as a set. Due to rounding, changes in R^2 may not equal the difference between the R^2 values. Regression weights are for the full model.
** = $p < .01$.

Table 11
Multiple Regression of Accomplishments on General Test, UGPA, and F-H 7-word Scores (n=192)

Independent Variable	R	R^2	Increment in R^2	Incremental F	p	Standardized Regression Weight
			Accomplishments Total Score			
1. GRE-V						.05
GRE-Q						-.03
GRE-A	.13	.02	.02	1.01	.39	-.07
2. UGPA	.35	.12	.10	22.00	.00	.33**
3. F-H (7)	.36	.13	.01	2.24	.14	.11
			Academic Subscore			
1. GRE-V						-.05
GRE-Q						.00
GRE-A	.16	.02	.02	1.60	.19	-.03
2. UGPA	.57	.33	.30	84.99	.00	.60**
3. F-H (7)	.58	.33	.00	.45	.51	-.04
			Leadership Subscore			
1. GRE-V						-.15
GRE-Q						.14
GRE-A	.14	.02	.02	1.28	.28	-.07
2. UGPA	.14	.02	.00	.01	.92	-.01
3. F-H (7)	.15	.02	.00	.57	.45	.06
			Linguistic Subscore			
1. GRE-V						.24*
GRE-Q						-.16
GRE-A	.22	.05	.05	3.19	.02	-.19
2. UGPA	.27	.07	.02	4.43	.04	.15*
3. F-H (7)	.27	.07	.00	.52	.47	.05
			Aesthetic Expression Subscore			
1. GRE-V						.28**
GRE-Q						-.22*
GRE-A	.25	.06	.06	4.28	.01	.00
2. UGPA	.27	.07	.01	2.10	.15	.09
3. F-H (7)	.31	.09	.02	3.85	.05	.15
			Science Subscore			
1. GRE-V						-.30**
GRE-Q						.21
GRE-A	.28	.08	.08	5.52	.00	.15
2. UGPA	.29	.08	.00	.73	.39	.06
3. F-H (7)	.29	.08	.00	.07	.80	.02
			Mechanical Subscore			
1. GRE-V						-.07
GRE-Q						.13
GRE-A	.12	.01	.01	.85	.47	-.12
2. UGPA	.13	.02	.00	.44	.51	-.06
3. F-H (7)	.15	.02	.01	1.40	.24	.09

Note. General Test scores were entered as a set. Due to rounding, changes in R^2 may not equal the difference between the R^2 values for any two steps. Regression weights are for the full model.
* = $p < .05$.
** = $p < .01$.

Table 12
Examinee Perceptions of the F-H Item Type and Its Computer Delivery

How easy was F-H?		
Too easy	4%	
About right	92%	
Too difficult	4%	n = 191
How adequate were time limits?		
Too little	15%	
About right	74%	
Too much	12%	n = 192
Which would you rather take: multiple-choice GRE-A questions or F-H?		
Regular multiple-choice	28%	
Formulating Hypotheses	51%	
No preference	22%	n = 192
Which question is a fairer indicator of your ability to undertake graduate study?		
Regular multiple-choice	16%	
Formulating Hypotheses	63%	
No preference	21%	n =189
Which would you rather take: a paper-and-pencil test or a computer-based test?		
Paper-and-pencil	32%	
Computer-based	44%	
No preference	25%	n =192
In the past year, how often have you used a computer?		
Never or almost never	5%	
About once a week	37%	
Daily or almost daily	59%	n = 192
When you have to write a paper for school, how do you usually do it?		
Pencil (or pen) and paper	7%	
Typewriter	4%	
Computer	89%	n = 187
How easy was it to use the computer to answer F-H items?		
Very easy	86%	
Somewhat difficult	13%	
Very difficult	2%	n = 189

Note. Questions were edited for tabular presentation.

computers: Ninety-five percent reported using them about once a week or more over the year before taking the F-H test and 89% indicated using them to write papers for school. Finally, most subjects found it easy to use the machine to answer F-H items (86%), although some examinees experienced difficulty. Those examinees who experienced difficulty were asked to indicate the reasons. Thirty-five reasons were checked, in several instances more than one reason by the same examinee. The most frequently checked reason was "not being a good typist," indicated 11 times.

Impressions offered in response to the questionnaire's "additional comments" item fell into three categories: F-H (138), computer delivery (98), and other (30). The most frequent statements about F-H characterized it as an interesting, enjoyable, or good idea (23) or as a better indicator of critical and/or creative thinking than the General Test (17). Thirteen comments expressed concern about the potential subjectivity of the scoring. In the computer-related category, the most common statements indicated the desire to go back to questions (20) or that the timing was not clearly communicated (10).

Because some examinees reported difficulty in using the computer to answer F-H problems, relationships among relevant Opinion Questionnaire items were computed, as were extension loadings of selected Opinion items on the F-H 7-word, F-H 15-word, and ideational fluency factors. Because the results are open to interpretation, their main value is in supporting the need for closer study of the role of computer familiarity in F-H test performance.

In general, the results can be viewed as consistent with the hypothesis that computer familiarity has an effect. For example, examinees who reported limited computer use tended to indicate more difficulty with the F-H interface than did experienced users ($r = -.22$, $t = -3.14$, $p < .01$). In addition, those who found the interface difficult to use tended to do worse on F-H than those who experienced no difficulty (loading for F-H 7-word items = -.18, $t = -2.46$, $p < .05$ and loading for F-H 15-word = -.22, $t = -3.09$, $p < .01$). Third, those who were occasional computer users were more likely than experienced users to feel that the F-H time limits were too short ($r = .21$, $t = 3.10$, $p < .01$). Finally, as would be expected, there was no relation between performance on the (paper-and-pencil) Ideational Fluency measure and either frequency of computer use (loading = .08, $t = 1.07$, $p > .05$) or difficulty with the F-H interface (loading = -.02, $t = -.28$, $p > .05$).

