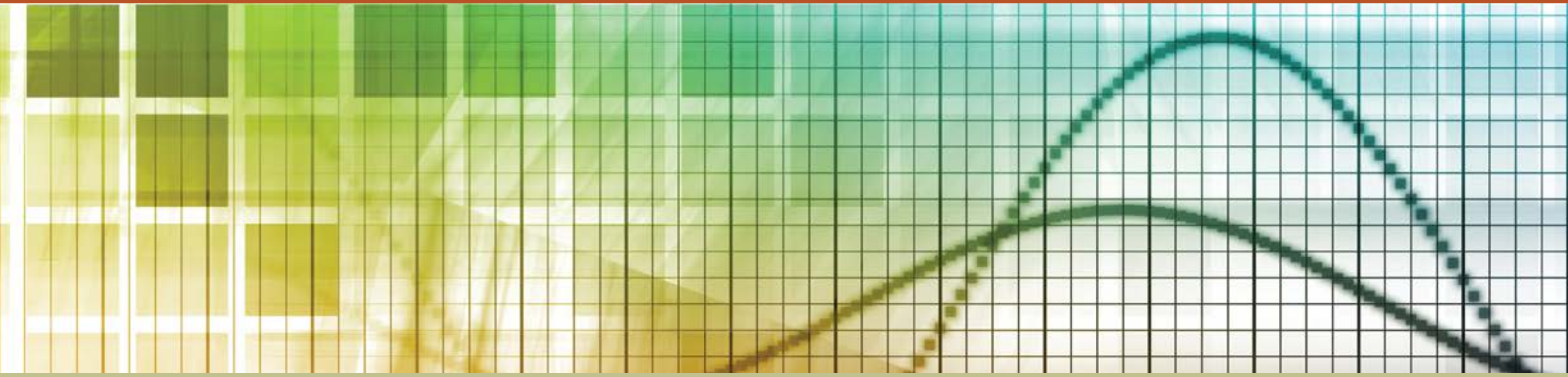
Assess Ability. Predict Performance.



The Research Foundation for the *GRE*® revised General Test: A Compendium of Studies

Cathy Wendler and Brent Bridgeman, Editors
With assistance from Chelsea Ezzo

The Research Foundation for the *GRE®* revised General Test:
A Compendium of Studies

Edited by Cathy Wendler and Brent Bridgeman
with assistance from Chelsea Ezzo

To view the PDF of *The Research Foundation for the* GRE® *revised General Test: A Compendium of Studies* visit

www.ets.org/gre/compendium

July 2014

Dear Colleague:

Since its introduction in 1949, the *GRE*® General Test has been an important part of graduate admissions as a proven measure of applicants' readiness for graduate-level work. GRE scores are used by the graduate and business school community to supplement undergraduate records, including grades and recommendations, and other qualifications for graduate-level study.

The recent revision of the GRE General Test was thoughtful and careful, with consideration given to the needs and practices of score users and test takers. A number of goals guided our efforts, such as ensuring that the test was closely aligned with the skills needed to succeed in graduate and business school, providing more simplicity in distinguishing performance differences between candidates, enhancing test security, and providing a more test-taker friendly experience.

As with other ETS assessments, the GRE General Test has a solid research foundation. This research-based tradition continued as part of the test revision. *The Research Foundation for the* GRE® *revised General Test: A Compendium of Studies* is a comprehensive collection of the extensive research efforts and other activities that led to the successful launch of the GRE revised General Test in August 2011. Summaries of nearly a decade of research, as well as previously unreleased information about the revised test, cover a variety of topics including the rationale for revising the test, the development process, test design, pilot studies and field trials, changes to the score scale, the use of automated scoring, validity, and fairness and accessibility issues.

We hope you find this compendium to be useful and that it helps you understand the efforts that were critical in ensuring that the GRE revised General Test adheres to professional standards while making the most trusted assessment of graduate-level skills even better. We invite your comments and suggestions.

Sincerely,

David Payne
Vice President & COO
Global Education
Educational Testing Service

Ida Lawrence
Senior Vice President
Research and Development
Educational Testing Service

Acknowledgments

We thank Yigal Attali, Jackie Briel, Beth Brownstein, Jim Carlson, Neil Dorans, Rui Gao, Hongwen Guo, Eric Hansen, John Hawthorn, Jodi Krasna, Lauren Kotloff , Longjuan Liang, Mei Liu, Skip Livingston, Ruth Loew, Rochelle Michel, Maria Elena Oliveri, Donald Powers, MaraGale Reinecke, Sharon Slater, John Young, and Rebecca Zwick for their help with previous versions of the Compendium. In particular, we acknowledge the time and guidance given by Kim Fryer and Marna Golub-Smith and thank them for their assistance.

