Examining the Validity of a Computer-Based Generating-Explanations Test in an Operational Setting

Randy Elliot Bennett and Donald A. Rock

July 1998
GRE Board Professional Report No. 93-01P
ETS Research Report 97-18

Educational Testing Service, Princeton, New Jersey 08541
Examining the Validity of a Computer-Based Generating-Explanations Test in an Operational Setting

Randy Elliot Bennett and Donald A. Rock

GRE Board Report No. 93-01P

July 1998

This report presents the findings of a research project funded by and carried out under the auspices of the Graduate Record Examinations Board.

Educational Testing Service, Princeton, NJ 08541
Researchers are encouraged to express freely their professional judgment. Therefore, points of view or opinions stated in Graduate Record Examinations Board Reports do not necessarily represent official Graduate Record Examinations Board position or policy.

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

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.
Abstract

Generating Explanations (GE) is a computer-delivered item type that presents a situation and asks the examinee to pose as many plausible reasons for it as possible. Previous research suggests that GE measures a divergent thinking ability largely independent of the convergent skills tapped by the GRE General Test. This study was conducted to determine if prior GE validity results generalized to the GRE candidate population, how population groups performed, what effects partial-credit modeling might have for validity, and what problems were associated with operational administration. Validity results showed that earlier findings were generally supported: GE was found to be reliable but only marginally related to the General Test and to make significant (but small) independent contributions to the explanation of relevant criteria. With respect to population groups, GE produced smaller gender and ethnic group differences than did the General Test and showed the same relations to outside criteria across groups, suggesting it was measuring similar skills in each population. Attempts to model GE responses on a partial-credit IRT scale succeeded but produced no improvement in relations with external criteria over those obtained by summing raw item scores. Finally, interviews conducted with examinees to detect potential delivery problems suggested that the directions needed to be shortened.
Acknowledgments

Many individuals contributed to this study. Jan Flaugher managed development and data collection; Kelli Boyles and Lois Frankel wrote items, created rubrics, assembled test forms, and scored a subset of responses; Holly Knott produced the test packages; Peggy Redman conducted examinee interviews; Eleanore DeYoung managed the accomplishments survey; and Phil Leung and Kathy Carberry prepared files for data analysis. Special thanks go to Eiji Muraki and Min hwei Wang for conducting Parscale analyses, and to Larry Stricker for use of his accomplishments questionnaire. Mary Enright, Manfred Steffen, and Bill Ward provided helpful comments on an earlier draft of this report. Finally, we would like to acknowledge the GRE Board for its continued and generous support.
Generating Explanations (GE), or Formulating-Hypotheses as it was originally called, is an item type created by Norman Frederiksen to measure skills similar to those used by a scholar in interpreting research findings (Frederiksen, 1959). The task presents a situation and asks the examinee to pose as many plausible reasons for it as possible. Situations may be contextualized to require discipline-specific knowledge or written to call only upon more general knowledge about the world.

The Graduate Record Examinations (GRE) Board began sponsoring GE research in the early 1970s to broaden the range of abilities measured by the General Test. Studies by Frederiksen and his colleagues found that, although GE was reliable, it correlated minimally with the General Test, differing from it and from an objectively scored version of the item type primarily in relating more strongly to a measure of the ability to generate ideas (Frederiksen & Ward, 1978; Ward, Frederiksen, & Carlson, 1980). In addition, the task predicted certain accomplishments in graduate school more effectively than the General Test. An attempt to make the task more operationally workable by casting it in the standard multiple-choice format failed in that this version was no longer psychometrically distinct, measuring the same ability as other General Test item types (Ward, Carlson, & Woisetschlaeger, 1983). In 1988, Carlson and Ward (1988) suggested a strategy for achieving the efficiencies needed to operationalize the task without sacrificing its uniqueness. Noting the growing availability of personal computers and advances in natural language processing, Carlson and Ward recommended that the open-ended version of the task be computer delivered and machine scored.

Carlson and Ward's (1988) recommendation led to the funding of a project in 1990 to develop a computer-delivered version of GE, examine its validity, and study the accuracy of a method for automatically scoring responses. For this project, a test composed of eight GE questions was administered to 192 graduate students. Questions were written to require only general knowledge about the world. Half of the GE items restricted examinee responses to 7 words per explanation and half allowed up to 15 words. Results confirmed that a computer-based version of GE could broaden the range of abilities measured by the General Test. As in paper-and-pencil studies, GE was highly reliable but only marginally related to the General Test; at the same time, it was strongly correlated with a measure of ideational fluency. Versions of GE based on different response limits tapped somewhat different abilities, with items employing the 15-word constraint appearing the more useful for graduate assessment. These items added to conventional measures like the General Test in explaining school performance and creative expression. Finally, the overwhelming majority of examinees found the GE interface easy to use, although some experienced difficulty, indicating the need to study further the role of computer familiarity in GE performance.

These results, detailed by Bennett & Rock (1995), suggest that GE would broaden the talent pool selected for graduate study if it were used in the admissions process. How this use would occur raises numerous issues, not the least of which involves developing an efficient scoring method. Two studies of the feasibility of automatic scoring have been completed (Burstein & Kaplan, 1995; Kaplan & Bennett, 1994), one using a semantic pattern-matching program developed by Kaplan (1992) and the other a statistically based method developed by Kud, Krupka, and Rau (1994). Results indicated that neither method was accurate enough to score the preponderance of responses automatically, suggesting the need for a semiautomatic method whereby responses that could not be accurately graded by machine would be routed to humans for analysis. Key to the success of such a semiautomatic method would be an automatic means for determining which responses could and could not be accurately machine scored.

While automatic scoring research continues, we can address other questions about the item type's functioning that will be important in evaluating whether and how to integrate GE into the GRE Program. In the current study, we delivered GE as the experimental section of the operational
computer-based General Test. This delivery permitted data to be gathered in a context and from a population as close to operational as possible. Using these data, we addressed the following questions:

1. Do the major results of the Bennett and Rock (1995) validity study generalize? In particular, is GE reliable but only minimally related to the General Test, does it differ from the General Test in its relations to grade and accomplishment criteria, and does it add to the explanation of these criteria beyond that provided by General Test performance?

2. How do population groups perform on GE? Are observed group differences simply a reflection of differences in the developed abilities measured by the General Test? Is the relationship between GE and established markers like the General Test and undergraduate grade-point-average (UGPA) the same for all groups?

3. What are the effects on validity of using more sophisticated methods to aggregate item data, in particular partial-credit item response theory (IRT) models?

4. What problems might be associated with operational administration, especially with presenting GE to the examinee and collecting the responses interactively?

Method

Participants

During summer and fall 1994, Generating Explanations appeared continuously as the experimental section of the computerized adaptive GRE General Test administered in approximately 200 computer-based test centers throughout the United States. This section was identified as experimental and examinees were told that their scores would not be reported (but no incentive was given for participation other than the chance to help improve the examination process). Each GE section contained one of four test forms, randomly assigned to examinees. The current study uses data from two of these forms, which were offered collectively to 9,657 examinees.

As results were returned they were screened to eliminate examinees who had chosen not to respond to the Generating Explanations section. From those who did respond, a subset was selected to receive a questionnaire requesting information on unusual accomplishments. Of the 3,674 questionnaires mailed, 2,167 (59%) were returned in usable form.

At the return cutoff date, all GE data were assembled and a thorough examination conducted. This examination showed that besides those who had not responded to the Generating Explanations test at all, some examinees spent minimal amounts of time on the items, the directions, or the interface tutorial; posed only a single response to each item; or entered nonsense. At the same time, many individuals appeared to take the test seriously, devoting reasonable time and offering multiple responses. As a consequence of these varying motivational levels, the distribution of responses to each item was distinctly bimodal, with many examinees entering only a single response and many others entering the maximum number permitted. This distribution contrasted with Bennett and Rock's (1995) study, which, using paid volunteers, lacked the large number of single-entry responses and produced a more normal distribution (but with a similar ceiling effect to that observed here).

To eliminate cases where motivation was questionable, examinees were removed if their records had any one of the following: (1) no response to one or more items, (2) only one response on all three items (suggesting an attempt to get through with minimal effort), (3) nonsense, (4) less than one minute on
the interface tutorial, (5) less than one minute reading the test directions, or (6) less than three minutes in total on the Generating Explanations questions. The time cutoffs were set low, probably somewhat under the minimums needed to understand and respond to the test effectively. These retention criteria produced a sample of 4,712 examinees with Generating Explanations scores, whose response distribution was similar to that found by Bennett & Rock (1995). The number in this sample with accomplishments data was 1,778.