Discussion

This study assessed generalizability, validity, and examinee perceptions for a computer-delivered Formulating-Hypotheses test. Generalizability results showed little variation across human scorers; a generalizability coefficient in the .80s would require a two- to four-question test scored by a single judge taking 20 to 40 minutes to administer. As in previous studies, validity analyses found F-H to be only weakly related to General Test scores and to differ from the General Test primarily in stronger relations to an Ideational Fluency measure. Versions of F-H based on different response limitations tapped somewhat different abilities, with the 15-word constraint producing a more promising result. This version had more positive relations with an accomplishments inventory than did the individual General Test scales; it also added incrementally over the General Test to explaining self-reported undergraduate grades, and beyond the General Test and grades in predicting creative expression. Finally, there was some suggestion that computer familiarity--in particular, typing skill--might constitute a source of construct-irrelevant variance.

These results confirm Frederiksen and Ward's (1978) finding that F-H can broaden the abilities measured by the General Test and possibly increase its predictive validity. The current study has extended this finding to the computer-delivered format and to a population that is more diverse in the disciplines represented. F-H would add to the General Test a measure of the ability to generate alternative explanations. This ability appears to be allied with the kind of creativity that underlies idea generation as well as artistic accomplishment. At the same time, this ability seems to have more global implications in that it adds incrementally to the explanation of undergraduate school performance.

Several important issues will need to be resolved in determining whether and how F-H might be included as a GRE Program offering. Some of these issues can be investigated by inserting F-H into the experimental section of an operational computer-based General Test administration. Collecting data in this manner would provide large sample sizes, an applicant population, and the administrative conditions needed for more confidently generalizing results. Other issues may require manipulations not amenable to an operational administration and will have to be accommodated through separate studies.

Data from administering F-H as an experimental General Test section might be used to confirm the source of the measurement differences between 15-word and 7-word items. Because the current study did not experimentally vary response constraint, it is impossible to know whether the two F-H versions operated differently because of word limits or because of some other characteristic that also happened to differentiate the two item types. Spiraling two forms of the test with common item stems but different response limits should resolve this question, as well as give a better indication of how important the observed measurement differences are.

A second use of data from an experimental section would be to study task generality across major fields. For example, does F-H measure the same dimension for students majoring in the arts and humanities as for those majoring in the natural sciences? If not, a more domain-specific approach might be tried in which items were placed in disciplinary contexts matched to student major.

Data from a large-scale administration might also help identify measurement models for assigning scores to items and for combining scores across items. In the present study, we added the number of acceptable hypotheses to form an item score and then summed the results. More informative scoring methods might take into account the fact that examinees run out of ideas: The more hypotheses one generates for an item, the harder it is to produce the next hypothesis. Also, some items are harder to generate explanations for and should be weighted accordingly.

A fourth question to address with large-scale administration data concerns predictive relations. In the present study, the school performance and accomplishments criteria summarized mostly achievement during the college years. Relating F-H performance to graduate grade-point average and accomplishments would provide more information about the value of the computer-based version of the task for graduate assessment. We should note, however, that based on the current study, the absolute contribution to prediction is likely to be relatively small. Although predictive relations may give secondary support, the primary argument for using the task probably will be in broadening the range of abilities tested.

Finally, data from the experimental section would permit a more thorough examination of subgroup differences. Our data showed no relation between F-H and sex group membership. Small sample sizes precluded a similar analysis for racial/ethnic groups, as well as an examination of potential differences in the meaning of F-H scores across populations.

Studies that may need to be conducted outside the purview of the experimental section relate primarily to construct-irrelevant variance. Perhaps the most pressing of these studies centers on computer familiarity. This study might relate computer-delivered versus paper-and-pencil F-H tests to typing proficiency. If the computer-delivered test produced higher relations, F-H might be offered in both forms so that examinees could choose the most appropriate mode. The paper-and-pencil offering would be temporary, however. Several firms are investing heavily in developing low-cost computers that achieve new levels of user friendliness, in some cases by eliminating the need for keyboarding skills altogether. The first generation of these "pen-based" computers, which recognize input written with an electronic stylus, is already on the market (Linderholm, Apiki, & Nadeau, 1992). Although recognition accuracy might not yet be high enough for testing applications, the technology is expected to improve very rapidly. Once it is perfected, testing programs could become all-electronic, eliminating the need for typing proficiency by offering both pen-based and keyboard options.

In addition to keyboarding skill, test-taking strategy is a potential source of construct-irrelevant variance that may require special study. One such strategy is to pose many instances of the same general idea (e.g., for the swing set item, state all the kinds of bad weather that could have prevented use of the playground equipment). A second strategy is--after exhausting good hypotheses--to generate vague ones that may get mistaken for poorly expressed, but creditable, explanations. A third approach is to routinely apply explanations that are likely to work across items (e.g., bad weather). The effectiveness of these strategies might be limited by test directions and rubrics that define response duplication in relatively broad terms, that require each response to contain a reasonably clear and full explanation, and that identify as unacceptable specific response classes that apply to many items. (Item stems might also be worded to disallow some of these "super explanations"). Finally, a penalty might be imposed for each duplicate response, thereby discouraging this form of gamesmanship. Obviously, how well these counter measures function will need to be empirically assessed, as will their effects on the meaning of F-H scores.

Several limitations of this study should be noted. First, a graduate student sample was used instead of one drawn from the less select examinee population. Second, the sample size was quite small. Finally, data were gathered in an experimental setting that differed from the operational testing context in important respects. These facts place clear restrictions on the generalizability of results. Still, the results agree fundamentally with those of Frederiksen and Ward (1978), who tested a large applicant sample in an operational setting, suggesting that the current findings may have wider applicability than these restrictions would otherwise imply.

