Relationships Among Multiple-Choice Reasoning Items and a Constructed-Response Generating-Explanations Task

Mary K. Enright, Donald A. Rock, and Randy Elliot Bennett

August 1998
GRE Board Professional Report No. 94-15P
ETS Research Report 98-41

Educational Testing Service, Princeton, New Jersey 08541
This report presents the findings of a research project funded by and carried out under the auspices of the Graduate Record Examinations Board.
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.

EDUCATIONAL TESTING SERVICE, ETS, the ETS logo, GRADUATE RECORD EXAMINATIONS, and GRE are registered trademarks of Educational Testing Service.

Copyright © 1998 by Educational Testing Service. All rights reserved.
Acknowledgments

We wish to express our gratitude to the many individuals who contributed to this study. Lois Frankel, Timothy Habick, and Carol Tucker developed the battery of experimental reasoning items. The production of the computer-based test was coordinated by Jan Flaugher with support from Holly Knott and Probal Tahbildar. The data were collected by graduate assistants who were supported by the Graduate Record Examinations Program. Special thanks go to Margaret Redman for her invaluable assistance in collecting data and coordinating the work of the graduate assistants. Brent Bridgeman, Timothy Habick, and Don Powers offered helpful comments on an earlier version of the report. Finally, we wish to thank the Graduate Record Examinations Board and its Research Committee for supporting this work.
Abstract

The potential benefits of computer-based testing include the ability to present a wider variety of item formats and to tailor tests for specific purposes. In this study we examined the relationships among revised reasoning items that had more varied formats than the traditional items, a constructed-response generating-explanations task, and the current GRE General Test. In addition we examined the relationship of these measures to other indicators of achievement. A computer-based test of the revised reasoning items and the generating-explanations task was administered to a sample of examinees who previously had taken a paper-and-pencil version of the GRE General Test. The revised reasoning item types had acceptable psychometric characteristics and were correlated highly with the current logical reasoning items as well as with items on the GRE verbal measure. The generating-explanations task was only marginally related to the GRE measures and to the revised reasoning items and was more strongly related to ideational fluency measures. Factor analytic results suggest that if the revised reasoning items were added to the analytical measure, the correlation between the verbal measure and the analytical measure would increase. A better fitting factor model resulted when some of the current and the revised reasoning items were included in the verbal measure. The relationship of the generating-explanations task and the analytical measure to other indicators of achievement varied for type of achievement and for broad major undergraduate fields of study. The implications of these results for understanding the nature of the skills assessed by different tasks and formats, and for long-term modifications in the GRE General Test, are discussed.
Executive Summary

In this study the potential effects of adding new variations of reasoning items and a generating-explanations task to the GRE General Test were evaluated. The impact of these item types on the internal structure of the test and their relationships to other, concurrent criteria of success and achievement were examined. An experimental, computer-based test with the new variations and the generating-explanations task was administered to a sample of 388 examinees who had previously taken a paper-and-pencil version of the GRE General Test. The two most significant outcomes of this study were that (a) adding the new variations of logical reasoning items to the analytical reasoning measure is likely to decrease its discriminant validity and that (b) the generating-explanations task is as distinct from the new variations of reasoning items as it is from verbal and quantitative reasoning.

The variations of reasoning items that were evaluated in this study included argument evaluation, logical functions, and analysis of explanations. The formats of the logical functions items and analysis of explanations differed from traditional multiple-choice items. For logical function items, examinees had to highlight a sentence in an argument that served a particular function. For analysis of explanations items, the examinees had to make yes/no decisions about a series of statements that might be relevant to some unstated explanation for an unexpected outcome. The generating-explanations task, a constructed-response item type, required the examinees to produce multiple explanations that might possibly account for the outcome of a situation.

Despite examinee unfamiliarity with the new item variations, the generating-explanations task, and the computer interface, the item types had reasonable psychometric characteristics. Scores for the logical functions and argument evaluation items were highly correlated with scores for the logical reasoning items currently on the analytical measure and for the verbal measure, but they correlated only moderately with scores for the analytical reasoning items. The correlation of analysis of explanations items with logical reasoning items and the verbal measure, although high, was lower than those for logical functions and argument evaluation. The correlations of the GRE measures and all the various reasoning item types with the generating-explanations task were consistently lower than the correlations among the GRE measures and the reasoning item types.

A number of confirmatory factor analyses were conducted. Those that compared the fit of different four-factor models are most informative with respect to the current practice of assigning the multiple-choice items in the GRE General Test to three separate measures. The best fitting four-factor model included a generating-explanations factor, a quantitative factor, an analytical factor defined only by the current analytical reasoning items, and a verbal factor defined by the current verbal and logical reasoning items and the new variations of reasoning items. A four-factor model that preserved the current organization of the GRE General Test by assigning analytical reasoning items, logical reasoning items, and the new variations to an analytical factor did not fit the data as well as a three-factor model did. This three-factor model consisted of a quantitative-analytical reasoning factor, a verbal-logical reasoning factor, and a generating-explanations factor.

The relationships of the current and revised versions of the GRE measures to external, concurrent measures of achievement were examined. An alternative model of revised verbal, quantitative, and analytical measures did not have a stronger relationship to self-reports of
undergraduate grade point average (UGPA) or accomplishments than did the current version. For the sample as a whole, neither the generating-explanations task nor the analytical measure added incrementally to the correlation of verbal and quantitative measures with UGPA. However, the generating-explanations task did add to the relationship with UGPA for a subsample of humanities and social science majors. Although both the analytical measure and the generating-explanations task added incrementally to the relationship of a verbal measure, a quantitative measure, and UGPA with accomplishments, the regression weight for the analytical measure was negative, making the nature of its contribution hard to evaluate.

These results have implications for future development of the GRE General Test. First, although increasing the diversity of item types on the GRE test may improve its content representativeness, we also need to carefully assess whether or not these variations have an impact on the constructs being measured. Item format effects, in particular, vary in ways that we do not yet fully understand. The results of this and other studies, however, suggest that format and domain may interact in such a way that constructed-response formats are more likely to assess unique constructs in verbal as opposed to quantitative domains. If increasing the variety of constructs being assessed is an important goal, more research on fluency tasks in the verbal domain would be worthwhile. Secondly, the results of this study with respect to disciplinary differences in the relationship of different test measures to external criteria of achievement suggest that taking advantage of the flexibility of computer-adaptive testing to tailor tests for specific disciplines is a direction worth pursuing.

Finally, the results of this and other studies indicate that item types on the analytical measure assess distinguishable dimensions of reasoning. Although these results add to our understanding of the variety of reasoning skills that can be assessed, the practice of combining tasks that assess diverse aspects of reasoning on the same measure reduces the coherence of the measure and increases its overlap with the verbal and quantitative measures. In addition, assessing these diverse aspects of reasoning does not increase the relationship with external concurrent measures of achievement when the contribution of verbal and quantitative measures is taken into account. One strategy that might lead to a more unified and distinctive measure would be to focus on either explanatory reasoning or formal-deductive reasoning. Furthermore, given our results on disciplinary differences, research is needed on what might be the best combination of measures for different disciplines.
Relationships among Multiple-Choice Reasoning Items and a Constructed-Response Generating-Explanations Task

In an attempt to broaden the range of cognitive skills assessed on the Graduate Record Examinations (GRE) General Test, the GRE Board has supported research on how to assess a wider variety of cognitive skills as well as the impact of different response formats. In both of these areas, a central issue is the degree to which different types of problems or formats assess similar or different cognitive skills. Studies of the GRE General Test, described below, are consistent in finding that a number of separable dimensions of reasoning can be distinguished. However, studies of the impact of response formats on the type of cognitive skills assessed present a mixed picture, as detailed in a subsequent section. In some situations, multiple-choice and constructed-response formats have been found to tap differentiated skills whereas in others no differences have been detected. In the current study, we examined the relationships among different item types, which varied in the aspects of cognitive skill assessed, as well as in format, to determine how test modules assessing different dimensions of cognition might best be organized.

Assessing Reasoning

In 1977 a test of reasoning, the analytical measure, was added to the GRE General Test. This measure was introduced to expand the range of reasoning skills assessed beyond those evaluated by the then-existing verbal and quantitative measures. Early research indicated that the analytical measure could be differentiated from the existing measures (Powers & Swinton, 1981) and that it added to the correlation of the existing measures with undergraduate grade point average (Wild, 1985). However, two of the four item types originally included in the measure were found to be susceptible to short-term practice and coaching (Powers & Swinton, 1984; Swinton & Powers, 1983). Therefore, since 1981 the operational GRE analytical measure has included only two types of items: logical reasoning and analytical reasoning. Logical reasoning items consist of a short verbal argument followed by a single question or a pair of questions assessing any one of a variety of critical reasoning skills, such as recognizing assumptions, analyzing evidence, or drawing conclusions. Analytical reasoning items include a brief scenario and a set of rules about how elements in the scenario can be combined, followed by a set of questions. The analytical reasoning item type emphasizes deductive reasoning skills. Until recently, 70% of the items on the analytical measure were analytical reasoning items and the remainder were logical reasoning items. Currently, the measure is evolving so the balance of analytical and logical reasoning items is more equal.

Concern about the present configuration of the analytical measure has arisen because of three factors. First, studies of whether analytical scores add to the relationship of verbal and quantitative scores with first-year graduate grades have reported mixed results (Kingston, 1985; Wilson, 1982). Second, the logical reasoning items are time-consuming and expensive to develop. This is primarily due to the difficulty of identifying appropriate real-life situations and then adapting them into rigorous logical structures. And third, problems with the convergent and discriminant validity of the measure have been noted. Logical reasoning items correlate more highly with verbal items than with analytical reasoning items, and analytical reasoning items correlate better with quantitative items than with logical reasoning items (Wilson, 1985). Other studies using full-information factor analysis (Schaeffer & Kingston, 1988) and confirmatory multidimensional item response theory (Kingston & McKinley, 1988) indicate a weak analytical factor defined by analytical reasoning items but not logical reasoning items. Rock, Bennett, and Jirele (1988) reported that a four-factor solution with logical reasoning and
analytical reasoning items constrained to load on separate factors fit better than the three-factor model on which the current General Test is based.

More recently, Emmerich, Enright, Rock, and Tucker (1991) evaluated additional item types for possible inclusion on the analytical measure. The item types included (a) pattern identification—a version of a number series item, (b) numerical logical reasoning—items that required the evaluation of the relevance of additional information to explaining or interpreting data presented in a graph or table, (c) analysis of explanations—items that required the evaluation of the relevance of additional information to explaining an unexpected outcome to a situation, and (d) contrasting views—items that required an analysis of alternative interpretations of a concept. In an exploratory factor analysis, Emmerich et al. found that these item types loaded variously on four separable though correlated dimensions of reasoning, which they labeled verbal, informal, formal-deductive, and quantitative reasoning. Items that best defined the informal reasoning factor included pattern identification, numerical logical reasoning, and analysis of explanations. Contrasting views items loaded on the verbal reasoning factor that was primarily defined by the item types on the verbal measure. Logical reasoning items loaded moderately on both the verbal reasoning factor and the informal reasoning factor. The formal-deductive factor was defined by analytical reasoning items and some of the item types on the quantitative measure. Finally, the item types currently on the quantitative measure defined the quantitative factor.

The Effects of Format

Response format variations do not appear to have equivalent effects on the types of reasoning items described above. Bridgeman and Rock (1993) explored the relationship between analytical reasoning items in either a multiple-choice or a constructed-response format. They found that a constructed-response version of the analytical reasoning item type did not measure anything different than did a multiple-choice format, even though some of the constructed-response items required cognitive flexibility in that the examinee was asked to generate multiple solutions. In another study, Bridgeman (1993) reported a high correlation between scores on multiple-choice and constructed-response versions of quantitative reasoning items.

The lack of format effects for formal-deductive and quantitative problems contrasts with the format effects found for a problem type known as formulating hypotheses or generating explanations. The formulating-hypotheses task was created by Frederiksen to measure skills similar to those used by a scholar in interpreting research (Frederiksen & Ward, 1978). The task presents a situation and asks the examinee to generate as many explanations as possible. Situations may be contextualized to require discipline-specific knowledge or written to call upon only more general knowledge about the world. The generating-explanations task is a computer-based version of the formulating-hypotheses task that draws upon general rather than domain-specific knowledge (Bennett & Rock, 1995). GRE Board-sponsored research on formulating hypotheses and, more recently, on generating explanations, has found scores on the task to be highly reliable, to have high correlations with measures of ideational fluency, to correlate only moderately with GRE General Test performance, and to add incrementally over the GRE General Test to the prediction of important criteria (Bennett & Rock, 1993; Frederiksen & Ward, 1978; Ward, Frederiksen, & Carlson, 1980). Presentations of this task in machine-scorable or multiple-choice formats (Ward, Carlson, & Woisetschlaeger, 1983; Ward et al., 1980) have not been successful in preserving its relationship to ideational fluency measures and may result in an increase in the relationship to scores on other multiple-choice reasoning items. Thus, the accumulated evidence suggests that the domain may determine whether format variations actually will result in the assessment of different skills (cf. Traub, 1993).
Description of the Current Study

In the current study we investigated the relationships among generating-explanations items, the item types on the current GRE General Test, and new variations of logical reasoning items and analysis of explanations items.

