Effects of Examinee Choice on a Test of Divergent Thinking

Donald E. Powers and Randy Elliot Bennett

July 1998
GRE Board Professional Report No. 93-20P
ETS Research Report 98-30

Educational Testing Service, Princeton, New Jersey 08541
Effects of Examinee Choice on a Test of Divergent Thinking

Donald E. Powers and Randy Elliot Bennett

GRE Board Final Report No. 93-20P

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.
Acknowledgments

Many individuals contributed to this study. Bob Mislevy advised us on study design; 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; Eleanor DeYoung helped edit data files; Phil Leung and Kathy Carberry prepared files for data analysis; and Ruth Yoder prepared the manuscript. Special thanks go to Eiji Muraki and Min hwei Wang for conducting analyses. Mary Enright and Bill Ward provided helpful comments on an earlier draft of this report, and Nancy Parr provided editorial suggestions. Finally, we would like to acknowledge members of the GRE Board and its Research Committee for their advice and support.
Abstract

Under some circumstances, allowing examinees to choose which items to respond to may increase test validity. In this study, we explored how choice affected examinee performance and related characteristics for a test of the ability to generate hypotheses about a given situation. Four forms of this experimental test were randomly spiraled within a large sample of students taking the computer-based GRE General Test, with two of the forms allowing examinee choice. Results suggested that items were, in fact, differentially attractive to gender, ethnic, undergraduate major, and ability groups, indicating that, in some form, choice was operating. Further, performance was significantly higher when a given item was chosen than when it was required. Effects of choice on subgroup performance were generally inconclusive. Finally, there was some indication that internal consistency was improved by the choice option.
Preface

Some 20 years ago, in an edition of the American Psychologist, industrial psychologist Robert Guion (1974) reported on a study of carton packers -- people who made their living by placing objects into containers. Among the objectives was to investigate the potential of a series of physical measurements for predicting job performance. The criterion was a brief work sample that entailed packing several cartons of golf balls. Success was measured in terms of the time taken to fill the cartons. Guion and his coinvestigator found that one particular examinee characteristic -- arm length, the number of inches in the outstretched limb -- was a significant predictor of performance on the task ($r = .46$): the longer the arm, the faster the packer.

On closer examination, the researchers discovered that arm length predicted packing speed only for women, not for men. It seems that the experimental conditions, as standardized by the investigators, constrained everyone to sit exactly 32 inches from the packing cartons. For most of the men in the study, this distance posed no problem. For most of the women, however, the reach was difficult.

The investigators repeated the study, this time decreasing the distance between seats and packing cartons to 24 inches. Under these new standardized conditions, the results were reversed: arm length was now a good predictor of performance for men, but not for women. It seems that the work space was now too cramped for many of the men, but about right for most of the women. From these results Guion reasoned that, if study participants were allowed to choose the appropriate distance for their chairs, seats could be placed at an optimum distance from packing cartons. The result of allowing this choice was that arm length was no longer predictive of performance for either men or women, and productivity was maximized.
Standardization -- presenting identical (or equivalent) tasks in the same way to every test taker -- has long been a valued aspect of many forms of testing, as it is thought to enable comparisons among examinees that are both fair and valid. Generally, this belief is strongly held by those who develop and administer such tests: the conditions of testing (e.g., the time allowed to complete a test, the content and format of questions, the test directions, and other features of the test and of the test administration) should be exactly the same for every test taker.

Sometimes, though, standardization is intentionally violated, as when visually impaired examinees are permitted to take Braille or large-print editions of a test. But currently, even these quite necessary modifications are typically noted by “flagging” test scores to indicate to test users that some aspect of standardization was breached. It is probably not surprising then that some test makers balk at the prospect of letting examinees choose the questions they will answer, for permitting such choices may amount to, as Wainer and Thissen (1994) characterize it, “allowing examinees to construct their own test forms” -- a far more serious transgression than those mentioned earlier. These authors cite Gulliksen’s (1950) contention that, because of the difficulties involved in equating tests that allow choice, “alternative questions should always be avoided” (p. 338). They conclude, therefore, that allowing examinee choice may be “anathema” to fair testing.

Yet there is a lesson to be learned, we think, from the example given in the preface -- that under certain circumstances valid measurement may be better served by selectively relaxing, rather than strictly observing, the conditions of standardization, provided that the conditions to be relaxed are in some sense artificial. In the example discussed above, breaching one aspect of standardization apparently eliminated an irrelevant, gender-related source of task difficulty, thus resulting in a fairer, more valid assessment of performance. In the example, think of packing speed as the ability to be assessed, and distance from the assembly line as one of the contextual variables under which ability is to be measured and on which the assessment is standardized. Arm length is an examinee characteristic that may affect measurement, but one that is not intrinsically related to the primary focus of measurement: packing speed. In some assessments (of writing ability, for example), distance might be analogous to the examiner’s choice of content or of topics on which the examinee is to write. Arm length would correspond to the examinee’s level of comfort or previous experience with the topic on which he or she is asked to write -- something that is of little interest to the examiner. One way to minimize any potential negative impact of largely irrelevant contextual variables might be to let test takers make certain choices for themselves.

Arguments for (and Against) Permitting Examinee Choice

The call for choice in testing has come from a number of quarters, most notably from the performance assessment arena. As Fitzpatrick and Yen (1995) asserted, offering choices appears to be “a desideratum of the authentic assessment movement” (p. 257). Gordon (1992), for instance, believes that allowing options and choices is a necessary condition for equitable assessment and has called for “self-designated tasks from examinee-generated inventories of
knowledge, skill, and understanding” (p. 7). Wiggins (1993) expressed a preference for giving examinees “an opportunity to show off” instead of having to undergo a “trial by question” (p. 141), believing that the former might be more likely than the latter to engage examinees.

Discussing learning environments, Snow and Jackson (1993) asserted that one of the factors affecting intrinsic motivation is the degree of control or choice that the learner has. Consistent with this view, Deci and Ryan (1985) provided evidence that when learners are allowed to choose, they exhibit greater motivation and are more likely to persevere at tasks. Other research also seems to support the common-sense notion that students work harder on tasks that are of interest to them (e.g., Hidi & Anderson, 1992). Quite conceivably, this situation might also apply to testing situations.

The call for examinee choice has also come from conventional testing circles. Even though computer adaptive testing tailors questions to examinees’ ability levels, it may not go far enough, according to some. Rocklin and his associates (Rocklin, O’Donnell, & Holst, 1995) implied that test questions could be tailored not only on the basis of difficulty level, but according to other, noncognitive factors as well. For instance, they suggest that permitting examinees to specify the difficulty level of the questions they will answer (self-adapted testing) results in higher test scores (even though adaptive testing should produce the same score for an examinee regardless of the specific questions presented). The researchers suggest that selecting questions so as to standardize the level of difficulty encountered by each examinee (that is, so every test taker answers the same proportion of items correctly) might not be the best strategy. To them, not allowing such choice seems inconsistent with the increasing encouragement that many students now receive to exercise choice as self-regulated learners.