Footnotes

1. These studies began with an investigation of four innovative item types collectively dubbed "The Tests of Scientific Thinking." Only the results for F-H are discussed here.

2. For UGPA, 16 examinees were missing data and were assigned the mean value.

3. Significant improvements were also found for the two-factor solution over the one-factor model (chi-square difference = 159.3, df difference = 1, $p < .01$) and for the one-factor solution over the null model (chi-square difference = 1502.6, df difference = 12, $p < .01$).

4. The low relationship between the Ideational Fluency marker and the General Test might suggest considering the former as a potential graduate admissions measure. Because ideas are to be generated within broad constraints, the Ideational Fluency test has no real substantive criteria for determining what is a correct reaponse. The measure works only as long as examinees try to generate ideas honestly from the stimulus, which in a low-stakes experimental situation, most examinees do. As such, the measure is better suited to the role of research marker than it is to a high-stakes test.

5. These figures are computed from R^2 values taken to four decimal places.

6. In some instances, the standardized regression weights for GRE-V or GRE-Q are significant and negative. In the case of the Aesthetic Expression and Science subscores, this may be because science majors in the sample were frequently non-native speakers with quantitative skills better than, and verbal skills worse than, those of individuals majoring in other areas.

7. Mean General Test and F-H scores were marginally higher (from .06 to .08 and .10 to .12 standard deviation units, respectively), and standard deviations somewhat lower, for the reduced sample (n=176) than for the full one. Examination of the resulting zero-order correlation matrix showed the relations between the General Test and the F-H scores to be lowered from four to eight points for the 15-word items and from five to ten points for the 7-word items; no other correlations were systematically affected.

References

Baird, L. L. (1976). Using self-reports to predict student performance (Research Monograph No. 7). New York: College Entrance Examination Board.

Bennett, R. E. (1993). On the meanings of constructed response. In R. E. Bennett & W. C. Ward (Eds.), Construction vs. choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment (pp. 1-27). Hillsdale, NJ: Erlbaum.

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

Bentler, P. M., & Bonnett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. Psychological Bulletin, 88, 588-606.

Carlson, S. B. (1985). Pilot study of an application of formulating hypotheses problems within the context of law school admissions (Final report to the LSAC/LSAS Test Battery Workgroup). Princeton, NJ: Educational Testing Service.

Carlson, S. B., & Ward, W. C. (1988). A new look at formulating hypotheses items (RR-88-12). Princeton, NJ: Educational Testing Service.

Ekstrom, R. B., French, J. W., & Harman, H. H. (1976). Kit of factor-referenced cognitive tests. Princeton, NJ: Educational Testing Service.

Educational Testing Service. (1991). GRE 1991-92 registration and information bulletin. Princeton, NJ: Educational Testing Service.

Educational Testing Service. (1992). GRE 1992-93 guide to the use of the Graduate Record Examinations program. Princeton, NJ: Author.

Frederiksen, N. (1959). Development of the test "Formulating Hypotheses": A progress report (Office of Naval Research Technical Report, Contract Nonr-2338(00)). Princeton, NJ: Educational Testing Service.

Frederiksen, N., & Evans, F. R. (1974). Effects of models of creative performance on ability to formulate hypotheses. Journal of Educational Psychology, 66, 67-82.

Frederiksen, N., & Ward, W. C. (1978). Measures for the study of creativity in scientific problem-solving. Applied Psychological Measurement, 2(1), 1-24.

Frederiksen, N., Ward, W. C., Case, S. M., Carlson, S. B., & Samph, T. (1981). Development of methods for selection and evaluation in undergraduate medical education (RR-81-4). Princeton, NJ: Educational Testing Service.

Kaplan, R. M. (1992). Using a trainable pattern-directed computer program to score natural language item responses (RR-91-31). Princeton, NJ: Educational Testing Service.

Kaplan, R. M., & Bennett, R. E. (in press). Using the Free-Response Scoring Tool to automatically score the Formulating-Hypotheses item (RR-xx-xx). Princeton, NJ: Educational Testing Service.

Klein, S. P., Frederiksen, N., & Evans, F. P. (1969). Anxiety and learning to formulate hypotheses. Journal of Educational Psychology, 60, 465-475.

Linderholm, O., Apiki, S., Nadeau, M. (1992). The PC gets more personal. Byte, 128-138.

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

McNemar, Q. (1962). Psychological statistics. New York: Wiley.

Powers, D. E., & Enright, M. K. (1987). Analytical reasoning skills in graduate study: Perceptions of faculty in six fields. Journal of Higher Education, 58, 658-682.

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.

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.

Thorndike, R. L. (1982). Applied psychometrics. Boston: Houghton Mifflin.

Torrance, E. P. (1974). Torrance Tests of Creative Thinking. Besenville, IL: Scholastic Testing Service, Inc.

Tucker, C. (1985). Delineation of reasoning processes important to the construct validity of the analytical test. Princeton, NJ: Educational Testing Service.

Wallach, M. A., & Kogan, N. (1965). Modes of thinking in young children: A study of the creativity-intelligence distinction. New York: Holt.

Ward, W. C., Carlson, S. B., & Woisetschlaeger, E. (1983). Ill-structured problems as multiple-choice items (RR-83-6). Princeton, NJ: Educational Testing Service.

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.

Appendix A

F-H Test Directions

Beginning

When finished reading directions click on the icon below

ETS ®

GRE Formulating Hypotheses Test

The purpose of this test is to measure your ability to solve problems that ask you to think of hypotheses that might explain a social phenomenon, the findings from a research study, or some other situation.