Table 1 describes the total population offered the Generating Explanations section, those judged to have provided valid data, and the subset of that latter group having usable questionnaire responses. For each variable, the sample results were compared to the population value via a two-tailed z-test. Because of the large sample sizes, this test is very sensitive, so only differences judged to be nontrivial are commented upon. Overall, the two samples differed only marginally from the total group. The main study sample performed slightly better on GRE verbal and analytical by .10 and .12 standard deviations, respectively (p < .05). The survey sample outperformed the total population by .12 and .15 standard deviations on these same sections (p < .05). In addition, the survey sample had a slightly higher proportion of females (62% vs. 56%), a lower proportion of minority group members (12% vs. 16%), and a higher percentage of life sciences majors (30% vs. 26%) (p < .05 in each instance).

Instruments

Generating Explanations. The primary instrument was a computer-delivered Generating Explanations test requiring no specific disciplinary knowledge (see Appendix A for test directions). Items for each of the two GE test forms were selected from those used in a previous study (Bennett & Rock, 1995), and varied in situational context (roughly classified as humanities, science, social science), the presence of graphical information in the stimulus, and the phenomenon being described (gradual change, sudden change, presence or absence of some object or event). These factors were balanced across forms to the degree possible.

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

Each GE item is scored on a 0-15 scale, with one point awarded for each plausible, unduplicated hypothesis. This scheme is based on earlier GE research suggesting that the number of hypotheses made for more meaningful relations with criterion measures than did scoring response quality (Frederiksen & Ward, 1978). Rubrics for each of the six items were adapted from Bennett and Rock (1995); see Appendix B for an example. Each rubric listed many specific categories into which correct responses might fall. In general, a response was considered creditable if it stated or implied a possible explanation that was readily apparent and did not duplicate another hypothesis generated by the student for that problem. Duplication was defined as more than one hypothesis falling into the same specific category. Thus, the rubric attempted to discredit instances in which an examinee generated a series of hypotheses that were conceptually identical. Aside from duplication, a response was not creditable if it directly contradicted the situation, if no plausible explanation was readily apparent, or if the response was based only on science fiction or the supernatural.
Table 1

Demographic Data

<table>
<thead>
<tr>
<th>Background Characteristic</th>
<th>Value</th>
<th>Population (n=9,657)</th>
<th>Offered Participation (n=9,657)</th>
<th>Main Study Survey Results (n=4,712)</th>
<th>Subsample with Survey Results (n=1,778)</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Test Performance</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verbal mean (SD)</td>
<td>494(114)</td>
<td>9,580</td>
<td>505(111)*</td>
<td>4.676</td>
<td>1.767</td>
</tr>
<tr>
<td>Quantitative mean (SD)</td>
<td>555(136)</td>
<td>9,550</td>
<td>561(134)*</td>
<td>4.666</td>
<td>1.757</td>
</tr>
<tr>
<td>Analytical mean (SD)</td>
<td>573(132)</td>
<td>9,213</td>
<td>589(136)*</td>
<td>4.523</td>
<td>1.705</td>
</tr>
<tr>
<td>UGPA</td>
<td>5.2(1.1)</td>
<td>4,569</td>
<td>5.2(1.1)</td>
<td>2.312</td>
<td>1.109</td>
</tr>
<tr>
<td>Major UGPA</td>
<td>5.7(1.0)</td>
<td>4,521</td>
<td>5.7(1.0)</td>
<td>2.288</td>
<td>1.097</td>
</tr>
<tr>
<td>Percentage Female</td>
<td>56%</td>
<td>9,609</td>
<td>58%*</td>
<td>4.694</td>
<td>62%*</td>
</tr>
<tr>
<td>Percentage Minority</td>
<td>16%</td>
<td>8,770</td>
<td>14%*</td>
<td>4.332</td>
<td>12%*</td>
</tr>
<tr>
<td>Percentage U.S. Citizen</td>
<td>92%</td>
<td>9,628</td>
<td>93%*</td>
<td>4.701</td>
<td>94%*</td>
</tr>
<tr>
<td>Percentage with Ph.D. Goal</td>
<td>44%</td>
<td>4,538</td>
<td>45%</td>
<td>2.293</td>
<td>42%</td>
</tr>
<tr>
<td>Undergraduate Major</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Life Sciences</td>
<td>26%</td>
<td>8,357</td>
<td>28%*</td>
<td>3,574</td>
<td>30%*</td>
</tr>
<tr>
<td>Physical Sciences</td>
<td>8%</td>
<td>8%</td>
<td>8%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Engineering</td>
<td>8%</td>
<td>8%</td>
<td>8%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social Sciences</td>
<td>21%</td>
<td>20%</td>
<td>19%*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Humanities/Arts</td>
<td>17%</td>
<td>17%</td>
<td>15%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Education</td>
<td>9%</td>
<td>9%</td>
<td>9%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Business</td>
<td>4%</td>
<td>4%</td>
<td>5%*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>8%</td>
<td>7%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intended Graduate Major</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Life Sciences</td>
<td>32%</td>
<td>4,195</td>
<td>33%</td>
<td>2,117</td>
<td>34%</td>
</tr>
<tr>
<td>Physical Sciences</td>
<td>8%</td>
<td>8%</td>
<td>8%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Engineering</td>
<td>7%</td>
<td>7%</td>
<td>7%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social Sciences</td>
<td>17%</td>
<td>16%</td>
<td>16%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Humanities/Arts</td>
<td>11%</td>
<td>12%</td>
<td>9%*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Education</td>
<td>14%</td>
<td>13%</td>
<td>16%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Business</td>
<td>1%</td>
<td>2%</td>
<td>2%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>9%</td>
<td>9%</td>
<td>9%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: UGPA and Major UGPA are self-reported on a 1-7 scale, with 1 denoting a D and 7 denoting an A. Percentages are based on the total number responding to each question. Percentage minority is for U.S. citizens only.

* p < .05, two-tailed z-test of sample value with population parameter.
According to the graph above, the rate of death (per 100,000 people) from infectious diseases in Alcadia declined steadily from 1900 to 1980.
The GRE CAT is a multiple-choice examination designed to measure broad, developed abilities required for success in graduate work. The test is preceded by a tutorial that teaches the examinee how to use the computer to enter responses and navigate the computer interface. As on the paper-and-pencil version, verbal, quantitative, and analytical sections are presented. Each section is administered adaptively, meaning that the particular test form given to an examinee is dynamically built based on the examinee's performance on the previous items and on various content constraints. This adaptive administration is intended to produce tests that are shorter and more closely targeted to the skill levels of individual examinees.

Because scores from the GRE computer and paper testing modes are thought to be comparable (Schaeffer, Reese, Steffen, McKinley, & Mills, 1993; Schaeffer, Steffen, Golub-Smith, Mills, & Durso, 1995), validity data from the paper test should be applicable to the GRE CAT. The psychometric characteristics of the paper-and-pencil General Test have been extensively studied. For example, factor analytic investigations have repeatedly supported the existence of distinguishable verbal and quantitative dimensions that are stable across population groups and related to demographic variables in predictable ways (Rock, Bennett, & Jirele, 1988; Rock, Werts, & Grandy, 1982; Stricker & Rock, 1987; Swinton & Powers, 1980). (Studies have typically determined the analytical section to be more factorially complex, however.) Predictive validity analyses have found correlations with first-year grades averaged across 1,038 graduate departments to be .30 for verbal, .29 for quantitative, .28 for analytical, and .34 for a weighted composite of the three (Educational Testing Service, 1992). The estimated KR-20 reliabilities for the computerized adaptive test were .90, .93, and .89 for verbal, quantitative, and analytical, respectively (Schaeffer, Steffen, Golub-Smith, Mills, & Durso, 1995).

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

Activities and Accomplishments Questionnaire. This 52-item paper-and-pencil measure was used by Bennett and Rock (1995) and is an adaptation from L. Stricker (personal communication, October 10, 1991). The measure asks the examinee to indicate whether or not a given accomplishment had been achieved and, if it had, to provide documentary information on that achievement. (See Appendix C for the questionnaire.) One point was awarded for each accomplishment. Scores were computed for the total questionnaire and for six subscales: academic achievement (5 items), leadership (5 items), linguistic (composed of 12 public speaking and writing questions), aesthetic expression (composed of 20 creative writing, art, music, and dramatics questions), science (5 items), and mechanical (5 items). Our use of the questionnaire is based on considerable evidence that (a) the best predictors of future, high-level accomplishment in science, writing, music, art, and leadership are similar (usually lower level) achievements in prior years and (b) past accomplishments can be reliably documented through self-reports (Baird, 1976, pp. 35-36).
Procedure

Examinees taking the GRE General Test were presented with a brief explanation of the GE test at the conclusion of their last operational GRE CAT section. If they elected to proceed with the experimental section, they were given an untimed tutorial that instructed them how to use the computer to answer GE items, followed by comprehensive directions (also untimed) explaining the GE task and how it would be scored. Last, one of two GE forms was presented. Each form had a time limit of 30 minutes, with no restriction on the time devoted to any one item. Revisiting a question after moving on to the next item was not allowed.