White (1994) portrayed a similar view of the assessment of writing, as held by some faculty members -- that to fix a topic is to constrain writers’ imagination. Underlying this conviction is an assumption that choosing a topic is itself so important a part of the construct of interest that it should be part of the task. But White disputed this view, feeling instead that this “supposed freedom” is in reality yet another constraint, requiring the writer to guess what the evaluator wants. Furthermore, as Wainer and Thissen (1994) asserted, not everyone is equally proficient at making wise choices, and this propensity seems to vary by gender and ethnicity. They cite numerous instances in which, when allowed to choose, examinees make less-than-optimal choices. The ability to choose well may also differ according to ability, for as White (1985) suggested, the most difficult (essay) questions are often the most interesting/challenging and therefore the most appealing to the most able examinees. This is in fact what Powers, Fowles, Farnum, and Gerritz (1992) found for a set of essay prompts developed for a beginning teacher certification test.

**Toward Defining the Circumstances Under Which Examinee Choice Is Appropriate**

Bridgeman, Morgan, and Wang (1995) argued that if the purpose of an (essay) assessment is to determine examinees’ ability to organize evidence and present a cogent argument on a topic
on which they are knowledgeable, choice may be necessary. On the other hand, if an assessment is intended to measure specific knowledge, the choice of questions is best left to the examiner.

Similarly, Burton (1993) expressed a belief that choice is acceptable in situations in which the ability to make good choices is central to the construct being measured, but that it is not acceptable when the objective is to make inferences about knowledge of a domain. She suggested that choice may be necessary, for example, when measuring low-level reading proficiency in a foreign language: unless examinees have some familiarity with the general content, a much higher level of proficiency is required. The same argument has been advanced with regard to measuring writing proficiency (Powers et al., 1992).

Wainer and Thissen (1994) believe that choice is permissible only when "...those aspects that characterize the choice are irrelevant to what is being tested" (p. 191). They also attributed to W. B. Schrader another justification for permitting examinee choice. Assuming that outstanding people are usually outstanding on a small number of qualities (and that the purpose of the exam is to identify such people), test takers ought to be allowed to select the tasks on which they can demonstrate their best performances. Similarly, Longford (1993) also suggested that when the goal is to determine strengths, choice may be appropriate, but when the objective is to diagnose weaknesses, permitting examinee choice is probably not an advisable measurement strategy.

Related Research

Several studies have investigated the role of choice in assessment situations. All of the studies cited here have been conducted relatively recently, indicating perhaps a heightened interest in the effects of allowing examinee choice. As suggested in Table 1, these studies have differed in their objectives, their designs, and the kinds of measures that have been investigated. Tests of reading and writing skills and achievement tests in chemistry and history have been the focus of the studies we were able to locate. Most of the studies were of tests that required students to construct responses. Most attempted to estimate the overall effect of allowing choice, most discussed examinee preferences for different content, and some examined the effects of choice on the validity of test scores. One study attempted to determine how examinees go about making their choices. A few addressed the question of who makes good choices.

As can be seen from Table 1, the results of the studies to date have been mixed with respect to the effects of permitting examinee choice. The most succinct summary we can provide is that

- the effects of choice may differ according to the nature of the test: the average effect may be negative for some tests and positive for others
- test takers do not uniformly make good choices: at least some examinees will apparently make poor choices regardless of the nature of the test
- the prevalence of poor choices may vary according to the nature of the test and the characteristics of examinees
Table 1

Summary of Studies Investigating Effects of Examinee Choice

<table>
<thead>
<tr>
<th>Measures/Subjects</th>
<th>Test Format/Scoring</th>
<th>Nature of Choice</th>
<th>Results</th>
<th>Possible Study Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advanced Placement Exams in U.S. and European History/Secondary school history students</td>
<td>Constructed response/ Holistic scoring on a 9-point scale</td>
<td>For each of two subject areas, students were presented with two topics (in all, four pairs were constituted). They were asked to read both topics and to choose the question that they felt best prepared to answer. They were then asked to answer both questions.</td>
<td>For each of the two subject areas, mean scores were higher on preferred topics by about .3 to .5 sds. For each area about 30% of students made the wrong choice, that is, got a higher score on the nonpreferred topic. Preferred essays correlated more strongly than did non-preferred essays with other measures.</td>
<td>Possible unintended effect of allowing students to choose and then requiring them to answer both questions.</td>
</tr>
<tr>
<td>Reading: ability to construct meaning from literary, informational, and instructional texts/3rd, 5th, and 8th grade students</td>
<td>Constructed response/ 2-, 3-, or 4-point ratings</td>
<td>Students received a combination of choice and no-choice reading passages.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Writing: composing persuasive essays/11th graders</td>
<td>Constructed response/ 4-point analytical scoring in 4 domains: content and organization, style, conventions, and sentence formation</td>
<td>Students were assigned to either a choice or a no-choice condition. Within the choice condition each student was given a pair of tasks and asked to choose one (60 different pairs were constituted).</td>
<td>Very small effects (.01 to .05 sds) were noted in favor of the no-choice condition.</td>
<td>Nonexperimental, confounding of passage content and choice condition, not designated to determine if students make good choices.</td>
</tr>
<tr>
<td>Writing: description, comparison, and argumentation/International students enrolled in an intensive English program at a U.S. university</td>
<td>Constructed response (30-minute essay)</td>
<td>Students were asked to choose one of three topics</td>
<td>Students appeared to have little if any difficulty in deciding which topic to write on: on average, little time was taken to choose, but there was significant variance. Students chose mainly on the basis of their familiarity with a topic (or their estimate of how much they could write about it).</td>
<td>Small, possibly unrepresentative sample. No analysis of test performance attempted.</td>
</tr>
<tr>
<td>Measures/ Subjects</td>
<td>Test Format/ Scoring</td>
<td>Nature of Choice</td>
<td>Results</td>
<td>Possible Study Limitations</td>
</tr>
<tr>
<td>---------------------</td>
<td>----------------------</td>
<td>------------------</td>
<td>---------</td>
<td>---------------------------</td>
</tr>
<tr>
<td>Advanced Placement exams in chemistry, U.S. and European history, comparative government and politics, and U.S. government and politics/ Secondary school students</td>
<td>Constructed response/ Holistic scoring on either a 9-point or 15-point scale</td>
<td>Amount of choice varied by exam -- either one of two, one of five, one of six, or two of four.</td>
<td>Choices differed by ability level (as defined by performance on multiple-choice questions or on mandatory free-response questions). The various choice questions were not of equal difficulty. The tests differed with respect to who was disadvantaged by allowing choice.</td>
<td>Results based on the assumption that multiple-choice and free-response questions are measures of the same construct.</td>
</tr>
<tr>
<td>Advanced Placement exams in chemistry/ Secondary school chemistry students</td>
<td>Multiple-choice/ Dichotomous (right/wrong)</td>
<td>Students were given three pairs of questions (within a 20-item test) and asked, for each pair, which one they would choose. After responding they were required to answer both questions.</td>
<td>Not all students chose the items on which they subsequently performed best. Item pairs differed with respect to the “wiseness” of choices made.</td>
<td>Few items studied, small sample size.</td>
</tr>
<tr>
<td>Advanced Placement exam in chemistry/ National population of secondary school AP chemistry test takers</td>
<td>Constructed response/ Holistic scoring on a 9-point scale</td>
<td>Three of five questions.</td>
<td>Correlation between popularity of topics and performance was -.60. Effect of choice was greater for middle-ability students than for high- and low-ability students. Each topic combination was chosen about equally often by students at different ability levels. Both teachers and test takers judged familiar topics to be easier.</td>
<td>Nonexperimental, no comparable choice and no-choice conditions.</td>
</tr>
</tbody>
</table>
Currently, it is possible, therefore (depending on the research one cites), to either strongly favor examinee choice or vigorously oppose it. As more becomes known about the impact of examinee choice under varying circumstances, it seems likely that extreme opinions will moderate, and the question will be phrased not in terms of whether or not choice should be allowed, but rather in terms of (a) under what circumstances choice is acceptable (or even necessary), (b) when it is desirable, and (c) when it is altogether unacceptable.