The problems do not require any special or technical knowledge. They involve situations similar to ones you might read about in a newspaper or magazine.

The test will consist of 8 problems. Answer each problem as completely as possible before moving on to the next problem. You will NOT be allowed to return to a problem after leaving it.

You will have 1 hour and 20 minutes to enter your hypotheses. You will need to

Dismiss Directions

Test Quit

?
Help

End

You will have 1 hour and 20 minutes to enter your hypotheses. You will need to keep track of the time and carefully pace yourself to answer each question. YOU MUST ANSWER EACH QUESTION TO RECEIVE PAYMENT FOR PARTICIPATION IN THE PROJECT.

When finished reading directions click on the icon below

Dismiss Directions

Test

Quit

?

Help

BEST COPY AVAILABLE

Beginning

When finished reading directions click on the icon below

Your task is to think of as many **different** plausible hypotheses as possible for the phenomenon presented in each problem. Remember, you are not looking for a single, right answer but for as many **different** plausible answers as you can specify within a limited time period.

You will be given 1 point for each **different** plausible hypothesis, with a maximum of 15 points for each problem. In some problems, you will be asked to limit each hypothesis to 7 words; in other problems you can use up to 15 words for each hypothesis. You will not be given credit for a hypothesis that is implausible or for one that duplicates the meaning of a hypothesis you have already given for that problem.

Look at the following two sample problems that illustrate plausible and implausible hypotheses.

SAMPLE PROBLEM 1

Jefferson City Public Transit Campaign

In an attempt to reduce air pollution levels, Jefferson City implemented a plan to make its public transit system more attractive and convenient to riders. Nonpolluting electric trains and buses with antipollution devices were introduced. Bus routes were extended to all parts of the city, and subway stations were cleaned and repainted. The improvements were

Dismiss Directions

Test Quit

Help

More Available

subway stations were cleaned and repainted. The improvements were funded through a federal grant and a fare increase. Nevertheless, 1 year after the system was improved, air pollution levels in Jefferson City were higher than ever before.

Think of hypotheses (possible explanations) for the increase in air pollution levels. Write each hypothesis as a separate answer of no more than **15 words.**

Examples of hypotheses that would be considered different plausible explanations:

1. Higher mass transit fares caused more people to drive cars than to take public transportation.

2. The improved transit system caused several factories to relocate to Jefferson City.

3. Antipollution devices installed on the buses did not work properly.

4. A transit workers' strike shut down the public transit system

5. A shift in wind currents caused a thermal inversion, which trapped pollution in the city.

When finished reading directions click on the icon below

Dismiss Directions

Test Quit

Help

BEST COPY AVAILABLE

More Available

the city.

Examples of hypotheses that would NOT be considered different plausible explanations:

1. Federal standards for automobile exhaust controls were made stricter.

2. The antipollution devices on the buses worked better than expected.

(These would not be plausible explanations because stricter controls for automobile exhaust and better-than-expected performance of the antipollution devices would lower rather than raise pollution levels.)

3. Two years later pollution levels in Jefferson City were lower than ever.

(This is not a plausible explanation because what happened two years later is irrelevant to what happened 1 year after the transit system was implemented.)

4. More people drove because it was cheaper.

(Although it is plausible, this is the same as the hypothesis above, "Higher mass transit fares caused more people to drive cars than to take public transportation.")

SAMPLE PROBLEM 2

When finished reading directions click on the icon below

Dismiss Directions

Test Quit

Help

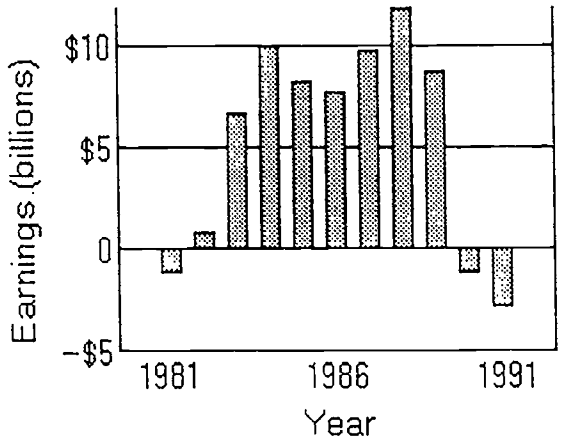
Directions

More Available

When finished reading directions click on the icon below

SAMPLE PROBLEM 2

Combined Earnings for the Three Largest Automobile Manufacturers in Country *X* From 1981 to 1991



From 1983-1989, the earnings of the three largest automobile manufacturers in Country *X* never fell below 5 billion dollars a year. In 1990, the companies suffered a combined loss of 1.2 billion dollars, and in 1991 the combined loss was over 3 billion dollars.

Think of hypotheses that might explain the losses in 1990 and 1991 in Country *X*. Write each hypothesis as a separate answer of no more than **7 words.**

Dismiss Directions

Test Quit

Help

More Available

When finished reading directions click on the icon below

Examples of hypotheses that would be considered different plausible explanations:

1. People stopped buying because of recession.

2. People bought cars from other countries.

3. An earthquake destroyed many of the factories.

4. Government subsidies were cut off after 1989.

Because of the word limit, you should include in each response only those details that cannot be readily inferred from your response. For example, the first response simply says that people stopped buying, meaning that they stopped buying cars made in Country *X* because of a recession in 1990-91. The fourth response, on the other hand, mentions a specific time frame, but implies that the government subsidies in question were for the auto industry in Country *X*.

Examples of hypotheses that would NOT be considered different plausible explanations:

1. Profits on cars went up in 1990-91.

(This response is not plausible because if profits on cars made by the 3 largest

Dismiss Directions

Test Quit

Time

Help

End

(This response is not plausible because if profits on cars made by the 3 largest automobile industries in Country *X* went up during this time frame, the manufacturers' profits also would have gone up rather than down.)