Data Analysis

In general, analyses were done using a single group collapsed across GE forms. Except for the factor analyses, we created this group by standardizing GE scores within test form and then combining the forms. This procedure assumes that the two forms were taken by random groups from the same population, which is consistent with the method used to assign forms to those offered participation. The two groups composing the main study sample do differ, however, suggesting that the forms may have been differentially attractive. Small, but statistically significant, differences between the two groups were found on educational level, the academic subscore of the Accomplishments measure, and the General Test. The groups diverged by about .10 standard deviations on the academic subscore and by .09 on GRE verbal, .08 on GRE quantitative, and .06 on GRE analytical. To determine the practical impact of these differences for the questions we wished to pursue, we conducted several analyses. First, we examined the simple correlations of the within-form GE Z-score with the General Test, UGPA, and the Accomplishments measure. Next, we looked at the equality of slopes of the within-form GE Z-score with each of the GRE sections. Finally, we ran regressions of the various outcomes on GRE and GE, with form inserted as one of the predictors. Each of these analyses suggested that combining the two groups would not materially change our results and, consequently, we combined the two groups.

Second, rather than manually scoring each of the roughly 70,000 Generating Explanations responses, we took advantage of a finding by Bennett and Rock (1995) showing that the number of responses submitted was almost perfectly correlated with the number credited by human judges. To be sure this result applied to the current sample, two test developers independently scored all responses generated by 30 of the examinees taking each form. The number submitted and the number credited by each judge were then standardized within form and the groups collapsed. For the total scores, results showed the judges to agree almost perfectly with one another (r = .99) and, for research purposes, the simple number submitted to be a reasonable surrogate for a human score (r = .95 with each judge). ¹

Generalizability of previous GE findings. To determine if the major results of the Bennett and Rock (1995) validity study extended to the General Test CAT population, we looked at GE’s discriminant validity with the General Test (including its reliability), and its ability to add to the explanation of important criteria beyond that provided by existing measures. Reliability was computed by standardizing each item score within test forms and computing coefficient alpha for each form separately. GE’s distinctiveness from the General Test was assessed in two ways. First, we contrasted the relations of the two measures with grade and Accomplishments indicators (using pairwise deletion of cases). Simple correlations were compared via a two-tailed t-test

¹ Even when two variables are highly related, their correlations with other variables can be considerably different. In our subsample of 60 examinees, we were able to compare the correlations for the number submitted versus. the number graded correct with (1) General Test scores, (2) UGPA, and (3) Accomplishments. In this subsample, number submitted and number correct had very similar relations to these criteria.
for the difference between correlations derived from the same sample (McNemar, 1962, p. 140). Second, we used confirmatory factor analysis.

The confirmatory factor analyses were conducted to estimate the disattenuated correlation between GE and the GRE General Test and to distinguish between the common factors underlying them through their relations with external criteria. These analyses included only those participants with GRE scores, Accomplishments data, and UGPA (n=1,050). For each analysis, the factor models were fitted across the samples taking the two GE forms under the assumption that the items were interchangeable; that is, that they measured the same factor, though not necessarily with equal difficulty or variance. We tested this assumption by also fitting the models separately in each sample, assuming that only the pattern of loadings was the same (but not the factor intercorrelations or the loadings themselves). Because this latter method produced only minimal improvements in fit over the former procedure, we report the results from the simpler analyses here.2

The factor analyses began by fitting a two-factor model to the sample correlation matrices using the maximum likelihood method as implemented in Lisrel 8 (Joreskog & Sorbom, 1993).3 The two-factor model was comprised of GE and GRE dimensions, in which the two factors were assumed to be correlated. Each factor was marked by three variables constrained to load only on that dimension: the three GE item raw scores for the GE factor and the verbal, quantitative, and analytical section scores for the GRE factor. Finally, the factor pattern, intercorrelation, and loadings were set to be equal across the two samples.

Next, extension loadings were computed by introducing Accomplishments scores and UGPA into the factor model and allowing them to load freely on each factor. For this run, the results of the initial solution were used to fix the loadings for the General Test sections and the GE items on their respective factors, as well as the correlation between the two factors.

Third, a three-factor model was computed to determine the disattenuated relations of GE and GRE with a common factor extracted from the Accomplishments measure. This model was comprised of GE, GRE, and a Linguistic/Aesthetic Expression dimension. Here, too, the factor pattern, intercorrelation, and loadings were set to be equal across the two samples.

The fit of the different factor models was assessed by examining their factor loadings, goodness-of-fit indicators, and factor intercorrelations, and by comparing them with a null model in which each marker was constrained to load only on its own factor. Several fit indices were used, each sensitive to different departures: the chi-square/degrees-of-freedom ratio; the goodness-of-fit index; the nonnormed fit index (Bentler & Bonnett, 1980) (an indicator of the proportion of reliable variance accounted for by the factor model); and the root mean square error of approximation (a replicability measure).

To assess GE’s incremental contribution, we used least-squares hierarchical linear multiple regression computed from the missing-data correlation matrix. For the first regression analysis, self-reported UGPA served as an indicator of current school achievement and was regressed on the three General Test scores (verbal, quantitative, and analytical entered as a set) and GE. For the second analysis, the Accomplishments scores were the outcome criteria and UGPA was returned to its traditional role as a predictor. Here, the Accomplishments total score and each of the subscale scores were

---

2 In one instance, the improvements were as follows: the chi-square/degrees of freedom ratio decreased from 2.8 to 2.4, and both the goodness-of-fit index and nonnormed fit index increased from .98 to .99. In a second instance, the chi-square/degrees of freedom ratio decreased from 2.4 to 2.2, the goodness-of-fit index remained unchanged at .98, and the nonnormed fit index increased from .97 to .98.

3 As a check on robustness, parallel analyses were run using the sample covariance matrices with substantively similar results.
regressed, in turn, on General Test scores, UGPA, and GE. For all analyses, GE was entered into the model last on the premise that it must demonstrate value over established measures to justify the added costs that would be associated with its use in admissions.

Functioning with population groups. Analysis of covariance was employed to assess GE’s functioning in population groups. Two sets of analyses were run, one for those ethnic groups with sample sizes larger than 100 and one for gender groups. Within each analysis, the first part indicated if the GE group means differed from one another after controlling for various measures of developed scholastic ability, with a separate analysis run for each of the following independent variables: GRE verbal; GRE quantitative; GRE analytical; the three GRE section scores entered together; and UGPA and the three GRE section scores together. The second part looked at whether GE was related in all groups to the individual General Test sections and to UGPA, and whether those relations were the same in all groups. These relationships were tested by examining the slope of GE on each independent variable across groups. For all analyses, we looked for differences between the reference group (Whites or males) and the focal groups.

Response modeling. Parscale (Muraki & Bock, 1991) was used to model GE item responses using the Generalized Partial-Credit Model (Muraki, 1992). The Generalized Partial-Credit Model (GPCM) essentially weights the contribution of items to an examinee’s total score differentially by parameters that characterize that item’s functioning (e.g., difficulty). GPCM models were fitted within each test form and then the scores collapsed across groups. Three models were fitted differing in the number of score categories used per item: (1) a 12-category model with examinees giving up to 11 responses placed in different categories but those giving 12-15 responses treated as one, (2) a 13-category model that collapsed those giving 11 and 12 responses and those giving 13 and 14 responses, and (3) a 15-category model. Estimated item parameters included the slope (discrimination), difficulty, and category (the location of the score category on the ability scale). (The guessing parameter was fixed at zero for all runs.) Model fit was evaluated by examining plots of standing on the underlying latent ability with probability of getting the item correct for each category on each item. For the best-fitting models, the correlations between theta estimates and external criteria were compared with the correlations of GE z-scores and those criteria to see if the model-based scores provided any improvement in validity.

Operational problems. To understand what problems might be associated with the test-center delivery of GE, we interviewed by telephone a random sample of 30 examinees who had taken the experimental section. In addition, 10 examinees at one test center were interviewed upon completing their examinations. Finally, we checked periodically with the ETS Computer-Based Test Center Network Group to see if test center administrators reported any problems associated with GE. Examinees were asked about the extent to which they had taken the experimental section seriously and about the test directions and the computer interface. Results are given only for those 22 examinees who reported taking the GE test seriously.

Results

Generalizability of Previous Findings

These analyses focused on (1) GE’s differences with the General Test and (2) whether GE adds to the explanation of important criteria beyond that provided by General Test scores.