Objectives

The overarching objective of the study reported here was to provide some additional information about the effects of allowing examinee choice, thereby helping to inform the issue of when choice may be appropriate and when it may not be. Specifically, we hoped to learn something about each of the following:

- the degree to which examinees (and subgroups of examinees) have preferences among alternative questions
- the effects of choice on level of performance for all examinees as well as for a number of subgroups
- the effects of choice on certain psychometric properties of test scores (e.g., reliability and relations with other variables).

The test studied here was a test of examinees' ability to generate plausible explanations or hypotheses for given situations. As will be discussed below, this test is quite different from other tests that have been investigated in studies of the effects of examinee choice.

Method

Study Participants

Participants were examinees who took the Graduate Record Examinations (GRE) General Test during the summer and fall of 1994 at approximately 200 computer-based testing centers throughout the United States. The version of the General Test that was administered during this time period included a final, 30-minute experimental section containing generating explanations (GE) items. The GE section was identified as experimental, and examinees were informed that their scores on this section of the test would not be reported.

Instruments

The measure -- "generating explanations" (formerly called "formulating hypotheses") -- was developed by Frederiksen (1959) and researched by Frederiksen and his colleagues (Frederiksen & Ward, 1978; Ward, Frederiksen, & Carlson, 1980). It can be characterized as a measure of divergent production -- by virtue of both its requirements (to generate multiple responses) and its relationship with measures of the ability to generate ideas. It contrasts, therefore, with most
traditional multiple-choice tests inasmuch as the latter tend to draw heavily on the ability to converge on a single correct answer. Performance on the measure, in fact, is only very modestly related to performance on the GRE General Test (Bennett & Rock, 1995; in press).

The measure was administered in a computerized format in which examinees typed, edited, and saved up to 15 hypotheses (each of 15 words or fewer) for each question. Examinees were given a description of the test at the end of the last operational section of the GRE General Test; an untimed tutorial that provided instructions on use of the computer; and information on how responses would be scored. In scoring, credit was awarded to each plausible, unduplicated hypothesis. Further details on scoring are available in Bennett & Rock (in press).

Six GE problems were used in the study. Each problem described, in approximately 50 to 150 words, a phenomenon to be explained. Some problems were accompanied by a graph or chart depicting the situation. Every problem concluded with a statement (e.g., “the sharp drop in the number of deer reported killed occurred because…” which the examinee was to complete by generating as many possible explanations as he or she could. An abbreviated description of each problem follows (abbreviations to be used henceforth are designated in bold):

A  The number of deer killed by automobiles remained relatively constant over the years, but showed a sharp drop in one year. The sharp drop in the number of deer reported killed occurred because...

B  Although there were far fewer major Dutch landscape painters than minor ones, of the paintings that have survived, those attributed to major artists greatly outnumber those attributed to 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...

C  The amount of garbage cleaned off beaches rose steadily over a 50-year period. The steady rise in the number of tons of garbage cleaned off of Datar’s beaches occurred because...

D  The total annual weight of codfish caught by ships sailing from a certain port remained constant over several years but dropped sharply in one year. The sharp drop in the annual tonnage of codfish occurred because...

E  “Daybooks” (accounts kept by 18th- and early 19th-century silversmiths) were consulted occasionally by scholars and collectors between 1930 and the 1980’s, but not from 1983 to 1988. In 1989 a visiting scholar was told that the books were unavailable. This occurred because...

F  Over a 20-year period the rate of death rate of police officers killed in the line of duty dropped significantly. The drop in the rate of police officers killed in the line of duty occurred because...
Design

Four different forms of the GE test were spiraled in the experimental section (i.e., they were randomly assigned to examinees). For two of the forms (no-choice), examinees were asked to complete all three problems presented in 30 minutes; for two other forms (choice) they were asked to complete two problems presented in 20 minutes. In a separately timed section they were then shown the final sentences (e.g., "The drop in the rate of police officers killed in the line of duty occurred because . . .") of three other problems. Examinees were then asked to read the directions and select one of the three alternative questions and to answer it. The level of choice is similar to that employed in operational test settings, for example, in the College Board's Advanced Placement Program, and was thus believed to allow a reasonable degree of latitude. The degree of choice allowed here was also believed to differ sufficiently from a no-choice situation to enable the detection of any practically significant effect of allowing choice.\(^1\) The basic design is shown in the following:

<table>
<thead>
<tr>
<th>Test Form</th>
<th>Choice</th>
<th>Problems</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>No</td>
<td>A, B, and C</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>D, E, and F</td>
</tr>
<tr>
<td>3</td>
<td>Yes</td>
<td>A, B, and (C, D, or E)</td>
</tr>
<tr>
<td>4</td>
<td>Yes</td>
<td>D, E, and (F, A, or B)</td>
</tr>
</tbody>
</table>

Study Design

The design resulted in each examinee taking one of several possible combinations of items. Examinees in the no-choice condition answered either questions A, B, and C or questions D, E, and F. Those in the choice condition answered (1) A, B, and C, (2) A, B, and D, (3) A, B, and E, (4) D, E and F, (5) D, E, and A, or (6) D, E, and B.

Results

Sample Description

Characteristics of the sample on which analyses are based are shown in Table 2. Slightly fewer than half of the examinees to whom the GE test was offered actually chose to respond to it.

\(^1\) In each of the choice conditions, the standard tutorial and the standard test directions were given, as in the no-choice conditions. Next, however, slightly modified section directions were presented, informing examinees that they would have 20 minutes to answer the next two required items. After they had completed these items, examinees were shown the three choice items and given up to four minutes to select one to answer. They had 10 minutes to answer the item chosen.
In addition, when all data were examined, it was clear that, although some examinees responded, they spent minimal time and effort attempting the items, reading the test directions, or studying the interface tutorial. Others either submitted only a single response (a maximum of 15 was possible) or entered humorous or nonsensical responses. Cases were deleted so that the final sample included only examinees who offered more than a single serious response and who spent at least a minimal amount of time on the tutorial, the directions, and in trying items. Additionally, examinees in the choice condition who failed to follow instructions regarding choice (e.g., they answered more than one of the alternative choice questions) were deleted. Further details (and rationale) for the exclusion of study participants are given by Bennett & Rock (in press).

Although the final choice and no-choice samples are similar with respect to several important characteristics, they do differ slightly, but significantly, with regard to average GRE verbal and quantitative scores. This difference results from the exclusion of examinees who did not comply with the directions for choosing, as these participants had slightly lower test scores on average.

Analysis of Examinee Choices

Table 3 shows that the items in each choice form were differentially attractive. One item ("police") appeared to be especially appealing when offered in combination with "deer" and "paintings," the latter being quite unpopular overall. As Tables 4-7 reveal, however, choices differed somewhat according to examinee characteristics. For instance, the item on "fish" was chosen less often by females than males relative to their choice of the item about "daybooks" (Table 4). There were also significant differences among ethnic groups with respect to their choice of items about "police," "deer," and "paintings" (Table 5). The patterns of choice were also significantly different according to undergraduate major field (Table 6) -- "paintings," for example, being relatively more attractive to humanities majors than to other groups. Examinees who chose to answer the "paintings" question were, on average, the most able of any groups according to their GRE scores, and those who chose "beaches" were the least able (Table 7).

