GRE®
RESEARCH

Psychometric Evaluation of the New GRE® Writing Assessment

Gary A. Schaeffer
Jacqueline B. Briel
Mary E. Fowles

April 2001

GRE Board Professional Report No. 96-11P

ETS Research Report 01-08

Princeton, NJ 08541

Psychometric Evaluation of the New GRE® Writing Assessment

Gary A. Schaeffer
CTB/McGraw-Hill

Jacqueline B. Briel and Mary E. Fowles
Educational Testing Service

*with an **appendix** by*

Gwyneth Boodoo
Henry Braun
Charles Lewis
Dorothy Thayer
Educational Testing Service

GRE Board Report No. 96-11P

April 2001

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

This project was conducted while the first author was at Educational Testing Service.

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
Princeton, NJ 08541

Abstract

This study analyzed examinee responses to two essay prompts being considered for the Graduate Record Examinations (GRE®) Writing Assessment: "Present your Perspective on an Issue" and "Analyze an Argument." Forty prompts (20 issue and 20 argument) were administered to over 2,300 students at 26 participating U.S. colleges and universities. Each student wrote two essays in response to either two issue prompts, two argument prompts, or one of each. Results show that the issue and argument writing tasks appear to assess relatively similar constructs, supporting the decision to include both types of prompts in the operational GRE Writing Assessment and to report a single, combined score. The results also support the random administration of prompts without equating adjustments, because within each task type, most of the prompts were comparable in difficulty and no important subgroup interactions with prompt classifications were detected. Finally, from a fairness perspective, the results show there were advantages to administering the issue prompt first and the argument prompt second. The GRE Program used the psychometric information provided by this study to make final design, delivery, and scoring decisions before introducing the operational assessment in the fall of 1999.

Key words:

Writing constructs
Higher education writing assessment
Relative difficulty of essay prompts
Order effects of prompts
Subgroup performance

Acknowledgements

The authors greatly appreciate the guidance provided by the Graduate Record Examinations (GRE®) Board, the GRE Research Committee, the GRE Technical Advisory Committee, and the GRE Writing Advisory Committee in defining this study and interpreting its results. We also wish to acknowledge the invaluable role played by our ETS colleagues. In particular, we wish to thank researchers Henry Braun, Don Powers, and Larry Stricker for their in-depth reviews; program administrators Robin Durso and Kathleen O'Neill for consultation in design and analyses; statisticians Feng Yu and Rosemary Reshetar for analyzing study data; test developers Kelly Boyles, Peter Cooper, and Cynthia Welsh for their help in classifying GRE essay questions, selecting the questions used in the study, and participating in essay-scoring sessions with the college faculty who served as GRE readers for this project; and administrative assistant Cynthia McAllister, who helped prepare the report.

Table of Contents

List of Tables

List of Figures

Introduction

The present study was designed to collect psychometric information about the new Graduate Record Examinations (GRE®) Writing Assessment before it was introduced in the fall of 1999. Two prompt types were evaluated: "Present Your Perspective on an Issue" (issue) and "Analyze an Argument" (argument). Participants wrote essay responses to two prompts -- either two issue prompts, two argument prompts, or one of each type. The results of this study helped the GRE Program finalize decisions about the structure of the GRE Writing Assessment and document its psychometric properties. Several major areas of concern were investigated.

Major Areas of Investigation

The four major areas of investigation for the present study were prompt difficulty, order effects, score distributions, and relationships between prompt scores. Other possible areas of investigation -- such as timing, choice of prompts, effects of predisclosing prompts, and test validity -- were not investigated in this study; they have been the subject of other GRE research studies, such as those by Powers and Fowles (for instance, 1997a & 1997b), and other studies.

Prompt difficulty. Examinees' scores should not, on average, be dependent on the relative difficulty levels of particular prompts administered within each type of writing task -- that is, examinees' scores on particular prompts should be representative of the scores that would have been obtained on any other prompt of the same type. In the operational assessment, it would not be feasible to continuously conduct equating studies to adjust for differences in prompt difficulty. It was imperative, therefore, to establish -- before the test became operational -- that prompts used in the GRE Writing Assessment are approximately equal in difficulty. Whereas the examination of prompt difficulty focused on the mean scores obtained for each prompt, the consistency of prompt scores was also investigated by examining the similarity of examinees' scores on two prompts of the same type (issue-issue or argument-argument).

Order effects. Scores should not be affected by the order in which the issue and argument prompts are administered. If, for example, a subgroup of examinees received higher total scores with the issue-argument order than with the argument-issue order, then this information would

influence the decision about the order in which the two types of prompts appear in the operational assessment.

Score distributions. Because the development of the GRE Writing Assessment has been guided by principles of fairness and equity, any differences in score distributions between gender or racial/ethnic subgroups on issue or argument prompts could affect the final content of the assessment. For example, if either the issue or the argument prompts resulted in much larger than expected mean score differences between two specified racial/ethnic groups, then for validity and fairness reasons, serious consideration would be given as to whether that prompt type could be used in the assessment.

Relationships between prompt scores. The magnitude of the relationship between issue and argument scores would affect whether two writing scores or a single combined score would be reported. A strong relationship between issue and argument scores would support reporting a single combined score. A weak relationship would suggest that the two prompt types were not measuring similar constructs, and that separate scores may need to be reported (assuming that the reliability of the individual scores was sufficient).

Prompt Types, Prompt Development

The GRE "Present Your Perspective on an Issue" prompt provides a brief quotation that makes a claim about an issue of general interest. Examinees are asked to write an essay that considers the complexity of the issue and develops their own perspective on the issue in a clear and logically compelling way. The time limit for this writing task is 45 minutes. (Student directions, a sample prompt, and the GRE scoring guide for the "Present Your Perspective on an Issue" writing task are provided in Appendix A.)

The "Analyze an Argument" task is more focused and specific. It presents an argument and then asks examinees to write an essay that addresses this question: "How logically convincing do you find the argument?" The time limit for this writing task is 30 minutes. (Student directions, a sample prompt, and the GRE scoring guide for the "Analyze an Argument" writing task are also provided in Appendix A.)

Because prompt content must be accessible to a testing population that is highly diverse with respect to cultural and academic background, GRE writing prompts are not content specific. The issues can be discussed from a variety of perspectives, and the arguments are embedded in generally familiar situations or scenarios. Together, the GRE issue and argument analytical writing tasks assess the examinee's ability to:

- articulate complex ideas clearly and effectively
- examine claims and accompanying evidence
- support ideas with relevant reasons and examples
- sustain a well focused, coherent discussion
- control the elements of standard, written English

University faculty and Educational Testing Service (ETS®) writing specialists wrote the GRE issue and argument prompts according to content specifications and classification schemes defined by the GRE Writing Advisory Committee.[1] Then, the prompts underwent a careful review process to help ensure that they would elicit strong evidence of analytical writing, that the content would be appropriate for all cultural groups, and that the language would be clear for all test takers -- including those for whom English is a second language. In addition, each prompt was classified according to specified characteristics, such as general subject matter, structural characteristics, specific reasoning features, and conceptual or pragmatic orientation.

Once the preliminary review process was completed, the GRE Program field tested hundreds of these newly written prompts by administering them to at least 100 examinees per prompt in a voluntary, identified research section of the operational GRE computer-based test. Examinees were asked to write on a single prompt, either issue or argument. College faculty and ETS writing assessment specialists read responses to the field-tested prompts and evaluated each prompt according to the GRE "Criteria for Approving Essay Prompts for Operational Use."[2]

[1] This committee -- comprised of representatives from different academic disciplines, colleges and universities, and ethnic groups -- guided all stages of the development of the GRE Writing Assessment.
[2] Readers evaluated the prompts according to fairness (e.g., Did examinees understand the vocabulary and the ideas presented? Does the topic discriminate on the basis of thinking and writing skills rather than content knowledge or cultural experience?), complexity (e.g., Does the prompt invite complexity of thought and a variety of cognitive

Prompts that survived field test analyses became available for operational use. However, because of limitations inherent in using the optional writing research section for data collection (including time restrictions that necessitated asking examinees to respond to only one prompt and the nonsystematic nature of voluntary participation), these data could provide only limited information about the psychometric properties of the test. Thus, the current study was designed so that each participant would respond to two prompts drawn from a subset of prompts that had survived the rigors of field testing.

Design, Recruitment, and Scoring

Study Design

A total of 40 GRE prompts (20 issue and 20 argument) were examined in the present study. Because data from the study prompts would be used to generalize to the total GRE pool of approximately 250 operational prompts, the study prompts were selected to be, as much as possible, representative of all of the writing prompts. For example, major classifications were proportionally represented in the study and, based on field-test data, prompts with relatively high, middle, and low means and standard deviations were included in the study.

The design used to administer two prompts to each participant is shown in Figure 1. The first letter (I or A) of each row and column heading in Figure 1 indicates the type of prompt that was administered -- issue or argument. The second letter (a-j, k-t) of each heading represents the sequence number for the 20 prompts of each type (though the prompts are listed in no particular order). Each row heading represents a prompt that was administered first, and each column heading represents a prompt that was administered second. Thus, cells labeled 'X' indicate the pairs of prompts that were administered, as well as the order in which they were administered. Empty cells in Figure 1 indicate pairs of prompts that were not administered together. Prompts were administered to examinees as follows:

---

approaches?), and scoring (e.g., Were the readers' expectations for this prompt more lenient or strict than for other GRE prompts of the same type? Are scores distributed reasonably across the GRE scale?).

One fourth wrote on two issue prompts.

- One fourth wrote on two argument prompts.
- One fourth wrote on an issue and then an argument prompt.
- One fourth wrote on an argument and then an issue prompt.

Participants were randomly assigned pairs of prompts to ensure the random equivalence of groups. Furthermore, the design was counterbalanced so that, for each pair of prompts, the two orders were administered to the same number of participants. Each prompt was administered in eight positions: four times in the first position and four times in the second position. Each prompt appeared with two other prompts of the same type (either issue or argument) and with two other prompts of the alternate type. The plan was to recruit a total of 4,000 participants, 25 for each prompt pair represented by an 'X' cell.