Contents

Overview to the Compendium

The decision to revise a well-established test, such as the *GRE*® General Test, is made purposively and thoughtfully because such a decision has major consequences for score users and test takers. Considerations as to changing the underlying constructs measured by the test, question types used on the test, the method for delivering the test, and the scale used to report scores must be carefully evaluated (see Dorans & Walker, 2013; Wendler & Walker, 2006). Changes in the test-taking population, the relationship of question types to the skills being measured, or expanding on the use of the test scores requires that a careful examination of the test be undertaken.

For the GRE General Test, efforts to evaluate possible changes to the test systematically began with approval from the *Graduate Record Examinations*® (GRE) Board. What followed was a decade of extensive efforts and activities that examined multiple question types, test designs, and delivery issues related to the test revision. Throughout the redesign process, the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council of Measurement in Education [NCME], 1999)[1] was used as guidance. The resulting GRE revised General Test is one that adheres to the *Standards*.

The Compendium provides chapters in the form of summaries as a way to describe the extensive development process. A number of these chapters are available in longer published documents, such as research reports, journal articles, or book chapters, and their original source is provided for convenience. The intention of the Compendium is to provide, in nontechnical language, an overview of specific efforts related to the GRE revised General Test. Other studies that were conducted during the time of the development of the GRE revised General Test are not detailed here. While these studies are important, only those that in some way contributed to decisions about the GRE revised General Test or contribute to the validity argument for the revised test are included in the Compendium.

The Compendium is divided into six sections, each of which contains multiple chapters around a common theme. It is not expected that readers will read the entire Compendium. Instead, the Compendium is designed so that readers may chose to review (or print) specific chapters or sections. Each section begins with an overview that summarizes the chapters found within the section. Readers may find it helpful to read each section overview and to use the overview as a guide to determine which particular chapters to read.

Section 1: Development of the GRE revised General Test

A test should be revised using planned, documented processes that include, among others, gathering data on the functioning of question types, timing allotted for the test or test sections, and performance differences for subpopulations. The focus of the first section is to outline the development efforts surrounding the GRE revised General Test and to show how the development process was deliberate, careful, and well documented. The first chapter in this section provides the rationale for revising the test, as well as an overview of the final test specifications. Other chapters in this section describe specific development and design efforts for the three measures—Verbal Reasoning, Quantitative Reasoning, and Analytical Writing. Information on the various pilots and field test activities for the revised test, evaluations of the impact of calculator availability, and additional foundational studies are provided in other chapters.

Section 2: Creating and Maintaining the Score Scales

The *Standards* (AERA, APA, & NCME, 1999) indicate that changes to an existing test may require a change in the reporting scale in order to ensure that the score reporting scale remains meaningful. Chapters in this section provide information on the new score scale created and being used with the Verbal Reasoning and Quantitative Reasoning measures. Included in this section are chapters on the considerations used in the decision to change the Verbal Reasoning and Quantitative Reasoning scales and the method used to define the revised scales. Also included are chapters on the processes used to maintain the scale on the Analytical Writing measure.

Section 3: Test Design and Delivery

The GRE revised General Test incorporates an innovative approach to computer-adaptive testing: that of a multistage adaptive test (MST) model. This section describes the specific efforts related to the decision to use the MST model with the GRE revised General Test. Chapters include an overview of practical considerations with computer-delivered tests, the methodology used to design the MST for the GRE revised General Test, and studies examining the impact of moving to the MST model to ensure scoring accuracy for all test takers.

Section 4: Understanding Automated Scoring

Automated scoring of essays from the Analytical Writing measure, in conjunction with human raters, was implemented prior to the release of the GRE revised General Test. However, much of the earlier research conducted also provides the foundation for the use of automated scoring with the GRE revised General Test. This work is critical to ensure that GRE essays continue to be scored accurately and fairly for test takers. An overview of automated scoring and its use with GRE essays is provided in the first chapter of the section. The remaining chapters detail the various studies that were completed that led to the decision to use automated essay scoring, including the functioning of *e-rater*® scoring engine, the automated scoring engine that is used; comparisons with scores by human raters; and comparisons with other indicators of writing proficiency.