We also attempted to describe examinee choices in a more multivariate fashion by trying to identify clusters of examinees making similar choices. To this end we attempted to segment the total sample into subgroups by using CHAID (Chi-squared automatic interaction detector); Magidson, 1993. The objective was to form the subgroups on the basis of variables that best predicted examinees' choices of items to answer. Sex, degree aspirations, GRE scores, and undergraduate major were used as the predictors. For each of the two choice forms, undergraduate major was the primary determinant of examinee clusters (i.e., the first variable selected by CHAID to classify examinees). Several other variables were instrumental in effecting further subclassifications. However, when subjected to verification by analysis on random halves of the samples, the results were not at all consistent for either form of the test. And, even though undergraduate major was, for each test form and each random sample of examinees, the variable that best defined choices, the groupings of majors were not consistent across samples.
Table 2
Characteristics of Choice Sample vs. No-Choice

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Choice Sample (n = 2429)</th>
<th>No-choice Sample (n = 4720)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex (% female)</td>
<td>57</td>
<td>58</td>
</tr>
<tr>
<td>Ethnicity (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>American Indian</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Black</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Asian American</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>White</td>
<td>85</td>
<td>84</td>
</tr>
<tr>
<td>Hispanic(^1)</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Undergraduate Major (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Life Science</td>
<td>25</td>
<td>28</td>
</tr>
<tr>
<td>Physical Science</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>Engineering</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Social Science</td>
<td>22</td>
<td>20</td>
</tr>
<tr>
<td>Humanities</td>
<td>16</td>
<td>17</td>
</tr>
<tr>
<td>Education</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>Business</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Other</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Undergraduate Grade Average (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean overall (sd)(^2)</td>
<td>5.3(1.1)</td>
<td>5.2(1.1)</td>
</tr>
<tr>
<td>Mean in major (sd)(^2)</td>
<td>5.7(1.0)</td>
<td>5.7(1.0)</td>
</tr>
<tr>
<td>GRE General Test</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean GRE verbal (sd)</td>
<td>512(111)</td>
<td>505(111)</td>
</tr>
<tr>
<td>Mean GRE quantitative (sd)</td>
<td>572(133)</td>
<td>561(135)</td>
</tr>
<tr>
<td>Mean GRE analytical (sd)</td>
<td>595(136)</td>
<td>589(136)</td>
</tr>
</tbody>
</table>

Note. Ns for each characteristic are based on variable numbers of subjects for whom data were available. Nearly 100% of the samples reported their gender, about 94% their ethnicity, about 88% their undergraduate major, about 49% their undergraduate grades. GRE scores were available for more than 99% of the sample.

\(^1\)Includes Mexican American, Puerto Rican, and other Hispanics.

\(^2\)Grades were reported on a 7-point scale, with A = 7, A- = 6, B = 5, B- = 4, C = 3, C- = 2, D or lower = 1.
Table 3

Percentages of Examinees Choosing Each Problem

<table>
<thead>
<tr>
<th>Test Form</th>
<th>N</th>
<th>Percentage Choosing Each Problem</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1242</td>
<td>Beaches (48) Fish (20) Books (32)</td>
</tr>
<tr>
<td>4</td>
<td>1187</td>
<td>Police (76) Deer (20) Paintings (4)</td>
</tr>
</tbody>
</table>

Table 4

Percentages of Females and Males Choosing Each Generating Explanations Items

<table>
<thead>
<tr>
<th>Group</th>
<th>Form/Item</th>
<th>Form 3</th>
<th></th>
<th>Form 4</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Beaches</td>
<td>Fish</td>
<td>Books</td>
<td>x²</td>
</tr>
<tr>
<td>Female (n=703)</td>
<td></td>
<td>45</td>
<td>16</td>
<td>39</td>
<td></td>
</tr>
<tr>
<td>Male (n=535)</td>
<td></td>
<td>51</td>
<td>25</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td>Female (n=674)</td>
<td></td>
<td>76</td>
<td>19</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Male (n=512)</td>
<td></td>
<td>76</td>
<td>20</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>
Table 5

Percentages Choosing Each Generating Explanations Item by Ethnicity

<table>
<thead>
<tr>
<th>Group</th>
<th>Form/Item</th>
<th>Form 3</th>
<th>Form 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Beaches</td>
<td>Fish</td>
</tr>
<tr>
<td>American Indian</td>
<td>(n=9)</td>
<td>67</td>
<td>0</td>
</tr>
<tr>
<td>Black</td>
<td>(n=48)</td>
<td>52</td>
<td>13</td>
</tr>
<tr>
<td>Asian American</td>
<td>(n=38)</td>
<td>37</td>
<td>26</td>
</tr>
<tr>
<td>White</td>
<td>(n=1012)</td>
<td>47</td>
<td>21</td>
</tr>
<tr>
<td>Hispanic</td>
<td>(n=84)</td>
<td>52</td>
<td>18</td>
</tr>
</tbody>
</table>
Table 6

Percentages Choosing Each Generating Explanations Item by Undergraduate Major Field

| Major Field          | Form/Item | Form 3 | | Form 4 | | x2 | p<.001 |
|----------------------|-----------|--------|---------|--------|-----|---------|
|                      |           | Beaches | Fish | Books | Police | Deer | Paintings | |
| Life Science (n=284) |           | 44      | 29    | 27    | 71     | 28   | 1        |      |
| Physical Science (n=97) |       | 45      | 28    | 27    | 72     | 25   | 3        |      |
| Engineering (n=84)   |           | 58      | 26    | 16    | 79     | 20   | 1        |      |
| Social Science (n=231)|         | 48      | 13    | 39    | 80     | 17   | 4        |      |
| Humanities (n=180)   |           | 38      | 18    | 44    | 68     | 17   | 15       |      |
| Education (n=90)     |           | 63      | 9     | 28    | 80     | 16   | 3        |      |
| Business (n=56)      |           | 41      | 20    | 39    | 89     | 9    | 3        |      |
| Other (n=84)         |           | 37      | 23    | 40    | 84     | 13   | 3        |      |
Table 7

GRE General Test Scores by Choice of Generating Explanations Items

<table>
<thead>
<tr>
<th>GRE General Test Score</th>
<th>Form/Item</th>
<th>Form 3</th>
<th>Form 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Beaches</td>
<td>Fish</td>
<td>Books</td>
</tr>
<tr>
<td>GRE verbal</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>496(112)</td>
<td>524(105)</td>
<td>532(111)</td>
<td>p&lt;.001</td>
</tr>
<tr>
<td>GRE quantitative</td>
<td>563(130)</td>
<td>603(119)</td>
<td>565(130)</td>
</tr>
<tr>
<td>GRE analytical</td>
<td>569(136)</td>
<td>629(129)</td>
<td>606(132)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GRE verbal</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>501(110)</td>
<td>524(104)</td>
<td>593(116)</td>
<td>p&lt;.001</td>
</tr>
<tr>
<td>GRE quantitative</td>
<td>564(140)</td>
<td>600(122)</td>
<td>601(142)</td>
</tr>
<tr>
<td>GRE analytical</td>
<td>585(138)</td>
<td>635(124)</td>
<td>656(123)</td>
</tr>
</tbody>
</table>

Note. Table entries are mean GRE scores with standard deviations in parentheses.
Performance on Choice and Required Items