The entire (approximately two-hour) testing session was administered on-screen. It began with a tutorial showing how to use the ETS word processor to compose essays. Participants then completed a brief background information questionnaire. Next, they wrote their two essays using the word processor and then answered a brief questionnaire that asked about the writing tasks they had completed.

Recruitment of Participants

Volunteers were recruited nationally from the 26 colleges and universities listed in Figure 2. Several campuses were selected because of their large minority population; thus the study was able to over-sample minority students. Participants were college undergraduates planning to take the operational GRE within a year of the study. Testing occurred from September through December 1997.

The initial compensation plan called for a total student payment of $15 for a two-hour testing commitment. As part of a related validity study, an additional $10 would also be paid to students who submitted two samples of writing from their coursework. Because the initial student response was low, the compensation was increased two weeks after the start of the study

in an effort to boost participation. The new payment of $40 plus the $10 for submitting the two course-related writing samples increased the number of participants. In addition, colleges and universities were offered summary data showing their students' performance on the writing prompts if they requested this information from ETS. The recruitment campaign included the following activities:

- distribution of flyers to students and departments
- distribution of posters across campus
- provision of press release for college papers
- provision of news release for college radio stations
- active on-campus recruitment by Test Center Administrators
- distribution of flyers to local chapters of Society of Hispanic Professional Engineers
- distribution of flyers to local chapters of National Society of Black Engineers

Scoring

Eighteen readers participated in this study. All were college faculty with a special interest and expertise in writing, and all had completed GRE reader training. The following procedures were used to score the study essays:

1. Each response was evaluated independently by two readers who assigned a holistic score on a six-point (1-6) scale, with 6 as the highest score. (GRE essay scoring guides are provided in Appendix A.)
2. Throughout the scoring process, no identifying information appeared on the essays, and the second reader did not know the first reader's score.
3. If the two readers' scores were identical or adjacent, the scores were averaged for a final task (issue or argument) score.
4. Essays with discrepant scores -- that is, scores greater than one point apart -- were adjudicated by a third reader, and the two closest scores were averaged for a final task score.

Analyses and Results

Participants who did not appear to be exerting appropriate effort on both prompts were excluded from the analysis sample. "Inappropriate effort" was defined as providing fewer than 500 bytes of text for each prompt -- the minimum amount of text that readers need to be able to score an essay response. On this criterion, approximately 12% of all participants were excluded from the analysis sample. Exclusion was fairly evenly distributed across all 40 prompts, indicating that there were no particular prompts to which participants were choosing not to respond with sufficient effort.

Table 1 shows the distribution of background characteristics of the total analysis sample and of groups defined by the order in which they were given the prompts. For comparison, background characteristics of examinees who took the operational GRE General Test during 1995-96 are also provided. A total of 2,326 study participants were included in the analysis sample. About 58% of participants were female, about 20% were African American, 20% were Asian, and 48% were White.

As expected, distributions of the participants' background variables for the four prompt-order groups were very similar because of the random assignment of participants to these groups. Compared with the operational 1995-96 GRE General Test population, the study group had a similar gender distribution, proportionally more African American and Asian participants, somewhat more participants for whom English was not their best language, slightly lower percentages of social science and humanities majors, and somewhat higher overall grade point averages. Except for greater minority group representation, the study participants were fairly similar to operational GRE examinees.

Because the actual total number of participants was well below the target of 4,000, not all planned interactions between participant background variables, prompt classifications, and prompt difficulty could be thoroughly investigated. The average number of participants per "X" cell in Table 1 was intended to be 25 participants, but it turned out to be about 15 participants. Nevertheless, a number of psychometric issues could be addressed.

Interreader agreement. Essay responses to the same prompts were grouped into folders (10 essays per folder) that were then randomly assigned to readers. Each essay was read by two readers; as noted earlier, a third reader adjudicated if scores from the first two readers differed by more than one point. For both issue and argument prompts, readers agreed -- that is, they did not differ by more than one score point -- about 98% of the time. Cohen's kappa (1960; exact or within 1 point) was computed for both issue and argument prompts in both first and second positions. The corresponding values were as follows: issue-first position (.95), issue-second position (.95), argument-first position (.97), and argument-second position (.96). According to the categorization developed by Landis and Koch (1977), these values fall within the highest ("almost perfect") category, which ranges from .81 to 1.00. The low reader disagreement was spread out over a number of prompts, indicating that no particular prompts led to rater disagreement.

Prompt difficulty. The prompts were randomly assigned to randomly equivalent participant groups. By design, the differences in score distributions among prompts could be attributed to the prompts themselves, not to the sample of participants who responded to a particular prompt. Table 2 and Table 3 show the mean scores for each issue prompt and each argument prompt, respectively. Three sets of means are provided in each table.

In Table 2, the first column identifies each issue prompt and the second column, "Issue given first," shows the means of these prompts when the issue prompt was given first -- regardless of whether it was followed by an argument prompt or another issue prompt. All prompts in the table are rank-ordered (I-01 to I-20) by the means in this second column. Because these prompts were administered first, their means are not influenced by any practice or fatigue effects. The remaining two columns show mean scores on the issue prompts when the issue prompt was given second (either after an issue prompt or after an argument prompt).

In the "Issue given first" column of Table 2, the range of prompt means is 1.0 (4.3-3.3). This range represents about a one-standard-deviation difference. An analysis of variance of all 20 issue prompts ($p = .0001$) indicated that there is a significant difference in difficulty among the 20 prompts. However, Duncan's Multiple Range Test (1955) showed that the 14 prompts in the middle range of difficulty did not differ significantly from each other at the 0.05 significance

level. The Duncan test also showed that, compared to the middle 14 prompts, the two easiest issue prompts (I-01 and I-02) and the four most difficult (I-17 through I-20) were significantly different in difficulty at the .05 level. An analysis of variance indicates that the middle 14 prompts did not differ significantly from each other at the 0.05 significance level.

The mean scores for argument prompts, shown in Table 3, indicate a pattern similar to that for the issue prompts: There is almost a one-point difference between the easiest and most difficult prompt. This range represents about a one-standard-deviation difference. An analysis of variance of all 20 argument prompts ($p = .0001$) indicates that there is a significant difference in difficulty among the 20 prompts. However, Duncan's Multiple Range Test (1955) showed that the 14 prompts in the middle range of difficulty did not differ significantly from each other at the 0.05 significance level. The Duncan test also showed that, compared to the middle 14 prompts , the three easiest argument prompts (A-01 to A-03) and the three most difficult (A-18 to A-20) were significantly different in difficulty at the .05 level.

A possible relationship between the prompt classifications and prompt difficulty was also investigated. Content experts carefully examined prompts at the extreme difficulty levels to ascertain whether there were any similarities in content, structure, or other features that would differentiate them from the other prompts. These investigations yielded no consistent explanations as to why these particular prompts proved more difficult (or easier) than the other prompts. The GRE Writing Advisory Committee also reviewed these prompts as part of a larger pool and did not identify them as potentially problematic. Moreover, the differences in prompt difficulty were not even consistent within the study: Different prompts were often at the high or low difficulty levels, depending on the administration order (see Table 2 and Table 3). Thus, the outlier prompts may have been statistical artifacts that resulted from small samples. The GRE Program will monitor the performance of all operational prompts and eliminate any that seem especially easy or difficult in relation to the total pool. (In addition, before the writing assessment became operational, the GRE Program implemented a test administration plan designed to mitigate possible differences in prompt difficulty.) [3]

[3] Based on field-test data and score data from this study, all operational GRE prompts were classified as either "relatively easy" or "relatively difficult." Now, during a testing session, examinees who choose a more difficult issue

The magnitudes of the differences among prompt mean scores were considered to be sufficiently small so that equating adjustments would not be necessary to make the scores interchangeable across prompts. Furthermore, there was no apparent relationship between prompt difficulty and prompt classifications. Although sample sizes were rather small, no apparent interactions between prompt difficulty and gender or racial/ethnic group membership were detected. These results suggest that randomly assigning prompts to examinees would be a fair method of prompt assignment in operational testing. At the request of the GRE Board, further analyses of these data were conducted; these results are reported in Appendix B.

Score distributions. Table 4 shows score distributions for subgroups of participants for six possible assessment formats: two issue prompts, one issue then one argument prompt, one argument then one issue prompt, two argument prompts, one issue prompt only, and one argument prompt only. The percentages of participants who received relatively high scores (i.e., an averaged score greater than or equal to 4 on the 6-point scale) were greater for issue only than for argument only prompts. More African American and Asian participants received high scores in the issue-argument order than in the argument-issue order. A chi-square test was conducted to compare the score distributions of the two orders (I-A and A-I) for both Asian and African American groups. Results confirm that there was a significant difference at the .05 level in terms of score distributions (for the Asian group: $x^2 = 7.93$, df = 3 [p-0.04]; for the African American group, $x^2 = 11.34$, df = 3 [p = 0.01]). Note that these distributions apply to participants who reported that English is their best language.

Relationship between prompt scores. Table 5 shows correlations between prompt scores for the four conditions of prompt administration for the total group and for selected subgroups. The highest correlations (.52 to .67) were observed when examinees wrote on two issue prompts. The lowest correlations (.33, .34) occurred when certain subgroups wrote on two different types of prompts (either issue-argument or argument-issue).

For the total group, the observed correlation between the two prompts in the issue-argument order was 0.54; the observed correlation of the two prompts in the argument-issue

---

prompt receive an easier argument prompt. Conversely, those who choose one of the easier issue prompts receive a more difficult argument prompt. When appropriate, prompts are reclassified on the basis of operational data.

order was 0.46. The magnitude of these correlations, when compared with the correlations between the same prompt types (0.62 for issue-issue and 0.51 for argument-argument; also see Table 6), suggests that, for the group of examinees as a whole, the two prompt types measure relatively similar writing constructs. In addition, the patterns of the correlations across the four conditions of prompt administration are generally similar for the subgroups.