Section 5: Validation Evidence

Test validation is an essential, if not the most critical, component of test design in that it ensures that appropriate evidence is provided to support the intended inferences being made with test results (AERA, APA, & NCME, 1999). Chapters in this section provide studies of the predictive validity of the GRE General Test, as well as studies related to long-term success in graduate school. Although many of the validity studies used data from the older version of the GRE General Test, the results are still relevant for and applicable to the revised test.

Section 6: Ensuring Fairness and Accessibility

All assessments should be designed, developed, and administered in ways that treat people equally and fairly regardless of differences in personal characteristics (AERA, APA, & NCME, 1999). With a revision to a test, it is important to conduct research on proposed new question types and test directions to understand the impact the revised test may have on particular groups of test takers. An overview of the definition of fairness and the processes used with the GRE General Test to ensure ongoing fairness for all test takers is provided in this section. Chapters in this section include information on field trials and studies for test takers with disabilities, the development of a computer-voiced version of the GRE revised General Test, and studies that examine other fairness concerns.

Summary

These chapters are intended to showcase the foundational psychometric and research work done prior to the launch of the GRE revised General Test. We hope they provide readers with an understanding of the efforts that were critical in ensuring that the GRE revised General Test was of the same high quality and as valid and accurate as its predecessor.

Cathy Wendler and Brent Bridgeman, Editors
With assistance from Chelsea Ezzo

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Dorans, N. J., & Walker, M. E. (2013). Multiple test forms for large-scale assessments: Making the real more ideal via empirically verified assessment. In K. F. Geisinger (Ed.), *APA handbook of testing and assessment in psychology* (Vol. 3, pp. 495–515). Washington, DC: American Psychological Association.

Wendler, C., & Walker, M. E. (2006). Practical issues in designing and maintaining multiple test forms for large-scale programs. In S. Downing & T. Haladyna (Eds.), *Handbook of test development* (pp. 445–467). Hillsdale, NJ: Erlbaum.

Notes

[1] Note that the development of the GRE revised General Test was guided by the 1999 version of the *Standards* (APA, AERA, & NCME). However, the test is also consistent with the newest version of the *Standards* published in 2014.

Section 1: Development of the *GRE*® revised General Test

The revision of a test used for high-stakes decisions requires careful planning, study, and various data collection efforts to ensure that the resulting test continues to serve all test takers and score users. The work done to revise the *GRE*® General Test exemplifies the careful planning and extensive evaluation needed to ensure that the final test was of the highest caliber. Chapters in this section describe many of the studies that provided foundational support for the GRE revised General Test, as well as specific design and development efforts for the three measures: Verbal Reasoning (Verbal), Quantitative Reasoning (Quantitative), and Analytical Writing.

- Chapter 1.1 discusses the rationale for revising the test and the primary goals of the test revision. It describes four main issues addressed during the revision: test content, test design, the score scales, and fairness and validity. As part of the revision, enhancements were made to the *test content* to better reflect the types of skills needed to succeed in graduate and business school programs. Changes were also made to the *design of the test* to support the goals of enhancing security, providing more test taker–friendly features, and ensuring validity and accuracy of the test scores. Although it was recognized that *changing the score scale* used with the Verbal and Quantitative measures would have significant impact on score users, the change was considered necessary given the revisions to the test content and test design. This change also adhered with the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council of Measurement in Education, 1999) and allowed more effective use of the entire score scale compared to the previous scale. Ensuring continued *fairness* for test takers and *validity* were critical aspects of the development of the GRE revised General Test. The resulting test is computer-delivered and composed of three measures: one Analytical Writing section containing two separately timed essay tasks (analyze an issue [issue] and analyze an argument [argument]), two Verbal Reasoning sections, and two Quantitative Reasoning sections. In addition, there are two unscored sections, one containing questions that are being tried out for use on future editions of the test, and one that is used for various research efforts. The final test specifications, provided in detail in this chapter, helped meet the goals defined as part of the revision of the GRE General Test.
- Chapter 1.2 describes a number of the pilot and field trials conducted over the last decade that support the development of the Verbal and Quantitative measures for the revised test. It traces the various efforts, providing a chronological look at how

the results of the pilots and field trials guided the various decisions about appropriate question types, section configurations, and ultimate test design. The chapter provides a brief description of various test designs (linear, computer-adaptive, and multistage) that were considered for the GRE revised General Test. The GRE revised General Test was initially conceived as a computer-delivered linear test, and this chapter describes the various pilots for the Verbal and Quantitative measures that were run to evaluate proposed new question types, various measure and test configurations, and psychometric characteristic; this work culminated in a large field trial that included all three GRE measures. When the decision was made to move to a multistage adaptive test (MST) design, additional studies were undertaken. The chapter describes the simulation work and additional pilots that were done using the MST design that resulted in the GRE revised General Test.