Table 8 shows the performance of all examinees on each of the two choice forms and on each of the two no-choice forms. The status of each item (i.e., whether it was required or whether it was an optional, choice item) is also indicated. As is clear (from the statistics from the no-choice forms), the items were differentially difficult. For instance, the difference between the easiest item (deer) and the most difficult (paintings) was about .7 sd units, a relatively large difference in terms of effect sizes (Cohen, 1977). Table 8 also suggests that the four items that were required on both the choice and the no-choice forms (deer, paintings, fish, and books) were easier for students who took the choice forms than for those who took the no-choice forms -- by, on average, approximately .1 sd units. These differences result almost entirely from our deletion from the sample of (the lower performing) examinees who failed to follow our instructions to answer only three items. Before the deletion of these noncomplying examinees, the choice and no-choice samples performed virtually identically on all but one of the items. Inexplicably, the deer item was slightly (.1 sd), but significantly (p<.001) easier for examinees taking the choice form. Of more interest, however, is that when comparisons are made within choice forms (where samples are comparable), performance was slightly but significantly (p<.001) higher (by .1 to .2 sd units) on each of the four items when examinees were allowed a choice of the item. On average, then, choice was beneficial to examinees.

This result was confirmed by applying a partial-credit item response model (Masters, 1982) to a portion of our data, specifically to item C in forms 1 and 3 and to item F in forms 2 and 4. The assumption underlying this analysis was that the groups taking the different forms (1 and 3 or 2 and 4) were equivalent in terms of latent trait distributions. (As we know, however, this assumption is questionable because we deleted the records of some examinees who took the choice forms.) It was also assumed that the slope parameter for all items was 1.0. This analysis too suggested that for each of these two items, allowing choice made the item slightly easier.

Popularity of Items Versus Performance

Table 9 reveals that, for test form 3, the relationship between the extent to which examinees chose an item and the difficulty of the item (as determined from performances under a no-choice condition) was not entirely consistent. Although the least popular item (paintings), chosen by only 4% of the sample, was also by far the most difficult item, there was a reversal for the other two items on this form. Some of the inconsistency between popularity and difficulty may be due, as shown later, to differences in the ability levels of examinees making particular choices. For the second form, however, the popularity of items was completely consistent with their difficulty, but in an unexpected direction: the harder the item (when required for all examinees), the more popular the item when choice was allowed.
Table 8
Performance on Generating Explanations Items by Test Form and Choice Condition

<table>
<thead>
<tr>
<th>Item</th>
<th>No Choice</th>
<th></th>
<th>Choice</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Form 1</td>
<td>Form 2</td>
<td>Form 3</td>
</tr>
<tr>
<td></td>
<td>Required</td>
<td>Required</td>
<td>Choice</td>
</tr>
<tr>
<td>Deer</td>
<td>m</td>
<td>6.3</td>
<td>6.6</td>
</tr>
<tr>
<td></td>
<td>sd</td>
<td>3.0</td>
<td>3.2</td>
</tr>
<tr>
<td></td>
<td>n</td>
<td>2351</td>
<td>1242</td>
</tr>
<tr>
<td>Paintings</td>
<td>m</td>
<td>4.3</td>
<td>4.8</td>
</tr>
<tr>
<td></td>
<td>sd</td>
<td>2.5</td>
<td>2.9</td>
</tr>
<tr>
<td></td>
<td>n</td>
<td>2351</td>
<td>1242</td>
</tr>
<tr>
<td>Beaches</td>
<td>m</td>
<td>5.3</td>
<td>6.0</td>
</tr>
<tr>
<td></td>
<td>sd</td>
<td>3.0</td>
<td>3.6</td>
</tr>
<tr>
<td></td>
<td>n</td>
<td>2351</td>
<td>592</td>
</tr>
<tr>
<td>Fish</td>
<td>m</td>
<td>—</td>
<td>6.1</td>
</tr>
<tr>
<td></td>
<td>sd</td>
<td>—</td>
<td>3.0</td>
</tr>
<tr>
<td></td>
<td>n</td>
<td>—</td>
<td>2369</td>
</tr>
<tr>
<td>Books</td>
<td>m</td>
<td>—</td>
<td>5.7</td>
</tr>
<tr>
<td></td>
<td>sd</td>
<td>—</td>
<td>3.1</td>
</tr>
<tr>
<td></td>
<td>n</td>
<td>—</td>
<td>2369</td>
</tr>
<tr>
<td>Police</td>
<td>m</td>
<td>—</td>
<td>6.0</td>
</tr>
<tr>
<td></td>
<td>sd</td>
<td>—</td>
<td>3.3</td>
</tr>
<tr>
<td></td>
<td>n</td>
<td>—</td>
<td>2369</td>
</tr>
</tbody>
</table>
Table 9

Relationships of Popularity of GE Items to Difficulty

<table>
<thead>
<tr>
<th>Item</th>
<th>Percentage Choosing Item on Choice Forms</th>
<th>Mean Score on No-choice Forms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Police</td>
<td>76</td>
<td>6.0</td>
</tr>
<tr>
<td>Deer</td>
<td>20</td>
<td>6.3</td>
</tr>
<tr>
<td>Paintings</td>
<td>4</td>
<td>4.3</td>
</tr>
<tr>
<td>Beaches</td>
<td>48</td>
<td>5.3</td>
</tr>
<tr>
<td>Books</td>
<td>32</td>
<td>5.7</td>
</tr>
<tr>
<td>Fish</td>
<td>20</td>
<td>6.1</td>
</tr>
</tbody>
</table>

Performance by Subgroups

Because of the small numbers of minority test takers choosing particular questions, we were unable to conduct an analysis for each ethnic classification. We did, however, attempt this analysis for Black and Hispanic examinees. (Table 10 displays the numbers of Black and Hispanic examinees. Women constituted about 58% of the total sample and were distributed across ethnic groups in roughly this proportion.) Even for these two groups the number of examinees was very small for some test items -- too small to generate meaningful statistics in all cases. For the sake of completeness, however, we have presented results for both of these groups for each test form.

Table 11 shows that men and women performed similarly on each of the items when they were administered without choice. There were two exceptions (deer and fish), however, on which men performed significantly (but only slightly) better -- by about .1 to .2 sd units. There is nothing in the data to suggest that allowing choice systematically benefited or disadvantaged women. For the four items that were administered to the choice sample under both required and choice conditions (and therefore yielded the most unambiguous comparisons), there was no substantial or consistent change, with respect to the differential performance of women and men, as a result of allowing examinee choice. Results of comparisons of the performances of Black and White examinees (Table 12) show that GE items were more difficult for Black than for White test takers (by .2 to .5 sd units), and that choice had no effect on decreasing the differential between Black and White examinees. For one item (books), the difference may have increased. The picture for Hispanic test takers (Table 13) is much the same. Two items (deer and fish) were only slightly (about .2 sds), but significantly, more difficult for Hispanic examinees. Small sample sizes precluded any firm conclusions about changes as a result of allowing choice, but on balance the evidence does not suggest that choice was instrumental in decreasing the differential between Hispanic and White test takers.
Table 10