Score consistency. One approach to assessing the consistency of test scores with these data is to estimate reliability based on the correlations of scores on alternate test designs. Table 6 presents means, standard deviations, reliability estimates, and standard errors of measurement for four different writing tests: issue then argument, argument then issue, one issue prompt only, and one argument prompt only. The reliability of a writing score based on one issue prompt was estimated to be equal to the correlation between the scores on two issue prompts -- that is, it is equal to the correlation between two alternate designs for the GRE writing assessment. The correlation of scores on any two issue prompts administered in the first position is 0.62, which is reported as the alternate-forms reliability of one issue prompt. Similarly, the alternate-forms reliability of a writing score based on one argument prompt is equal to the correlation between the scores on two first-position argument prompts (0.51).

Score reliabilities from the issue-argument and argument-issue prompt-orders were estimated by treating them as composite scores and employing a version of the Spearman-Brown formula that assumes that the different tasks are congeneric (Feldt & Brennan, 1989; equation 11). The standard errors of measurement were computed using the traditional classical test theory formula: $\text{sem} = \text{sd}\,(1 - \text{reliab})^{1/2}$. The estimated reliability was higher in the issue-argument order than in the argument-issue order (0.70 vs. 0.63), and the standard errors of measurement were not very different for the two orders (0.49 vs. 0.52).

If the assessment is sufficiently reliable, then it should not matter which particular prompts are administered to individuals; very similar score distributions should be obtained regardless of which prompts are administered. Table 7 and Table 8 show cumulative percents of difference scores for participants who took two issue prompts and two argument prompts, respectively. For the total group and for most subgroups, about 84% of participants had difference scores for two issue prompts that were less than or equal to one point, and about 82%

had difference scores for two argument prompts that were less than or equal to one point. (An exception was that about only 75% of Hispanic participants had difference scores for two issue prompts that were less than or equal to one point; however, there were only 44 Hispanic participants in the issue-issue condition.)

About 98% of participants had difference scores less than or equal to two points (on a 1-to-6 scale) for both the two-issue and two-argument conditions. These results indicate that for the vast majority of participants, there was not a big difference in their scores on two prompts of the same type. Note also that these data were collected in the same session, so practice, fatigue, and/or motivation effects also could have differentially impacted scores on the two prompts. Thus, these results may underestimate the true degree of similarity of scores on two prompts of the same type.

Table 7.1 and Table 8.1 show that, compared to the argument prompts, the issue prompts are considerably easier for all groups. Mean issue scores range from 3.2 to 4.1, whereas mean argument scores range from 2.9 to only 3.7. Also, correlations among issue prompts are considerably higher (.52 to .67) than correlations among argument prompts (.36 to .56). These results apply consistently across study groups, regardless of whether the issue or argument prompts were administered in the first or second position. To some extent, the differences between issue and argument means may be due to examinees' relative unfamiliarity with the argument writing task. When the GRE Writing Assessment becomes operational, and examinees have had a chance to obtain GRE test preparation materials and review specific strategies for responding to each type of prompt, the differences may diminish.[4]

Table 9 and Table 10 show distributions of difference scores for participants who wrote essays in the issue-argument order and argument-issue order, respectively. The difference score was computed as the score on the second essay minus the score on the first essay. In Table 9, then, a positive difference indicates that the issue score was higher for the subgroup than the

[4] This was the case for the Analytical Writing Assessment of the Graduate Management Admission Test (GMAT®) program, which administers similar types of prompts. As GMAT examinees became more familiar with the argument item type, GMAT issue and argument means moved closer together -- although the mean score for GMAT issue prompts is still somewhat higher than the mean for GMAT argument prompts.

argument score; conversely, a negative difference indicates that the argument score was higher. In Table 10, however, a positive difference indicates that the argument score was higher for the subgroup than the issue score, while a positive difference indicates that the issue score was higher.

More participants received higher issue scores than argument scores in both administration orders. In both conditions, however, Hispanic participants received relatively higher scores on the first prompt than they did on the second prompt. Overall, the distribution of difference scores for gender and racial/ethnic subgroups were similar, suggesting that there was not an interaction between prompt type and these background variables. (Further analyses of these data are provided in Appendix B.)

Score differences among subgroups. The analysis provided in Table 11 is limited to examinees for whom English is their best language. The table shows standardized, total-score mean differences among selected subgroups for the four study conditions, as well as standardized scores for all issue prompts administered in the first position and for all argument prompts administered in the first position -- regardless of whether the first-position prompt was followed by an argument prompt or an issue prompt. The numbers in parentheses are the p-values derived from t-tests between two corresponding groups. For example, women scored slightly higher than men in all six presentation designs, but only the issue-issue design was statistically significant. The standardized differences between African American participants and White participants, and between Hispanic participants and White participants, were smaller than the differences found between these groups on other GRE measures[5]. In general, the standardized mean differences were smaller when an issue prompt was administered first and larger when an argument prompt was given first.

Standardized score differences for gender and racial/ethnic subgroups across a number of other writing assessments are summarized by Breland, Bridgeman, and Fowles (1999). As expected, the magnitude of the differences observed in the current study are similar to those found in similar tests, such as the Analytical Writing Assessment of the Graduate Management

[5] See, for example, Sex, Race, Ethnicity, and Performance on the GRE General Test, 1998-99 (Graduate Record Examinations Board, 1998).

Admission Test. The lack of consistency across all tests is likely due to a variety of factors, such as the tasks themselves, the reliability of the tests, the characteristics and size of test examinee groups, the motivation of test takers, and scoring methods. When the GRE Writing Assessment becomes operational, subgroup score differences will continue to be closely monitored.

Table 12 presents score summary statistics by undergraduate major field. Mean scores for humanities majors were consistently higher than mean scores for other undergraduate majors. This finding is not surprising, given the amount of writing that is generally required of humanities majors. No major field consistently received the lowest mean score.

Order effects. The scores of participants who were administered one issue prompt and then one argument prompt suggest that an order effect was present. As noted above, the differences between White and minority-group mean scores were smaller when the two prompts were administered in the issue-argument order than in the argument-issue order. As also indicated earlier, the correlation of issue and argument scores is higher in the issue-argument order, and the estimated reliability is somewhat higher in the issue-argument order. There is no clear explanation for the apparent order effect. However, these subgroup and score consistency results suggest that it would be beneficial to operational examinees to administer the prompts in the issue-argument order (which the GRE Program decided to do).

Score reporting. With two different prompt types included in the operational test, there were two methods of score reporting to consider: Report two separate scores (issue and argument) or report one combined score. The content and structure of the tasks, as well as correlations between issue and argument scores, suggest that the two prompt types assess similar constructs. Based in part on this information, the GRE Program decided to report a single score that is the average of the two prompt scores rounded up to the nearest 0.5 level for the operational test.

Essay questionnaire data. A questionnaire asking participants about the writing assessment was administered at the end of the testing session. Table 13 shows the questions asked as well as percentages of responses for the total group and for the four study conditions.

Almost all participants responded. No subgroup data are presented here because there were no meaningful variations by subgroup.

According to the questionnaire data, participants were fairly positive about some aspects of the writing tasks. The vast majority of participants indicated that they thought the topics were at least somewhat interesting, that their performance was at least somewhat indicative of their writing ability, and that they were comfortable using the computer to compose their essays.

Respondents were less positive about other aspects of the tasks. About 20% of the participants felt rushed, and nearly 10% said that they did not finish. Timing may be less of a problem on the operational test for several reasons: Examinees will be aware of time limits well before the test session and will have had the opportunity to practice under timed conditions before taking the test. Also, the time spent becoming familiar with the task directions and prompts should be reduced, since examinees will have access to all of this material before the testing session. In addition, GRE readers are trained not to penalize an essay merely because it is not quite finished; an unfinished essay that is otherwise very well developed, well organized, and well written can earn the highest score.

About one-third of the participants believed that their essays would have been of higher quality if they had handwritten them. However, for the operational GRE Writing Assessment, examinees can use test preparation software to practice using the word processor. They also have the option of handwriting their essays.

On some questions, the study examinees had different opinions about issue and argument prompts. For example, a chi-square test based on all the responses of study examinees who answered both types of prompts (issue and argument, regardless of order) showed that the study participants considered the issue topics significantly more interesting than the argument topics ($p = .0001$). They also believed that more specific knowledge was required for responding to the issue prompts, a result that supports the definition of the writing constructs assessed by the two tasks. For the issue task, examinees must draw on their own experiences, observations, and readings to provide the content (i.e., reasons, examples, details) to support their views on the issue. Because there are no "right answers," the content varies according to the examinees' own

experiences and interests. For the argument task, the content is provided in the prompt itself. The examinee's job is to analyze the line of reasoning in the argument, not to discuss the content of the argument. (As noted earlier, instructions and essay scoring guides for the two GRE writing tasks are provided in Appendix A.)

Conclusions

This study provided psychometric information to help inform decisions about the design, delivery, and scoring of the new GRE Writing Assessment. Within each task type -- argument and issue -- prompts were found to be reasonably similar in difficulty. No subgroup interactions with prompt difficulty and prompt type were identified, and the issue and argument prompts appear to assess relatively similar writing constructs. The results of this study support administering the tasks in the issue-argument order, selecting prompts for individual examinees at random from a large pool, and reporting a single score based on the examinee's average performance on the two prompts.

Beyond this study, the GRE Program is conducting predictive validity studies to determine the relationship between GRE writing scores and other graduate school indicators of writing performance. Further analyses of the issue and argument prompts will be carried out when sufficient operational data become available.

References

Breland, H., Bridgeman, B., and Fowles, M. (1999). Writing assessment in admission to higher education: Review and framework (College Board Report No. 99-3; GRE Research Report No. 96-12R). New York: College Entrance Examination Board.

Cohen, J. A. (1960). Coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20, 37-46.

Duncan, D. B. (1955). Multiple-range and multiple-F tests. Biometrics, 11, 1-42.

Feldt, L. S., and Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), Educational measurement (3rd ed., p.112). Washington, DC: American Council on Education.

Graduate Record Examinations Board. (1998). Sex, race, ethnicity, and performance on the GRE General Test, 1998-99. Princeton, NJ: Educational Testing Service.