- Chapter 1.3 focuses on an exploration done using a text analysis tool that allows for the efficient development of Verbal question types. The chapter describes the use of the tool to enhance the development of the paragraph reading question type used on the GRE revised General Test. This question type consists of a short passage followed by two, three, or four questions. Two approaches are described in the chapter. The first approach focuses on the passage development side of the question type, and the second approach focuses on the question development side. Results indicated that use of this tool efficiently increased the percentage of acceptable passages located, as well as helped test developers write questions based on a passage at required difficulty levels.
- Chapter 1.4 reports on a study that explored whether test takers could transfer strategies they use to solve certain Quantitative questions to other questions that were very similar (referred to as *question variants*). Applying these strategies inappropriately could impact the validity of the test. Three types of questions were examined: (a) matched questions that addressed the same content area (e.g., algebra) and had the same context (e.g., word problem), (b) close variants that were essentially the same question mathematically but had altered surface features (e.g., different names or numbers), and (c) appearance variants that were similar in surface features but required different mathematical operations to solve. Results indicated that appearance variants were always more difficult than close variants and generally more difficult than matched variants. Close variants were generally easier than matched questions. Having seen a question with the same mathematical structure seemed to enhance performance, but having seen a question that appeared to be similar but had a different mathematical structure degraded performance.

- Chapter 1.5 describes a study that looked at the impact of calculator availability on the Quantitative measure. The study determined if an adjustment was needed for Quantitative questions that were pretested without a calculator and evaluated the effects of calculator availability on overall scores. Two Quantitative question types were examined: standard multiple-choice and quantitative comparisons. Results indicated that there was only a minimal calculator effect on most questions in that a greater percentage of test takers who used a calculator did not get the question correct when compared to test takers who did not use a calculator. However, questions that were categorized as being calculator sensitive were generally answered more quickly by students who used a calculator. Results also indicated that the use of a calculator seemed to have little impact on test takers' scores on the Quantitative measure.
- Chapter 1.6 reports on a study that explored calculator usage on the GRE revised General Test. The study examined the relationship of test-taker characteristics (ability level, gender, and race/ethnicity) and question characteristics (difficulty level, question type, and content) with calculator use. It also explored whether response accuracy was related to calculator use. Results indicated that the calculator was used by most students; it was used slightly more by test takers who were female, White, or Asian. The highest (top 5%) and lowest (bottom 10%–20%) scoring test takers used the calculator less frequently than other test takers. Analyses also showed that calculator usage was higher on easier questions and that questions with higher calculator usage required more time to answer. Finally, results indicated that for most questions, but especially for easier questions, test takers who used the calculator were more likely to answer the question correctly than test takers with the same score on the Quantitative measure who did not use the calculator.
- Chapter 1.7 investigates the alignment of the skills measured by the Analytical Writing measure with those writing tasks thought to be important for academic success at both the master's and doctoral level. Data were gathered using a survey of writing tasks statements. The survey was completed by 720 graduate faculty members across six disciplines: English, education, psychology, natural sciences, physical sciences, and engineering. Results indicated that faculty who taught master's level students ranked the statements, on average, as *moderately important* to *very important*. Faculty who taught doctoral level students ranked the statements, on average, as *moderately important* to *extremely important*. In addition, those skills thought to be necessary to score well on the GRE Analytical Writing measure were judged to be important for successfully performing on the

writing tasks statements. The findings of this study provided foundational support for the continuation of the Analytical Writing measure.