The new variations of the logical reasoning items—argument evaluation and logical functions—were designed to increase the variety of items available to assess skills in analyzing arguments and to improve efficiency in developing these kinds of items. Argument evaluation and logical functions items are standardized in that they each have one invariant stem (question) and a large set of standardized options from which appropriate ones are chosen for each item. Thus, unique options do not have to be developed for each item, one reason why these item types are more efficient to develop. Both argument evaluation and logical functions items require a critical analysis and evaluation of verbal arguments and have much in common with existing logical reasoning items.

The analysis of explanations item type, based on Peirce’s (1931-1958) ideas about hypothesis formation and evaluation, was developed for the original version on the analytical measure but was dropped because it was found susceptible to coaching (Wild, Swinton, & Wallmark, 1982). Nevertheless, interest in this item type has persisted because graduate faculty believe skills such as “generating valid explanations” (p. 7, Powers, 1987) and “formulating alternative possibilities of conceptualization, classification, and explanation” (p. 11, Tucker, 1985) are important for success in graduate study. In an analysis of explanations problem, a situation is described in a passage and a result is stated that seems paradoxical in terms of the situation. Analysis of explanations problems have been evaluated with a variety of response formats. In its original form, this problem type was presented in a fixed response format with a tree-like decision structure. Examinees had to make a sequence of decisions as to whether a statement (a) was inconsistent with the fact situation, (b) provided an explanation for the result, (c) was necessarily true given the fact situation, (d) weakened or supported a possible (but unstated) explanation for the result, or (e) was irrelevant to an explanation of the result.

Bridgeman and Rock (1993) presented analysis of explanations problems in a computerized version of this format that led the examinee through a series of yes/no decisions, and Emmerich et al. (1991) pilot tested two paper-and-pencil multiple-choice versions of analysis of explanations problems with options that were unique to each item. In exploratory factor analyses in both of these studies, analysis of explanations was found to load on a factor distinguishable from a verbal reasoning factor, thereby supporting its potential for improving the divergent validity of the analytical measure. This is consistent with what Powers and Swinton (1981) found for the original version of the analytical measure, which included analysis of explanations in a fixed response format.

Among the item types considered in this study, generating explanations is clearly the most unique. These items differ from the GRE General Test measures in that they are related to ideational fluency. Of the three measures on the GRE General Test at present, the analytical measure has been found to have the lowest correlation with generating-explanations items as well as with measures of ideational fluency (Bennett & Rock, 1993). However, the GRE analytical measure as it is currently constituted emphasizes formal-deductive reasoning. Attempts to reduce this emphasis on formal-deductive reasoning by introducing new reasoning item types may change the relationship between generating explanations and the analytical measure. For example, analysis of explanations items can be presented in such a way that examinees are required to generate multiple explanations for outcomes.
Thus, this item type may have a higher correlation with generating explanations than do either formal-deductive items or items concerned with the analysis of verbal arguments.

The two major questions to be addressed in this study are (a) what are the relationships among the new variations of reasoning items and generating explanations, and (b) how might these new item types be best combined with current item types into relatively independent test sections. If less formal reasoning item types, such as logical functions, argument evaluation, and analysis of explanations, are added to the analytical measure, the magnitude of its relationship to the verbal measure could increase. Emmerich et al. (1991) reported that the correlations corrected for attenuation between the verbal measure and the current analytical measure increased from .69 to a range of .76 to .82 when simulated analytical measures, composed of various combinations of current and new logical reasoning item types, were evaluated.

Method

Instruments

Computer-Based Test

Three linear computer-based test sections were prepared and assembled into a package. The first and third sections were composed of six-item blocks of argument evaluation, analysis of explanations, and logical functions items (see Table 1). The order of the item blocks was reversed in these two sections, with analysis of explanations always the middle block. The order of the first and third sections was counterbalanced across participants. The four generating-explanations items always were presented in the second section. Forty-five minutes was allocated for each test section. Tutorials on how to use a mouse, how to scroll, and how to use the computer-based testing tools were presented in an untimed section at the beginning of the testing session, and a separate tutorial on how to answer generating-explanations items was presented at the beginning of the second section. For the reasoning items, participants reviewed a copy of the instructions for answering the different reasoning item types and practice items prior to beginning the computer-based test; the instructions were repeated during computer testing prior to each block of items. Copies of the generating-explanations tutorial and the instructions for the reasoning item types are in Appendix A.

Argument Evaluation. The argument evaluation item type (see Figure 1) tests the ability to judge the degree to which particular arguments succeed in drawing their conclusions on the basis of supporting material or to identify the particular reason why certain arguments are flawed or only weakly support their conclusions.

Logical Functions. The logical functions item type (see Figure 2) is intended to test the ability to understand arguments by identifying the role played by a specified part of an argument. Logical functions questions can be administered in two ways:

1. by underlining a sentence or phrase in an argument and asking examinees to select from five possible answer choices the role played in the argument by that underlined portion, or

2. by presenting an argument and asking examinees to identify which sentence or sentence part in the argument performs a specified function.
<table>
<thead>
<tr>
<th>Section</th>
<th>Item Type</th>
<th>Item No.</th>
<th>Item Type</th>
<th>Item No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Logical Functions</td>
<td>1 - 6</td>
<td>Argument Evaluation</td>
<td>19 - 24</td>
</tr>
<tr>
<td></td>
<td>Analysis of Explanations</td>
<td>7 - 12</td>
<td>Analysis of Explanations</td>
<td>25 - 30</td>
</tr>
<tr>
<td></td>
<td>Argument Evaluation</td>
<td>13 - 18</td>
<td>Logical Functions</td>
<td>31 - 36</td>
</tr>
<tr>
<td>2</td>
<td>Generating Explanations</td>
<td>1 - 4</td>
<td>Generating Explanations</td>
<td>1 - 4</td>
</tr>
<tr>
<td>3</td>
<td>Argument Evaluation</td>
<td>19 - 24</td>
<td>Logical Functions</td>
<td>1 - 6</td>
</tr>
<tr>
<td></td>
<td>Analysis of Explanations</td>
<td>25 - 30</td>
<td>Analysis of Explanations</td>
<td>7 - 12</td>
</tr>
<tr>
<td></td>
<td>Logical Functions</td>
<td>31 - 36</td>
<td>Argument Evaluation</td>
<td>13 - 18</td>
</tr>
</tbody>
</table>
Research has proved that eating lots of fish greatly decreases the risk of developing heart disease. The key factor providing protection has been identified as omega-3 fatty acids, a family of fatty acids found in fish oils. Many people think that they can decrease their risk of developing heart disease without eating fish merely by taking dietary supplements of pure omega-3 fatty acids in capsule form. However, as things stand, they must be wrong since the effectiveness of omega-3 fatty acids depends on their being eaten together with other substances found only in fish.

Which of the following most precisely characterizes the reasoning in the argument?

- The reasoning is conclusive: that is, the conclusion cannot be false if the statements offered in its support are true.
- The reasoning is strong but not conclusive: if the statements offered in support of the conclusion are true, they provide adequate grounds for that conclusion, though it is possible that additional information might weaken the argument.
- The reasoning is weak: the statements offered in support of the conclusion, though relevant to it, by themselves provide at best inadequate grounds for the conclusion.
- The reasoning is flawed in that the argument treats one event’s occurring before another event as proof that the earlier event is the cause of the later event.
- The reasoning is flawed in that the argument treats popular opinion regarding the conclusion as evidence in support of the conclusion.

Figure 1. An argument evaluation item.
Click on a word in the sentence or sentence part that plays the specified role.

Role: Stating the position to be refuted by the argument.

It is widely believed that the notion that teachers are generally apathetic about microcomputer technology is at least dated and probably false; a recently published survey indicates that 86 percent of the 5,000 teachers who responded to survey questionnaires expressed a high level of interest in microcomputers.

No sentence or sentence part plays the specified role.

Figure 2. A logical functions item.
The second administration method can be used only in computer administrations by using a mouse to click on the desired sentence or sentence part in the argument. This was the method used in the current study.

**Analysis of Explanations.** Each group of analysis of explanations items (see Figure 3) consists of a description of a situation and a result that seems paradoxical given the situation and so requires explanation. The examinee is then asked to consider a number of independent statements and decide whether each statement is relevant to some possible adequate explanation of the result. For each of the four situations described, nine possibly relevant statements were presented in three groups of three. Note that in this version of analysis of explanations, possible explanations were not explicitly stated but had to be generated by the examinee, who also had to consider multiple explanations for each situation.

**Generating Explanations.** Four generating-explanations items that had been used in previous research were presented (Bennett & Rock, 1995; Bennett & Rock, in press). Although these items required no specific disciplinary knowledge, they varied by general context (humanities, social science, science) and by whether they included a graph.

A sample item is presented in Figure 4. The item and the prompt are presented in the left-hand panel on the computer screen. The boxes in the right-hand panel are used to enter, save, or edit a list of possible explanations. Explanations are typed in the lower box in the right-hand panel, and moved to the list in the upper box when the “save” button is clicked. An entry in the upper box can be modified by highlighting it, clicking on the “edit” button, and moving it back to the typing box. A limit of 15 explanations can be entered per item.

**Paper-and-Pencil Instruments**

**Ideational Fluency Items.** A paper-and-pencil instrument to measure ideational fluency was also administered. This instrument consisted of four items. One, from the Kit of Factor-Referenced Cognitive Tests (Ekstrom, French, & Harman, 1976), required examinees to suggest as many ideas as they could about a topic (a man climbing a ladder). The second item, from the Torrance Tests of Creative Thinking (Torrance, 1974), asked examinees to suggest questions that could be asked about an object, for example, a cardboard box. The third and fourth items were “pattern meaning” tasks and required examinees to generate ideas about what an unfinished drawing could possibly be (Wallach & Kogan, 1965). Each of these tasks was scored in terms of the number of responses generated.

**Activities and Accomplishments Questionnaire.** In addition, the participants were asked to complete an Activities and Accomplishments Questionnaire consisting of 52 items that probed for evidence of achievement in six areas. The areas of achievement covered were academic, leadership, practical language, aesthetic expression, scientific, and mechanical (Stricker & Rock, 1996). A copy of the questionnaire is in Appendix B.

**GRE General Test and Background Questionnaire.** Finally, the participants’ responses to the items on the December 1995 GRE General Test and to the Biographical Information Questionnaire (part of the registration form for the GRE tests) were extracted from GRE files.
Figure 3. An analysis of explanations item.
Figure 4. A generating-explanations item.
Participants and Procedures

Examinees who had taken the December 1995 GRE General Test (paper-and-pencil version) at test centers located near 10 data collection sites throughout the continental United States were sent letters offering them $40 to participate in a study of a computer-based test of some new item types. The letters were sent to 4,330 examinees, and 966 indicated a willingness to participate. Of these respondents, 408 were eventually tested, and usable data were collected from 388 participants. Participants were tested in individual sessions by graduate assistants at nine universities and at ETS. Data from 20 participants were excluded from the analyses because of (a) computer failure ($n = 2$), (b) failure to complete one or more of the ideational fluency or generating-explanations items within the allotted time ($n = 17$), and (c) evidence of only minimal effort on the generating-explanations section (only 1 response per item, $n = 1$).

The background characteristics of the study sample and the GRE population are presented in Table 2. Compared with the GRE population, the study sample had mean GRE analytical scores that were about $1/3$ of a standard deviation higher, although their mean GRE verbal and quantitative scores were only about $1/10$ of a standard deviation higher. In the study sample, there were about 10% more women, 10% more U.S. citizens, and a slightly higher percentage of non-Whites than in the GRE population.

Data Analyses

Basic descriptive statistics for the experimental measures and the reasoning item types on the current GRE were computed and inspected. Correlations among the experimental measures, the current GRE measures, and other concurrent criterion measures were examined. The relationships among the experimental measures and the current measures were explored further through exploratory and confirmatory factor analyses. Finally, the relationships of measures based on multiple-choice items and the generating-explanations tasks to other indicators of achievement were evaluated.