Table 2 presents summary statistics. Of note is that scores for several of the Accomplishments subscales—which are intended to measure unusual achievements—were substantially skewed. For all subsequent analyses, we
Table 2

Summary Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Scale</th>
<th>n</th>
<th>Mean</th>
<th>SD</th>
<th>Skewness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generating Explanations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Form 1 raw score</td>
<td>1-45</td>
<td>2,345</td>
<td>15.9</td>
<td>7.4</td>
<td>1.1</td>
</tr>
<tr>
<td>Form 2 raw score</td>
<td>1-45</td>
<td>2,367</td>
<td>17.8</td>
<td>8.3</td>
<td>.9</td>
</tr>
<tr>
<td>Z-score</td>
<td>--4.0--4.0</td>
<td>4,712</td>
<td>0.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Time on test (sec)</td>
<td>180-1799</td>
<td>4,712</td>
<td>1060</td>
<td>451</td>
<td>.1</td>
</tr>
<tr>
<td>Time on tutorial (sec)</td>
<td>60-</td>
<td>4,712</td>
<td>276</td>
<td>161</td>
<td>2.0</td>
</tr>
<tr>
<td>Time on directions (sec)</td>
<td>60-</td>
<td>4,712</td>
<td>171</td>
<td>80</td>
<td>1.6</td>
</tr>
<tr>
<td>GRE General Test</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verbal</td>
<td>200-800</td>
<td>4,676</td>
<td>505</td>
<td>111</td>
<td>.3</td>
</tr>
<tr>
<td>Quantitative</td>
<td>200-800</td>
<td>4,666</td>
<td>561</td>
<td>134</td>
<td>-.2</td>
</tr>
<tr>
<td>Analytical</td>
<td>200-800</td>
<td>4,523</td>
<td>589</td>
<td>136</td>
<td>-.4</td>
</tr>
<tr>
<td>Accomplishments</td>
<td>0-52</td>
<td>1,778</td>
<td>4.4</td>
<td>3.5</td>
<td>1.4</td>
</tr>
<tr>
<td>Academic</td>
<td>0-5</td>
<td>1,778</td>
<td>1.3</td>
<td>1.1</td>
<td>.6</td>
</tr>
<tr>
<td>Aesthetic Expression</td>
<td>0-5</td>
<td>1,778</td>
<td>.7</td>
<td>1.4</td>
<td>3.1</td>
</tr>
<tr>
<td>Linguistic</td>
<td>0-12</td>
<td>1,778</td>
<td>.9</td>
<td>1.4</td>
<td>2.0</td>
</tr>
<tr>
<td>Leadership</td>
<td>0-20</td>
<td>1,777</td>
<td>.8</td>
<td>1.0</td>
<td>1.4</td>
</tr>
<tr>
<td>Mechanical</td>
<td>0-5</td>
<td>1,778</td>
<td>.3</td>
<td>.8</td>
<td>2.9</td>
</tr>
<tr>
<td>Science</td>
<td>0-5</td>
<td>1,776</td>
<td>.4</td>
<td>.8</td>
<td>2.4</td>
</tr>
<tr>
<td>UGPA</td>
<td>1-7 (D-A)</td>
<td>2,312</td>
<td>5.2</td>
<td>1.1</td>
<td>-.3</td>
</tr>
<tr>
<td>Major UGPA</td>
<td>1-7 (D-A)</td>
<td>2,288</td>
<td>5.7</td>
<td>1.0</td>
<td>-.5</td>
</tr>
</tbody>
</table>

Note: The minimum possible GE times for the sample are greater than zero because examinees with times below these values were excluded from the study. Accomplishments are in untransformed raw-score units.
transformed each Accomplishments total and subscale score by adding a constant and taking the log of the result.

**Discriminant validity.** Observed correlations of GE with the criterion variables are in Table 3. Internal consistency reliabilities appear on the diagonal. Several results deserve comment. First, whereas both the GE scores and the General Test scores are reasonably reliable, the two measures were only marginally correlated (GE \( r = .22 \) with GRE verbal, .13 with GRE quantitative, and .17 with GRE analytical). Second, GE's relations with the criterion variables were uniformly small but statistically significant for UGPA, major UGPA, Accomplishments total score, and the Aesthetic Expression and Linguistic subscores. Because of the large sample sizes, GE differed from the individual GRE sections in its relations to most of the criterion variables. In comparison to each GRE section, GE was less related to UGPA, major UGPA, Accomplishments total score, the Academic subscore, and the Science subscore (\( t \)-range = -2.21 to -8.50, \( p < .05 \)). In addition, it was more related to the Linguistic subscore than either GRE quantitative (\( t = 3.59, p < .01 \)) or analytical (\( t = 2.42, p < .05 \)), but slightly less related to this subscore than the verbal section (\( t = -2.46, p < .05 \)). Compared to GRE verbal, GE was less associated with Aesthetic Expression (\( t = -2.05, p < .05 \)) but more tied to Leadership (\( t = 2.23, p < .05 \)). Finally, it was more affiliated with Aesthetic Expression than was GRE quantitative (\( t = 3.19, p < .01 \)) but less linked to Mechanical accomplishments (\( t = -3.78, p < .01 \)).

A second approach to estimating the relationship of GE to external criteria is through confirmatory factor analysis. Table 4 gives the loadings for a model composed of GRE General Test and GE factors. Although the uniqueness for GRE verbal indicated that considerable variance in this marker was not accounted for by the model, the overall fit was acceptable for most indices: a goodness-of-fit index of .98, nonnormed fit index of .98, a root mean square error of approximation of .04, and model chi-square/degrees-of-freedom ratio of 2.8 (versus 92.8 for the null model). The common factors underlying GE and the General Test were correlated at .22 (\( p < .01 \)).

Table 5 gives extension loadings for UGPA and Accomplishments scores on the two factors. As can be seen, the GRE factor is clearly more scholastic in nature, being related to UGPA and to the Academic and Science subscores of the Accomplishments scale. There is also a suggestion that the GE factor might be marginally oriented toward Linguistic and Aesthetic Expression.

To test this suggestion more rigorously, we computed a factor model with a dimension defined by these two Accomplishments subscales. Table 6 gives the loadings and the uniquenesses. Fit indices for this model were .98 for the goodness-of-fit index, .97 for the nonnormed fit index, .04 for the root mean square error of approximation, and 2.4 for the model chi-square/degrees-of-freedom ratio (versus 53.2 for the null model). The GE and GRE factors were related at .22 (\( p < .01 \)). The Linguistic/Aesthetic factor was related to GE at .19 (\( p < .01 \)) and to GRE at .13 (\( p < .01 \)). To test the difference between these two latter correlations, we recomputed the model fixing the correlations to be equal. A hierarchical test of the constrained and unconstrained models produced a nonsignificant result (\( \chi^2 = 1.29, df = 1, p > .05 \)).

**Incremental validity.** Table 7 shows the effect of regressing UGPA on the General Test and GE in turn. GE did not add to explanation beyond the GRE, which accounted for 11% of the variance in grade-point average.

Table 8 depicts the regression of the Accomplishments scores on the General Test, UGPA, and GE. In all but one instance, the General Test contributed significantly to explaining the dependent variable and, in three of seven, so did UGPA. GE also added significantly for three of the seven criteria: Accomplishments total score, Aesthetic Expression subscore, and Linguistic subscore. As a percentage of the total variance accounted for,
Table 3

Observed Correlations of GE and General Test Scores with Criterion Variables (n range = 1,096-4,676)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>GE-Z</td>
<td>.83, .86</td>
<td>.22*</td>
<td>.13*</td>
<td>.17*</td>
<td>.09*</td>
<td>.09*</td>
<td>.12*</td>
<td>.03</td>
<td>.09*</td>
<td>.06</td>
<td>.13*</td>
<td>.04</td>
<td>.02</td>
</tr>
<tr>
<td>GRE-V</td>
<td>.90</td>
<td>.49*</td>
<td>.58*</td>
<td>.28*</td>
<td>.25*</td>
<td>.23*</td>
<td>.17*</td>
<td>.14*</td>
<td>.00</td>
<td>.20*</td>
<td>.04</td>
<td>.12*</td>
<td></td>
</tr>
<tr>
<td>GRE-Q</td>
<td>.93</td>
<td>.69*</td>
<td>.29*</td>
<td>.20*</td>
<td>.18*</td>
<td>.15*</td>
<td>.10*</td>
<td>.05</td>
<td>.14*</td>
<td>.24*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GRE-A</td>
<td>.89</td>
<td>.29*</td>
<td>.25*</td>
<td>.19*</td>
<td>.19*</td>
<td>.05*</td>
<td>.07*</td>
<td>.07*</td>
<td>.05</td>
<td>.16*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UGPA</td>
<td>---</td>
<td>.71*</td>
<td>.27*</td>
<td>.50*</td>
<td>.02</td>
<td>.07*</td>
<td>.04</td>
<td>-.05</td>
<td>.11*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MajUGPA</td>
<td>---</td>
<td>.25*</td>
<td>.49*</td>
<td>.02</td>
<td>.05</td>
<td>.05</td>
<td>-.08*</td>
<td>.09*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accomp</td>
<td>.73</td>
<td>.50</td>
<td>.54</td>
<td>.53</td>
<td>.63</td>
<td>.32</td>
<td>.35</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acad</td>
<td>.54</td>
<td>.05*</td>
<td>.11*</td>
<td>.08*</td>
<td>-.03</td>
<td>.09*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AE</td>
<td>.69</td>
<td>.21*</td>
<td>.34*</td>
<td>.06*</td>
<td>-.01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lead</td>
<td>.50</td>
<td>.30*</td>
<td>.01</td>
<td>.09*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ling</td>
<td>.63</td>
<td>.16*</td>
<td>.06*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mech</td>
<td>.66</td>
<td>.14*</td>
<td>.63</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sci</td>
<td>.63</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Coefficient alpha internal consistency reliability estimates are on the main diagonal. General Test estimates are taken from Schaeffer, Steffen, Golub-Smith, Mills, and Durso, (1995). Correlations between Accomplishments and its subscales are part-whole correlations.