- Chapter 1.8 describes efforts related to timing issues for the Analytical Writing measure. It first summarizes the results of a field trial that provided preliminary input for the revised timing configuration of the Analytical Writing measure. As part of the field trial, three possible timing configurations were tried out. While this study faced a number of challenges, the results still provided sufficient evidence to support the final timing configuration of 30 minutes for each of the two essay prompts for further development and eventual operational implementation. The chapter also provides information on the continuity of the measure on the GRE revised General Test. A comparison of the psychometric properties of Analytical Writing measure before and after the launch of the revised test is given. Results indicated that, in general, the psychometric proprieties of the revised Analytical Writing measure are similar to those of the original measure.
- Chapter 1.9 reports on a study that examined four psychometric aspects of the Analytical Writing measure when it was first introduced in 1999. The first, *prompt difficulty*, looked at test takers' scores on a number of prompts to see if they were representative of the scores obtained on other prompts of the same type. The impact of the order that prompt types were given on test scores, or o*rder effects*, was the second aspect analyzed. *Score distributions* by race/ethnicity and gender groups for each of the two prompt types were also examined. Finally, relationships among the scores from the issue and argument writing tasks were examined to determine whether two writing scores or a single combined score would be reported. Results guided the decisions made about the configuration and scoring of the Analytical Writing measure.
- Chapter 1.10 describes a study that explored issues related to essay variants. Essay variants are created from the same prompt; a specific prompt (parent) is used as the basis for one or more variants that specify different writing tasks in response to the parent prompt. The study examined the comparability of score distributions across Analytical Writing prompts and their variants, differential difficulty of variant types across subgroups, and the consistency of reader scores across prompts and variants. Results indicated that for both issue and argument variants the average differences were quite small, no significant interaction with race/ethnicity or gender was seen, and no variant type appeared to have more or less rater reliability than the other.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

1.1 Revisiting the *GRE*® General Test

Jacqueline Briel and Rochelle Michel

In August 2011, the *GRE*® program launched the GRE revised General Test. While there have been a number of changes to the GRE General Test since its introduction in 1949, this revision represents the largest change in the history of the GRE program. Previous changes included test content changes such as the introduction of the Analytical Reasoning measure in 1985 and the introduction of the Analytical Writing measure in 2002, which replaced the Analytical Reasoning measure. Changes to test delivery included the transition of the GRE General Test from a paper-based test (PBT) to a computer-based test (CBT) in 1992, followed by the introduction of the computer adaptive test (CAT) design that was introduced in 1993. The launch of the GRE revised General Test in 2011 included major changes to test content, a new test design, and the establishment of new score scales for the Verbal Reasoning and Quantitative Reasoning measures. The *Graduate Record Examinations*® (GRE) Board,[1] which consists of graduate deans and represents the graduate community, was instrumental in guiding the development of the test and related policies.

Four primary goals shaped the revising of the GRE General Test:

- More closely align with the skills needed to succeed in graduate and business school
- Provide more simplicity in distinguishing performance differences between candidates
- Provide more test taker–friendly features for an enhanced test experience
- Enhance test security

Test Content

As was the case with the GRE General Test prior to August 2011, the GRE revised General Test focuses on the types of skills that have been identified as critical for success at the graduate level—verbal reasoning, quantitative reasoning, critical thinking, and analytical writing—regardless of a student's field of study. However, enhancements have been made to the content of the test to better reflect the types of reasoning, critical thinking, and analysis that students will face in graduate and business school programs and to align with the skills that are needed to succeed.

The Verbal Reasoning measure assesses reading comprehension skills and verbal and analytical reasoning skills, focusing on the ability to analyze and evaluate written material. The measure was revised to place a greater emphasis on complex reasoning skills with more text-based materials, such as reading passages, and less dependency on vocabulary out of context.

As a result, the antonyms and analogies on the prior test were removed from the Verbal Reasoning measure to reduce the effects of memorization, and they were replaced with new question types, including those that take advantage of new computer-enabled tasks, such as highlighting a relevant sentence to answer a question.

The Quantitative Reasoning measure assesses problem-solving ability, focusing on basic concepts of arithmetic, algebra, geometry, and data analysis. The revised measure places a greater emphasis on quantitative reasoning skills and has an increased proportion of questions involving real-life scenarios and data interpretation. An on-screen calculator was added to this measure to reduce the emphasis on computation. The Quantitative Reasoning measure also takes advantage of new question types and new computer-enabled tasks, such as entering a numerical answer rather than selecting from the options presented.

The Analytical Writing measure assesses critical thinking and analytical writing skills, specifically the ability to articulate complex ideas clearly and effectively. Although the Analytical Writing measure has not changed dramatically from the prior version, test takers are now asked to provide more focused responses to questions, reducing the possibility of reliance on memorized materials.