Estimating Missing Data

For 36 examinees, responses for single generating-explanations items were not recorded because of a programming problem. In addition, undergraduate grade point averages (UGPA) were missing for 37 examinees, and 10 examinees omitted one or more items on the accomplishments questionnaire. These missing data were estimated using MISS (RJS Software, 1990). First, maximum likelihood procedures were used to estimate the means and covariance matrix of the complete data set via the EM algorithm (Little & Rubin, 1987). After the covariance matrix and means had been estimated for the complete data set, the missing data points were imputed. The imputations were generated from a set of pseudorandom numbers with mean equal to the expected mean of the missing observations, given the estimated mean and covariance matrix of the complete data set and the non-missing observations, and covariance matrix equal to the expected covariance matrix of the missing observations, given the mean and covariance matrix of the complete data set and the nonmissing observations.

Item Analysis

Item scoring. Because previous research on generating-explanations items has established that the number of responses submitted and those deemed acceptable by human judges are almost perfectly
Table 2
Demographic Data

<table>
<thead>
<tr>
<th>Background Characteristics</th>
<th>Examinee Population</th>
<th>Study Sample (n=388)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>GRE verbal mean (SD)</td>
<td>485 (123)</td>
<td>497 (117)</td>
</tr>
<tr>
<td>GRE quantitative mean (SD)</td>
<td>561 (142)</td>
<td>576 (131)</td>
</tr>
<tr>
<td>GRE analytical mean (SD)</td>
<td>534 (129)</td>
<td>580 (120)</td>
</tr>
<tr>
<td>Percent Women</td>
<td>52%</td>
<td>62%</td>
</tr>
<tr>
<td>Percent Non-White</td>
<td>15%</td>
<td>19%</td>
</tr>
<tr>
<td>Percent U.S. Citizen</td>
<td>74%</td>
<td>84%</td>
</tr>
<tr>
<td>Percent English Best Language</td>
<td></td>
<td>86%</td>
</tr>
<tr>
<td>Undergraduate Major Field</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Life Sciences</td>
<td>15%</td>
<td>22%</td>
</tr>
<tr>
<td>Physical Sciences</td>
<td>9%</td>
<td>8%</td>
</tr>
<tr>
<td>Engineering</td>
<td>11%</td>
<td>9%</td>
</tr>
<tr>
<td>Social Sciences</td>
<td>16%</td>
<td>23%</td>
</tr>
<tr>
<td>Humanities</td>
<td>13%</td>
<td>16%</td>
</tr>
<tr>
<td>Education</td>
<td>8%</td>
<td>5%</td>
</tr>
<tr>
<td>Business</td>
<td>3%</td>
<td>1%</td>
</tr>
<tr>
<td>Other or Missing</td>
<td>25%</td>
<td>17%</td>
</tr>
</tbody>
</table>


*The percentages for non-White and U.S. citizen are based on n's of 334 and 384, respectively.
correlated (Bennett & Rock, 1995; in press), these items were scored in terms of the number of intelligible responses given, as were the ideational fluency items.

Two ways of scoring analysis of explanation items were possible. These “items” consisted of three statements, each of which was evaluated as relevant or not relevant to some possible explanation by the examinee. For these items the response to each statement could be scored separately. A “super item” could also be scored as correct only if all three “yes/no” decisions were correct. The first method was used in this study because it seemed more apt to assess ideational fluency; that is, examinee ability to generate one explanation would be assessed independently of the ability to generate a different explanation.

Factor Analyses

Parcel definition. Multiple-choice items were grouped into item type parcels to better approximate the assumption of multivariate normality underlying linear factor models (Cattell & Burdsal, 1975). Each of the item types included in the December 1995 verbal, quantitative, and analytical measures was divided into two parcels, yielding eight parcels of items from the verbal measure, four from the quantitative measure, and four from the analytical measure. Each of the reasoning item types from the experimental battery was divided into two parcels. Items were assigned to parcels in an alternating, sequential pattern. Parcels within an item type were inspected and adjusted, if necessary, to ensure that the parcels were roughly equivalent in terms of mean item difficulty.

Exploratory factor analysis. A principal components analysis was conducted on the item parcels to determine the number of factors to extract. The factors were then subjected to a Varimax rotation.

Confirmatory factor analyses. Confirmatory factor analyses were used to evaluate how different configurations of reasoning item types affected the convergent and discriminant validity of potential GRE measures (cf. Bridgeman & Rock, 1993; Emmerich et al., 1991). Models were fit to the correlation matrix with the LISREL program, using maximum likelihood factor estimation procedures (Jöreskog & Sörbom, 1985). Initially, an eight-factor model with each of the experimental item types loading on a separate factor was evaluated. Then a number of alternative models that differed in both the number and composition of factors were compared with the eight-factor model in terms of convergent and discriminant validity. Convergent validity of individual factors and their marker variables was assessed by inspecting the size and significance of the indicators (factor loading coefficients). The overall convergent validity of the total factor model was evaluated by examining a number of goodness-of-fit indicators supplied by the LISREL program (Jöreskog & Sörbom, 1985). These indicators include the nonnormed fit or Tucker-Lewis index (Tucker & Lewis, 1973), the root mean square error of approximation, the root mean square standardized residual, and the chi-square/degrees of freedom ratio. Discriminant validity was evaluated by inspecting the correlations among factors in the model. Factor intercorrelations in the range of .85 and up were considered to be redundant from a discriminant validity point of view.

Relationship to Other Criteria

The relationships between test-based measures and other indicators of achievement and ideational fluency were ascertained through factor extension and least-squares linear multiple regression. The other indicators of achievement included self-reports of UGPA and accomplishments.
In one set of analyses, UGPA was regressed on generating explanations and either the current GRE measures or a set of revised GRE measures. These analyses were conducted for the sample as a whole and separately for examinees from two broad undergraduate major fields. Similar analyses were conducted for accomplishment scores. For the analyses in which UGPA was the dependent variable, verbal and quantitative measures were first entered as a set, and the incremental validity of either a reasoning measure or a generating-explanations measure was evaluated in a second step. For the analyses in which accomplishment scores were the dependent variables, UGPA was also entered as an independent variable in the first step.

Results

Summary Statistics

Descriptive statistics for the logical and analytical reasoning items from the GRE General Test, the computer-administered item types, and criterion measures are presented in Table 3. Compared with the current reasoning items presented on the GRE, reasoning items on the experimental test were slightly less reliable. Reliability for the generating-explanations items was very high.

The distributions of responses for the ideational fluency items tended to be bimodal rather than normal. Many examinees entered as many responses as the answer sheet permitted, so the greatest permissible response count had a high frequency of occurrence. Therefore, the response counts for the ideational fluency items were converted to quartile scores to produce a rectilinear distribution.

Practice effects. Because the order of the two experimental sections of reasoning item types was reversed for half the participants, practice effects for these item types could be evaluated. Table 4 lists the mean order effect per item, an effect size measure, and the F value for the order by item set interaction in a 2 x 2 repeated measures analysis of variance. The mean order effect per item is the average difference in proportion correct when an item appeared in the first section from when it appeared in the final section, and the effect size is this difference divided by the standard deviation of the proportion correct when items were presented in the first section. A positive value indicates a practice effect—performance was better for these items when they were presented in the final section rather than in the first section. The order x item set interaction was significant for both analysis of explanations and logical functions but the effect sizes for all three item types were small by conventional standards (Cohen, 1977). Thus, although there were significant practice effects for the new item variations, these effects were not large or of much practical importance.

Correlational Analyses

Correlations among the experimental item types and scores on the GRE General Test measures are presented in Table 5. A number of effects are worth noting in this table. First, all three new reasoning item types correlate more with the current verbal measure than with the analytical reasoning items. Second, the correlations between the generating-explanations task and the GRE measures and the new reasoning item types are smaller than the correlations between the GRE measures and the new reasoning items. This suggests that the generating-explanations task measures skills different from those measured by the current GRE measures and new reasoning item types. Finally, the correlation between analysis of explanations items and the generating-explanations task is of the same magnitude as correlations between generating explanations and other types of reasoning items. Thus, analysis of explanations items do not appear to have more in common with generating explanations than do other reasoning item types.
### Table 3

**Summary Statistics (n = 388)**

<table>
<thead>
<tr>
<th>Item Type</th>
<th>n of items</th>
<th>Range</th>
<th>Mean</th>
<th>SD</th>
<th>Observed Alpha</th>
<th>Alpha for 30 items</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Generating Explanations</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item 1</td>
<td>1-15</td>
<td>6.9</td>
<td>2.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item 2</td>
<td>1-15</td>
<td>7.2</td>
<td>3.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item 3</td>
<td>1-15</td>
<td>7.6</td>
<td>3.2</td>
<td></td>
<td>.92</td>
<td></td>
</tr>
<tr>
<td>Item 4</td>
<td>1-15</td>
<td>6.0</td>
<td>2.9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>4</td>
<td>6-60</td>
<td>27.7</td>
<td>10.8</td>
<td></td>
<td>.92</td>
</tr>
<tr>
<td><strong>Ideational Fluency</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.82</td>
<td></td>
</tr>
<tr>
<td>Item 1</td>
<td>1-4</td>
<td>2.4</td>
<td>1.1</td>
<td>.74</td>
<td>.85</td>
<td></td>
</tr>
<tr>
<td>Item 2</td>
<td>1-4</td>
<td>2.4</td>
<td>1.1</td>
<td>.74</td>
<td>.85</td>
<td></td>
</tr>
<tr>
<td>Item 3</td>
<td>1-4</td>
<td>2.4</td>
<td>1.1</td>
<td>.74</td>
<td>.85</td>
<td></td>
</tr>
<tr>
<td>Item 4</td>
<td>1-4</td>
<td>2.3</td>
<td>1.1</td>
<td>.74</td>
<td>.85</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>4</td>
<td>4-16</td>
<td>9.6</td>
<td>3.6</td>
<td>.82</td>
<td></td>
</tr>
<tr>
<td><strong>Logical Reasoning</strong></td>
<td>15</td>
<td>0-15</td>
<td>8.6</td>
<td>3.3</td>
<td>.74</td>
<td>.62</td>
</tr>
<tr>
<td><strong>Analytical Reasoning</strong></td>
<td>35</td>
<td>7-35</td>
<td>23.2</td>
<td>6.1</td>
<td>.85</td>
<td>.83</td>
</tr>
<tr>
<td><strong>Argument Evaluation</strong></td>
<td>12</td>
<td>0-12</td>
<td>6.3</td>
<td>2.4</td>
<td>.57</td>
<td>.77</td>
</tr>
<tr>
<td><strong>Analysis of Explanations (yes/no)</strong></td>
<td>36</td>
<td>11-36</td>
<td>24.9</td>
<td>5.0</td>
<td>.77</td>
<td>.77</td>
</tr>
<tr>
<td><strong>Logical Functions</strong></td>
<td>12</td>
<td>0-12</td>
<td>7.5</td>
<td>2.2</td>
<td>.57</td>
<td>.70</td>
</tr>
<tr>
<td><strong>Accomplishments</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Academic</td>
<td>0-5</td>
<td>1.4</td>
<td>1.2</td>
<td>.52</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leadership</td>
<td>0-5</td>
<td>0.9</td>
<td>0.1</td>
<td>.52</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linguistic</td>
<td>0-12</td>
<td>0.9</td>
<td>2.0</td>
<td>.52</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aesthetic Expression</td>
<td>0-20</td>
<td>0.8</td>
<td>1.7</td>
<td>.52</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Science</td>
<td>0-5</td>
<td>0.6</td>
<td>0.1</td>
<td>.52</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mechanical</td>
<td>0-5</td>
<td>0.3</td>
<td>0.0</td>
<td>.52</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>0-21</td>
<td>4.9</td>
<td>3.6</td>
<td>.52</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>UGPA</strong></td>
<td>2-7</td>
<td>5.4</td>
<td>1.1</td>
<td>.52</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 4
Order Effects for Types of Reasoning Items from the Experimental Battery

<table>
<thead>
<tr>
<th>Type of Reasoning Item</th>
<th>Order Effect/Item</th>
<th>Effect Size</th>
<th>F(1,386)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Argument Evaluation</td>
<td>.02</td>
<td>.13</td>
<td>1.70</td>
</tr>
<tr>
<td>Analysis of Explanations</td>
<td>.04</td>
<td>.18</td>
<td>30.22**</td>
</tr>
<tr>
<td>Logical Functions</td>
<td>.03</td>
<td>.16</td>
<td>7.25*</td>
</tr>
</tbody>
</table>