* = \( p < .05 \)
Table 4
Loadings and Uniquenesses for the Two-Factor Model (n=1,050)

<table>
<thead>
<tr>
<th>Variable</th>
<th>GRE Factor</th>
<th>GE Factor</th>
<th>Uniqueness</th>
</tr>
</thead>
<tbody>
<tr>
<td>GRE Verbal</td>
<td>.65</td>
<td>--</td>
<td>.55</td>
</tr>
<tr>
<td>GRE Quantitative</td>
<td>.78</td>
<td>--</td>
<td>.39</td>
</tr>
<tr>
<td>GRE Analytical</td>
<td>.90</td>
<td>--</td>
<td>.19</td>
</tr>
<tr>
<td>GE item #1</td>
<td>--</td>
<td>.77</td>
<td>.32</td>
</tr>
<tr>
<td>GE item #2</td>
<td>--</td>
<td>.87</td>
<td>.22</td>
</tr>
<tr>
<td>GE item #3</td>
<td>--</td>
<td>.84</td>
<td>.33</td>
</tr>
</tbody>
</table>

Note: All loadings were significant at $p < .01$.

Table 5
Extensions of Outside Variables on the Two-Factor Solution (n = 1,050)

<table>
<thead>
<tr>
<th>Outside Variable</th>
<th>GRE General Test</th>
<th>GE</th>
</tr>
</thead>
<tbody>
<tr>
<td>UGPA</td>
<td>.37**</td>
<td>.00</td>
</tr>
<tr>
<td>Accomplishments Total</td>
<td>.22**</td>
<td>.07*</td>
</tr>
<tr>
<td>Academic</td>
<td>.28**</td>
<td>-.07*</td>
</tr>
<tr>
<td>Aesthetic Expression</td>
<td>.03</td>
<td>.11**</td>
</tr>
<tr>
<td>Leadership</td>
<td>.06</td>
<td>.07*</td>
</tr>
<tr>
<td>Linguistic</td>
<td>.08*</td>
<td>.12**</td>
</tr>
<tr>
<td>Mechanical</td>
<td>.06</td>
<td>.02</td>
</tr>
<tr>
<td>Science</td>
<td>.24**</td>
<td>-.04</td>
</tr>
</tbody>
</table>

Note: Extension loadings are in the correlational metric. The standard error for each extension loading is .03.
* $p < .05$
** $p < .01$
Table 6
Loadings and Uniquenesses for a Three-Factor Model (n=1,050)

<table>
<thead>
<tr>
<th>Factor</th>
<th>Linguistic/</th>
<th>Variable</th>
<th>GRE</th>
<th>GE</th>
<th>Aesthetic</th>
<th>Uniqueness</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Aesthetic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Uniqueness</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GRE Verbal</td>
<td>.65</td>
<td>--</td>
<td>--</td>
<td>.55</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GRE Quantitative</td>
<td>.78</td>
<td>--</td>
<td>--</td>
<td>.39</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GRE Analytical</td>
<td>.90</td>
<td>--</td>
<td>--</td>
<td>.19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GE item #1</td>
<td>--</td>
<td>.77</td>
<td>--</td>
<td>.32</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GE item #2</td>
<td>--</td>
<td>.87</td>
<td>--</td>
<td>.23</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GE item #3</td>
<td>--</td>
<td>.84</td>
<td>--</td>
<td>.34</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aesthetic Expression</td>
<td>--</td>
<td>--</td>
<td>.50</td>
<td>.73</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linguistic</td>
<td>--</td>
<td>--</td>
<td>.68</td>
<td>.51</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: All loadings were significant at $p < .01$.

Table 7
Multiple Regression of UGPA on General Test and GE Z-Scores (n=2,218)

<table>
<thead>
<tr>
<th>Independent Variable</th>
<th>R</th>
<th>R^2</th>
<th>Increment in R^2</th>
<th>Incremental F</th>
<th>p</th>
<th>Regression Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. GRE-V</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.14*</td>
</tr>
<tr>
<td>GRE-Q</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.16*</td>
</tr>
<tr>
<td>GRE-A</td>
<td>.34</td>
<td>.11</td>
<td>.11</td>
<td>94.13</td>
<td>.00</td>
<td>.10*</td>
</tr>
<tr>
<td>2. GE z-Score</td>
<td>.34</td>
<td>.11</td>
<td>.00</td>
<td>1.21</td>
<td>.27</td>
<td>.02</td>
</tr>
</tbody>
</table>

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

Table 8

Multiple Regression of Accomplishments on General Test, UGPA, and GE Z-Scores (n=1,109)

<table>
<thead>
<tr>
<th>Independent Variable</th>
<th>R</th>
<th>R^2</th>
<th>Increment in R^2</th>
<th>Incremental R^2</th>
<th>F</th>
<th>p</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accomplishments</td>
<td>.13**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Academic Subscore</td>
<td>.18**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aesthetic Expression Subscore</td>
<td>.23**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linguistic Subscore</td>
<td>.23**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leadership Subscore</td>
<td>.07</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mechanical Subscore</td>
<td>.01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Science Subscore</td>
<td>.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: General Test scores were entered as a set. Regression weights are for the full model.
* = p < .05.
** = p < .01.
the increments in each case were trivial. However, with the exception of the Academic subscore, the amount of the total variance accounted for by the other independent variables was also relatively minimal (10% or less), so that the increment explained by GE was proportionally more consequential. Regarding the Linguistic subscore, for example, the General Test accounted for about 4% of the variance. Adding in GE brought just under another point, or 21% more.

In terms of independent contributions, GE had just under half the weight of GRE verbal and under three-tenths the weight of UGPA for explaining Accomplishments. For Aesthetic Expression, GE had about a third the power of GRE verbal and over 70% of the power of GRE quantitative (which had a negative weight). Finally, GE had slightly more than 40% the weight of GRE verbal for explaining the Linguistic subscore.

To see if GE’s contribution to the explanation of the Linguistic subscore was affected by including students with limited English skill, we regressed Linguistic subscore on General Test scores, UGPA, a dummy variable identifying whether or not English was reported as the best language, and GE (entered in that order). The dummy variable added nothing to the explanation of the Linguistic subscore and GE continued to contribute as before.

Functioning in Population Groups

These analyses were intended to determine how population groups perform on GE, whether any group differences were simply a reflection of differences in the developed abilities measured by the General Test and UGPA, and if the relationship between GE and established markers was the same for all groups. Because our study sample contained a smaller proportion of minority group members and of males than the CAT population and because our sample had slightly higher General Test scores, we first compared the General Test scores of the accepted and rejected samples within population group. As expected, those who gave acceptable data generally had significantly higher General Test scores than those who did not (see Table 9). However, this tendency differentially affected population groups: on all three sections, less able Asian examinees were more likely than their White counterparts to drop out, as were less able males versus females; the same pattern characterized Black versus White examinees on the Verbal section. Thus, in some instances our analyses may underestimate population group score differences.

Table 10 shows the observed means and standard deviations by ethnic group and separately by gender for GE, the General Test, and UGPA. On GE, the largest majority-minority difference was for Black examinees, who scored just over .4 standard deviations below the White mean (in the units of the latter group) and the smallest was for “Other” examinees at under .1 standard deviations below the mean. Group differences on the General Test were considerably larger, running up to 1.2 standard deviations for Black examinees on GRE analytical, .6 standard deviations above the White mean for Asians on GRE quantitative, and .5 standard deviations below the mean for Other examinees on GRE analytical. Group differences on UGPA were larger than those on GE for the Black-White comparison (.7 below the White mean) and for the Other-White comparison (.2 below the White mean).

For gender groups, the difference on GE was trivial, less than one tenth of a standard deviation. Differences on the General Test were larger, running from .2 to .7 standard deviations. The mean UGPAs were essentially the same across groups.