Landis, J. D., and Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33, 159-174.

Powers, D., and Fowles, M. (1997a). Effects of applying different time limits to a proposed GRE writing test (GRE Board Report No. 93-26cR; ETS Research Report No. 96-28). Princeton, NJ: Educational Testing Service.

Powers, D., and Fowles, M. (1997b). Effects of disclosing essay topics for a new GRE writing test (GRE Board Report No. 93-26R; ETS Research Report No. 96-26). Princeton, NJ: Educational Testing Service.

Powers, D., Fowles, M., & Boyles, K. (1996). Validating a writing test for graduate admissions (GRE Board Report No. 93-26b; ETS Research Report No. 96-27). Princeton, NJ: Educational Testing Service.

Tukey, J. W. (1977). Exploratory data analysis. Reading, MA: Addison-Wesley.

Figures

	Ia	Ib	Ic	Id	Ie	If	Ig	Ih	Ii	Ij	Aa	Ab	Ac	Ad	Ae	Af	Ag	Ah	Ai	Aj
Ia		X								X	X									X
Ib	X		X								X	X								
Ic		X		X								X	X							
Id			X		X								X	X						
Ie				X		X								X	X					
If					X		X								X	X				
Ig						X		X								X	X			
Ih							X		X								X	X		
Ii								X		X								X	X	
Ij	X								X										X	X
Aa	X	X										X								X
Ab		X	X								X		X							
Ac			X	X								X		X						
Ad				X	X								X		X					
Ae					X	X								X		X				
Af						X	X								X		X			
Ag							X	X								X		X		
Ah								X	X								X		X	
Ai									X	X								X		X
Aj	X									X	X								X	

Note. First letter 'I' indicates issue prompt, and first letter 'A' indicates argument prompt; second letter (a - j, k - t) identifies individual prompts. Prompts appearing as row headings were administered first; prompts appearing as column headings were administered second. Cells with an 'X' indicate pairs of prompts that were administered.

Figure 1. Prompt administration design.

	Ik	Il	Im	In	Io	Ip	Iq	Ir	Is	It	Ak	Al	Am	An	Ao	Ap	Aq	Ar	As	At
Ik		X								X	X									X
Il	X		X								X	X								
Im		X		X								X	X							
In			X		X								X	X						
Io				X		X								X	X					
Ip					X		X								X	X				
Iq						X		X								X	X			
Ir							X		X								X	X		
Is								X		X								X	X	
It	X								X										X	X
Ak	X	X										X								X
Al		X	X								X		X							
Am			X	X								X		X						
An				X	X								X		X					
Ao					X	X								X		X				
Ap						X	X								X		X			
Aq							X	X								X		X		
Ar								X	X								X		X	
As									X	X								X		X
At	X									X	X								X	

Note. First letter 'I' indicates issue prompt, and first letter 'A' indicates argument prompt; second letter (a - j, k - t) identifies individual prompts. Prompts appearing as row headings were administered first; prompts appearing as column headings were administered second. Cells with an 'X' indicate pairs of prompts that were administered.

Figure 1. Prompt administration design, continued.

Arizona State University
California State University -- San Bernardino
Clemson University
Fayetteville State University
Florida A&M
Florida International University
Howard University
New Jersey Institute of Technology
North Carolina A&T State University
North Carolina Central University
Rutgers University
San Diego State University
Savannah State University
State University of West Georgia
Texas A&M University at College Station
University of Florida at Gainesville
University of Houston -- Victoria
University of Maryland at College Park
University of Miami
University of Minnesota
University of Texas at Austin
University of Texas -- Pan American
University of Utah
University of Wisconsin at Oshkosh
Virginia Polytechnic Institute and State University
Winthrop University

Figure 2. Colleges and universities that participated in the study.

Tables

Table 1. Distribution of Background Characteristics of GRE General Population and Study Participants

	GRE 1995-96*	Study Total	Issue-issue	Issue-argument	Argument-issue	Argument-argument
GENDER						
Female	58	58	59	57	56	59
Male	42	42	41	43	44	41
ETHNICITY						
African American	8	20	21	19	19	20
Asian	4	20	21	19	21	19
Hispanic	5	7	7	7	6	8
White	79	48	46	49	49	47
Other	4	5	5	6	5	5
ENGLISH BEST						
Yes	94	83	82	83	85	81
No	6	17	18	17	15	19
INTENDED GRAD MAJOR						
Natural science	28	26	25	26	25	27
Engineering	11	14	16	14	13	12
Social science	20	14	14	15	12	13
Humanities	13	6	7	6	6	6
Education	7	8	7	9	6	8
Business	3	10	11	9	11	10
Other	8	13	10	14	13	13
No response	10	10	10	8	14	10
OVERALL GPA						
A	12	19	20	17	20	21
A-	23	28	28	25	30	27
B	35	30	30	31	26	32
B-	14	13	13	15	13	10
C or lower	8	9	8	11	10	9
No response	8	1	1	1	1	1
WRITING GPA						
A	**	26	28	26	26	24
A-	**	31	30	31	32	33
B	**	28	26	27	27	31
B-	**	7	7	9	7	6
C or lower	**	4	5	4	5	4
No response	**	4	4	3	5	3
TOTAL NUMBER	261,970	2,326	618	598	552	558

Note. The total number of study participants was 2,636, but 310 (12%) were excluded from the analyses because they did not have scores for two prompts.
* U.S. citizens only. ** Not available.

Table 2. Mean Scores on Issue Prompts

Prompt*	Issue given first	Issue-issue**	Argument-issue**
I-01	4.3 (50)	3.4 (31)	3.5 (32)
I-02	4.2 (60)	3.9 (33)	3.9 (27)
I-03	4.1 (54)	3.4 (26)	3.6 (26)
I-04	4.1 (64)	3.7 (37)	4.1 (26)
I-05	4.1 (59)	4.0 (23)	4.3 (22)
I-06	4.0 (56)	4.3 (26)	4.0 (34)
I-07	4.0 (57)	3.9 (34)	3.8 (27)
I-08	4.0 (63)	3.4 (35)	3.7 (34)
I-09	3.9 (67)	3.8 (37)	3.9 (25)
I-10	3.9 (57)	3.7 (24)	3.8 (36)
I-11	3.8 (65)	3.5 (22)	3.5 (32)
I-12	3.8 (64)	3.7 (43)	4.1 (35)
I-13	3.8 (62)	3.1 (30)	3.7 (30)
I-14	3.7 (54)	3.8 (33)	3.9 (22)
I-15	3.7 (59)	3.2 (31)	3.5 (31)
I-16	3.7 (68)	3.5 (35)	3.9 (23)
I-17	3.6 (71)	3.5 (29)	3.7 (23)
I-18	3.5 (66)	3.8 (28)	3.2 (23)
I-19	3.5 (68)	3.7 (31)	3.7 (24)
I-20	3.3 (52)	3.7 (30)	3.4 (20)
All prompts	3.8	3.6	3.8
Total n	1,216	618	552

Note. The average standard deviation of scores across prompts and within each condition was 1.0. The range of standard deviations within each condition was about 0.8-1.2. An analysis of variance on the issue prompts given first indicated that the means of 14 prompts did not differ from each other at the 0.05 significance level; prompts I-01 and I-02 had significantly higher means, and prompts I-17 through I-20 had significantly lower means.

* Prompts are rank ordered by the means of "Issue given first" prompts.

** Mean score for issue task administered second.

Table 3. Mean Scores on Argument Prompts

Prompt*	Argument given first	Argument-argument**	Issue-argument**
A-01	3.8 (60)	3.6 (32)	3.7 (27)
A-02	3.6 (58)	3.5 (28)	3.6 (32)
A-03	3.6 (60)	3.4 (24)	3.5 (39)
A-04	3.5 (46)	3.5 (23)	3.6 (38)
A-05	3.4 (55)	4.0 (32)	3.4 (33)
A-06	3.4 (49)	3.4 (29)	3.6 (27)
A-07	3.4 (56)	3.3 (29)	3.7 (32)
A-08	3.4 (50)	3.4 (28)	3.3 (28)
A-09	3.3 (59)	2.9 (26)	3.1 (22)
A-10	3.3 (56)	2.8 (21)	3.3 (32)
A-11	3.3 (63)	3.8 (33)	3.6 (40)
A-12	3.3 (57)	3.7 (20)	3.0 (38)
A-13	3.3 (49)	3.3 (27)	3.4 (22)
A-14	3.2 (61)	3.1 (32)	3.1 (24)
A-15	3.2 (56)	2.9 (30)	3.1 (28)
A-16	3.1 (60)	3.2 (32)	3.2 (25)
A-17	3.1 (45)	3.1 (29)	3.1 (26)
A-18	3.0 (48)	3.3 (29)	3.2 (35)
A-19	3.0 (53)	3.1 (27)	3.4 (18)
A-20	2.9 (69)	3.0 (27)	2.8 (32)
All prompts	3.3	3.3	3.4
Total n	1,110	558	598

Note. The average standard deviation of scores across prompts and within each condition was 1.0. The range of standard deviations within each condition was about 0.8-1.2. An analysis of variance on the argument prompts given first indicated that the means of the middle 14 prompts did not differ from each other at the 0.05 significance level; prompts A-01, A-02, and A-03 had significantly higher means, and prompts A-18, A-19, and A-20 had significantly lower means.

* Prompts are rank-ordered by the means of "Argument given first" prompts.