*p< .01.  **p< .001
Table 5

Correlations Among GRE Measures and Experimental Item Types

<table>
<thead>
<tr>
<th>V</th>
<th>Q</th>
<th>LR</th>
<th>AR</th>
<th>AE</th>
<th>AX</th>
<th>LF</th>
<th>GE</th>
</tr>
</thead>
<tbody>
<tr>
<td>GRE Verbal</td>
<td>1</td>
<td>.48</td>
<td>.70</td>
<td>.43</td>
<td>.52</td>
<td>.56</td>
<td>.55</td>
</tr>
<tr>
<td>GRE Quantitative</td>
<td>.46</td>
<td>1</td>
<td>.38</td>
<td>.60</td>
<td>.29</td>
<td>.35</td>
<td>.31</td>
</tr>
<tr>
<td>GRE Logical Reasoning Items</td>
<td>.92</td>
<td>.47</td>
<td>1</td>
<td>.47</td>
<td>.53</td>
<td>.50</td>
<td>.52</td>
</tr>
<tr>
<td>GRE Analytical Reasoning Items</td>
<td>.51</td>
<td>.69</td>
<td>.57</td>
<td>1</td>
<td>.38</td>
<td>.31</td>
<td>.34</td>
</tr>
<tr>
<td>Argument Evaluation</td>
<td>.78</td>
<td>.42</td>
<td>.83</td>
<td>.53</td>
<td>1</td>
<td>.46</td>
<td>.48</td>
</tr>
<tr>
<td>Analysis of Explanations</td>
<td>.71</td>
<td>.41</td>
<td>.68</td>
<td>.39</td>
<td>.72</td>
<td>1</td>
<td>.44</td>
</tr>
<tr>
<td>Logical Functions</td>
<td>.85</td>
<td>.44</td>
<td>.84</td>
<td>.50</td>
<td>.88</td>
<td>.74</td>
<td>1</td>
</tr>
<tr>
<td>Generating Explanations</td>
<td>.40</td>
<td>.23</td>
<td>.37</td>
<td>.30</td>
<td>.36</td>
<td>.33</td>
<td>.31</td>
</tr>
</tbody>
</table>

Note. Observed correlations are above the diagonal. Disattenuated correlations are below the diagonal. All p < .01.
Correlations of the GRE measures and experimental item types with a number of criterion measures are presented in Table 6. All the GRE measures and the experimental item types have significant correlations with UGPA, a traditional concurrent measure of achievement. However, the correlational patterns for the test-based measures are differentiated with respect to the subscales of the accomplishments instrument and the ideational fluency measure. The largest relationship of the academic accomplishments scale is with measures and item types that tap verbal reasoning skills (GRE verbal, logical reasoning, argument evaluation). The aesthetic and linguistic accomplishment subscores are most associated with the generating-explanations task. The GRE quantitative measure is most strongly correlated with scientific and mechanical accomplishments. Finally, although all the test measures except quantitative have significant correlations with the ideational fluency task, the magnitude of this relationship was much greater for the generating-explanations task.

Factor Analyses

Exploratory Analysis

In the initial principal components analysis, four factors had eigenvalues greater than 1, but the next two factors were close (.99 and .86). Thus, it was decided to examine a six-factor model. The results of a Varimax rotation are presented in Table 7. Although the factors are readily interpretable, a number of the parcels load on more than one factor.

The first factor, labeled verbal, is defined by all the item types that now comprise the verbal measure. The logical reasoning parcels and one each of the analysis of explanations and the logical functions parcels also loaded on this factor. One notable feature of this factor is that discrete items, which test vocabulary knowledge with minimal context, have the strongest loadings and item types that assess comprehension and inference have smaller loadings, a pattern also reported by Emmerich et al. (1991). The second factor, quantitative, is defined by the item types from the quantitative measure and analytical reasoning. The third factor is defined only by the generating-explanations items.

The last three factors appear to separate out various aspects of reasoning, including analysis of arguments (factor 4), formal-deductive reasoning (factor 5), and analysis of explanations (factor 6). Note that two of the experimental item types, argument evaluation and logical functions, load on the same factor but analysis of explanations loads on a separate factor. Although this exploratory analysis suggests that a number of reasoning dimensions may be observed, it does not clarify the relationships among these dimensions and how they might best be combined into reasonably independent measures. For example, no exploratory rotation, whether orthogonal or oblique, will yield a unique set of intercorrelations among the factors. Only a confirmatory solution will provide a unique answer to this discriminant validity question. Therefore, a series of confirmatory analyses were conducted to evaluate more parsimonious models, with evidence for or against discriminant validity. The criteria for evaluating these models include goodness-of-fit measures and the pattern of correlations among the factors.
Table 6

Correlations of Test Measures with Criterion Measures

<table>
<thead>
<tr>
<th>Measures</th>
<th>Acad</th>
<th>Aest</th>
<th>Lang</th>
<th>Leader</th>
<th>Mech</th>
<th>Scient</th>
<th>Total</th>
<th>UGPA</th>
<th>Fluency</th>
</tr>
</thead>
<tbody>
<tr>
<td>GRE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verbal</td>
<td>.15**</td>
<td>.10*</td>
<td>.12*</td>
<td>.03</td>
<td>.11*</td>
<td>-.01</td>
<td>.16**</td>
<td>.22**</td>
<td>.22**</td>
</tr>
<tr>
<td>Quantitative</td>
<td>.10*</td>
<td>.00</td>
<td>-.00</td>
<td>-.09</td>
<td>.15**</td>
<td>.20**</td>
<td>.09</td>
<td>.29**</td>
<td>.09</td>
</tr>
<tr>
<td>Analytical</td>
<td>.12*</td>
<td>.02</td>
<td>-.02</td>
<td>-.01</td>
<td>.06</td>
<td>-.03</td>
<td>.03</td>
<td>.26**</td>
<td>.21**</td>
</tr>
<tr>
<td>Logical Reasoning</td>
<td>.14**</td>
<td>.03</td>
<td>.08</td>
<td>.04</td>
<td>.11*</td>
<td>-.09</td>
<td>.09</td>
<td>.26**</td>
<td>.20**</td>
</tr>
<tr>
<td>Analytical Reason.</td>
<td>.10</td>
<td>.01</td>
<td>-.02</td>
<td>.00</td>
<td>.01</td>
<td>-.00</td>
<td>.02</td>
<td>.23**</td>
<td>.17**</td>
</tr>
<tr>
<td>Experimental Item Types</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Argument Evaluation</td>
<td>.16**</td>
<td>.03</td>
<td>.07</td>
<td>-.01</td>
<td>.09</td>
<td>-.03</td>
<td>.08</td>
<td>.25**</td>
<td>.10*</td>
</tr>
<tr>
<td>Analysis of Explan.</td>
<td>.08</td>
<td>.05</td>
<td>.13*</td>
<td>.01</td>
<td>.10*</td>
<td>-.02</td>
<td>.10*</td>
<td>.14**</td>
<td>.16**</td>
</tr>
<tr>
<td>Logical Functions</td>
<td>.10*</td>
<td>.02</td>
<td>.06</td>
<td>.01</td>
<td>.09</td>
<td>.07</td>
<td>.10</td>
<td>.26**</td>
<td>.11*</td>
</tr>
<tr>
<td>Generating Explan.</td>
<td>.08</td>
<td>.16**</td>
<td>.19**</td>
<td>.03</td>
<td>.09</td>
<td>.02</td>
<td>.19**</td>
<td>.15**</td>
<td>.55**</td>
</tr>
<tr>
<td>Criterion</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UGPA</td>
<td>.54**</td>
<td>.06</td>
<td>.04</td>
<td>-.06</td>
<td>-.12</td>
<td>.09</td>
<td>.19**</td>
<td>1.00</td>
<td>.17**</td>
</tr>
<tr>
<td>Idea. Fluency</td>
<td>.07</td>
<td>.16**</td>
<td>.05</td>
<td>-.01</td>
<td>-.01</td>
<td>-.00</td>
<td>.09</td>
<td>.17**</td>
<td>1.00</td>
</tr>
</tbody>
</table>

*p<.05. **p<.01.
Table 7

Factor Loadings for the Exploratory Factor Analysis

<table>
<thead>
<tr>
<th>Parcels</th>
<th>Factors</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Analogies</td>
<td>A</td>
<td>.68</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>.77</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Antonyms</td>
<td>A</td>
<td>.81</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>.78</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reading Comprehension</td>
<td>A</td>
<td>.49</td>
<td></td>
<td>.49</td>
<td>.32</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>.53</td>
<td></td>
<td>.47</td>
<td></td>
</tr>
<tr>
<td>Sentence Completion</td>
<td>A</td>
<td>.63</td>
<td></td>
<td>.35</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>.56</td>
<td></td>
<td>.34</td>
<td></td>
</tr>
<tr>
<td>Problem Solving</td>
<td>A</td>
<td>.81</td>
<td>.85</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>.85</td>
<td></td>
<td>.85</td>
<td></td>
</tr>
<tr>
<td>Quantitative Comparisons</td>
<td>A</td>
<td>.83</td>
<td>.88</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>.88</td>
<td></td>
<td>.88</td>
<td></td>
</tr>
<tr>
<td>Analytical Reasoning</td>
<td>A</td>
<td>.50</td>
<td>.69</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>.49</td>
<td></td>
<td>.71</td>
<td></td>
</tr>
<tr>
<td>Logical Reasoning</td>
<td>A</td>
<td>.49</td>
<td>.33</td>
<td>.47</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>.54</td>
<td></td>
<td>.40</td>
<td></td>
</tr>
<tr>
<td>Argument Evaluation</td>
<td>A</td>
<td>.66</td>
<td></td>
<td>.61</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>.66</td>
<td></td>
<td>.61</td>
<td></td>
</tr>
<tr>
<td>Analysis of Explanations</td>
<td>A</td>
<td>.32</td>
<td></td>
<td>.86</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>.33</td>
<td></td>
<td>.67</td>
<td></td>
</tr>
<tr>
<td>Logical Functions</td>
<td>A</td>
<td>.39</td>
<td></td>
<td>.43</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>.75</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Generating Explanations</td>
<td>A</td>
<td>.85</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>.88</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>.89</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>.88</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. Principle Components Analyses with varimax rotation. Loadings less than .30 have been omitted.
Confirmatory Analyses

The eight-factor model. Initially, an eight-factor model was fit to the data as the starting point for developing and evaluating more parsimonious models. In the eight-factor model the item types that presently comprise the verbal and quantitative measures were constrained to load on their respective factors and the two item types from the analytical measure and the four item types from the experimental test were constrained to load on separate factors. The goodness-of-fit indicators for this model are presented in Table 8. The correlations among the factors are the disattenuated correlations presented in Table 5. The generating-explanations factor had the lowest correlations with other factors, and its correlation with analysis of explanations was of the same magnitude as its correlations with other reasoning factors. All of the factors for the three experimental reasoning item types had higher correlations with the verbal factor and the logical reasoning factor than they did with the analytical reasoning factor. However, the correlations between the analysis of explanations factor and the verbal and logical reasoning factors were slightly lower than those for argument evaluation and logical functions.

Six-factor models. Based on the correlations among the factors, two alternative six-factor models were fit to the data. In one, the logical reasoning, argument evaluation, and logical functions parcels were constrained to load on the same factor. In the second, the logical reasoning parcels were added to the verbal factor and the argument evaluation and logical functions parcels remained on a separate factor. As can be seen in the fit measures reported in Table 8, these models fit the data as well as the eight-factor model did. However, as shown in Table 9, the correlations between factors 1 (verbal) and 3 (argument analysis) in both these models were high enough to suggest that further reduction in the number of factors was warranted.

Four-factor models. In the next step three alternative four-factor models were evaluated. These four-factor models preserved the current division of the multiple-choice GRE General Test into three measures and included an additional generating-explanations measure. The alternative models varied in the composition of the verbal and analytical factors. The goodness-of-fit measures for these models are included in Table 8 and the correlations among factors are presented in Table 10.

The organization of the first model, 4.A, is similar to that of the current the GRE General Test in that analytical reasoning and logical reasoning parcels were constrained to load on the same factor and the reasoning item types from the experimental test were also placed on this factor. The goodness-of-fit measures for this model are borderline, and the correlation between the verbal factor and the analytical factor is unacceptably high. This suggests that adding new variations of logical reasoning items to the analytical measure will further reduce the already weak discriminant validity of the measure.