2. Development of teleportation devices made automobiles obsolete.

(This response is not plausible. As a general rule, any response that depends on science fiction, supernatural, or other generally disputed phenomena will not be given credit.)

3. Poor economic conditions slowed car sales.

(Although it is plausible, this is the same as the hypothesis above, "People stopped buying because of recession.")

PLEASE TAKE A MOMENT TO REVIEW THIS PROBLEM TYPE.

When finished reading directions click on the icon below

Dismiss Directions

Test Quit

? Help

When finished reading directions click on the icon below

Try to complete the following 4 problems in 40 minutes, using no more than 10 minutes per problem, so that you will have an equal amount of time for the remaining 4 problems. Do not spend too much time on any one problem. The timer will flash every 10 minutes as an aid in pacing through the test.

For the following 4 problems, you will be limited to 7 **words** for each hypothesis.

Dismiss Directions

Test Quit

Time

? Help

When finished reading directions click on the icon below

Note the amount of time you have left for the remaining 4 problems and pace yourself accordingly. Do not spend too much time on any one problem.

For the following 4 problems, you will be limited to **15 words** for each hypothesis.

Dismiss Directions

Test Quit | Time | Help

Testing Completed

You have completed the computer questions.

Please raise your hand so the center staff can give you the remaining paper and pencil materials.

Appendix B

An F-H Scoring Rubric

1. SWING AND SEESAW EQUIPMENT NOT USED

GENERAL CATEGORY	SPECIFIC CATEGORY
A) CHILDREN NOT AT SCHOOL	1. Played hooky 2. Teacher strike 3. Teacher conference 4. Emergency/disaster 5. Away on field trip, on tour, etc.
B) NOT ALLOWED TO PLAY	1. Being punished 2. Required to attend assembly, special event, special visitor 3. Religious observance - no playing 4. Being tested, tryouts, rehearsals 5. Must not get dirty or messy, picture day 6. Bomb Scare / Fire Drill
C) PREOCCUPIED ELSEWHERE	1. Playing a game, on other equipment, etc. 2. Parade - accident - eclipse, etc., distracts them 3. Something new, toy, pet, etc., engages them
D) CONDITION OF CHILDREN	1. Too tired to play 2. Hurt, injured 3. Ill
E) CHILDREN'S FEELINGS / ATTITUDE	1. Very sad, upset over a death, tragedy, etc. 2. Bored, don't feel like playing 3. Afraid to play - someone hurt 4. Refuse to play - Boycott
F) NO ACTIVITY PERIOD TODAY	1. Early dismissal 2. Another event ran into overtime
G) PROBLEM WITH TEACHERS / ADULTS	1. Lack of Supervision 2. Teachers forgot activity period 3. Adults using the equipment
H) CONDITION OF SWING / SEESAW	1. Off limits, being inspected, recent accident 2. Dangerous (splinters, sharp edges) 3. Broken 4. Being Repaired 5. Being Painted 6. Removed - loaned 7. Stolen 8. Knocked over by storm 9. Nasty graffiti
I) CONDITION ON / NEAR PLAYGROUND	1. Puddles, mud, too wet, snow, ice on equipment 2. Being paved, seeded 3. Downed power line 4. Dangerous people (crooks, molester, drug dealer) 5. Dangerous animals (bees, mad dog) 6. Fire outside 7. Gate locked, no key 8. Broken glass. etc. 9. Unpleasant smell, sight (vomit, dead animal)
J) BAD WEATHER	1. Rain, Lightning, Snow, Winds, Cold, Heat
K) ECONOMIC REASONS	1. Liability insurance expired
L) OBSERVATION ERRORS	1. Observation was made too early/not activity period now 2. Children are between turns on the equipment

Appendix C

Ideational Fluency Measure

PLEASE COMPLETE THIS FORM BEFORE TAKING THE COMPUTER-BASED TEST

GRE Research: Paper and Pencil Divergent Thinking Problems

Name: ____

Social Security Number: ______________________________

This test asks you to list as many ideas as you can about a topic or a picture. The directions at the beginning of each task provide specific information about what you are to do. You will have a total of 25 minutes to complete this test. Remember: to receive compensation you must answer all questions, so please pace yourself accordingly.

GO ON TO NEXT PAGE

Part I

TOPICS

Directions

These items test how many ideas you can think of about a topic. Be sure to list all the ideas you can about the topic whether or not they seem important to you. You are not limited to one word. Instead you may use a word or phrase to express each idea.

Here is a sample topic, "A train journey." Two examples are given below of ideas about the topic. Look at these examples. Now go ahead and fill in the blanks with more ideas about this topic.

Number of miles

Catching the train

Your score will be the number of appropriate ideas that you write.

PLEASE DO NOT TURN THIS PAGE UNTIL ASKED TO DO SO

BY THE TEST ADMINISTRATOR

The topic is: "A man going up a ladder."

List all the ideas you can about <u>a man going up a ladder</u>.

GO ON TO NEXT PAGE

PART II

UNUSUAL QUESTIONS

In this item you are to think of as many questions as you can about boxes. These questions should lead to a variety of different answers and might arouse interest and curiosity in others concerning boxes. Try to think of questions about aspects of cardboard boxes which people do not usually think about.