Numbers of Examinees Upon Whom Item Score Comparisons Are Based

<table>
<thead>
<tr>
<th>Item</th>
<th>Group</th>
<th>Required (No-choice Form)</th>
<th>Required (Choice Form)</th>
<th>Choice (Choice Form)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deer</td>
<td>Black</td>
<td>115</td>
<td>48</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Hispanic</td>
<td>116</td>
<td>84</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>White</td>
<td>1,891</td>
<td>1,012</td>
<td>201</td>
</tr>
<tr>
<td>Paintings</td>
<td>Black</td>
<td>115</td>
<td>48</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Hispanic</td>
<td>116</td>
<td>84</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>White</td>
<td>1,891</td>
<td>1,012</td>
<td>37</td>
</tr>
<tr>
<td>Beaches</td>
<td>Black</td>
<td>115</td>
<td>—</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>Hispanic</td>
<td>116</td>
<td>—</td>
<td>44</td>
</tr>
<tr>
<td></td>
<td>White</td>
<td>1,891</td>
<td>—</td>
<td>472</td>
</tr>
<tr>
<td>Fish</td>
<td>Black</td>
<td>124</td>
<td>47</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Hispanic</td>
<td>149</td>
<td>54</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>White</td>
<td>1,867</td>
<td>954</td>
<td>209</td>
</tr>
<tr>
<td>Books</td>
<td>Black</td>
<td>124</td>
<td>47</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>Hispanic</td>
<td>149</td>
<td>54</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>White</td>
<td>1,867</td>
<td>954</td>
<td>331</td>
</tr>
<tr>
<td>Police</td>
<td>Black</td>
<td>124</td>
<td>—</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>Hispanic</td>
<td>149</td>
<td>—</td>
<td>46</td>
</tr>
<tr>
<td></td>
<td>White</td>
<td>1,867</td>
<td>—</td>
<td>716</td>
</tr>
</tbody>
</table>
Table 11

Mean Generating Explanations Item Scores by Gender and Choice Condition

<table>
<thead>
<tr>
<th>Item</th>
<th>Females/Males</th>
<th>Required (No-choice Form)</th>
<th>Required (Choice Form)</th>
<th>Choice (Choice Form)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Deer</td>
<td>F</td>
<td>6.1(3.0)</td>
<td>6.4(3.2)</td>
<td>6.7(3.3)</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>6.5(3.1)</td>
<td>6.9(3.2)</td>
<td>7.5(3.4)</td>
</tr>
<tr>
<td></td>
<td>d</td>
<td>-.12*</td>
<td>-.18**</td>
<td>-.25</td>
</tr>
<tr>
<td>Paintings</td>
<td>F</td>
<td>4.3(2.5)</td>
<td>4.8(2.9)</td>
<td>5.5(2.0)</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>4.3(2.6)</td>
<td>4.9(2.9)</td>
<td>4.5(1.8)</td>
</tr>
<tr>
<td></td>
<td>d</td>
<td>-.02</td>
<td>-.03</td>
<td>+.56</td>
</tr>
<tr>
<td>Beaches</td>
<td>F</td>
<td>5.3(3.0)</td>
<td>—</td>
<td>6.1(3.5)</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>5.3(3.1)</td>
<td>—</td>
<td>5.9(3.7)</td>
</tr>
<tr>
<td></td>
<td>d</td>
<td>+.01</td>
<td>—</td>
<td>+.06</td>
</tr>
<tr>
<td>Fish</td>
<td>F</td>
<td>6.0(3.1)</td>
<td>6.2(3.2)</td>
<td>7.0(3.3)</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>6.4(3.0)</td>
<td>6.8(3.2)</td>
<td>6.8(3.3)</td>
</tr>
<tr>
<td></td>
<td>d</td>
<td>-.13**</td>
<td>-.19**</td>
<td>+.07</td>
</tr>
<tr>
<td>Books</td>
<td>F</td>
<td>5.6(3.2)</td>
<td>5.8(3.2)</td>
<td>6.3(3.7)</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>5.7(3.1)</td>
<td>6.2(3.2)</td>
<td>7.1(3.5)</td>
</tr>
<tr>
<td></td>
<td>d</td>
<td>-.02</td>
<td>-.12</td>
<td>-.21</td>
</tr>
<tr>
<td>Police</td>
<td>F</td>
<td>5.9(3.2)</td>
<td>—</td>
<td>6.4(3.5)</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>6.1(3.4)</td>
<td>—</td>
<td>6.7(3.5)</td>
</tr>
<tr>
<td></td>
<td>d</td>
<td>-.07</td>
<td>—</td>
<td>-.07</td>
</tr>
</tbody>
</table>

$d = \text{mean of female test takers minus mean of male test takers divided by the within-group sd.}$

$sds \text{ are in parentheses.}$

*$p<.05, **p<.01$
Table 12
Mean Generating Explanations Item Scores by Ethnicity (Black vs. White) and Choice Condition

<table>
<thead>
<tr>
<th>Item</th>
<th>Black/White</th>
<th>Choice Status</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Required</td>
<td>Required</td>
<td>Choice</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(No-choice Form)</td>
<td>(Choice Form)</td>
<td>(Choice Form)</td>
<td></td>
</tr>
<tr>
<td>Deer</td>
<td>B</td>
<td>5.2(3.1)</td>
<td>5.1(3.5)</td>
<td>4.8(2.2)</td>
<td>-0.39***</td>
</tr>
<tr>
<td></td>
<td>W</td>
<td>6.4(3.0)</td>
<td>6.7(3.1)</td>
<td>7.2(3.4)</td>
<td>-0.49**</td>
</tr>
<tr>
<td></td>
<td>d</td>
<td>-0.39***</td>
<td>-0.49**</td>
<td></td>
<td>-0.71</td>
</tr>
<tr>
<td>Paintings</td>
<td>B</td>
<td>3.8(2.9)</td>
<td>3.6(2.6)</td>
<td></td>
<td>n.a.</td>
</tr>
<tr>
<td></td>
<td>W</td>
<td>4.3(2.5)</td>
<td>4.9(2.9)</td>
<td>5.5(2.0)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>d</td>
<td>-0.20</td>
<td>-0.44**</td>
<td></td>
<td>n.a.</td>
</tr>
<tr>
<td>Beaches</td>
<td>B</td>
<td>4.7(3.1)</td>
<td></td>
<td></td>
<td>4.1(3.4)</td>
</tr>
<tr>
<td></td>
<td>W</td>
<td>5.4(3.0)</td>
<td></td>
<td></td>
<td>6.1(3.5)</td>
</tr>
<tr>
<td></td>
<td>d</td>
<td>-0.23*</td>
<td></td>
<td></td>
<td>-0.55*</td>
</tr>
<tr>
<td>Fish</td>
<td>B</td>
<td>4.8(2.9)</td>
<td>5.7(3.5)</td>
<td>3.8(2.6)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>W</td>
<td>6.3(3.0)</td>
<td>6.5(3.1)</td>
<td>7.0(3.3)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>d</td>
<td>-0.49***</td>
<td>-0.24</td>
<td></td>
<td>-0.97</td>
</tr>
<tr>
<td>Books</td>
<td>B</td>
<td>4.3(2.7)</td>
<td>5.3(3.0)</td>
<td>4.8(3.0)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>W</td>
<td>5.8(3.1)</td>
<td>6.0(3.2)</td>
<td>6.7(3.5)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>d</td>
<td>-0.50***</td>
<td>-0.23</td>
<td></td>
<td>-0.54*</td>
</tr>
<tr>
<td>Police</td>
<td>B</td>
<td>4.8(2.9)</td>
<td></td>
<td></td>
<td>6.6(4.0)</td>
</tr>
<tr>
<td></td>
<td>W</td>
<td>6.1(3.3)</td>
<td></td>
<td></td>
<td>6.6(3.5)</td>
</tr>
<tr>
<td></td>
<td>d</td>
<td>-0.40***</td>
<td></td>
<td></td>
<td>0.02</td>
</tr>
</tbody>
</table>

d = mean of Black test takers minus mean of White test takers divided by the within group sd.