In model 4.B, the reasoning item types-logical reasoning, argument evaluation, and logical functions-that had very high correlations with the verbal factor were removed from the analytical factor and placed on the verbal factor. When compared with model 4.A, three of the four goodness-of-fit measures were improved. The one measure that worsened was the root mean square residual. When logical reasoning, argument evaluation, and logical functions were removed from the analytical factor, the loadings of the analytical reasoning parcels increased from .58 - .59 to .87 and the loadings of the analysis of explanations items decreased from .52 - .66 to .33 - .43. (The magnitude of the loadings of the logical reasoning, argument evaluation, and logical functions parcels, ranging from .45 to .76, did not change when they were moved from the analytical factor to the verbal factor.) The correlations among the factors were also different for models 4.A and 4.B (see Table 10). The
### Table 8

**Goodness-of-fit Measures for Alternative Confirmatory Factor Models**

<table>
<thead>
<tr>
<th>Model</th>
<th>Chi Sq/df</th>
<th>RMSEA</th>
<th>RMR</th>
<th>NNFI</th>
</tr>
</thead>
<tbody>
<tr>
<td>8-Factor</td>
<td>1.89</td>
<td>0.048</td>
<td>0.038</td>
<td>0.95</td>
</tr>
<tr>
<td>V; Q; LR; LF; AE; AX; AR; GE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6-Factor.A</td>
<td>1.89</td>
<td>0.048</td>
<td>0.040</td>
<td>0.95</td>
</tr>
<tr>
<td>V; Q; LR+ LF+ AE; AX; AR; GE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6-Factor.B</td>
<td>1.90</td>
<td>0.048</td>
<td>0.040</td>
<td>0.95</td>
</tr>
<tr>
<td>V+ LR; Q; LF+ AE; AX; AR; GE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4-Factor.A</td>
<td>3.24</td>
<td>0.076</td>
<td>0.063</td>
<td>0.88</td>
</tr>
<tr>
<td>V; Q; LR+ AR+ AE+ AX+ LF; GE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4-Factor.B</td>
<td>2.84</td>
<td>0.059</td>
<td>0.083</td>
<td>0.90</td>
</tr>
<tr>
<td>V+ LR+ AE+ LF; Q; AX+ AR; GE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4-Factor.C</td>
<td>2.17</td>
<td>0.055</td>
<td>0.044</td>
<td>0.94</td>
</tr>
<tr>
<td>V+ LR+ AE+ LF+ AX; Q; AR; GE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3-Factor</td>
<td>2.94</td>
<td>0.071</td>
<td>0.057</td>
<td>0.89</td>
</tr>
<tr>
<td>V+ LR+ AE+ LF+ AX; Q+ AR; GE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note.* Chi sq/df = chi-square/degrees of freedom ratio, RMSEA = root mean square error of approximation, RMR = root mean square standardized residual, NNFI = nonnormed fit index.
<table>
<thead>
<tr>
<th></th>
<th>Model 6.A</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. V</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Q</td>
<td>.46</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. LR, LF, AE</td>
<td>.93 .48</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. AX</td>
<td>.72 .42 .74</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. AR</td>
<td>.51 .69 .58 .39</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. GE</td>
<td>.40 .23 .38 .33 .30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Model 6.B</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. V, LR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Q</td>
<td>.47</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. LF, AE</td>
<td>.87 .45</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. AX</td>
<td>.72 .41 .76</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. AR</td>
<td>.53 .69 .54 .39</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. GE</td>
<td>.40 .23 .35 .33 .30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Table 10

**Correlations Among Factors for the Four-Factor Models**

<table>
<thead>
<tr>
<th></th>
<th>Model 4.A</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. V</td>
<td></td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>2. Q</td>
<td></td>
<td>.46</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. LR, AR, AE, AX, LF</td>
<td>.91</td>
<td>.59</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. GE</td>
<td></td>
<td>.40</td>
<td>.23</td>
<td>.40</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Model 4.B</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. V, LR, AE, LF</td>
<td></td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>2. Q</td>
<td></td>
<td>.47</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. AR, AX</td>
<td></td>
<td>.60</td>
<td>.71</td>
<td></td>
</tr>
<tr>
<td>4. GE</td>
<td></td>
<td>.40</td>
<td>.23</td>
<td>.32</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Model 4.C</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. V, LR, AE, AX, LF</td>
<td></td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>2. Q</td>
<td></td>
<td>.48</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. AR</td>
<td></td>
<td>.54</td>
<td>.69</td>
<td></td>
</tr>
<tr>
<td>4. GE</td>
<td></td>
<td>.40</td>
<td>.23</td>
<td>.30</td>
</tr>
</tbody>
</table>
correlation between the verbal and the analytical factor decreased from .91 to .60 and the correlation between the quantitative and the analytical factor increased from .59 to .71.

Finally, in model 4.C, analysis of explanations, which had moderately high correlations with the verbal factor, was also placed on the verbal factor so that the analytical factor was defined only by analytical reasoning parcels. The goodness-of-fit measures for this model are almost as good as those for the eight- and six-factor models.

The correlational pattern among the factors also is better for this model than for the other four-factor models. Thus, this model appears the best of those considered with respect to parsimony, convergent validity, and discriminant validity. The one drawback of this model is that the analytical factor is very narrow.

**Three-factor model.** A three-factor model consisting of verbal-logical reasoning, quantitative-analytical reasoning, and generating-explanations factors was also fit to the data. The goodness-of-fit measures were marginal, although better than the fit of the four-factor model where the analytical factor is similar to the current measure. The correlations between the generating-explanations factor and the verbal and quantitative factors were .40 and .25 respectively, and the correlations between the verbal and quantitative factors was .52.

**Relationships to Other Criteria**

**Factor extensions.** The relationships of the factors from the best fitting four-factor model to other criteria were investigated by estimating the extension loadings for UGPA, the accomplishment scales, and the ideational fluency measures. These results are presented in Table 11. As expected, the generating-explanations factor differs from the GRE factors in that it has a very strong relationship to the fluency measures. It also relates to achievements in linguistic and aesthetic areas, a result that replicates the findings of Bennett and Rock (in press). The verbal factor is related to overall academic achievement and to linguistic achievement in particular. The quantitative factor also is related to a overall academic achievement but has a negative relationship to the fluency tasks. Finally, the results for the analytical factor, which consisted of analytical reasoning items only, are noteworthy. There was a positive loading on the analytical factor for the fluency tasks that required the examinees to generate ideas about what an unfinished drawing could be. This finding is interesting because there is a spatial or figural component to analytical reasoning items, for which it is often useful to draw diagrams. However, linguistic, scientific, and total accomplishment scores had significant, negative loadings on this factor. This may be the result of one or more of the factors acting as a suppressor.

**Regression analyses.** In these analyses, the criterion measures of self-reported UGPA and accomplishments were regressed on GRE measures and generating explanations. For analyses where UGPA was the criterion, the GRE verbal and quantitative scores were entered as a set and then either an analytical reasoning score or a generating-explanations score was entered on the second step to determine the incremental validity of the additional information. Similar analyses were conducted with the accomplishment scores as the criterion except that UGPA was entered with the verbal and quantitative scores on the first step. The analyses were carried out for two different versions of the GRE General Test. For one set of analyses the participants’ scores on the current GRE measures were used. For a second set of analyses, simulated versions of the GRE measures were created based on the best fitting four-factor model (4C) described above. The revised verbal measure includes the participants’ scores on the four item types currently included in the verbal measure as well as their scores on the logical reasoning, logical functions, argument evaluation, and analysis of explanations.
Table 11

Extensions of Criterion Measures on Factor Model 4 C

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Verbal (V, LR, AE, AX, LF)</th>
<th>Quantitative</th>
<th>Analytical Reasoning (AR only)</th>
<th>Generating Explanations</th>
</tr>
</thead>
<tbody>
<tr>
<td>UGPA</td>
<td>.18**</td>
<td>.15*</td>
<td>.04</td>
<td>.05</td>
</tr>
<tr>
<td>Accomplishments</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Academic</td>
<td>.18**</td>
<td>.06</td>
<td>-.03</td>
<td>.01</td>
</tr>
<tr>
<td>Leadership</td>
<td>.11</td>
<td>-.09</td>
<td>-.04</td>
<td>.02</td>
</tr>
<tr>
<td>Linguistic</td>
<td>.17**</td>
<td>.11</td>
<td>-.27**</td>
<td>.18**</td>
</tr>
<tr>
<td>Aesthetic Expression</td>
<td>.06</td>
<td>.00</td>
<td>-.09</td>
<td>.19**</td>
</tr>
<tr>
<td>Science</td>
<td>-.11</td>
<td>.46</td>
<td>-.28**</td>
<td>.05</td>
</tr>
<tr>
<td>Mechanical</td>
<td>.11</td>
<td>.27</td>
<td>-.26</td>
<td>.07</td>
</tr>
<tr>
<td>Total</td>
<td>.15*</td>
<td>.16*</td>
<td>-.21**</td>
<td>.17**</td>
</tr>
<tr>
<td>Ideational Fluency</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verbal Stimulus</td>
<td>.09</td>
<td>-.16**</td>
<td>.03</td>
<td>.57**</td>
</tr>
<tr>
<td>Figural Stimulus</td>
<td>-.07</td>
<td>-.12*</td>
<td>.18*</td>
<td>.50**</td>
</tr>
</tbody>
</table>

* = p < .05. ** = p < .01.
item types. The revised analytical measure consisted of only analytical reasoning items, and the quantitative measure was unchanged.

Multiple regression models were fit to the sample as a whole and to two subsamples consisting of participants who reported undergraduate majors in the sciences (engineering, physical and life sciences) or in the humanities and social sciences.

The results for the multiple regressions for UGPA and for the total accomplishments score are summarized in Tables 12 and 13, respectively. Overall, there is little difference between the current GRE measures and the simulated revised measures in their relationship to either criterion. For both versions, the verbal and quantitative scores combined accounted for about 10% of the variance in UGPA. Verbal, quantitative, and UGPA combined accounted for about 6% of the variance in the accomplishment scores. The magnitude of the correlations between the GRE General Test measures and UGPA reported in Table 11 is equivalent to those found in many studies of the prediction of first-year graduate grade point average (see summary by Briel, O'Neill, & Scheuneman, 1993).

It is important to note the differences between the broad undergraduate major fields in the size of the regression coefficients that are evident in Tables 12 and 13. The combined set of verbal and quantitative scores has a stronger association with UGPA for humanities and social science majors than for science majors, but the opposite is true for the accomplishment scores.

These broad major field differences are also important in evaluating the degree to which either an analytical measure or generating explanations contributes incrementally to the relationship with the criteria. Neither of these two measures improved the relationship with UGPA for the sample as a whole, but generating explanations added significantly to the correlation with UGPA for humanities and social science majors. The latter results are very similar to those reported by Bennett and Rock (1995) for a sample of graduate students, more than 60% of whom reported majors in the humanities and social sciences. The results for the group as a whole resemble those found for a large sample of GRE examinees of whom fewer than 40% reported majors in these disciplines (Bennett & Rock, in press).