1. ______________________________
2. ______________________________
3. ______________________________
4. ______________________________
5. ______________________________
6. ______________________________
7. ______________________________
8. ______________________________
9. ______________________________
10. ______________________________
11. ______________________________
12. ______________________________
13. ______________________________
14. ______________________________
15. ______________________________
16. ______________________________
17. ______________________________
18. ______________________________
19. ______________________________
20. ______________________________

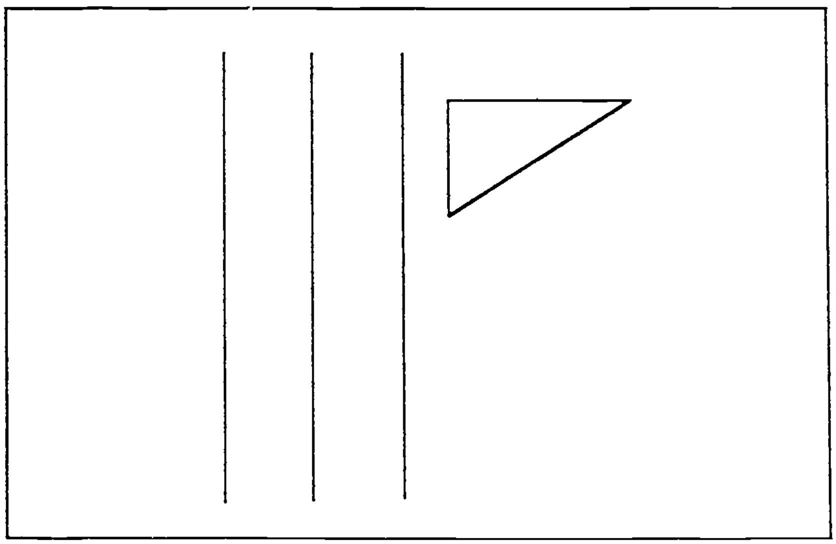
GO ON TO NEXT PAGE

Part III

PATTERN MEANINGS

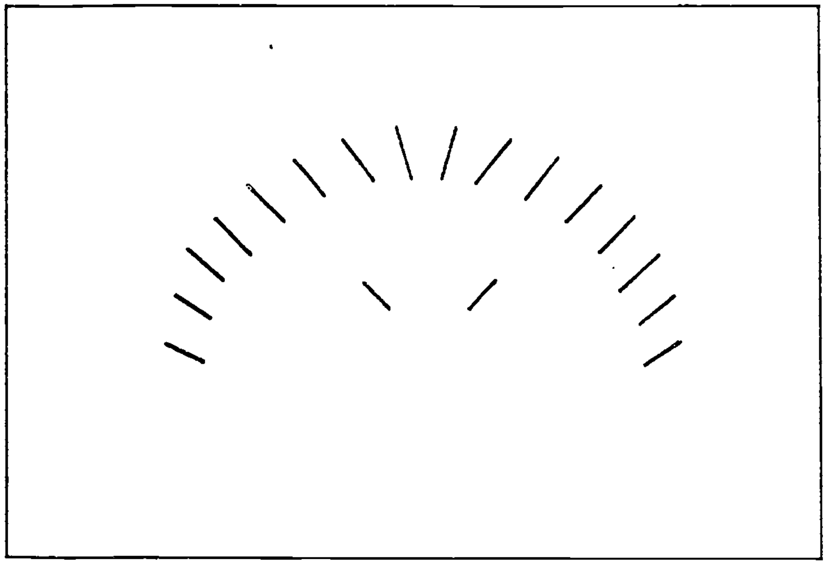
In the items that follow, you will be asked to list ideas about some unfinished drawings. In the spaces provided, list all of the things you can think of that the finished drawing might represent. Drawings can be rotated in any direction.

For example, if the drawing below were finished, it could represent a bird, a sailboat on the water, part of a map or a mask.



For the next two items, write as many things as you can think of that the drawing could represent if it were finished.

GO ON TO NEXT PAGE



Remember: The drawing can be rotated in any direction.

1. ____________________

2. ____________________

3. ____________________

4. ____________________

5. ____________________

6. ____________________

7. ____________________

8. ____________________

9. ____________________

10. ____________________

11. ____________________

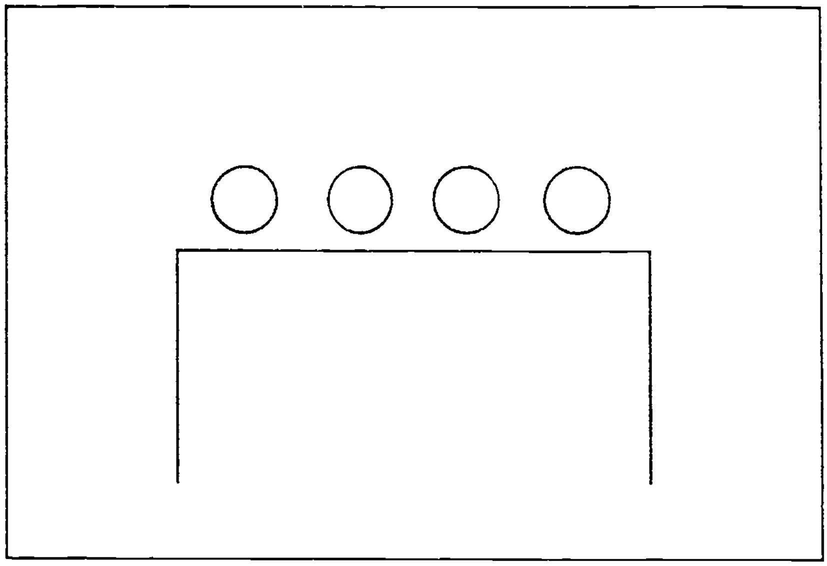
12. ____________________

13. ____________________

14. ____________________

15. ____________________

GO ON TO NEXT PAGE



Remember: The drawing can be rotated in any direction.