** Mean score for argument task administered second.

Table 4. Percentage of English-Best-Language Examinees at or Above Specified Score Levels

Subgroup	Score	Issue-issue	Issue-argument	Argument-issue	Argument-argument	Issue first*	Argument first*
Afr. Amer.	6	1	0	0	0	1	0
(103)	5	7	6	1	1	9	1
	4	45	32	19	16	50	15
	3	83	75	70	61	80	59
	2	99	99	99	98	96	97
Asian	6	1	0	1	0	3	1
(71)	5	12	12	8	2	19	3
	4	62	48	28	21	64	24
	3	86	82	79	69	86	62
	2	97	95	96	100	96	97
Hispanic	6	3	0	0	0	8	2
(36)	5	21	10	12	3	24	11
	4	56	50	56	24	68	27
	3	87	90	76	68	94	68
	2	97	100	100	100	97	100
White	6	2	2	2	1	7	2
(247)	5	24	17	18	14	29	17
	4	72	64	60	56	73	50
	3	93	94	97	91	93	85
	2	99	100	100	100	99	99
Female	6	2	1	2	0	5	2
(280)	5	19	13	13	7	23	10
	4	64	53	49	39	70	37
	3	93	88	89	80	91	76
	2	99	99	100	100	98	99
Male	6	3	1	1	1	5	2
(199)	5	18	14	13	10	23	11
	4	61	54	41	38	62	36
	3	84	88	84	77	86	72
	2	99	99	98	99	97	97
Total	6	2	1	1	1	5	2
	5	18	13	13	8	22	10
	4	63	53	46	39	66	35
	3	89	88	86	79	89	73
	2	99	99	99	99	97	98

Note. Examinees were distributed evenly across the four study conditions; mean sample sizes for each subgroup are shown in parentheses.

* Data in this column are based on combined samples from the two issue-first conditions (issue-issue and issue-argument) or the two argument-first conditions (argument-issue and argument-argument). Thus, the mean subgroup-sample sizes in this column are twice as large as those shown (in parentheses) for each of the four study conditions.

Table 5. Correlations Between Prompt Scores

	Issue-issue	Issue-argument	Argument-issue	Argument-argument	Average
TOTAL GROUP	.62 (618)	.54 (598)	.46 (552)	.51 (558)	582
GENDER					
Female	.56 (363)	.52 (338)	.46 (309)	.48 (329)	335
Male	.66 (254)	.57 (258)	.47 (242)	.54 (226)	245
ETHNICITY					
African American	.52 (130)	.47 (111)	.34 (104)	.36 (110)	116
Asian	.67 (130)	.58 (115)	.47 (114)	.47 (106)	117
White	.56 (285)	.49 (293)	.35 (272)	.45 (264)	279
Hispanic	.61 (44)	.33 (44)	.38 (32)	.38 (45)	41
ENGLISH BEST					
YES	.60 (508)	.52 (494)	.46 (467)	.50 (452)	480
NO	.59 (108)	.57 (104)	.40 (85)	.49 (104)	100

Note. Sample size is provided in parentheses.

Table 6. Summary Statistics, Reliability Estimates, and Standard Errors of Measurement

Test Format	Mean	SD	N	Reliability estimate	Standard error of measurement
Issue-argument*	3.6	.90	598	0.70	0.49
Argument-issue*	3.5	.86	552	0.63	0.52
One issue	3.8	1.05	618	0.62	0.65
One argument	3.3	.98	558	0.51	0.69

* For the reliability estimation, the standard deviations of these tests were multiplied by 2because these test scores were computed as an average of the two parts.

Table 7. Cumulative Percents of Difference Scores for Two Issue Prompts

Difference*	Total	Female	Male	African American	Asian	Hispanic	White
0.0	29	31	27	35	23	30	30
0.5	60	60	59	62	56	52	61
1.0	84	85	82	85	84	75	86
1.5	92	92	92	91	95	89	93
2.0	98	98	99	97	99	100	99
2.5	99	99	100	99	100		100
3.0	100	100		100			
Number of participants	618	363	254	130	130	44	285

* Difference is the absolute value of difference between scores for two issue prompts.

Table 7.1. Mean, Standard Deviation, and Correlation for Two Issue Prompts

	Female	Male	African American	Asian	Hispanic	White
Mean (position 1)	3.9	3.7	3.5	3.4	3.9	4.1
Std. (position 1)	0.99	1.12	1.00	1.06	1.11	0.95
Mean (position 2)	3.8	3.5	3.4	3.2	3.6	4.0
Std. (position 2)	0.99	1.09	0.92	1.08	1.17	0.92
Correlation between 1 & 2	0.56	0.66	0.52	0.67	0.61	0.56
Number of participants	363	254	130	130	44	285

Table 8. Cumulative Percents of Difference Scores for Two Argument Prompts

Difference*	Total	Female	Male	African American	Asian	Hispanic	White
0.0	24	23	27	26	28	27	23
0.5	58	57	59	57	59	60	58
1.0	82	82	83	82	84	87	81
1.5	93	93	94	96	93	96	92
2.0	98	97	99	99	99	98	96
2.5	99	99	100	99	100	100	99
3.0	99	99		100			99
3.5	100	100					100
Number of participants	558	329	226	110	106	45	264

* Difference is the absolute value of difference between scores for two argument prompts.

Table 8.1. Mean, Standard Deviation, and Correlation for Two Argument Prompts

	Female	Male	African American	Asian	Hispanic	White
Mean (position 1)	3.4	3.3	2.9	2.9	3.1	3.7
Std. (position 1)	0.97	0.99	0.78	0.86	0.88	0.97
Mean (position 2)	3.4	3.3	3.0	3.0	3.1	3.7
Std. (position 2)	0.97	0.98	0.86	0.89	0.78	0.98
Correlation Between 1 & 2	0.48	0.56	0.36	0.47	0.38	0.45
Number of participants	329	226	110	106	45	264

Table 9. Cumulative Percents of Difference Scores for Issue-Argument Order

Difference*	Total	Female	Male	African American	Asian	Hispanic	White
-3.5	1	1	0	0	0	2	0
-3.0	1	1	2	1	1	4	1
-2.5	3	2	4	1	5	9	2
-2.0	11	11	9	9	11	16	10
-1.5	21	22	19	22	17	32	21
-1.0	43	44	40	43	41	57	40
-0.5	60	60	60	56	58	73	59
0	81	81	80	80	80	86	81
0.5	90	90	89	88	88	93	90
1.0	96	96	95	96	96	96	96
1.5	98	98	99	98	99	96	99
2.0	99	99	100	100	100	98	100
2.5	100	100				100	
Number of participants	598	338	258	111	115	44	293

* Positive differences indicate that scores on the argument issue prompt were higher.

Table 10. Cumulative Percents of Difference Scores for Argument-Issue Order

Difference*	Total	Female	Male	African American	Asian	Hispanic	White
-2.0	2	1	3	1	3	6	1
-1.5	4	3	6	3	7	6	3
-1.0	12	9	16	11	12	19	11
-0.5	24	18	31	24	25	34	21
0	44	42	47	40	47	47	43
0.5	60	60	60	50	65	59	61
1.0	78	76	81	78	83	78	75
1.5	90	88	92	90	95	91	86
2.0	95	95	95	98	98	94	93
2.5	98	98	98	100	99	97	97
3.0	99	99	99		100	100	99
3.5	99	99	100				99
4.0	100	100					100
Number of participants	552	309	242	104	114	32	272

* Positive differences indicate that scores on the argument prompt were higher.

Table 11. Standardized Total Score Mean Differences Among Selected Subgroups for English–Best-Language Examinees

	Issue-issue	Issue-argument	Argument-issue	Argument-argument	Issue first*	Argument first**
African American/White	-0.76 (0.001)	-0.79 (0.001)	-1.14 (0.001)	-1.14 (0.001)	-0.76 (0.001)	-1.14 (0.001)
Asian/White	-0.40 (0.002)	-0.49 (0.001)	-0.78 (0.001)	-0.89(0.001)	-0.44 (0.001)	-0.79 (0.001)
Hispanic/White	-0.25 (0.128)	-0.27 (0.120)	-0.28 (0.146)	-0.85 (0.001)	-0.26 (0.030)	-0.62 (0.001)
Female/Male	0.19 (0.043)	0.05 (0.587)	0.10 (0.298)	0.07 (0.485)	0.12 (0.05)	0.07 (0.277)

Note. Mean score differences are in standard deviation units. Sample size ranges for each group across the four conditions were as follows: African American: 93 -121; Asian: 66 - 82; Hispanic: 25 - 40; White: 232 - 258; Female: 261 - 301; and Male: 178 -207.
* First prompt was issue (whether followed by an issue or an argument prompt).
** First prompt was argument (whether followed by an issue or an argument prompt).

Table 12. Score Summary Statistics by Undergraduate Major

Undergraduate major	Issue-argument			Argument-issue			Issue only			Argument only		
	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N
Natural sciences	3.5	0.9	164	3.5	0.9	159	3.8	1.0	320	3.3	1.0	321
Engineering	3.5	0.9	101	3.3	0.8	84	3.6	1.1	217	3.0	0.8	169
Social sciences	3.7	0.9	106	3.6	0.9	104	3.9	1.1	206	3.4	1.0	197
Humanities	4.0	0.8	60	3.8	0.9	55	4.3	1.0	122	3.7	1.1	113
Education	3.9	0.7	28	3.4	0.6	22	4.0	0.9	66	3.2	0.9	62
Business	3.4	0.9	47	3.7	0.8	47	3.6	1.0	109	3.4	1.0	91

Note. The "Issue-argument" and "Argument-issue" columns list averages of the two prompts given in these orders. "Issue only" is for issue prompt scores in the two conditions when the issue task was given first. "Argument only" is for argument prompt scores when the argument task was given first.