Test Design

In addition to the test content, the overall design of the test was revised to support the goals of enhancing security, introducing more test taker–friendly features, and ensuring access to testing, as well as enhancing the validity and the measurement characteristics of the test scores. The availability and the power of the Internet have presented increased opportunities to memorize and share information. These challenges were mitigated with the split-test administrations,[2] but a goal of the revision was to eliminate the need for this alternative testing model. The test was revised to reduce the effects of memorization by eliminating single-word verbal questions and reducing the possibility of nonoriginal essay responses. In addition, ETS incorporated security features in the test design to further enhance the existing security measures.

Given these test design goals, consideration was given as to whether the GRE revised General Test would continue to be delivered as a CAT or move to a linear test form delivery model. While the CAT design has a number of advantages (i.e., efficiency, measurement precision), a linear form model offers a less complex transition to a revised test with new content, new question types, and new score scales.

A linear form model was initially explored and significant research was conducted as described in this compendium. However, a relatively small number of large, fixed test administrations did not meet the goals of the program to provide frequent access to testing and provide convenient opportunities for candidates to take the test where and when they chose to do so. While a linear form test delivery model that significantly increased the test administration opportunities was considered, the testing time required for a linear test form model and the sustainability of such a model in the long term were considered less than desirable. Since linear forms were deemed impractical in a continuous testing environment, the GRE program explored other testing models.

Building on the significant research that had been conducted on the linear form model, a multistage adaptive model (MST), in which blocks of preassembled questions are delivered by an adaptive algorithm, was explored. The MST design represented a compromise between the question-level CAT design and a linear test design and met the test design goals for the revised test. After considerable research, it was determined that the use of an MST design would be preferable for the GRE revised General Test (Robin & Steffen, Chapter 3.3, this volume).

Score Scales

The GRE Board and GRE program recognized early on that changes to the score scales would have a significant impact on the score user community. However, the mean scores for the Verbal Reasoning and Quantitative Reasoning measures had shifted away from the midpoint of the scale and were no longer in alignment, the current population had changed significantly from the original reference group on which the scale was based, and a number of content and scoring changes were made to the test (Golub-Smith & Wendler, Chapter 2.1, this volume). Given these factors, the *Standards for Educational and Psychological Testing* required a change in the score scales (American Educational Research Association [AERA], American Psychological Association [APA], and the National Council of Measurement in Education [NCME], 1999).

The changes to the score scales also provided an opportunity to make more effective use of the entire score scale than the previous scale did, and since candidates are more spread out on the scale, each point is more meaningful. The new scales were also intended to make more apparent the differences between candidates and to facilitate more appropriate comparisons.

A number of scaling solutions were considered and seven scaling goals were defined prior to the launch of the GRE revised General Test (Golub-Smith & Moses, Chapter 2.2, this volume). The new scales were selected to balance the changes in content, new question types, the new psychometric model, and test length, and they successfully met the established scaling goals.

The decision to change the score scales was not made lightly, and the GRE Board and the GRE program had many discussions about the nature of the changes and the extensive communications plan that would be required to ease the transition to a new scale as much as possible. For example, since GRE scores are valid for 5 years, a decision was made to provide estimated scores on the new scales on GRE score reports for those scores earned prior to the launch of the GRE revised General Test.

Fairness and Validity

Throughout the entire development process of the GRE revised General Test, the GRE program has been diligent in continuing to uphold the ETS commitment to fairness and access. A number of steps were undertaken to ensure that the GRE revised General Test would continue to address the needs of all test takers. Staff worked with outside, independent experts and other contributors who represent diverse perspectives and underrepresented groups to provide input on a range of test development issues, from conceptualizing and producing frameworks for the assessments to designing test specifications and writing or reviewing questions. Multiple pilots, field trials, and field tests were held to evaluate the proposed changes for all groups (see chapters in Section 6, this volume; Wendler, Chapter 1.2, this volume).

Ongoing validation is done for all GRE tests to evaluate whether the test is measuring the intended construct and providing evidence for the claims that can be made based on candidates' test results. This ongoing validation process provides evidence that what is measured is in fact what the test intends to measure, in consideration of the skills and abilities that are important for graduate or business school. In addition, the GRE program continues to provide products and services to improve access to graduate education, such as free test preparation materials, fee reductions for individuals who demonstrate financial need and for programs that work with underrepresented populations, and special accommodations for test takers who have disabilities to ensure that they have fair access to the test.