1. ______________________________

2. ______________________________

3. ______________________________

4. ______________________________

5. ______________________________

6. ______________________________

7. ______________________________

8. ______________________________

9. ______________________________

10. ______________________________

11. ______________________________

12. ______________________________

13. ______________________________

14. ______________________________

15. ______________________________

WHEN YOU HAVE FINISHED PLEASE RAISE YOUR HAND TO ALERT THE TEST ADMINISTRATOR THAT YOU ARE READY FOR THE COMPUTER-BASED TEST.

Appendix D

Accomplishments Questionnaire

COMPLETE THIS QUESTIONNAIRE RIGHT AFTER THE COMPUTER-BASED TEST

Name: ____________________ Social Security #: ____________________

GRE RESEARCH: ACTIVITIES AND ACCOMPLISHMENTS QUESTIONNAIRE

Descriptions of a variety of activities and accomplishments in school, in volunteer work, or in part-time or full-time jobs are listed below. Please read each description, and then indicate whether you engaged in the activity or achieved the accomplishment since high school by checking the "YES" or "NO" box next to the description. If you check the "YES" box, also fill in the requested information in the blank below the description. Many of the activities and accomplishments are relatively uncommon ones that you may not have engaged in or achieved.

REMEMBER: To receive compensation, you must answer all questions.

YES NO

[ ] [ ] 1. Was in an independent study program for outstanding students in college.
If YES: ____________________
Program and School

[ ] [ ] 2. Was on the Dean's list in college.
If YES: ____________________
Year and School

[ ] [ ] 3. Was elected to Phi Beta Kappa or an equivalent honor society in college.
If YES: ____________________
Society and School

[ ] [ ] 4. Graduated from college with honors (e.g., cum laude).
If YES: ____________________
Honors and School

[ ] [ ] 5. Was the valedictorian or salutatorian in college.
If YES: ____________________
School

[ ] [ ] 6. Served on a student-faculty committee in college.
If YES: ____________________
Position, Organization, and School

[ ] [ ] 7. Was appointed or elected to a school-wide student group, such as student council or student senate, in college.
If YES: ____________________
Position, Organization, and School

[ ] [ ] 8. Was elected to a major class office (e.g., president, vice president, treasurer) in college.
If YES: ____________________
Position, Class, and School

[ ] [ ] 9. Was appointed or elected an officer in a club, sorority, professional society, or other organized interest group.
If YES: ____________________
Position and Organization

[ ] [ ] 10. Star a club, sorority, professional society, or other organized group.
If YES: ____________________
Organization

[ ] [ ] 11. Was a member of a school-wide debating team in college.
If YES: ____________________
Team and School

YES NO

[ ] [ ] 12. Made a formal speech at a large public gathering (i.e., over 100 people), other than graduation ceremonies.
If YES: ______________________________
Subject and Sponsoring Organization

[ ] [ ] 13. Was a winner or runner-up of a prize or award for public speaking from a statewide, regional, or national organization.
If YES: ______________________________
Award and Organization

[ ] [ ] 14. Was a master or mistress of ceremonies at a large banquet, awards ceremony, or show (i.e., over 100 people).
If YES: ______________________________
Gathering and Sponsoring Organization

[ ] [ ] 15. Appeared regularly on a radio or television program in a non-performing role (e.g., announcer, disc jockey, host, correspondent).
If YES: ______________________________
Position, Duties, and Broadcasting Organization

[ ] [ ] 16. Was a paid spokesperson or press aide for a company or other organization.
If YES: ______________________________
Position, Duties, and Organization

[ ] [ ] 17. Wrote a "letter to the editor" that was published.
If YES: ______________________________
Subject and Publication

[ ] [ ] 18. Wrote a feature article, column, or editorial that was published.
If YES: ______________________________
Type of Material, Subject, and Publication

[ ] [ ] 19. Was on the editorial staff of a publication or a radio or television station.
If YES: ______________________________
Position, Duties, and Organization

[ ] [ ] 20. Wrote a speech for someone else that was given at a large public gathering (i.e., over 100 people).
If YES: ______________________________
Speaker, Subject, Gathering, and Sponsoring Organization

[ ] [ ] 21. Wrote advertising or public relations material, for pay, for a company or other organization.
If YES: ______________________________
Position, Duties, and Organization

[ ] [ ] 22. Wrote technical manuals or other instructional material, for pay, for a company or other organization.
If YES: ______________________________
Position, Duties, and Organization

[ ] [ ] 23. Wrote poetry, fiction, or essays that were published.
If YES: ______________________________
Type of Writing and Publication

[ ] [ ] 24. Wrote a play that was publicly performed or a screenplay for a film that was publicly shown.
If YES: ______________________________
Play or Film and Theater or Film Organization

[ ] [ ] 25. Wrote the script for a dramatic or comedy show for radio or television that was publicly broadcast.
If YES: ______________________________
Show and Broadcasting Organization

YES NO

[ ] [ ] 26. Invited to participate in a writer's workshop sponsored by a statewide, regional, or national organization.
If YES: ______________________________
Workshop and Organization

[ ] [ ] 27. Was a winner or runner-up of a prize or award for creative writing from a statewide, regional, or national organization.
If YES: ______________________________
Type of Writing, Award, and Organization

[ ] [ ] 28. Designed the scenery or costumes for a play or dance that was publicly performed or a film that was publicly shown.
If YES: ______________________________
Activity, Play or Dance, and Theater or Film Organization

[ ] [ ] 29. Created artwork (i.e., painting, photography, sculpture) that was exhibited.
If YES: ______________________________
Type of Art and Exhibition

[ ] [ ] 30. Created artwork (e.g., painting, photography, sculpture) that was sold to a gallery or dealer or that was sold by a gallery or dealer to someone else.
If YES: ______________________________
Type of Art and Gallery or Dealer