Table 13. Summary of Responses to Exit Questions

Question	All examinees	Issue-issue	Issue-argument	Argument-issue	Argument-argument
1a. Were the topics interesting to write about? Topic 1	(2,296)	(612)	(591)	(546)	(547)
Very interesting	20	27	27	11	14
Interesting	36	36	36	35	38
Somewhat interesting	32	27	27	38	36
Not interesting	12	10	11	16	12
Topic 2	(2,278)	(607)	(585)	(541)	(545)
Very interesting	27	32	18	38	19
Interesting	38	36	39	36	41
Somewhat interesting	27	24	32	20	32
Not interesting	9	8	12	6	8
2. How sufficient was the time allocated? Topic 1	(2,298)	(612)	(592)	(547)	(547)
I had enough time to finish without feeling very rushed.	71	71	70	71	70
I finished but felt very rushed	20	19	21	19	21
I did not have time to finish.	9	10	9	10	9
Topic 2	(2,296)	(610)	(591)	(548)	(547)
I had enough time to finish without feeling very rushed.	79	82	78	78	78
I finished but felt very rushed.	15	14	15	14	19
I did not have time to finish.	6	5	7	8	4
3. How well did your performance indicate your writing ability? Topic 1	(2,294)	(612)	(590)	(546)	(546)
Very well	13	15	15	11	9
Fairly well	47	46	47	47	48
Somewhat	33	31	29	34	37
Not at all	8	8	8	8	6
Topic 2	(2,297)	(611)	(592)	(547)	(547)
Very well	12	11	10	16	10
Fairly well	45	44	44	45	46
Somewhat	36	38	37	32	39
Not at all	7	8	9	7	5
4. How much effort did you make while writing? Topic 1	(2,296)	(612)	(592)	(547)	(545)
About the same effort if test had counted for admissions	46	49	47	43	43
Somewhat less effort than if test had counted	45	42	45	45	49
Considerably less effort than if test had counted	9	9	8	12	8
Topic 2	(2295)	(612)	(592)	(547)	(544)
About the same effort if test had counted for admissions	40	39	37	44	40
Somewhat less effort than if test had counted	49	50	51	46	51
Considerably less effort than if test had counted	11	12	12	10	9
5. Do you believe specific knowledge is required to perform well? Topic 1	(2,297)	(612)	(592)	(547)	(546)
Yes	37	42	41	32	32
No	63	58	59	68	68
Topic 2	(2,297)	(612)	(592)	(547)	(546)
Yes	37	42	31	41	34
No	63	59	69	59	66

Note. Entries are percentages. Sample size is in parentheses.

Table 13. Summary of Responses to Exit Questions (continued)

Question	All examinees	Issue-issue	Issue-Argument	Argument-issue	Argument-argument
6. On which writing task do you believe you performed better?	(2,297)	(593)	(574)	(528)	(535)
Better on first task	39	46	46	31	33
Better on second task	38	33	32	48	42
About the same on both tasks	23	22	23	21	25
7. How do you think the activity of writing on the first task affected your performance on the second task?	(2,280)	(608)	(587)	(541)	(544)
It had a positive effect on my performance on the second task.	34	32	26	29	50
It had a negative effect on my performance on the second task.	13	18	15	10	8
It had very little or no effect.	53	51	59	61	42
8. What is your opinion about the order in which the two writing tasks were presented to you?	(2,268)	(605)	(584)	(539)	(540)
The order was fine.	82	83	76	84	87
I would prefer to have started with the second task.	18	17	24	16	13
9. If I had handwritten the essays, they probably would have been	(2,268)	(605)	(584)	(541)	(538)
Higher quality	31	30	29	31	33
Lower quality	28	28	31	27	27
About the same quality	41	42	40	42	40
10. Overall, how would you rate the writing you did for the writing tasks you were given today? Topic 1	(2,297)	(612)	(592)	(546)	(547)
Seriously deficient	2	3	2	2	1
Weak	9	9	10	9	8
Limited	23	22	22	25	23
Clearly competent	35	34	34	37	38
Well developed	24	25	24	23	24
Thorough and insightful	6	8	7	4	5
Topic 2	(2,298)	(612)	(592)	(547)	(547)
Seriously deficient	2	2	2	2	1
Weak	9	11	10	7	9
Limited	23	27	24	22	19
Clearly competent	38	36	40	34	40
Well developed	22	18	19	25	25
Thorough and insightful	7	6	6	10	6
11. How comfortable are you with using a computer to write?	(2,245)	(599)	(573)	(536)	(537)
Very comfortable	59	58	61	57	58
Comfortable	24	24	23	25	22
Somewhat comfortable	14	13	13	14	15
Not at all comfortable	4	5	3	4	5

Note. Entries are percentages. Sample size is in parentheses.

Appendix A

GRE Writing Assessment Task Directions and Essay Scoring Guides

PRESENT YOUR PERSPECTIVE ON AN ISSUE

DIRECTIONS

Present Your Perspective on an Issue

45 minutes

In this section, you will have a choice between two Issue topics. Each topic will appear as a brief quotation that states or implies an issue of general interest. You are free to accept, reject, or qualify the claim made in the topic, so long as the ideas you present are clearly relevant to the topic you select. Support your views with reasons and examples drawn from such areas as your reading, experience, observations, or academic studies.

Before making a choice, read each topic carefully. Then decide on which topic you could write a more effective and well-reasoned essay. College and university faculty will read your essay and evaluate its overall quality, based on how well you:

- consider the complexities and implications of the issue
- organize, develop, and express your ideas on the issue
- support your ideas with relevant reasons and examples
- control the elements of standard written English

To indicate your choice, click on one of the labeled icons: TOPIC 1 or TOPIC 2. Once you have confirmed your choice, you will not be able to go back to the other topic.

Timing will begin when you click on the Proceed icon below.
You will have 45 minutes to plan and compose an essay that presents your perspective on the topic you selected. An essay on any topic other than the one you selected is not acceptable.

You may want to take a few minutes to think about the issue and to plan a response before you begin writing. Be sure to develop your ideas fully and organize them coherently, but leave time to read what you have written and make any revisions that you think are necessary.

SAMPLE PROMPTS

"The media (books, film, music, television, for example) tend to create rather than reflect the values of a society."

"Both the development of technological tools and the uses to which humanity has put them have created modern civilizations in which loneliness is ever increasing."

GRE ESSAY SCORING GUIDE

"PRESENT YOUR PERSPECTIVE ON AN ISSUE"

Score

6 A 6 paper presents a cogent, well-articulated analysis of the complexities of the issue and demonstrates mastery of the elements of effective writing.

A typical paper in this category

-- develops a position on the issue with insightful reasons and/or persuasive examples
-- sustains a well-focused, well-organized discussion
-- expresses ideas clearly and precisely
-- uses language fluently, with varied sentence structure and effective vocabulary
-- demonstrates superior facility with the conventions (grammar, usage, and mechanics) of standard written English but may have minor flaws

---

5 A 5 paper presents a well-developed analysis of the complexities of the issue and demonstrates a strong control of the elements of effective writing.

A typical paper in this category

-- develops a position on the issue with well-chosen reasons and/or examples
-- is focused and generally well organized
-- expresses ideas clearly and well
-- uses varied sentence structure and appropriate vocabulary
-- demonstrates facility with the conventions of standard written English but may have minor flaws

---

4 A 4 paper presents a competent analysis of the issue and demonstrates adequate control of the elements of writing.

A typical paper in this category

-- develops a position on the issue with relevant reasons and/or examples
-- is adequately organized
-- expresses ideas clearly
-- demonstrates adequate control of language but may lack sentence variety
-- demonstrates control of the conventions of standard written English but may have some flaws

GRE SCORING GUIDE: ISSUE -- PAGE 2

Score

3 A 3 paper demonstrates some competence in its analysis of the issue and in its control of the elements of writing but is clearly flawed.

A typical paper in this category exhibits one or more of the following characteristics:

-- is vague or limited in developing a position on the issue
-- is weak in the use of relevant reasons or examples
-- is poorly focused and/or poorly organized
-- has problems expressing ideas clearly
-- uses language imprecisely and/or lacks sentence variety
-- contains occasional major errors or frequent minor errors in grammar, usage, and mechanics

---

2 A 2 paper demonstrates serious weaknesses in analytical writing.

A typical paper in this category exhibits one or more of the following characteristics:

-- is unclear or seriously limited in developing a position on the issue
-- provides few, if any, relevant reasons or examples
-- is unfocused and/or disorganized
-- has serious and frequent problems in the use of language and sentence structure
-- contains numerous errors in grammar, usage, or mechanics that interfere with meaning

---

1 A 1 paper demonstrates fundamental deficiencies in analytical writing skills.

A typical paper in this category exhibits one or more of the following characteristics:

-- provides little evidence of the ability to develop or organize a coherent response to the topic
-- has severe and persistent errors in language and sentence structure
-- contains a pervasive pattern of errors in grammar, usage, and mechanics that severely interferes with meaning

---

0 Off topic, in a foreign language, merely copies the topic, consists of only keystroke characters, or is illegible, blank, or nonverbal.

Developed with university faculty at GRE Writing Test Committee meetings and essay scoring sessions.

ANALYZE AN ARGUMENT

DIRECTIONS

Analyze an Argument

30 minutes

You will have 30 minutes to plan and write a critique of an argument presented in the form of a short passage. A critique of any other argument is not acceptable.

Analyze the line of reasoning in the argument. Be sure to consider what, if any, questionable assumptions underlie the thinking and, if evidence is cited, how well it supports the conclusion.

You can also discuss what sort of evidence would strengthen or refute the argument, what changes in the argument would make it more logically sound, and what additional information might help you better evaluate its conclusion. *Note that you are NOT being asked to present your own views on the subject.*

College and university faculty will read your critique and evaluate its overall quality, based on how well you:

- identify and analyze important features of the argument
- organize, develop, and express your critique of the argument
- support your critique with relevant reasons and examples
- control the elements of standard written English

Before you begin writing, you may want to take a few minutes to evaluate the argument and to plan a response. Be sure to develop your ideas fully and organize them coherently, but leave time to read what you have written and make any revisions that you think are necessary.

Timing for this section will begin when you click on the Dismiss Directions icon.

SAMPLE PROMPT

"Hospital statistics regarding people who go to the emergency room after roller-skating accidents indicate the need for more protective equipment. Within this group of people, 75 percent of those who had accidents in streets or parking lots were not wearing any protective clothing (helmets, knee pads, etc.) or any light-reflecting material (clip-on lights, glow-in-the-dark wrist pads, etc.). Clearly, these statistics indicate that by investing in high-quality protective gear and reflective equipment, roller skaters will greatly reduce their risk of being severely injured in an accident."

GRE SCORING GUIDE

"ANALYZE AN ARGUMENT"

SCORE

6 A 6 paper presents a cogent, well-articulated critique of the argument and demonstrates mastery of the elements of effective writing.