4 The group designated “Other” includes those choosing the “Mexican American or Chicano,” “Puerto Rican,” “other Hispanic or Latin American,” or “Other” classifications.
Table 9

<table>
<thead>
<tr>
<th>Group</th>
<th>GRE Verbal</th>
<th>GRE Quantitative</th>
<th>GRE Analytical</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accepted</td>
<td>Rejected</td>
<td>Stand Diff.</td>
</tr>
<tr>
<td>Ethnic</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>515(106)</td>
<td>499(108)</td>
<td>.14*</td>
</tr>
<tr>
<td>Asian</td>
<td>485(133)</td>
<td>433(126)</td>
<td>.40*</td>
</tr>
<tr>
<td>Black</td>
<td>427(103)</td>
<td>398(97)</td>
<td>.29*</td>
</tr>
<tr>
<td>Other</td>
<td>472(116)</td>
<td>458(117)</td>
<td>.12</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>520(113)</td>
<td>493(119)</td>
<td>.23*</td>
</tr>
<tr>
<td>Female</td>
<td>495(109)</td>
<td>478(110)</td>
<td>.15*</td>
</tr>
</tbody>
</table>

Note: Standardized differences were calculated by subtracting the Rejected from the Accepted means using the pooled within-group standard deviations.

* Difference between Accepted and Rejected means significant at p < .05.
Table 10

<table>
<thead>
<tr>
<th>Group</th>
<th>n</th>
<th>GE z-score</th>
<th>Observed</th>
<th>GRE</th>
<th>GRE</th>
<th>GRE</th>
<th>UGPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ethnic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>3,750</td>
<td>.04 (.100)</td>
<td>515 (106)</td>
<td>563 (128)</td>
<td>601 (127)</td>
<td>5.2 (1.1)</td>
<td></td>
</tr>
<tr>
<td>Asian</td>
<td>174</td>
<td>-.22 (.91)</td>
<td>485 (133)</td>
<td>635 (122)</td>
<td>591 (151)</td>
<td>5.3 (.9)</td>
<td></td>
</tr>
<tr>
<td>Black</td>
<td>239</td>
<td>-.38 (.100)</td>
<td>427 (103)</td>
<td>432 (140)</td>
<td>454 (140)</td>
<td>4.5 (1.1)</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>265</td>
<td>-.08 (.99)</td>
<td>472 (116)</td>
<td>533 (139)</td>
<td>537 (151)</td>
<td>5.0 (1.2)</td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>1,965</td>
<td>.04 (.100)</td>
<td>520 (113)</td>
<td>611 (129)</td>
<td>611 (134)</td>
<td>5.2 (1.1)</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>2,729</td>
<td>-.03 (.99)</td>
<td>495 (109)</td>
<td>526 (126)</td>
<td>574 (135)</td>
<td>5.2 (1.1)</td>
<td></td>
</tr>
</tbody>
</table>

Note: Listed n's are for GE z-score. GRE n's are slightly smaller. The n's for UGPA were 1,761 Whites, 87 Asians, 96 Blacks, 110 Others, 901 males, and 1,287 females.
Table 11 shows the extent of GE score differences after controlling for performance on each of the different General Test sections and the test as a whole. Included here are only those examinees with valid scores on all three GRE sections, a subset of those presented in Table 10. Differences between the adjusted means were significant for the Black-White and Asian-White comparisons on all sections \((p < .01)\), although the adjusted differences between Blacks and Whites moderated substantially from the observed ones. The Other-White and the male-female differences were nonsignificant.

In Table 12 are GE scores after controlling for UGPA, and for UGPA and GRE combined, for examinees with complete data on all variables. As the table shows, controlling for UGPA alone had little effect. With this adjustment, mean differences were significant for the Black-White and Asian-White comparisons \((p < .05)\), but not for the Other-White or male-female contrasts. When UGPA and the General Test were combined, the adjustments were more substantial, although the pattern of significant differences remained unchanged.

Table 13 shows the correlations between GE and the external markers for the population groups. Overall significance tests were performed on the slopes, which are less sensitive than the correlation coefficients to differences in variance across groups. The slopes associated with all correlations were significantly different from zero, indicating that GE is related to each of the General Test sections and to UGPA. Further, for the overall statistical test, the slopes were not significantly different among the ethnic groups or between the gender groups. Thus, GE appears to have a similar relationship to the General Test and to UGPA from one to another population group.

Response Modeling

Response modeling was conducted to determine the effect of using more sophisticated methods to aggregate information across GE items than simply summing the raw item scores. Plots for the 13-category model are given in Appendix D. Plots for the 12-category model were similar. For each item and score category within each item, the plots show two things: (1) the predicted relationship between the underlying ability (theta) and the probability of getting a particular score category on the given item and (2) the actual response data—that is, how examinees at each ability level did on average. Judging from the plots, both models fit the data reasonably well, with the observed data points generally tracking the predicted response functions. (There were, of course, divergences but with small samples in some response categories, "wobble" in the observed data is to be expected.) The 15-category model did not converge because of the small numbers of students in the higher score categories.

Item parameters for the 13-category model are given in Table 14; those for the 12-category model were not very different. The item slope parameter represents how sensitive an item is to ability differences, with higher slopes indicating better discriminating power. Thus, on both forms, items 2 and 3 appear to be better discriminators than item 1. The difficulty parameter is roughly similar across the three items composing form 2. On form 1, item 1 (concerning a drop in the number of deer killed by cars) appears considerably "easier" than item 2 (about the disproportionate number of Dutch landscape paintings attributed to major artists).

Both models produced similar correlations with the external criteria. Table 15 shows the correlations for the 13-category model. As can be seen, the validity benefit derived from modeling was minimal, largely because the modeled scores were themselves highly correlated with GE \(z\)-score \((r = .97\) for each model).
Table 11

GE Performance After Controlling for GRE General Test Scores

<table>
<thead>
<tr>
<th>Group</th>
<th>n</th>
<th>Observed GE z-score</th>
<th>GRE Verbal</th>
<th>GRE Quant.</th>
<th>GRE Anal.</th>
<th>GRE-V+Q+A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ethnic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>3,568</td>
<td>-.04</td>
<td>.03</td>
<td>.04</td>
<td>.03</td>
<td>.02</td>
</tr>
<tr>
<td>Asian</td>
<td>162</td>
<td>-.22</td>
<td>-.19</td>
<td>-.30</td>
<td>-.22</td>
<td>-.20</td>
</tr>
<tr>
<td>Black</td>
<td>221</td>
<td>-.41</td>
<td>-.26</td>
<td>-.27</td>
<td>-.25</td>
<td>-.22</td>
</tr>
<tr>
<td>Other</td>
<td>246</td>
<td>-.08</td>
<td>-.02</td>
<td>-.06</td>
<td>-.02</td>
<td>-.01</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>1,860</td>
<td>.03</td>
<td>.00</td>
<td>-.02</td>
<td>.01</td>
<td>.00</td>
</tr>
<tr>
<td>Female</td>
<td>2,598</td>
<td>-.03</td>
<td>-.01</td>
<td>.01</td>
<td>-.01</td>
<td>-.01</td>
</tr>
</tbody>
</table>

Table 12

GE Performance After Controlling for UGPA and General Test Scores

<table>
<thead>
<tr>
<th>Group</th>
<th>n</th>
<th>Observed GE z-score</th>
<th>UGPA</th>
<th>GRE-V+Q+A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ethnic</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>1,761</td>
<td>.01</td>
<td>.01</td>
<td>-.01</td>
</tr>
<tr>
<td>Asian</td>
<td>87</td>
<td>-.23</td>
<td>-.24</td>
<td>-.25</td>
</tr>
<tr>
<td>Black</td>
<td>96</td>
<td>-.48</td>
<td>-.42</td>
<td>-.25</td>
</tr>
<tr>
<td>Other</td>
<td>110</td>
<td>-.17</td>
<td>-.16</td>
<td>-.09</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>901</td>
<td>-.02</td>
<td>-.02</td>
<td>-.07</td>
</tr>
<tr>
<td>Female</td>
<td>1,287</td>
<td>-.05</td>
<td>-.05</td>
<td>-.01</td>
</tr>
</tbody>
</table>

Table 13

Correlations of GE Z-Score with External Markers for Population Groups

<table>
<thead>
<tr>
<th>Group</th>
<th>n</th>
<th>GRE Verbal</th>
<th>GRE Quant.</th>
<th>GRE Anal.</th>
<th>UGPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ethnic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>3,568</td>
<td>.18</td>
<td>.13</td>
<td>.13</td>
<td>.09</td>
</tr>
<tr>
<td>Asian</td>
<td>162</td>
<td>.22</td>
<td>.12</td>
<td>.23</td>
<td>-.12</td>
</tr>
<tr>
<td>Black</td>
<td>221</td>
<td>.11</td>
<td>.04</td>
<td>.17</td>
<td>-.07</td>
</tr>
<tr>
<td>Other</td>
<td>246</td>
<td>.32</td>
<td>.24</td>
<td>.33</td>
<td>.20</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>1,860</td>
<td>.24</td>
<td>.11</td>
<td>.20</td>
<td>.09</td>
</tr>
<tr>
<td>Female</td>
<td>2,598</td>
<td>.19</td>
<td>.13</td>
<td>.15</td>
<td>.09</td>
</tr>
</tbody>
</table>