With respect to the accomplishments measure, both the analytical measure and generating explanations added significantly to correlation of verbal, quantitative, and UGPA with the total score. However, the regression weights for the analytical measure were negative, suggesting a suppressor effect. The incremental contribution of generating explanations approached significance for science majors but not for humanities and social science majors; the contribution of the analytical measure was not significant for either of the two broad fields.
Table 12

Multiple Regression of UGPA on Current and Revised General Test Measures and Generating Explanations for All Participants and for Broad Undergraduate Major Fields

<table>
<thead>
<tr>
<th>Independent Variables</th>
<th>R</th>
<th>R^2</th>
<th>Change in R^2</th>
<th>F for Increment</th>
<th>( V )</th>
<th>( Q )</th>
<th>( A )</th>
<th>( GE )</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Current GRE Measures and Generating Explanations</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All Participants (n= 388)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GRE V,Q</td>
<td>.30</td>
<td>.09</td>
<td>18.46**</td>
<td>.14*</td>
<td>.21**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GRE A</td>
<td>.30</td>
<td>.09</td>
<td>.00</td>
<td>1.44</td>
<td>.09</td>
<td>.17*</td>
<td>.09</td>
<td></td>
</tr>
<tr>
<td>GE</td>
<td>.30</td>
<td>.09</td>
<td>.00</td>
<td>1.70</td>
<td>.11*</td>
<td>.20**</td>
<td>.07</td>
<td></td>
</tr>
<tr>
<td>Science (n= 142)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GRE V,Q</td>
<td>.28</td>
<td>.08</td>
<td>5.80**</td>
<td>-.00</td>
<td>.28*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GRE A</td>
<td>.29</td>
<td>.08</td>
<td>.01</td>
<td>1.10</td>
<td>-.06</td>
<td>.23*</td>
<td>.12</td>
<td></td>
</tr>
<tr>
<td>GE</td>
<td>.30</td>
<td>.09</td>
<td>.01</td>
<td>1.90</td>
<td>.03</td>
<td>.29**</td>
<td>-.12</td>
<td></td>
</tr>
<tr>
<td>Human./SocSci (n= 146)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GRE V,Q</td>
<td>.34</td>
<td>.12</td>
<td>9.45**</td>
<td>.26*</td>
<td>.11</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GRE A</td>
<td>.34</td>
<td>.12</td>
<td>.00</td>
<td>.24</td>
<td>.24*</td>
<td>.09</td>
<td>.06</td>
<td></td>
</tr>
<tr>
<td>GE</td>
<td>.40</td>
<td>.16</td>
<td>.04</td>
<td>6.93**</td>
<td>.21*</td>
<td>.08</td>
<td>.22**</td>
<td></td>
</tr>
<tr>
<td><strong>Revised GRE Measures and Generating Explanations</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All Participants (n= 388)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V (revised), Q</td>
<td>.32</td>
<td>.10</td>
<td>21.49**</td>
<td>.18**</td>
<td>.19**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A (revised)</td>
<td>.32</td>
<td>.10</td>
<td>.00</td>
<td>.59</td>
<td>.17**</td>
<td>.17**</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td>GE</td>
<td>.32</td>
<td>.10</td>
<td>.00</td>
<td>1.059</td>
<td>.16*</td>
<td>.19**</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td>Science (n= 142)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V (revised), Q</td>
<td>.26</td>
<td>.07</td>
<td>5.13*</td>
<td>.02</td>
<td>.27**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A (revised)</td>
<td>.26</td>
<td>.07</td>
<td>.00</td>
<td>.55</td>
<td>.00</td>
<td>.22*</td>
<td>.07</td>
<td></td>
</tr>
<tr>
<td>GE</td>
<td>.28</td>
<td>.08</td>
<td>.01</td>
<td>1.88</td>
<td>.05</td>
<td>.26**</td>
<td>-.11</td>
<td></td>
</tr>
<tr>
<td>Human./SocSci (n= 146)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V (revised), Q</td>
<td>.36</td>
<td>.13</td>
<td>10.80**</td>
<td>.30**</td>
<td>.09</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A (revised)</td>
<td>.36</td>
<td>.13</td>
<td>.00</td>
<td>.04</td>
<td>.30**</td>
<td>.10</td>
<td>-.02</td>
<td></td>
</tr>
<tr>
<td>GE</td>
<td>.40</td>
<td>.17</td>
<td>.04</td>
<td>6.01*</td>
<td>.23*</td>
<td>.07</td>
<td>.20*</td>
<td></td>
</tr>
</tbody>
</table>

* = \( p < .05 \), ** = \( p < .01 \)

Note. The verbal and quantitative measures were entered as a set in step one and then either the analytical measure or the generating-explanations task was entered in step 2.
Table 13

Multiple Regression of Accomplishment Scores on UGPA, Current and Revised General Test Measures, and Generating Explanations for All Participants and for Broad Undergraduate Major Fields

<table>
<thead>
<tr>
<th>Independent Variables</th>
<th>R</th>
<th>R²</th>
<th>Change F for in R²</th>
<th>Increment V</th>
<th>Q</th>
<th>UGPA</th>
<th>A</th>
<th>GE</th>
</tr>
</thead>
</table>

### Current GRE Measures and Generating Explanations

<table>
<thead>
<tr>
<th>All participants (n= 388)</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>GRE V, Q, UGPA</td>
<td>.25</td>
<td>.06</td>
<td>8.58**</td>
<td>.13*</td>
<td>-.02</td>
<td>.19**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GRE A</td>
<td>.29</td>
<td>.08</td>
<td>7.76**</td>
<td>.23**</td>
<td>.07</td>
<td>.20**</td>
<td>-.22*</td>
<td></td>
</tr>
<tr>
<td>GE</td>
<td>.28</td>
<td>.08</td>
<td>7.35**</td>
<td>.09</td>
<td>-.34</td>
<td>.18**</td>
<td>.14**</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Science (n= 141)</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>GRE V, Q, UGPA</td>
<td>.35</td>
<td>.12</td>
<td>6.15**</td>
<td>.24**</td>
<td>.18*</td>
<td>.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GRE A</td>
<td>.37</td>
<td>.14</td>
<td>2.36</td>
<td>.33**</td>
<td>.25*</td>
<td>.03</td>
<td>-.17</td>
<td></td>
</tr>
<tr>
<td>GE</td>
<td>.38</td>
<td>.14</td>
<td>3.19+</td>
<td>.20*</td>
<td>.17+</td>
<td>.04</td>
<td>.15+</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Human./SocSci (n= 147)</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>GRE V, Q, UGPA</td>
<td>.30</td>
<td>.09</td>
<td>4.54**</td>
<td>.22*</td>
<td>-.24*</td>
<td>.22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GRE A</td>
<td>.31</td>
<td>.10</td>
<td>.99</td>
<td>.26*</td>
<td>-.18</td>
<td>.22*</td>
<td>-.13</td>
<td></td>
</tr>
<tr>
<td>GE</td>
<td>.31</td>
<td>.10</td>
<td>1.43</td>
<td>.20*</td>
<td>.17+</td>
<td>.04</td>
<td>.11</td>
<td></td>
</tr>
</tbody>
</table>

### Revised GRE Measures and Generating Explanations

<table>
<thead>
<tr>
<th>All Participants (n= 388)</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>V (revised), Q, UGPA</td>
<td>.25</td>
<td>.06</td>
<td>8.69**</td>
<td>.12*</td>
<td>.01</td>
<td>.18**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A (revised)</td>
<td>.27</td>
<td>.08</td>
<td>4.82*</td>
<td>.16**</td>
<td>.08</td>
<td>.19**</td>
<td>-.14**</td>
<td></td>
</tr>
<tr>
<td>GE</td>
<td>.28</td>
<td>.08</td>
<td>6.87**</td>
<td>.08</td>
<td>.01</td>
<td>.18**</td>
<td>.14**</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Science (n= 141)</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>V (revised), Q, UGPA</td>
<td>.35</td>
<td>.12</td>
<td>6.33**</td>
<td>.26**</td>
<td>.17+</td>
<td>.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A (revised)</td>
<td>.36</td>
<td>.13</td>
<td>.76</td>
<td>.28**</td>
<td>.20*</td>
<td>.02</td>
<td>-.09</td>
<td></td>
</tr>
<tr>
<td>GE</td>
<td>.38</td>
<td>.15</td>
<td>3.21+</td>
<td>.21*</td>
<td>.15+</td>
<td>.04</td>
<td>.15+</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Human./SocSci (n= 147)</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>V (revised), Q, UGPA</td>
<td>.26</td>
<td>.07</td>
<td>3.46*</td>
<td>.11</td>
<td>-.18+</td>
<td>.24**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A (revised)</td>
<td>.28</td>
<td>.08</td>
<td>1.11</td>
<td>.14</td>
<td>-.12</td>
<td>.24**</td>
<td>-.12</td>
<td></td>
</tr>
<tr>
<td>GE</td>
<td>.28</td>
<td>.08</td>
<td>1.63</td>
<td>.09</td>
<td>-.19+</td>
<td>.21*</td>
<td>.11</td>
<td></td>
</tr>
</tbody>
</table>

Note. The verbal and quantitative measures and UGPA were entered as a set in step one and then either the analytical measure or the generating-explanations task was entered in step 2.

+ = p < .10. * = p < .05. ** = p < .01
Discussion

Summary

In this study the potential effects of adding new variations of reasoning items and a generating-explanations task to the GRE General Test were evaluated with respect to the convergent and discriminant validity of alternative measures and to their incremental validity. An experimental, computer-based test of the new variations and the generating-explanations task was administered to a sample of 388 examinees who had previously taken a paper-and-pencil version of the GRE General Test. The two most significant outcomes of this study were that (a) adding new variations of logical reasoning to the analytical reasoning measure is likely to decrease its discriminant validity and that (b) generating explanations is as distinct from new variations of reasoning items as it is from verbal and quantitative reasoning.

The variations of reasoning items that were evaluated in this study included argument evaluation, logical functions, and analysis of explanations. The formats of the logical functions items and analysis of explanations differed from those of traditional multiple-choice items. For logical functions items, examinees had to highlight a sentence or sentence part in an argument that served a particular function. For analysis of explanations items, the examinees had to make yes/no decisions about a series of statements that might be relevant to some unstated explanation for an unexpected outcome. Despite examinee unfamiliarity with the new variations, the generating-explanations task, and the computer interface, the item types had reasonable psychometric characteristics. Scores for the logical functions and argument evaluation items were highly correlated with scores for the logical reasoning items currently on the analytical measure and for the verbal measure but only moderately correlated with scores for the analytical reasoning items. The correlation of analysis of explanations items with logical reasoning items and the verbal measure, although high, was smaller than those for logical functions and argument evaluation. The correlations between generating explanations and the GRE measures and all various reasoning item types were consistently lower than the correlations among the GRE measures and the reasoning item types.

A number of confirmatory factor analyses were conducted. Those that compared the fit of different four-factor models are most informative with respect to the current practice of assigning the multiple-choice items in the GRE General Test to three separate measures. The best fitting four-factor model included a generating-explanations factor, a quantitative factor, an analytical factor defined only by the current analytical reasoning items, and a verbal factor defined by the current verbal and logical reasoning items and the new variations of reasoning items. A four-factor model that preserved the current organization of the GRE General Test by assigning analytical reasoning items, logical reasoning items, and the new variations to an analytical factor did not fit the data as well as a three-factor model did. The three-factor model consisted of a quantitative-analytical reasoning factor, a verbal-logical reasoning factor, and a generating-explanations factor.

The relationships of the current GRE measures and revised measures based on the best fitting four-factor model to other criteria were examined. The revised model of verbal, quantitative, and analytical measures did not have a stronger relationship to UGPA or accomplishments than did the current model. For the sample as a whole, neither generating explanations nor the analytical measure added incrementally to the relationship with UGPA. However, generating explanations did increase the regression coefficient for UGPA for a subsample of humanities and social science majors.
Although both the analytical measure and generating explanations added incrementally to the regression for accomplishments, the regression weight for analytical reasoning was negative, making the nature of its contribution hard to interpret. These results have implications for theories of reasoning and intelligence, for understanding when item format will interact with the construct being measured, and for future development of the GRE General Test. The implications for theories of reasoning and intelligence will be discussed first because consideration of these issues provides a basis for understanding why it has been so difficult to unify the analytical measure, and for understanding the effects of item format.

**Theoretical Implications**

Two findings of theoretical importance in this study are that (a) generating explanations defines a dimension of ability distinct from those defined by various reasoning tasks and (b) the various reasoning tasks define ability dimensions that are distinct from each other. These findings replicate the results of a number of previous studies and are consistent with the extensive body of factor-analytic research on cognitive abilities. The results of the current study are informative, however, not just because of their consistency with previous work. These results also have implications for understanding (a) the organization of the domain of abilities covered by the tasks in this study and (b) what cognitive processes distinguish generating explanations from reasoning tasks. Our findings can be interpreted both in relation to long-standing factor-analytic descriptions of cognition and to more recent philosophical discussions of the nature of reasoning.

Two factor-analytic descriptions of the domain of cognitive abilities are relevant to the results of this study. Guilford (1967) in his “Structure of the Intellect” model classified factors according to the dimensions of operations, contents, and products. He identified three cognitive operations that are of possible relevance to the current study: divergent production, convergent production, and evaluation. Divergent production reflects generation of information where the emphasis is on variety and quantity rather than on quality of responses. Convergent production involves drawing out logically necessary conclusions or generating answers that satisfy a set of specifications. Guilford contrasts divergent and convergent in terms of the strictness of the criteria for success and the distinction that divergent thinking deals with possibilities but convergent thinking deals with logical necessities. Evaluation involves comparing information with criteria and deciding if criteria are satisfied. Two of the four content areas that Guilford discussed, symbolic and semantic, are also relevant to the present study.