[ ] [ ] 31. Did artwork (i.e., painting, photography, sculpture), for pay, for a company or other organization.
If YES: ______________________________
Position, Duties, and Organization

[ ] [ ] 32. Was a winner or runner-up of an award or prize for art (e.g., painting, photography, sculpture) from a statewide, regional, or national organization.
If YES: ______________________________
Type of Art, Award, and Organization

[ ] [ ] 33. Sang as a soloist or member of a group at a public performance.
If YES: ______________________________
Activity or Group and Theater or Hall

[ ] [ ] 34. Played a musical instrument as a soloist or member of a group at a public performance.
If YES: ______________________________
Activity or Group and Theater or Hall

[ ] [ ] 35. Conducted a band, orchestra, or vocal group at a public performance.
If YES: ______________________________
Group and Theater or Hall

[ ] [ ] 36. Composed or arranged music that was publicly performed.
If YES: ______________________________
Type of Music, Performer, and Theater or Hall

[ ] [ ] 37. Was a winner or runner-up of an award or prize for composing or performing music from a statewide, regional, or national organization.
If YES: ______________________________
Activity, Award, and Organization

[ ] [ ] 38. Acted in a play that was publicly performed or a film that was publicly shown.
If YES: ______________________________
Play or Film and Theater or Film Organization

[ ] [ ] 39. Acted in a radio or television show that was publicly broadcast.
If YES: ______________________________
Show and Broadcasting Organization

[ ] [ ] 40. Directed a play that was publicly performed or a film that was publicly shown.
If YES: ____________________
Play or Film and Theater or Film Organization

[ ] [ ] 41. Directed a dramatic or comedy show for radio or television that was publicly broadcast.
If YES: ____________________
Show and Broadcasting Organization

[ ] [ ] 42. Was a winner or runner-up of a prize or award for acting or directing from a statewide, regional, or national organization.
If YES: ____________________
Activity, Award, and Organization

[ ] [ ] 43. Was a research assistant on a scientific project in college.
If YES: ____________________
Position, Duties, Project, and School

[ ] [ ] 44. Authored or co-authored a paper that was presented at a scientific meeting.
If YES: ____________________
Subject and Meeting

[ ] [ ] 45. Authored or co-authored an article that was published in a scientific journal.
If YES: ____________________
Subject and Publication

[ ] [ ] 46. Received a grant for scientific research from a foundation or government agency.
If YES: ____________________
Subject and Granting Agency

[ ] [ ] 47. Was a winner or runner-up of an award or prize for science from a statewide, regional, or national organization.
If YES: ____________________
Activity, Award, and Organization

[ ] [ ] 48. Designed machinery or equipment, for pay, for a company or other organization.
If YES: ____________________
Position, Duties, and Organization

[ ] [ ] 49. Built or maintained machinery or equipment, for pay, for a company or other organization.
If YES: ____________________
Position, Duties, and Organization

[ ] [ ] 50. Operated machinery or equipment, other than standard office machines, for pay, for a company or other organization.
If YES: ____________________
Position, Duties, and Organization

[ ] [ ] 51. Designed new buildings or the renovation of old ones, for pay, for a company or other organization.
If YES: ____________________
Position, Duties, and Organization

[ ] [ ] 52. Constructed, renovated, or maintained buildings, for pay, for a company or other organization.
If YES: ____________________
Position, Duties, and Organization

[PLE]ASE BEGIN WORK ON THE FINAL PACKET LABELED OPINION QUESTIONNAIRE.

Appendix E

Opinion Questionnaire

COMPLETE THIS FORM *AFTER* THE ACTIVITIES & ACCOMPLISHMENTS QUESTIONNAIRE

NAME:________________________ SS#:________________

GRE Research: OPINION QUESTIONNAIRE

Please answer each of these questions by circling the letter next to the phrase that best characterizes your opinion. Please remember to answer ALL questions.

1. How easy were the computer-based Formulating Hypotheses items?

   a. Too easy
   b. About right
   c. Too difficult

2. How adequate was the time allowed for answering the computer-based Formulating Hypotheses questions?

   a. Too little
   b. About right
   c. Too much

3. Which kind of test question would you rather take: multiple-choice questions like those on the analytical section of the GRE General Test or questions like Formulating Hypotheses?

   a. Regular multiple-choice
   b. Formulating Hypotheses
   c. No preference

4. Which kind of question do you think is a fairer indicator of your ability to undertake graduate study: multiple-choice questions like those on the analytical section of the GRE General Test or questions like Formulating Hypotheses?

   a. Regular multiple-choice
   b. Formulating Hypotheses
   c. No preference

5. Which kind of test would you rather take: a paper-and-pencil test or a computer-based one?

   a. Paper-and-pencil
   b. Computer-based
   c. No preference

6. In the past year, how often have you used a computer?

   a. Never or almost never
   b. About once a week
   c. Daily or almost daily

7. When you have to write a paper for school, how do you usually do it?

a. Pencil (or pen) and paper
b. Typewriter
c. Computer

8. How easy was it to use the computer to answer the Formulating Hypotheses items?

a. Very easy
b. Somewhat difficult
c. Very difficult

9. If you found it "somewhat difficult" or "very difficult" to use the computer, why was that? (Check all that apply)

a. The tutorial program didn't do a good job explaining how to use the computer
b. The computer screens were confusing
c. The sequence of commands was not clear
d. The mouse was hard to use
e. I am not a good typist
f. Other:______________________________

____

10. Additional Comments:

____
____
____
____
____
____
____
____
____

WHEN YOU HAVE COMPLETED THIS FORM, RAISE YOUR HAND TO ALERT THE TEST ADMINISTRATOR THAT YOU ARE THROUGH.

THANK YOU FOR PARTICIPATING IN THIS STUDY!

54020 10243 • Y113M.7 • 209082