A typical paper in this category

-- clearly identifies important features of the argument and analyzes them insightfully
-- develops ideas cogently, organizes them logically, and connects them with clear transitions
-- effectively supports the main points of the critique
-- demonstrates control of language, including diction and syntactic variety
-- demonstrates facility with the conventions of standard written English but may have minor flaws

---

5 A 5 paper presents a well-developed critique of the argument and demonstrates good control of the elements of effective writing.

A typical paper in this category

-- clearly identifies important features of the argument and analyzes them in a generally thoughtful way
-- develops ideas clearly, organizes them logically, and connects them with appropriate transitions
-- sensibly supports the main points of the critique
-- demonstrates control of language, including diction and syntactic variety
-- demonstrates facility with the conventions of standard written English but may have occasional flaws

---

4 A 4 paper presents a competent critique of the argument and demonstrates adequate control of the elements of writing.

A typical paper in this category

-- identifies and analyzes important features of the argument
-- develops and organizes ideas satisfactorily but may not connect them with transitions
-- supports the main points of the critique
-- demonstrates sufficient control of language to convey ideas with reasonable clarity
-- generally follows the conventions of standard written English but may have flaws

GRE SCORING GUIDE: ARGUMENT -- PAGE 2

SCORE

3 A 3 paper demonstrates some competence in analytical writing skills and in its control of the elements of writing but is plainly flawed.

A typical paper in this category exhibits one or more of the following characteristics:

-- does not identify or analyze most of the important features of the argument, although some analysis of the argument is present
-- mainly analyzes tangential or irrelevant matters, or reasons poorly
-- is limited in the logical development and organization of ideas
-- offers support of little relevance and value for points of the critique
-- does not convey meaning clearly
-- contains occasional major errors or frequent minor errors in grammar, usage, or mechanics

---

2 A 2 paper demonstrates serious weaknesses in analytical writing skills.

A typical paper in this category exhibits one or more of the following characteristics:

-- does not present a critique based on logical analysis, but may instead present the writer's own views on the subject
-- does not develop ideas, or is disorganized and illogical
-- provides little, if any, relevant or reasonable support
-- has serious and frequent problems in the use of language and in sentence structure
-- contains numerous errors in grammar, usage, or mechanics that interfere with meaning

---

1 A 1 paper demonstrates fundamental deficiencies in analytical writing skills.

A typical paper in this category exhibits more than one of the following characteristics:

--provides little evidence of the ability to understand and analyze the argument
--provides little evidence of the ability to develop an organized response
--has severe and persistent errors in language and sentence structure
--contains a pervasive pattern of errors in grammar, usage, or mechanics that results in incoherence

---

0 Off topic, in a foreign language, merely copies the topic, consists of only keystroke characters, or is illegible, blank, or nonverbal.

Developed with university faculty and approved by the GRE Writing Test Committee.

Appendix B

Further Analyses of Data Collected for the GRE Psychometric Evaluation Study

by
Gwyneth Boodoo
Henry Braun
Charles Lewis
Dorothy Thayer

Educational Testing Service

Introduction

Further analyses of data from this study were conducted by Gwyneth Boodoo, Henry Braun, Charles Lewis, and Dorothy Thayer (personal communication, April 28, 1999):

(1) A components of variance study was carried out to estimate reliability as well as the magnitude of particular sources of error.

(2) A further investigation was conducted on the relationship between differences in difficulty among prompts and prompt ordering.

(3) Prompt difficulty was examined for argument prompts that seemed more variable for some racial/ethnic subgroups (e.g., Hispanic test takers) than others.

Empirical Bayes Analysis[6]

An empirical Bayes (EB) analysis of the data was conducted. The purpose was to obtain estimates of the variability among prompt means, eliminating as much as possible the contributions of random variation among observed scores within each set of responses to a prompt.

The prompts selected for the study were a systematic stratified sample from a large pool of pre-operational prompts and intended to represent major substantive classifications as well as the range of observed score difficulties. Consequently, the assumptions underlying the EB analysis were not strictly satisfied.

In this analysis, true prompt means were assured to follow a normal distribution about a grand mean μ with variance σ_{μ}^{2}. EB estimates of these parameters were obtained from various sets of data. Table B1a displays the results for issue essays and Table B1b for argument essays.

[6] These analyses were conducted by Charles Lewis and Dorothy Thayer.

Table B1a

EB Estimates of Hyperparameters for Issue Prompts

	I* I, I* A	I I*	A I*
$\hat{\mu}$	3.8500	3.6500	3.7600
$\hat{\sigma}^2_{\mu}$	0.0470	0.0510	0.0240

Table B1b

EB Estimates of Hyperparameters for Argument Prompts

	A* A, A*I	A A*	I A*
$\hat{\mu}$	3.3100	3.3600	3.3500
$\hat{\sigma}^2_{\mu}$	0.0390	0.0320	0.0360

In Table B1a, there are three sets of estimates depending on whether the Issue prompts of interest (starred) were administered first, second following another issue prompt, or second following an argument prompt. In the following discussion, as in the tables, I refers to "issue" prompt and A indicates an "argument" prompt.

Note: The number of candidates who responded to a particular pair of prompts in a specified order is quite small, ranging from 9 to 25 with a median around 15. The results for a set like A I* or I I* are based on 40 pairs of prompts. The results from the set I* I,I* A are based on twice as many pairs of prompts and, consequently, are more stable.)

For issue essays, the estimates for the average difficulty (μ) are consistent across the three sets, while there is an apparent anomaly in the variance estimates σ^2_{μ} for the set A I*. There is no obvious explanation for this except that the combination of an argument essay followed by an issue essay has less attractive psychometric properties throughout (see further below). For argument essays, estimates of μ and σ^2_{μ} are consistent across all three sets. Attempts to use pre-operational data on these 40 prompts were not successful due to a variety of difficulties with the database.

For purposes of comparison, the typical variance between candidates within prompts varies from about .6 to 1.5 – at least an order of magnitude greater than the variance between prompts. Table B2a provides EB estimates of specific issue prompt means for two conditions and the differences between them. There is moderate "shrinkage" from the observed mean to the EB estimate. Substantial differences between observed means in the two sets are somewhat attenuated by EB, as one would expect.

Table B2a

Empirical Bayes Method: Issue Prompts

	Conditions I,I and I,A		Condition A,I		
Issue prompt	Mean 1	EB est. 1	Mean 2	EB est. 2	EB est.1-2
I – 20	3.3269	3.4734	3.4250	3.6527	-0.1793
I – 19	3.4632	3.5742	3.7292	3.7479	-0.1737
I – 18	3.4773	3.5670	3.2174	3.5784	-0.0114
I – 17	3.6338	3.6886	3.7174	3.7427	-0.0541
I – 16	3.7353	3.7633	3.8696	3.7984	-0.0351
I – 15	3.7373	3.7725	3.5484	3.6823	0.0902
I – 14	3.7407	3.7648	3.8864	3.8096	-0.0448
I – 13	3.7742	3.7875	3.6500	3.6997	0.0878
I – 12	3.8125	3.8229	4.0857	3.9238	-0.1009
I – 11	3.8154	3.8227	3.5312	3.6597	0.1630
I – 10	3.8596	3.8560	3.7639	3.7613	0.0947
I – 09	3.8806	3.8713	3.8800	3.8033	0.0680
I – 08	3.9603	3.9329	3.6618	3.7124	0.2205
I – 07	4.0000	3.9563	3.8148	3.7846	0.1717
I – 06	4.0268	3.9743	3.9706	3.8268	0.1475
I – 05	4.0763	3.9977	4.2955	4.0139	-0.0162
I – 04	4.1094	4.0462	4.0962	3.9434	0.1028
I – 03	4.1481	4.0651	3.6346	3.7024	0.3627
I – 02	4.2333	4.1404	3.8889	3.8068	0.3336
I – 01	4.3000	4.1408	3.5000	3.6819	0.4589

Table B2b

Empirical Bayes Method: Argument Prompts

	Conditions A,A and A,I		Condition I,A		
Argument prompt	Mean 1	EB est. 1	Mean 2	EB est. 2	EB est.1 - 2
A – 20	2.9275	3.0027	2.8438	3.0552	-0.0525
A – 19	2.9811	3.0920	3.3889	3.3668	-0.2748
A – 18	3.0000	3.1098	3.2286	3.2804	-0.1706
A – 17	3.0556	3.1410	3.1154	3.2531	-0.1121
A – 16	3.1083	3.1722	3.2000	3.2900	-0.1178
A – 15	3.1696	3.2022	3.1071	3.2251	-0.0229
A – 14	3.2377	3.2584	3.0625	3.2173	0.0411
A – 13	3.2551	3.2710	3.4318	3.3864	-0.1154
A – 12	3.2807	3.2889	3.0395	3.1667	0.1222
A – 11	3.2937	3.2973	3.6250	3.5144	-0.2171
A – 10	3.3036	3.3046	3.2656	3.2971	0.0075
A – 09	3.3220	3.3179	3.0909	3.2152	0.1027
A – 08	3.3500	3.3381	3.2857	3.3109	0.0272
A – 07	3.3929	3.3698	3.7188	3.5700	-0.2002
A – 06	3.4184	3.3788	3.6111	3.4819	-0.1031
A – 05	3.4364	3.3956	3.3939	3.3734	0.0222
A – 04	3.5000	3.4325	3.6053	3.5044	-0.0719
A – 03	3.6000	3.4888	3.4872	3.4338	0.0550
A – 02	3.6466	3.5458	3.5781	3.4768	0.0690
A – 01	3.8417	3.6951	3.7407	3.5785	0.1166

Table B2b presents similar results for argument essays. Here the differences between observed means are not as large as they were for issue essays, confirming the implications of the results in Tables B1a and B1b for σ^2_μ. Differences between observed means are further attenuated by EB, and are quite small.

The likely implications of the anomalous results in Table B1 are that, to the extent that the prompts selected for this study are indeed representative of the current universe of prompts, then the control of prompt difficulty has been more successful with argument prompts than with issue prompts.