Alternatively, Carroll (1993, p. 584), in a recent reanalysis and integration of a multitude of factor-analytic studies, identifies four higher-order factors that are relevant. These include

1. fluency, a “broad retrieval factor” that involves the “ready production (retrieval) of a variety of responses from long-term memory storage”

2. fluid intelligence, the “basic intellectual processes of manipulating abstractions, rules, generalizations, and logical relationships”

3. crystallized intelligence, reflecting the role of learning

4. the indeterminate combination of fluid and crystallized intelligence.
The constructs of divergent thinking and fluency described by Guilford (1967) and Carroll (1993) are clearly applicable to the generating-explanations factor that emerged in the present study and in other investigations of this problem type. The relevance of their descriptions of other aspects of cognition to the reasoning factors in the current study is less readily apparent, particularly with respect to verbal reasoning. In Guilford’s system analytical reasoning items and quantitative reasoning items might be characterized as involving convergent production and symbolic content. In Carroll’s analysis they might be said to require fluid intelligence; however, they are also likely to require crystallized intelligence about formal operations and well-established mathematical constraints. It is more difficult to describe the variety of item types that were on the “verbal reasoning” factor in this study using the constructs from factor-analytic systems. It is likely that they would involve a number of Guilford’s operations on semantic content or, in Carroll’s terms, “the indeterminate combination of fluid and crystallized intelligence.”

The results of the factor analysis of the complex tasks used in this study suggest an organization of the reasoning domain that is consistent with philosophical distinctions about the modes of reasoning. Toulmin (1958) distinguished between (a) analytical arguments, which are concerned with determining if conclusions follow from assertions, assuming a particular truth value for assertions, and (b) substantial arguments, where aspects of the argument over and above its logical form that need to be evaluated might include determining what the assertions mean, whether or not assertions are believable, or what assertions are relevant and appropriate in a situation.

Similarly, some psychologists and educators have distinguished between formal and informal reasoning in describing how people respond to ill-structured and to everyday-kinds of problems (Galotti, 1989; Kuhn, 1991; Voss, Perkins, & Segal, 1991). Voss, Blais, Means, Greene, and Ahwesh (1989) note that both formal and informal reasoning center around the evaluation of arguments. In formal reasoning, the procedures typically involved include converting propositions to symbolic form, combining these propositions to deduce new information, and determining whether symbolic relationships are in accord with the rules of the system. This is the kind of reasoning applied in formal deductive systems such as logic and mathematics and is quite similar to the kinds of processes people use to solve analytical reasoning problems (Enright, Tucker, & Katz, 1995). In contrast, informal arguments consist of conclusions or hypotheses supported by reasons and are evaluated in terms of their soundness. Among the procedures involved in assessing the soundness of nondeductive arguments are evaluation of (a) whether information is relevant to conclusions, (b) whether and to what degree information supports a conclusion, and (c) whether all relevant information that could support an alternative conclusion has been taken into account.

Emmerich et al. (1991), in their study of additional item types for the GRE analytical measure, used the constructs of formal-deductive reasoning and informal reasoning to describe two of four factors that emerged in an exploratory analysis. The formal-deductive factor consisted of analytical reasoning items, and the items that loaded on the informal reasoning factor included a version of analysis of explanations, numerical reasoning items (another item type involving explanations), and a number series item. Subsequently, Enright et al. (1995) using protocol analysis, found that there were many commonalities in the way examinees reasoned about some of the item types that loaded on the verbal and informal reasoning factors, and these processes were quite distinct from the way people solved analytical reasoning items. In terms of representational tactics and the ordering of the processes of evaluation and justification, items that loaded on the verbal and informal reasoning factors were very similar. However, they differed with respect to the nature of the justifications offered, focusing respectively on similarities in meaning and on practical implications and suppositions. In contrast, the solution of the analytical reasoning items that loaded on the formal-deductive factor differed from the
solution of the other item types with respect to all the processes examined. These results suggest that two factors labeled verbal reasoning and informal reasoning in the Emmerich et al. study actually draw on similar reasoning processes but may be distinguished in terms of the role of intensional and pragmatic content. Bejar (1991) discussed this distinction in a study of GRE verbal analogies. Intensional relationships are those that are based solely on attributes intrinsic to meaning, while pragmatic relationships invoke knowledge of the world that goes beyond the meaning of words and reflects use and function. Thus, the verbal reasoning and informal reasoning factors might be better identified, respectively, as “informal intensional reasoning” and “informal pragmatic reasoning.” In the current study, the new versions of logical reasoning items appear to have strong intensional characteristics as suggested by their correlations with traditional verbal reasoning items. The slightly smaller correlation between the analysis of explanations factor and the verbal factor might reflect the pragmatic content of these items. Finally, the evidence that performance on analytical reasoning items is substantially correlated with performance on quantitative items suggests that many of these item types involve similar formal-deductive reasoning processes, but differ with respect to the degree that they draw on quantitative knowledge or knowledge of how to manipulate arbitrary relations and rules.

In sum, the distinction between fluency tasks and reasoning tasks is a well-established psychometric fact although the critical features of fluency tasks are not clearly documented. As discussed below, the contrast between the generating-explanations task and the analysis of explanations task provides some insight into this issue. With respect to the relationships among reasoning tasks, factor-analytic descriptions are less informative, perhaps because of the complexity of the tasks in the current study. Nevertheless, factor-analytic studies of reasoning do have important implications for the organization of the GRE measures that are discussed in the final section.

Implications for Format Effects

In a recent discussion of whether the traits assessed by multiple-choice and constructed-response formats are equivalent, Traub (1993) notes that there is currently little sound evidence to answer this question. The existing evidence suggests, however, that the answer will vary by domain. Research on format effects for formulating-hypotheses or generating-explanations tasks is suggestive of what task characteristics are likely to contribute to format effects and how these task characteristics may interact with domain effects. This research suggests that the degree to which a task requires evaluation or the application of strict criteria may be an important characteristic affecting whether format effects will be observed.

Ward, Frederiksen, and Carlson (1980) documented format effects for formulating-hypotheses tasks with behavioral science content. The tasks were cast in either a machine-scorable or a constructed-response format. Both versions of the task included a description of a research investigation and the results. For the machine-scorable version, the participants were provided with a list of nine possible explanations that they had to rank order from best to worst. For the constructed-response version, the participants had to generate possible explanations themselves and indicate which explanation they thought was best. Performance on the different versions of the task was not highly correlated. Scores on the two versions of the task were extended onto cognitive factors based on a battery of other cognitive tasks. Scores for the number and unusualness of hypotheses generated in the constructed-response version of the task loaded on expressional and ideational fluency factors. However, scores for the quality of the hypotheses generated or scores on the machine-scorable version of the task did not load on these factors.
In the Ward et al. (1980) study, the generative aspects of the constructed-response version of the task were clearly absent from the machine-scorable version where possible explanations were provided to the participants. The analysis of explanations item type in the current study was designed to have a generative or productive component in that possible explanations were not listed for the participants. Rather, the participants had to evaluate the relevance of additional information to unstated explanations. This item type was also designed to have fluency and flexibility components in that different pieces of additional information were relevant to different alternative explanations. Nevertheless, the analysis of explanations task in the current study did not have a stronger relationship to fluency measures or to the generating-explanations task than did other types of reasoning items.

The results of the current study and those of Ward et al. suggest that the degree to which evaluative criteria constrain the construction of possible responses contributes to the detection of format differences, as evidenced by correlational data. In the Ward et al. study, when constructed responses were scored in terms of quality rather than quantity, they loaded on the same factors as did the machine-scorable tasks. In the current study, the explanations to be generated for the analysis of explanations task were constrained by being related to additional information. These notions may be related to Guilford’s (1967) distinctions between divergent and convergent thinking as related to possibilities versus logical necessities and the strictness of applicable criteria. Future research on analysis of explanations and generating-explanations tasks could be designed to assess the contribution of generative or evaluative components on these tasks. A recent study by Janssen and De Boeck (1996) is relevant here. Janssen and De Boeck reasoned that a constructed-response synonym task involved a response-generation component as well as an evaluation component, whereas a multiple-choice version of the task involved only the evaluation component. They collected data about performance on a verbal fluency task, a constructed-response synonym task, and a multiple-choice synonym task. Multiple regression with latent variables was used to evaluate their hypothesis. The “explanation” tasks in the current study can be cast in versions that involve generation only, generation and evaluation, or evaluation only. Examining the relationships among such tasks would clarify the cognitive processes involved in different versions of the tasks.

Our analysis of the reasons for the format differences found in the current study and in the Ward et al. (1980) study suggests a basis for the interaction between format and domain differences. It is difficult to imagine formal problems where criteria are relaxed so that a wide range of possible solutions might be generated prior to evaluation. For formal problems, the solutions generated are constrained by strict rules. Thus, it may be much more difficult to design constructed-response problems in, for instance, quantitative domains that tap a different dimension of individual differences than do multiple-choice responses.

Implications for the GRE General Test

Among the potential benefits of computer-adaptive testing are (a) increased opportunity for varying item formats, (b) time savings that will allow more skills to be tested, and (c) the ability to tailor tests for specific purposes. These potential benefits raise many questions about how item variety and format might affect the validity of the GRE General Test. The results of this study are relevant to these issues.

The logical functions and analysis of explanations tasks in the present study are two examples of some format variations that with little difficulty could be incorporated successfully into a computer-based GRE General Test. Increasing the diversity of the tasks on the General Test will improve its content representativeness. For example, Powers and Enright (1986) found five dimensions underlying
faculty ratings of important reasoning skills. These included (a) critically evaluating arguments, (b) generating, supporting, and evaluating conclusions and explanations, (c) analyzing formal problems, (d) inductive reasoning, and (e) generating alternative hypotheses and explanations. The logical reasoning, argument evaluation, and logical functions items are designed to assess argument evaluation or critical thinking skills. The analysis of explanations tasks assess the ability to generate and evaluate explanations, and the generating-explanations tasks assess the ability to generate a wide variety of explanations. Finally, the analytical reasoning tasks assess the ability to analyze formal problems.

Nevertheless, the results of this study as well as of a number of others indicate that including tasks that assess diverse aspects of reasoning on the same measure reduces the coherence of the measure and increases its overlap with either the verbal or quantitative measure but adds little in the way of incremental validity. The results of this study suggest that a strategy that might lead to a more unified and distinctive reasoning measure would be to focus either on explanatory reasoning or on formal-deductive reasoning and to combine argument analysis with other indicators of verbal reasoning.

More research would be needed before generating explanations could be incorporated into an operational test because an automated scoring system for this task has not yet been perfected (Kaplan & Bennett, 1994). Furthermore, there are a number of issues about how test-taking strategies might affect the validity of generating-explanations tasks in high-stakes settings (Bennett & Rock, 1995). Notwithstanding these problems, it would be worthwhile to continue research on generating explanations or on other versions of verbal fluency tasks because they clearly measure aspects of talent that are distinguishable from those assessed by other GRE measures and that, at least in the case of generating explanations, add to the prediction of important criteria.

Finally, one way to capitalize on the flexibility associated with computer-based testing is to increase the diversity of measures available to users so that different configurations of measures can be used for different disciplines. For example, a writing test and a mathematical reasoning test that presumes college-level mathematics will be introduced as part of the GRE General Test in the near future. Our results support this new direction because they indicate that the contribution of different measures to the prediction of success varies with broad major fields and with different criteria. This suggests that evaluating the usefulness of new item types or measures for different fields should be an important direction for future research. For example, our finding that generating explanations added incrementally to the prediction of UGPA in the humanities and social sciences is consistent with Powers and Enright’s (1987) report that faculty members in psychology departments rated skills in generating alternative hypotheses as very important for success in their discipline. Continuing to include item types that do not add incrementally to the prediction of currently available criteria of success in the population as a whole will permit future research to identify which measures are most appropriate for different disciplines.
References


Appendix A

Generating-Explanations Tutorial and

Instructions for Reasoning Items
How to Answer

Click on the icon on the right to continue.
This type of problem will require you to generate a list of explanations to account for a given phenomenon. Look at the sample screen below.

The problem will be presented on the left side of the screen. You will enter each explanation in the box on the lower right. As you save each one, it will be added to the list in the box on the upper right.

Click on one of the icons on the right.

The problem will appear here.

Explaination List — Your saved explanations will appear here.

... because ...

Typing Box — You will type your explanation here.
Look at the typing directions below. These will help you enter and edit answers.

As you type answers, you will need to use the keys listed below and know how to insert text. Be sure to read all of the typing directions thoroughly before continuing. The directions will also be available in Help during the test.

The next screen will let you practice typing. Remember, you can’t start typing until you’ve clicked inside the Typing Box to make a blinking cursor appear.

When finished reading, click on one of the icons on the right.

**Typing Directions**

To begin typing, click inside Typing Box to make cursor appear.

To insert text, position cursor in the appropriate place, click, and type.