*p<.05, **p<.01, ***p<.001
Table 13

Mean Generating Explanations Item Scores by Ethnicity (Hispanic vs. White) and Choice Condition

<table>
<thead>
<tr>
<th>Item</th>
<th>Hispanic/White</th>
<th>Required (No-choice Form)</th>
<th>Required (Choice Form)</th>
<th>Choice (Choice Form)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deer</td>
<td>H</td>
<td>5.8(2.9)</td>
<td>6.3(3.7)</td>
<td>8.5(4.4)</td>
</tr>
<tr>
<td></td>
<td>W</td>
<td>6.4(3.0)</td>
<td>6.7(3.1)</td>
<td>7.2(3.4)</td>
</tr>
<tr>
<td></td>
<td>d</td>
<td>-.18*</td>
<td>-.13</td>
<td>.39</td>
</tr>
<tr>
<td>Paintings</td>
<td>H</td>
<td>4.1(2.2)</td>
<td>4.4(2.8)</td>
<td>3.8(1.0)</td>
</tr>
<tr>
<td></td>
<td>W</td>
<td>4.3(2.5)</td>
<td>4.9(2.9)</td>
<td>5.5(2.0)</td>
</tr>
<tr>
<td></td>
<td>d</td>
<td>-.09</td>
<td>-.15</td>
<td>-.89***</td>
</tr>
<tr>
<td>Beaches</td>
<td>H</td>
<td>5.2(3.1)</td>
<td>—</td>
<td>6.3(4.1)</td>
</tr>
<tr>
<td></td>
<td>W</td>
<td>5.4(3.0)</td>
<td>—</td>
<td>6.1(3.5)</td>
</tr>
<tr>
<td></td>
<td>d</td>
<td>-.04</td>
<td>—</td>
<td>-.07</td>
</tr>
<tr>
<td>Fish</td>
<td>H</td>
<td>5.7(2.8)</td>
<td>6.9(3.6)</td>
<td>6.1(3.6)</td>
</tr>
<tr>
<td></td>
<td>W</td>
<td>6.3(3.0)</td>
<td>6.5(3.1)</td>
<td>7.0(3.3)</td>
</tr>
<tr>
<td></td>
<td>d</td>
<td>-.20*</td>
<td>-.13</td>
<td>-.26</td>
</tr>
<tr>
<td>Books</td>
<td>H</td>
<td>5.5(3.3)</td>
<td>6.3(3.4)</td>
<td>5.3(3.8)</td>
</tr>
<tr>
<td></td>
<td>W</td>
<td>5.8(3.1)</td>
<td>6.0(3.2)</td>
<td>6.7(3.5)</td>
</tr>
<tr>
<td></td>
<td>d</td>
<td>-.11</td>
<td>.09</td>
<td>-.40</td>
</tr>
<tr>
<td>Police</td>
<td>H</td>
<td>6.0(3.6)</td>
<td>—</td>
<td>6.9(3.5)</td>
</tr>
<tr>
<td></td>
<td>W</td>
<td>6.1(3.3)</td>
<td>—</td>
<td>6.6(3.5)</td>
</tr>
<tr>
<td></td>
<td>d</td>
<td>-.03</td>
<td>—</td>
<td>.09</td>
</tr>
</tbody>
</table>

d = mean of Hispanic test takers minus mean of White test takers divided by the within group sd.

*p < .05, **p < .01, ***p < .001
The final subgroup analysis of GE items was by undergraduate major area. An analysis of variance revealed that, for all but two of the items (fish and police), there were significant differences among undergraduate major fields when examinees were required to answer on the no-choice test forms. As suggested by F ratios, the differences among groups were smaller (and nonsignificant) on the choice forms -- both when items were required and when they were optional. The lack of any significant differences among groups on the choice forms may have been, however, largely a function of our having deleted examinees who did not follow directions. A group's performance was more often higher when an item could be chosen instead of when it was required, but the patterns were not consistent from item to item.

**Correlations With Other Variables**

Correlations of GE item scores with GRE scores and with undergraduate grade averages are shown in Table 14. When examinees were required to answer items -- either on the choice forms or on the no-choice forms -- the correlations were modest, ranging from .05 to .22 on the no-choice forms and from .04 to .22 on the choice forms. Allowing examinee choice did not seem to have any dramatic effect on these correlations. (We note, however, that correlations were slightly higher for three of four items when examinees chose them than when they were required, but the reverse was true for the fourth item (paintings), for which we had comparatively little data and for which performance was considerably more restricted in range.)

It was also possible to examine, in a limited way, the effects of choice on the internal consistency of GE measures. This was accomplished by correlating the combined performance on the first two, required items (deer and paintings) with the third item (beaches), which was required on the no-choice form and optional on the choice form. This correlation was significantly higher ($p<.001$) when examinees were allowed to choose the third item instead of when it was required ($r = .71 \text{ vs. } .78$). For the other form of the test (in which fish and books were the required items and police the third item), the correlation was also slightly higher under the choice condition than in the required one ($r = .74 \text{ vs. } .77$), but not significantly different statistically. Thus there was, at the least, no evidence that allowing choice disturbed the internal consistency of the GE measures, and there was some indication that it may have improved them.

**Timing**

By analyzing the amount of time that examinees took to answer questions under choice and no-choice conditions, we hoped to provide corroborating evidence of the effects of choice. To this end, the amount of time that examinees took to answer each question was captured. Using the two choice forms, we compared times for the four items that were choosable in one form and required in the other. The results revealed that for three of the four items the average amount of time taken to answer the item was significantly less ($p < .01$) when the item was chosen than when it was required. For the fourth item, there was no significant difference. These results were rendered equivocal, however, by the fact that there appeared to be an
Table 14

Correlations of Generating Explanations Items with GRE Scores and Undergraduate Grade Point Average (by Choice Condition)

<table>
<thead>
<tr>
<th>Form</th>
<th>Item</th>
<th>Deer</th>
<th>Paintings</th>
<th>Beaches</th>
<th>Fish</th>
<th>Books</th>
<th>Police</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. (No choice)</td>
<td>GRE-V</td>
<td>.22</td>
<td>.18</td>
<td>.13</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>GRE-Q</td>
<td>.15</td>
<td>.09</td>
<td>.06</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>GRE-A</td>
<td>.19</td>
<td>.14</td>
<td>.09</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>UGPA</td>
<td>.07</td>
<td>.09</td>
<td>.09</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>2. (No choice)</td>
<td>GRE-V</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>.21</td>
<td>.22</td>
<td>.16</td>
</tr>
<tr>
<td></td>
<td>GRE-Q</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>.15</td>
<td>.12</td>
<td>.12</td>
</tr>
<tr>
<td></td>
<td>GRE-A</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>.16</td>
<td>.19</td>
<td>.15</td>
</tr>
<tr>
<td></td>
<td>UGPA</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>.05</td>
<td>.07</td>
<td>.09</td>
</tr>
<tr>
<td>3. (Choice)</td>
<td>Required</td>
<td>GRE-V</td>
<td>.22</td>
<td>.15</td>
<td>.10</td>
<td>.26</td>
<td>.29</td>
</tr>
<tr>
<td></td>
<td>Choice</td>
<td>GRE-Q</td>
<td>.21</td>
<td>.12</td>
<td>.06ns</td>
<td>.16</td>
<td>.22</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GRE-A</td>
<td>.19</td>
<td>.12</td>
<td>.05ns</td>
<td>.19</td>
<td>.28</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UGPA</td>
<td>.13</td>
<td>.10</td>
<td>.08ns</td>
<td>.13ns</td>
<td>.01ns</td>
</tr>
<tr>
<td>4. (Choice)</td>
<td>Choice</td>
<td>GRE-V</td>
<td>.24</td>
<td>.10ns</td>
<td>—</td>
<td>.19</td>
<td>.20</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GRE-Q</td>
<td>.24</td>
<td>-.04ns</td>
<td>—</td>
<td>.13</td>
<td>.15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GRE-A</td>
<td>.27</td>
<td>.11ns</td>
<td>—</td>
<td>.10</td>
<td>.15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UGPA</td>
<td>.20</td>
<td>-.27ns</td>
<td>—</td>
<td>.04ns</td>
<td>.04ns</td>
</tr>
</tbody>
</table>