The Final Design

For more than 60 years, GRE scores have been a proven measure of graduate-level skills. As a result of the redesign, the Verbal Reasoning, Quantitative Reasoning, and Analytical Writing measures are even better measures of the kinds of skills needed to succeed in graduate and business school. As of this writing, the GRE revised General Test is administered in a secure testing environment at about 700 ETS-authorized test centers in more than 160 countries. In most regions of the world, the computer-based GRE revised General Test is administered on a continuous basis throughout the year. In areas of the world where the computer-based test is not available, the test is administered in a paper-based format up to three times per year.

General Design

The computer-based GRE revised General Test contains one Analytical Writing measure with two separately timed tasks, two Verbal Reasoning measures, and two Quantitative Reasoning measures. In addition, some questions on the GRE General Test are being tried out (i.e., pretested) for possible use in future editions of the test. These questions are included in an unidentified, unscored section of the test. Other questions may also appear in identified but unscored research sections. Answers to pretest and research questions are not used in the calculation of scores for the test. Total testing time is approximately 3 hours and 45 minutes.

The Analytical Writing measure is always the first section in the test. The Verbal Reasoning, Quantitative Reasoning, and pretest/research sections may appear in any order following the Analytical Writing measure.

The Verbal Reasoning and Quantitative Reasoning measures of the computer-based GRE revised General Test use an MST design, meaning that the test is adaptive at the section level. This test design allows test takers to move freely within any timed section, allowing them to use more of their own personal test-taking strategies and providing a friendlier test-taking experience. Specific features include preview and review capabilities within a section, *mark* and *review* features to tag questions so that test takers can skip them and return later if they have time remaining in the section, the ability to change/edit answers within a section, and an on-screen calculator for the Quantitative Reasoning measure.

The Verbal Reasoning and Quantitative Reasoning measures each have two operational sections. Overall, the first operational section is of average difficulty. The second operational section of each of the measures is administered based on a test taker's overall performance on the first section of that measure. Verbal Reasoning and Quantitative Reasoning scores are each reported on a scale from 130 to 170, in one-point increments. A single score is reported for the Analytical Writing measure on a 0 to 6 score scale, in half-point increments.

Verbal Reasoning

The Verbal Reasoning measure is composed of two sections, 20 questions per section. Students have 30 minutes per section to complete the questions. The Verbal Reasoning measure assesses the ability to analyze and draw conclusions from discourse and reason from incomplete data; understand multiple levels of meaning, such as literal, figurative, and author's intent; and summarize text and distinguish major from minor points. In each test edition, there is a balance among the passages across three different subject matter areas: humanities, social sciences (including business), and natural sciences. There is an emphasis on complex reasoning skills, and this measure contains new question types and new computer-enabled tasks.

There are three types of questions used on the Verbal Reasoning measure: reading comprehension, text completion, and sentence equivalence. Reading comprehension passages are drawn from the physical sciences, the biological sciences, the social sciences, the arts and humanities, and everyday topics, and they are based on material found in books and periodicals, both academic and nonacademic. The passages range in length from one paragraph to four or five paragraphs. There are three response formats used with the reading comprehension questions. The multiple-choice select-one-answer-choice questions are the traditional multiple-choice questions with five answer choices from which a test taker must select one. The multiple-choice select-one-or-more-answer-choices questions provide test takers with three answer choices and ask them to select all that are correct; one, two, or all three of the answer choices may be correct. To gain credit for these questions, a test taker must select all the correct answers and only those; there is no credit for partially correct answers. The select-in-passage questions ask the test taker to click on the sentence in the passage that meets a certain description. To answer the question, the test taker chooses one of the sentences and clicks on it (clicking anywhere on the sentence will highlight the sentence).

Text completion questions include a passage composed of one to five sentences with one to three blanks. There are three answer choices per blank or five answer choices if there is a single blank. There is a single correct answer, consisting of one choice per blank. Test takers receive no credit for partially correct answers.

Finally, sentence equivalence questions consist of a single sentence, one blank, and six answer choices. The sentence equivalence questions require test takers to select two of the answer choices. Test takers receive no credit for partially correct answers.

Quantitative Reasoning

The Quantitative Reasoning measure is composed of two sections, 20 questions per section. Students have 35 minutes per section to complete the questions. The Quantitative Reasoning measure assesses basic mathematical concepts of arithmetic, algebra, geometry, and data analysis. The measure tests the ability to solve problems using mathematical models, understand quantitative information, and interpret and analyze quantitative information. There is an emphasis on quantitative reasoning skills, and this measure contains new question types and new computer-enabled tasks. An on-screen calculator is provided in the Quantitative Reasoning measure to reduce the emphasis on computation.