G-Theory Analyses[7]

Ordinarily, generalizability theory (G-theory) has been used to model sources of error in measurement quantitatively and to obtain, among other things, estimates of reliability under different conditions. In the present setting, the most reasonable approach seemed to be to analyze separately the data for examinees who took a particular pair of prompts in a specific order and then aggregate the results appropriately. As was pointed out above, the number of

[7] These analyses were conducted by Gwyneth Boodoo with the assistance of Duanli Yan.

candidates in a particular data set ranged from 8 to 25, so that individual results are quite variable. However, some credibility attaches to the summary statistics obtained from the aggregation over a collection of data sets.

The generalizability coefficients (estimates of reliability) were generated for a student-by-rater-by-item random effects design, assuming two raters for each prompt and two prompts per examinee. The model underlying the G-theory analyses is presented below. Throughout, we assume a subject-by-prompt-by-rater random effects design and employ the following notation:

σ_s^2 variance component due to subjects

σ_i^2 variance component due to prompts

σ_r^2 variance component due to raters

Variance components due to higher order effects are denoted by $\sigma_{si}^2, \sigma_{sr}^2, \sigma_{ir}^2, \sigma_{sri}^2$ in the usual manner. Note that σ_{sri}^2 is confounded with random error. Carets denote estimated quantities. $E\hat{p}^2$ denotes the estimated generalizability coefficient. We take

$$E\hat{p}^2 = \hat{\sigma}_s^2 / (\hat{\sigma}_s^2 + \sigma_s^2)$$

where

$$\hat{\sigma}_\delta^2 = \frac{\hat{\sigma}_{si}^2}{n_i} + \frac{\hat{\sigma}_{sr}^2}{n_r} + \frac{\hat{\sigma}_{sri}^2}{n_i n_r} .$$

In our calculations, we set $n_i = 2$ and $n_r = 2$.

Consider the I I condition where candidates were presented with two issue prompts in a particular order. There are 20 issue prompts and each appeared first with one of two other issue prompts appearing second. Moreover, for each pair of prompts appearing together, one group of candidates took them in one order and another group took them in reverse order. Consequently, these are 20 x 2 = 40 data sets to be analyzed in the I I condition with a particular pair of prompts contributing either no data or two sets of data.

The five-number summaries (Tukey, 1977; p. 33) for the 40 reliability estimates in the I I, AA, I A, and A I conditions are given below:

Table B3

Five-Number Summaries of Reliability Estimates in Four Conditions

Key to format of reliability estimate summaries		
	Median value	
Estimate at 25th percentile Lowest estimate		Estimate at 75th percentile Highest estimate

II # = 40	.74	.67 .17	.82 .95
AA # = 40	.70	.49 .11	.82 .94
IA # = 40	.70	.53 .04	.83 .96
AI # = 40	.62	.46 .00	.79 .89

In all conditions, it is evident that there is considerable variability among the 40 estimates. Overall, the results for I I are best followed by the A A and I A conditions (which are nearly identical in terms of the summary statistics), with the results for the A I condition being noticeably poorer. It is definitely heartening that the "typical" reliability in three of the conditions (including the I A condition that corresponds to the actual GRE Writing Assessment) is about 0.70. On the other hand, it is somewhat disturbing that an appreciable number of estimates fall below 0.50. However, the sample sizes for particular prompt pairs can be quite small and most of the lower reliability estimates do occur with the smaller data sets.

Below is a five-number summary of the sample sizes for the 40 data sets in the I A condition:

IA
= 40

	.14	
.12		.18
.08		.25

A detailed review of the G-theory analyses indicates that the contribution of the variance component due to raters is quite negligible in all conditions. In the I I and A A conditions, the variance component due to prompts is generally quite a bit smaller than the variance component due to subjects by prompts. In fact, for the I I condition, the five-number summaries are:

$\hat{\sigma}_i^2$

I I
= 40

	.01.	
.00		.06
.00		.21

$\hat{\sigma}_{si}^2$

I I
= 40

	.29	
.18		.40
.02		.92

The median value of 0.01 for $\hat{\sigma}_i^2$ is quite a bit smaller than the value of 0.05 obtained through the EB analysis reported above. It is not obvious why this has occurred. The results for the A A condition are:

$\hat{\sigma}_i^2$

A A
= 40

	.04	
.00		.08
.00		.22

$\hat{\sigma}_{si}^2$

A A
= 40

	.32	
.20		.46
.00		.82

Here the median values of 0.035 for $\hat{\sigma}_i^2$ accords quite well with the value of 0.032 obtained through the EB analysis reported above.

In the I A condition, the results are somewhat different. The relevant five-number summaries are:

$\hat{\sigma}_i^2$

I A # = 40		
	.08	
.02		.18
.00		.74

$\hat{\sigma}_{si}^2$

I A # = 40		
	.36	
.18		.51
.04		.98

We see that the distribution of $\hat{\sigma}_i^2$ is clearly shifted to the right here in comparison with the distributions in the I I and A A conditions. This is expected, given that I and A represent different types of writing demands. On the other hand, the distribution of $\hat{\sigma}_{si}^2$ is quite comparable to those in the I I and A A conditions. Indeed, in a number of datasets in the I A condition, $\hat{\sigma}_i^2$ exceeds $\hat{\sigma}_{si}^2$.

The results for the A I condition are somewhat poorer than the I A condition. The relevant five-number summaries are:

$\hat{\sigma}_i^2$

I A # = 40		
	.09	
.00		.20
.00		.71

$\hat{\sigma}_{si}^2$

I A # = 40		
	.40	
.25		.57
.14		1.13

Viewed from a purely psychometric perspective, the I I condition would be preferred to the other conditions by virtue of its somewhat higher estimated reliability. On the other hand, from a construct representation perspective, an I A (or A I) condition would likely be preferred since it taps into two different aspects of the construct. The latter statement, of course, assumes that the sets of issue prompts and argument prompts are appropriately designed. Poor psychometric characteristics may indicate a problem on that score. The results presented here do not suggest such a problem. Indeed, the GRE Writing Advisory Committee considered the skills assessed in both prompt types (argument and issue) to be highly relevant to advanced study in many different academic disciplines.

If we consider building a writing instrument as an exercise in test design, then the choice of I A over I I, say, represents a reasonable tradeoff between construct representation and reliability. Clearly our results indicate the superiority of I A over A I. It only remains to examine the evidence for differential variability among prompts for different subgroups of candidates.

Subgroup Analyses

Table 7 and Table 8 in the main report address the issue of differential subgroup unreliability by providing cumulative percents of difference scores for candidates taking two issue prompts and candidates taking two argument prompts. The results are reported by gender and by race/ethnicity. The distributions of difference scores are relatively uniform across groups. The exception is the Hispanic group, which had a slightly more favorable distribution on argument

prompts and a slightly less favorable distribution on issue prompts. However, there were only 45 Hispanic candidates in each case, less than half the sample size for African Americans and Asians.

The following procedure was carried out to accomplish G-theory analyses. In the I I condition, for example, pairs of prompts that were taken by at least eight minority group members (in either order) were identified. Analyses were then run for that data set and for the parallel data set -- White candidates who took the same pair of prompts (in either order).

Table B4a and Table B4b present the data for Asian candidates, and Table B5a and Table B5b present the data for African American candidates. It was not possible to carry out comparable analyses for Hispanic candidates because of the small sample sizes.

Table B4a

G-Theory Analyses Comparing Asian and White Candidates on Issue Prompts

Pair	Sample size	Estimated reliability	Subgroup
1	10	.73	Asian
	13	.73	White
2	12	.79	Asian
	7	.00	White
3	13	.82	Asian
	16	.80	White
4	9	.94	Asian
	7	.94	White
5	8	.60	Asian
	14	.69	White

Table B4b

G-Theory Analyses Comparing Asian and White Candidates on Argument Prompts

Pair	Sample size	Estimated Reliability	Subgroup
1	8	.87	Asian
	13	.55	White
2	8	.93	Asian
	15	.94	White
3	8	.79	Asian
	7	.40	White
4	10	.79	Asian
	10	.43	White

Table B4a presents data for 52 Asian candidates (57 White candidates) who took pairs of issue prompts, and Table B4b presents data for 34 Asian candidates (45 White candidates) who took pairs of argument prompts.

Table B5 presents data for 50 African American candidates (68 White candidates) who took pairs of issue prompts and 57 African American candidates (88 White candidates) who took pairs of argument prompts.

Table B5a

G-Theory Analyses Comparing African American and White Candidates on Issue Prompts

Pair	Sample size	Estimated Reliability	Subgroup
1	9	.12	African American
	12	.56	White
2	10	.69	African American
	15	.57	White
3	9	.00	African American
	17	.47	White
4	14	.81	African American
	14	.82	White
5	8	.96	African American
	10	.46	White

Table B5b

G-Theory Analyses Comparing African American and White Candidates on Argument Prompts

Pair	Sample size	Estimated reliability	Subgroup
1	8	.47	African American
	13	.50	White
2	9	.15	African American
	11	.68	White
3	9	.35	African American
	15	.69	White
4	12	.15	African American
	10	.43	White
5	10	.62	African American
	18	.77	White
6	9	.89	African American
	21	.72	White

The results are summarized in Table B6a and Table B6b. We see there that, for issue prompts, the median estimated reliability is higher for both Asians and African Americans than for their respective comparison groups of Whites. For argument prompts, the same is true when comparing Asians and Whites, but not when comparing African Americans and Whites.

Table B6a

Summary Data on G-Theory Analyses Comparing Groups on Issue Prompts

# of prompts	Total sample size	Median estimated reliability	Subgroup
5	52	.79	Asian
	57	.73	White
5	50	.69	African American
	68	.56	White

Table B6b

Summary Data on G-Theory Analyses Comparing Groups on Argument Prompts

# of prompts	Total sample size	Median estimated reliability	Subgroup
4	34	.83	Asian
	45	.49	White
6	57	.35	African American
	88	.69	White

Overall, there is no substantial evidence, one way or another, about differential reliability for White and other candidates. These results are quite fragile given the small sample sizes available and, of course, we have not had access to comparable data for Hispanic candidates.

54020-007951 • Y51P.4 • Printed in U.S.A.
I.N. 991532