Note. All correlations except those designated by n.s. are significant at the .05 level or beyond.
order effect: In each of the no-choice forms, examinees took progressively less time on each subsequent item — possibly, we speculate, because of either a practice effect or waning motivation. Because the choice items were always given last, the order of administration and choice condition were confounded and, therefore, no conclusions were possible from our data about the effects of choice on time needed to answer questions.

Discussion

Summary

The GE items that were the focus of our study varied dramatically with respect to the extent to which they were chosen by (attractive to) examinees. The implication of this finding is that examinees would probably welcome having a choice of items. Although they might appreciate having a choice of questions to answer, our results do not suggest that, as a group, examinees would make optimal choices. For instance, there was no consistent tendency for examinees to choose easier items over more difficult ones. In fact, except for one very unpopular, difficult item, there appeared to be a tendency to choose difficult items more often than easier ones.

There was also ample evidence in our data that items were differentially attractive to different subgroups of examinees, particularly according to undergraduate area of study. Test takers' choices were also associated with their ability levels, as indexed by performance on the GRE General Test, which measure largely unrelated constructs. In particular, there was some indication, as noted in another earlier effort (Powers et al., 1992), that the most able students were more likely to choose unpopular alternatives than were their less able counterparts, possibly because of a perceived greater challenge.

Whether or not every examinee made optimal choices, each item was, on average, somewhat easier when it was chosen instead of required. The effect varied somewhat from item to item, but on average was about .1-.2 sds — an effect that can be considered small but noteworthy by most standards.

Subgroups differed with respect to their average performances on GE items. Although there was very little difference in the scores of men and women on the items, average scores did differ by ethnicity. Hispanic test takers performed very nearly as well, on average, as White examinees, but Black examinees did slightly less well on average than their White counterparts. Performances also differed significantly by undergraduate major area of study.

The evidence was less clear with regard to the effects of allowing choice on subgroups of test takers. For men and women, the ability to choose their questions did little or nothing, on average, to change their (nearly equal) standing relative to one another. Because of smaller sample sizes and less stable estimates, the results for ethnic subgroup comparisons were less conclusive. Overall, differences in favor of White over Black test takers tended to be at least as
large when choice was offered as when it was not. For Hispanic examinees, the results were inconsistent, but, overall, choice did not seem to have any differential effect on their test performance relative to that of White test takers. The analysis of subgroups by undergraduate major was inconclusive.

Finally, there was some evidence that the reliability (internal consistency) of GE items may have increased when choice was permitted. This in turn may have accounted for the slightly, but not consistently, higher correlations of GE performance with indicators of academic ability (GRE scores and grades). To the extent that performance is more stable over items, choice may render the GE measure a better indicator of the construct of interest.

Implications

Perhaps the most consistent finding was that performance was higher, on average, for every item studied when examinees were allowed to choose an item from among a set than when they were required to answer an item. Apparently, when allowed a choice, at least a portion of examinees were able to avoid answering an item on which their performance would have been worse (or succeeded in choosing an item on which their performance would be better) than that of their fellow test takers, thus increasing the average performance of the group. A major implication of this finding stems from the apparent movement toward more complex assessments of performance. It appears that educational assessment in general, and the GRE Program in particular, is moving away from traditional multiple-choice formats to methods that are more performance based in nature (Aschbacher, 1991). The GRE Program, for instance, has invested relatively heavily in the development of measures that require examinees to construct responses, and most recently has endorsed the development of a GRE writing measure, which will require examinees to develop an extended response in the form of an essay. Many believe that these more complex assessments offer a number of advantages over conventional multiple-choice tests. Despite their advantages, however, these assessments often lack some of the desirable qualities (e.g., breadth of content coverage) that have made multiple-choice measures so prevalent.

One major drawback of many performance assessments is their limited generalizability from one task to another. That is, performance is often highly dependent on the particular task given: Typically, questions are not interchangeable. Moreover, because of their complexity, these assessments are usually rather time-consuming, thus restricting the number of tasks that can be posed. These limitations have been noted with some frequency (e.g., Bennett, 1993; Dunbar, Koretz, & Hoover, 1991; Linn, 1993) in a wide variety of contexts: writing assessments (Breland, Camp, Jones, Morris, & Rock, 1987), hands-on science exercises (Shavelson, Baxter, & Pine, 1992), performance tasks in a military setting (Shavelson, Mayberry, Li, & Webb, 1990), and clinical measures in medical licensure exams (Swanson, Norcini, & Grosso, 1987). Linn (1993) concluded that

... low levels of generalizability across tasks limit the validity of inferences about performances for a domain and pose serious problems regarding comparability and
fairness to individuals who are judged against performance standards on a small number, and perhaps different set, of tasks (p.13).

Limited generalizability is to some degree apparent also in the measures being developed for the GRE Program. In a recently reported GRE Board-sponsored study, for example, Bennett and Rock (1995) found that, for each of two different formats of a test of the ability to formulate hypotheses (the kind of item we have referred to as "generating explanations" in our current study), a significant portion of variation (22% to 24%) was due to a person-by-question interaction. In practical terms, this means that an examinee's performance depends to a practically significant degree on the particular items administered. To the extent that examinee choice enables at least some examinees to sidestep questions (or content) that are, by chance, outside the realm of their experience, examinee choice may be a legitimate testing tool. Vigilance will be needed, however, to ensure that choice does not inadvertently disadvantage some test takers.

Conclusion

This study has, we believe, added modestly to the available information about the potential effects of allowing examinee choice. It is clear, however, that considerably more work will be needed to understand (a) the role that examinee choice might play in standardized testing and (b) the effects it may have on examinees and on the psychometric quality of the tests they take. Further explication will be needed to sort out the circumstances for which choice may enhance the psychometric properties of a test as well as the conditions under which choice might instead constitute an irrelevant source of variation, and possibly unfairness, in test performance.

Next steps in research on examinee choice might be (a) investigations involving a variety of different sorts of tests, (b) studies varying the degree or latitude of examinee choice, (c) documentation of the bases on which examinees make choices (and validation of the extent to which these factors facilitate sound choices), and (d) further effort to assess the effects of examinee choice on test validity (when a variety of informative criteria are available).
References


Burton, N. (1993, August 3). Notes from our session on test standardization. (Unpublished memorandum.)