Note: The slopes associated with all correlations are significantly different from zero at p < .01. Listed n's are for GE z-score and GRE. The n's for UGPA were 1,761 Whites, 87 Asians, 96 Blacks, 110 Others, 901 males, and 1,287 females.
Table 14
Parscale Item Parameters for the 13-Category Model

<table>
<thead>
<tr>
<th>Item</th>
<th>Form 1</th>
<th></th>
<th>Form 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Slope</td>
<td>Difficulty</td>
<td>Slope</td>
<td>Difficulty</td>
</tr>
<tr>
<td>1</td>
<td>.29</td>
<td>.11</td>
<td>.41</td>
<td>.21</td>
</tr>
<tr>
<td>2</td>
<td>.59</td>
<td>.98</td>
<td>.71</td>
<td>.44</td>
</tr>
<tr>
<td>3</td>
<td>.62</td>
<td>.54</td>
<td>.54</td>
<td>.28</td>
</tr>
<tr>
<td>Mean</td>
<td>.50</td>
<td>.54</td>
<td>.56</td>
<td>.31</td>
</tr>
</tbody>
</table>

Table 15
Correlations of Model-Based and Model-Free GE Scores with Criterion Measures (n range = 1,775 to 4,676)

<table>
<thead>
<tr>
<th>GE z-score</th>
<th>GE Theta (13-Category Model)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GRE verbal</td>
<td>.21*</td>
</tr>
<tr>
<td>GRE quantitative</td>
<td>.13*</td>
</tr>
<tr>
<td>GRE analytical</td>
<td>.17*</td>
</tr>
<tr>
<td>UGPA</td>
<td>.09*</td>
</tr>
<tr>
<td>Major UGPA</td>
<td>.09*</td>
</tr>
<tr>
<td>Accomplishments</td>
<td>.12*</td>
</tr>
<tr>
<td>Academic</td>
<td>.03</td>
</tr>
<tr>
<td>Aesthetic Expression</td>
<td>.09*</td>
</tr>
<tr>
<td>Leadership</td>
<td>.06*</td>
</tr>
<tr>
<td>Linguistic</td>
<td>.13*</td>
</tr>
<tr>
<td>Mechanical</td>
<td>.04</td>
</tr>
<tr>
<td>Science</td>
<td>.02</td>
</tr>
</tbody>
</table>

* p < .05.
Operational Problems

Examinees were interviewed in person or by telephone to get a sense of the potential problems associated with administering GE operationally. Also, the ETS Computer-Based Test Center Network Group was periodically contacted to see if test center administrators reported problems with GE. No such problems were reported.

The interviews concentrated on problems in understanding test directions and using the computer interface to respond. Whereas 13 of 22 examinees thought the length of the directions was appropriate, 9 found them problematic (long, tedious, redundant, too many examples). The large majority (20) found the directions clear and helpful, and the description of what was meant by "plausible" and "distinct" understandable. The "because" leader was helpful to 15 examinees, had no effect for 3, and was not remembered by the remainder. Most (19) thought it was clear that they were not to restate the problem when typing their response, but several felt that the 15-word limitation was too constraining. Finally, none indicated losing a response by accidentally saving another response or a blank on top of it.

Discussion

This study was conducted to determine if prior GE validity results generalized to the GRE candidate population, how population groups performed, what effects partial-credit modeling might have for validity, and what problems were associated with operational administration.

Table 16 compares the validity results of the current study with those of Bennett and Rock (1995). The samples differed considerably in size, presumed motivation, and similarity to the GRE CAT population, so some differences are inevitable. Even so, Bennett and Rock’s general findings were supported. In both studies, GE was found to be reliable but only marginally related to the General Test, confirming that the two tests measure distinct attributes. In addition, GE made significant (but small) independent contributions to the explanation of relevant criteria, particularly overall accomplishments and artistic achievement, after controlling for General Test scores. At the same time, several important differences were apparent. One was in Bennett and Rock’s finding a strong relation between GE and ideational fluency, an association with a type of creative thinking that was not tested here because of the time constraints imposed by the GRE experimental section. A second difference was in the considerably closer kinship in the current study of the General Test with academic criteria.

With respect to population group functioning, observed differences on GE were considerably smaller than those for the General Test. GE group differences were reduced further by controlling for the developed abilities measured by the GRE and by UGPA. After these adjustments, the only significant GE differences were for Black and Asian examinees: each group scored about .20 standard deviations below their White counterparts. Additionally, GE was related to the General Test sections and to UGPA in the same way across groups, suggesting that GE was measuring similar skills in each population. However, because of the differential loss of examinees from the Asian group in particular, these results should be considered preliminary.

As to IRT scaling, GE responses were modeled successfully with Parscale (Muraki & Bock, 1991). Such modeling provides a more sophisticated means than simple summation for aggregating item scores by weighting each item’s contribution to total score according to its functioning in the test-taking population. Even so, in this instance, the modeling had little discernible effect on GE’s relations with external criteria.

Finally, our interviews with examinees suggested few potential operational problems in the test-center delivery of GE. The problem of most
Table 16
A Comparison of Two GE Validity Studies

<table>
<thead>
<tr>
<th>Study Aspect</th>
<th>Bennett &amp; Rock (1995)</th>
<th>Current Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample</td>
<td>192 paid volunteers enrolled in first year of graduate school, disproportionately drawn from social sciences and humanities/arts majors relative to the General Test population.</td>
<td>4,712 examinees taking operational General Test CAT, constituting about half of those offered participation but who still matched the characteristics of the CAT population fairly well.</td>
</tr>
<tr>
<td>Discriminant validity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GE Reliability</td>
<td>.90 for a four-item test</td>
<td>.83 for one three-item test form; .86 for the other three-item form</td>
</tr>
<tr>
<td>Relations with GRE</td>
<td>Minimal ($r = .27, .31, .19$ for GRE verb., GRE quant., and GRE anal.)</td>
<td>Minimal ($r = .22, .13, .17$ for GRE verb., GRE quant., and GRE anal.)</td>
</tr>
<tr>
<td>Differences from GRE</td>
<td>1. GE much more strongly related to ideational fluency than the General Test.</td>
<td>1. No ideational fluency measure included due to GRE time limitations.</td>
</tr>
<tr>
<td></td>
<td>2. GE more highly related than individual General Test sections to some Accomplishments measures (e.g., to Aesthetic Expression than GRE quant. or to Linguistic than GRE quant. or GRE anal.).</td>
<td>2. All General Test sections more highly related than GE to UGPA, Major UGPA, Accomplishments total score, and Academic and Science subscores. GE more highly related to Aesthetic Expression than GRE quant., and to Linguistic than GRE quant. or GRE anal.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3. General Test common factor more highly related to academic measures like UGPA, academic accomplishments, and scientific accomplishment.</td>
</tr>
<tr>
<td>Incremental validity</td>
<td>After controlling for General Test scores, GE made small but significant independent contributions to the explanation of UGPA and aesthetic expression. GE's contribution to explaining overall accomplishments just missed significance.</td>
<td>After controlling for General Test scores, GE made small but significant independent contributions to the explanation of overall accomplishments, Aesthetic Expression, and linguistic accomplishments.</td>
</tr>
</tbody>
</table>

Note. Similarities across studies are denoted in italics.
concern related to the directions, which some of those interviewed felt were too long.

What are the implications of these findings for the GRE Program? First, the best argument for using GE would lie in demonstrating its ability to tap skills that (1) are not adequately captured by conventional measures, (2) are required for success in graduate education, and (3) by virtue of incorporation in the admissions process would diversify the composition of the admitted applicant pool. To be sure, the evidence to support this argument is incomplete. That GE measures something different from the General Test and UGPA is a persistent finding. However, it may be that other, more structured logical reasoning item types (e.g., Analysis of Explanations) can add similar breadth with less logistical difficulty. Work to investigate this possibility is underway (Enright & Bennett, in progress). Also in need of strengthening is GE's link with the skills required for graduate education. Powers and Enright (1987) provided a high-level connection with their study of graduate faculty, finding that professors in psychology and education rated the ability to "generate alternatives" as among the reasoning dimensions most differentiating marginal from successful students. How well this ability is captured in our version of GE—and whether it might be better captured in more subject-specific implementations—is an open question.