**Keys to Use**

- Backspace – removes text to left of cursor
- Delete – removes text to right of cursor
- Home – moves cursor to beginning of line
- End – moves cursor to end of line
- Arrows – moves cursor up, down, left, or right

...because...
The highlighted line shows where your answer will appear in the Explanation List. The next line becomes highlighted each time you save an answer.

To begin typing, click inside the box on the lower right (Typing Box) and a blinking cursor will appear. Practice typing each of these samples and click on the SAVE icon after typing each one:

```
this is a sample explanation
this is another explanation
```

When finished reading, click on one of the icons on the right.

**Typing Directions**

To begin typing, click inside Typing Box to make cursor appear.
To insert text, position cursor in the appropriate place, click, and type.

**Keys to Use**

- Backspace – removes text to left of cursor
- Delete – removes text to right of cursor
- Home – moves cursor to beginning of line
- End – moves cursor to end of line
- Arrows – moves cursor up, down, left, or right

...because...
The EDIT icon will display in the typing box what is highlighted in the Explanation List. To edit an explanation from the list, first highlight it by clicking on it, then click on the EDIT icon. Your explanation will appear in the typing box so you can edit it.

Click on the SAVE icon to replace the old explanation with your newly edited one. Be careful—if you click on EDIT instead of SAVE, the highlighted text from the list will replace what you’ve just typed in the typing box.

Highlight #2 in the list. Click on the EDIT icon and practice editing and saving it.

When finished practicing, click on one of the icons on the right.

**Typing Directions**
To begin typing, click inside Typing Box to make cursor appear.
To insert text, position cursor in the appropriate place, click, and type.

**Keys to Use**
- Backspace — removes text to left of cursor
- Delete — removes text to right of cursor
- Home — moves cursor to beginning of line
- End — moves cursor to end of line
- Arrows — moves cursor up, down, left, or right
The SAVE icon will save what is in the typing box to the line highlighted in the list. You could lose (or intentionally delete) a highlighted explanation by either
- saving an empty typing box to the same highlighted line, OR
- saving new text in the typing box to the highlighted line.
A message will appear warning you about either action.

Try saving an empty typing box. When the message appears, click on the SAVE icon and watch what happens to the highlighted line in the list.

When finished practicing, click on one of the icons on the right.

**Typing Directions**
To begin typing, click inside Typing Box to make cursor appear.
To insert text, position cursor in the appropriate place, click, and type.

**Keys to Use**
- Backspace — removes text to left of cursor
- Delete — removes text to right of cursor
- Home — moves cursor to beginning of line
- End — moves cursor to end of line
- Arrows — moves cursor up, down, left, or right

1. this is a sample explanation
2. this is another explanation
3.
4.
5.
6.
7.

... because...

[Editorial note: The text in the boxes on the right side of the page is not transcribed here as it is not clearly visible in the image. The instructions suggest that users should practice saving and editing text in the typing box.]
Each explanation will be limited to 15 words. If you exceed the word limit, a warning message will appear and the last word will be deleted. Make sure to include a space between each word. To see what happens, enter the following explanation:

this very long sample explanation has more than 15 words, which exceeds the word limit allowed by the system

When the message appears, dismiss it. Shorten your explanation, and then save it.

When finished practicing, click on one of the icons on the right.

**Typing Directions**
To begin typing, click inside Typing Box to make cursor appear.
To insert text, position cursor in the appropriate place, click, and type.

**Keys to Use**
Backspace — removes text to left of cursor
Delete — removes text to right of cursor
Home — moves cursor to beginning of line
End — moves cursor to end of line
Arrows — moves cursor up, down, left, or right

... because ...

![Edit button]
![Save button]
Note: Since screen space is limited, only the first few words of an explanation will appear in the list. For example, look at explanation #1 below. It is too long to fit on the screen. The entire explanation can be viewed by highlighting it and clicking on the EDIT icon.

Here is your chance to practice all the skills you've been taught (typing, saving, editing). Use the directions below to insert text and try using all the keys.

When finished practicing, click on one of the icons on the right.

### Typing Directions
To begin typing, click inside Typing Box to make cursor appear.
To insert text, position cursor in the appropriate place, click, and type.

#### Keys to Use
- **Backspace** - removes text to left of cursor
- **Delete** - removes text to right of cursor
- **Home** - moves cursor to beginning of line
- **End** - moves cursor to end of line
- **Arrows** - moves cursor up, down, left, or right

1. this very long sample explanation

2.
3.
4.
5.
6.
7.

...because...
You have just learned how to enter explanations. Review typing directions below.

Tips to remember (these will be available in Help during the test):

- Highlighted line shows where answer will appear in list
- Clicking inside Typing Box makes cursor appear so you can type
- EDIT icon always displays in typing box what is highlighted in list
- SAVE icon always saves what is in typing box to the line highlighted in list
- Always SAVE after typing

Click on one of the icons on the right.

**Typing Directions**

To begin typing, click inside Typing Box to make cursor appear.

To insert text, position cursor in the appropriate place, click, and type.

**Keys to Use**

Backspace — removes text to left of cursor
Delete — removes text to right of cursor
Home — moves cursor to beginning of line
End — moves cursor to end of line
Arrows — moves cursor up, down, left, or right
You have finished this section.

You may repeat it if you wish by clicking on the icon below.

Note: Once you leave this section, you will not be able to come back to it.

Click on one of the icons on the right.
For each problem, your task is to enter as many **plausible and distinct** explanations as possible (up to 15). Each distinct, plausible explanation will receive 1 point.

As you generate explanations, keep the following points in mind.

A **plausible** answer is one that gives a sound and reasonably likely explanation of the phenomenon described. Credit will not be given for an explanation that:

- fails to explain the phenomenon adequately;
- contradicts any part of the overall situation described in the problem;
- offers a farfetched explanation, such as the intervention of extraterrestrial life; OR
- is unclear.

Moreover, all of your explanations should be **distinct** from each other: you should NOT enter an explanation that merely repeats the essential elements of another explanation in your list.
The following sample problem provides examples of acceptable and unacceptable explanations.

**SAMPLE PROBLEM**

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

As the graph above indicates, from 1983 to 1990 the combined earnings of the three largest automobile manufacturers in Country X never fell below 5 billion dollars a year. In 1991, this trend changed dramatically: the same companies suffered a combined loss of 1.5 billion dollars.
suffered a combined loss of 1.5 billion dollars.

The dramatic change in 1991—i.e., the companies’ sudden combined loss after several years of substantial combined earnings—occurred because...

NOTE: Phrase each explanation so that it completes the bold-faced phrase appearing below the passage. You need not repeat any part of the bold-faced phrase.

Acceptable

The following explanations would be accepted as distinct and plausible.

... because...

1. people stopped buying due to a recession
2. people bought more cars from other countries
3. an earthquake destroyed many of the factories
4. the most popular models produced by all three had safety recalls in 1991
5. another domestic company perfected a cheap imitation of the top-selling car for all three
Each answer above is acceptable because it clearly describes a logical and reasonably plausible explanation for the drop in 1991 and is distinct from the others.

An explanation need only be explicit enough so that its meaning can be readily inferred. However, any essential information that cannot be readily inferred must be included in your response. For example, response 5 would NOT be sufficiently clear if it read “they perfected an imitation of the top-selling car type for all three.” “They” could refer either to the three car manufacturers in question or to other car manufacturers, so it is not clear that the answer explains the drop.

Unacceptable

The following explanations would not receive credit.

Explanations 6, 7, and 8 each duplicate the essential elements of one of the explanations in 1 through 5 above.

6. residents of X bought more imports than before

Explanation 6 is redundant because it is essentially a rewording of the explanatory elements of 2, “people bought more cars from other countries.”
7. economic concerns affected people's buying behavior
Explanation 7 duplicates 1 ("people stopped buying due to a recession") because it is merely a more general expression of the basic explanation offered by 1—that some money-related concern reduced the number of cars being bought—without significant variation.

8. a landslide destroyed many of the factories
Explanation 8 duplicates 3 ("an earthquake destroyed many of the factories"), although in a slightly different way. The main point—that the factories were destroyed by a sudden, catastrophic natural occurrence—should be made only once. No trivial variations will receive credit.

Explanation 9 does not account for the drop:

9. X has one of the worst economies in the world
The relative poorness of X's economy is not relevant to an explanation of the drop in 1991.

Explanations 10 and 11 contradict the information provided in the passage:

10. the three companies each suffered a major drop in earnings in 1987 and never recovered
The passage indicates that there was no drop in 1987; thus 10 cannot explain the
The passage indicates that there was no drop in 1987; thus 10 cannot explain the drop.

Explanation 11 contradicts the passage in a slightly different way:

11. the information collected was inaccurate and then the numbers on the graph were miscalculated.

In the passage, the financial loss is described as a fact; furthermore, the bold-faced phrase requires you to explain the loss itself. If the bold-faced phrase had asked instead for an explanation of “the report that there was a sudden loss,” 11 would have been an acceptable answer.

Explanation 12 is too farfetched:

12. time travelers ruined the companies’ reputations by exposing them as participants in an illegal scheme.

Explanation 12 is farfetched because it assumes the reality of time travel, something that is widely thought of as impossible.
GRE Experimental Test

Supplementary Instructions

This packet contains prints of the direction screens and a few sample items similar to those in sections 1 and 3 of this experimental test. (Directions for section 2 are presented via computer.)

After you have reviewed this packet, tell the test administrator that you are ready to begin.

The test administrator can answer questions about how to take the test but will not provide explanations about correct and incorrect answers.
TIMING AND BREAKS

This GRE Experimental Test has three 45-minute sections. During the time allowed for each section, you will NOT be able to access any other section. Between sections there will be optional 10-minute breaks. At that time, notify the administrator if you wish to leave the room.

If you wish to leave your seat at any other time during the test, please raise your hand—section timing will not stop for this type of break.

ANSWERING INFORMATION

You must answer a question before you move to the next screen even if you have to guess. There is no penalty for guessing wrong. If time expires while you are answering a question, you will be given a chance to record any answer already entered on the screen. You will not be able to change the answer or enter a new answer.
Reasoning Section

This 45-minute section has 3 parts, each containing a different type of reasoning problem. You will be given specific instructions on how to answer each type of question as it is introduced. For each type, be sure that you understand the directions before answering the questions. After you finish each set of directions, six questions will be administered. There is a total of 18 questions in this section.
Analysis of Explanations

You will be presented with a description of a situation and a result and asked to evaluate a number of statements in relation to the situation and result. Use the scroll bar in the center of the screen to read the entire situation and result. Preceding each group of two or three statements is a question to which you are to answer “Yes” or “No” for each statement in the group. Consider each of these statements independently from one another. Do not assume either that there are more “Yes” answers or that there are more “No” answers in this section of the test. Click on either the “Yes” or “No” box following each statement.

To receive credit, you need to give correct answers for all of the statements in a group. When you are satisfied with your answers for all of the statements in the group, click on the NEXT and then the ANSWER CONFIRM testing tool to enter your answers and move on to the next group.
Evaluating Arguments

You will be presented with a number of brief arguments. You will be asked to judge the degree to which an argument succeeds in drawing its conclusion on the basis of supporting material, or to identify the particular reason why a certain argument is flawed or only weakly supports its conclusion.

A question and a number of answer choices follow each argument. To answer the question, use the mouse to click on your choice. You can mark your choice by clicking on either the words or the oval preceding them. When you click, the oval becomes filled.

There are two ways to change an answer. They are
1) click on a different choice, OR
2) click on your selected answer again to cancel it, then click on a different choice.
Logical Functions

You will be presented with a description of a role that a sentence or sentence part can play in an argument. This description will be followed by a brief argument. You should identify the sentence or sentence part, if any, that plays the specified role in the argument given.

In some cases, the specified role will actually be played by a sentence or sentence part in the argument.

If what you wish to identify is stated in a complete sentence, click on any word in the sentence in order to highlight it. If what you wish to identify is only a part of a sentence, click on any word in that part to highlight it.

If clicking on a word does not result in highlighting the entire sentence or sentence part you have identified, then you may assume that the sentence or sentence part you identified is not a correct answer.

In some cases, more than one sentence or sentence part will each play the specified role; you can receive credit for a correct answer by highlighting any one of the possible correct answers.

In other cases, no sentence or sentence part plays the specified role.

If you believe a given argument contains no sentence or sentence part that plays the specified role, you should click on any word in the sentence "No sentence or sentence part plays the specified role." This sentence
“No sentence or sentence part plays the specified role.” This sentence appears near the bottom of your screen.

Before confirming your selection, you may make a different selection at any time by clicking on any word in a different sentence or sentence part.
Appendix B

Activities and 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. Started 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
[ ] [ ] 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
[ ] [ ] 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