In the current study we did not detect any strong indication that GE tapped aspects of undergraduate attainment not already captured by conventional measures. It is possible that the ability to generate alternatives plays little role in undergraduate success and does not come prominently into play until the graduate level. Indeed, the low relation of GE to UGPA and General Test score leaves room for a more substantial increment over these measures in predicting graduate outcomes.

To maximize the chances for incremental validity (and domain relevance) at the graduate level, we might try creating implementations of GE that were more subject specific and locating graduate criteria that were also discipline based. Many cognitive psychologists argue that the kind of reasoning used in a domain is strongly influenced by the domain's structure (Brown, Collins, & Duguid, 1989). Thus, GE tests tailored to different broad subject matters might be more successful in tapping reasoning that is reflected in graduate GPA (or other field-based indicators) but not captured by the General Test. So, for example, one might conceive of separate tests for those applying for admission to programs in (1) physical sciences and engineering, (2) social sciences and education, and (3) humanities/arts. The challenge in creating such tests, of course, would be to write problems that required reasoning characteristic of the broad domain but that did not depend on knowledge that was too specific to a particular field (e.g., in the case of the physical sciences, biology).

To be effective, the development of these tests would need to be firmly grounded in research that elucidated the types of reasoning common across related fields. This research should serve as the basis for cognitive-domain specifications for GE, including a more complete description of the underlying construct. These cognitive-domain specifications would be the blueprint for test design. Protocol analyses of tests developed according to these specifications could then be conducted to verify that the reasoning processes examinees employ to solve GE items are similar to those actually used in graduate study.

Finally, any justification for GE would require examining the effect of different combinatorial rules on the diversity of the admitted applicant pool. Weighting GE by its (slight) relation to academic criteria would have little effect on changing the pool. Allowing high GE performance to compensate for a somewhat lower score on one of the General Test sections than would normally be required might have more impact.
A second major implication regards operational delivery. We discovered no serious impediments to delivering GE in computer-based test centers. All the same, significant questions not considered in this study remain, including whether items can be efficiently developed on a large scale, how sensitive the item type is to coaching, and whether responses can be scored economically. With respect to scoring, it is clear that GE responses can be accurately graded by humans but less so by computer (Bennett & Rock, 1995; Burstein & Kaplan, 1995; Kaplan & Bennett, 1994). More sophisticated computer-based tools are being explored and perhaps these tools can be combined with humans to create a workable semiautomatic approach. As for the new GRE writing measure, such an approach would likely involve some delay in scoring, as opposed to the instantaneous grading now available in the CAT environment.

A third implication of this study concerns methods for gathering data on new computer-based item types. Researchers and GRE Program staff should reexamine the idea of collecting such data using experimental sections that follow the operational test. In the current investigation about half of those offered participation either skipped the section or provided unusable data (which might be understandable as no tangible reward was offered). Precisely differentiating those giving a half-hearted response from those who were motivated proved a difficult task, one we only approximated at best. Unless we can find better ways to encourage participation and to measure effort, the evaluation of new item formats might be more effectively done through special data collections conducted outside the operational program (where incentives might be more easily provided).

Several limitations should be noted. The most important one was the sample bias introduced by the large number of examinees who did not participate or who gave inadequate data. Even so, this study's main findings were consistent with those of earlier investigations, suggesting that the bias was limited in its impact. A second limitation was the absence of a divergent thinking measure. Although previous studies have detected moderate-to-strong relations with GE (e.g., Bennett & Rock, 1995; Ward, Frederiksen, & Carlson, 1980), the absence of this indicator made it impossible to confirm in the CAT population that GE is related to measures of other like constructs. Finally, as suggested above, all of our external validation criteria were concurrent. GE might show more positive relations to graduate GPA and accomplishments than it did to the undergraduate indicators used in this study.
References


Appendix A

GE Test Directions
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.

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—with no 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 and 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 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.
Appendix B

A GE Item and Scoring Rubric
Dutch Landscape Paintings

Many seventeenth-century Dutch landscape paintings survive today. The genre was very popular: a number of artists made their living by producing such paintings, and a few major artists received very high commissions for their work.

To determine which artist produced a particular painting, scholars usually identify the signature appearing on the work. In the absence of a signature, documentary evidence, such as a bill of sale from a seventeenth-century art dealer, may serve as the basis on which a painting is attributed to a particular artist. In some cases, the artist is identified solely on the basis of a painting’s stylistic characteristics.

For most of the seventeenth-century Dutch landscape paintings that survive, the artists have been identified, but the identifications reveal an apparent disproportion: although far fewer major artists than minor artists painted in this genre, paintings attributed to those few major artists greatly outnumber paintings attributed to the minor artists.

The apparent disproportion -- the fact that paintings attributed to the relatively few major artists far outnumber attributions to the minor artists -- exists because...

1. MAJOR LANDSCAPE ARTISTS PRODUCED MORE LANDSCAPES THAN MINOR ARTISTS DID
   a) More committed, more professional
   b) More talented, could work faster
   c) Had assistants/minor artists to work for them
   d) Had commissions, money to paint
   e) Had money for longer, healthier life
   f) Controlled the art scene
   g) Major artists produced more art (GENERAL)
   h) Majors more into painting landscapes/ Minors concentrated more in another genre

3. MAJOR WORKS SURVIVED/LASTED LONGER
   a) Worth more money, more valuable, in safer environment
   b) Have more prestige, fame
   c) More beautiful, interesting, therefore saved
   d) Major artists used better materials
   e) Minor works destroyed intentionally--used for firewood, painted over
   f) Minor works were accidentally destroyed (fire, etc.)
   g) More works of major artists preserved (GENERAL)

4. MORE EVIDENCE OF ATTRIBUTION
   a) Major artists more likely to sign (minors didn't sign or used pseudonyms)
   b) Major artists more likely to have documented records or records more likely to be saved
   c) Have distinct styles, easier to identify (minor styles less familiar)
   d) Scholars more interested in attributing major, not minor
   e) Collectors demanded, needed evidence
   f) Other evidence--e.g., handwriting, materials--makes majors' paintings easier to recognize
g) Minor artists work thought to be from a different time period
h) More evidence of attribution (GENERAL)
i) Minor artists less likely to sell works so generated fewer bills of
sale as evidence

5. STYLISTIC PROBLEMS
   a) Experts don't know styles for certain/make guesses
   b) Minor artists copied major artists
   c) Some/more/less of the artists used similar styles
   d) Minor artists' work incorrectly attributed to major artists
   e) Some artists changed their styles, harder to identify

6. FORGERY, FAKES/FALSIFICATION
   a) Minors faked major signatures, to earn more
   b) Majors put their names on minors works
   c) Dealers, others faked documents, works

7. REPORTING/DOCUMENT PROBLEMS
   a) Many unattributed works still exist, and they could be by minors
   b) People ignore minors--don't care about their work
   c) Documents or other evidence hard to read or unavailable
   d) Definition of genre has changed
   e) When evidence sparse, works attributed to major artists
   f) Fewer minor works readily available for attribution or study
   g) Inaccurate records
   h) Minor artists misdescribed as major
   i) Problems with definition of major and minor artists
   j) Set of minor artists, by definition, includes many now-unknown
      individuals
   k) Some minor works are undiscovered

8. INADEQUATE RESPONSE
   a) Vague
   b) Incoherent
   c) Explains the reverse situation
   d) Nonexplanatory response
   e) Supernatural, conspiracy, luck
   f) Incomplete
   g) Contradicts information given
Appendix C

Accomplishments Questionnaire

Note: Subscores are composed as follows: Academic (1-5), Leadership (6-10), Linguistic (11-22), Aesthetic Expression (23-42), Science (43-47), Mechanical (48-52).
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
YES NO

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

THANK YOU FOR YOUR TIME IN COMPLETING THIS QUESTIONNAIRE!
Appendix D
Plots for the 13-Category Partial-Credit Model
Form 1: Item 1 (13 Category)

Score = 13

![Graph showing probability distribution with theta on the x-axis and probability on the y-axis. The graph includes a curve and some data points.](image-url)
Form 1: Item 2 (13 Category)

Score = 1

Score = 2

Score = 3

Score = 4

Score = 5

Score = 6
Form 1: Item 2 (13 Category)

Score = 13
Form 1: Item 3 (13 Category)

Score = 13
Form 2: Item 1 (13 Category)

Score= 7

Score= 8

Score= 9

Score= 10

Score= 11

Score= 12
Form 2: Item 1 (13 Category)

Score = 13
Form 2: Item 2 (13 Category)

Score = 1

Score = 2

Score = 3

Score = 4

Score = 5

Score = 6
Score = 13
Form 2: Item 3 (13 Category)

Score = 7

Score = 8

Score = 9

Score = 10

Score = 11

Score = 12
Form 2: Item 3 (13 Category)

Score = 13

![Graph showing probability vs. theta]