There are four content areas covered on the Quantitative Reasoning measure: arithmetic, algebra, geometry, and data analysis. The content in these areas includes high school mathematics and statistics at a level that is generally no higher than a second course in algebra; it does not include trigonometry, calculus, or other higher-level mathematics. There are four response formats that are used on the Quantitative Reasoning measure: quantitative comparison, multiple-choice select one answer, multiple-choice select one or more answer choices, and numeric entry. Quantitative comparison questions ask test takers to compare two quantities and then determine whether one quantity is greater than the other, if the two quantities are equal, or if the relationship cannot be determined from the information given. Multiple-choice select-one-answer-choice questions ask the test taker to select only one answer choice from a list of five choices. Multiple-choice select-one-or-more-answer-choices questions ask test takers to select one or more answer choices from a list of choices. A question may or may not specify the number of choices to select. Numeric entry questions ask test takers either to enter their answer as an integer or a decimal in a single answer box or to enter their answer as a fraction in two separate boxes, one for the numerator and one for the denominator.

Analytical Writing

The Analytical Writing measure consists of two separately timed analytical writing tasks: a 30-minute analyze an issue (issue) task and a 30-minute analyze an argument (argument) task. The Analytical Writing measure assesses the ability to articulate and support complex ideas, support ideas with relevant reasons and examples, and examine claims and accompanying evidence. The issue task presents an opinion on an issue of general interest, followed by specific instructions on how to respond to that issue. Test takers are required to evaluate the issue, consider its complexities, and develop an argument with reasons and examples to support their views. The argument task requires test takers to evaluate a given argument according to specific instructions. Test takers need to consider the logical soundness of the argument rather than agree or disagree with the position it presents. The two task types are complementary in that one requires test takers to construct their own argument by taking a position and providing evidence supporting their views on an issue, and the other requires test takers to evaluate someone else's argument by assessing its claims and evaluating the evidence it provides. The measure does not assess specific content knowledge, and there is no single best way to respond. The task directions require more focused responses, reducing the possibility of reliance on memorized materials.

In the Analytical Writing measure of the computer-based GRE revised General Test, an elementary word processor developed by ETS is used so that individuals familiar with specific commercial word processing software are not advantaged or disadvantaged. This software contains the following functionalities: inserting text, deleting text, cutting and pasting, and undoing the previous action. Tools such as spelling checker and grammar checker are not available in the software, in large part to maintain fairness with those examinees who handwrite their essays on the paper-based GRE revised General Test.

Conclusion

The goals of designing a revised test that is more closely aligned with the skills needed to succeed in graduate and business school, allows score users to more appropriately distinguish performance differences between candidates, provides enhanced test security, and presents a more test taker–friendly experience were all met in the redesign of the GRE revised General Test. Test volumes are strong, and feedback about the test revisions from the graduate community and test takers alike has been extremely positive.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Notes

[1] The GRE Board was formed in 1966 as an independent board and is affiliated with the Association of Graduate Schools (AGS) and the Council of Graduate Schools (CGS). The GRE Board establishes policies for the GRE program and consists of 18 appointed members.

[2] The GRE General Test was offered in two parts in some regions. The Analytical Writing measure was offered on computer; the Verbal Reasoning and Quantitative Reasoning measures were offered at a paper-based administration a few times per year.

1.2 A Chronology of the Development of the Verbal and Quantitative Measures on the *GRE®* revised General Test

Cathy Wendler

The exploration of possible enhancements to the Verbal Reasoning and Quantitative Reasoning measures of the *GRE®* General Test began with discussions with the *Graduate Record Examinations®* (GRE) Board in 2002 (for Verbal) and 2003 (for Quantitative). Briel and Michel (Chapter 1.1, this volume) provide more detail on the events leading to the decision to revise the GRE General Test.

The goal of these explorations was to ensure continued validity and usefulness of scores. A number of objectives were considered as part of this exploration, including, among others, (a) eliminating question types that did not reflect the skills needed to succeed in graduate school; (b) providing maximum testing opportunities, the availability of a computer-delivered test, and a more *friendly* testing experience overall for all test takers; (c) allowing the use of appropriate technology, such as a calculator; and (d) providing the highest level of test security.

The results of many of these explorations were documented in internal and unpublished papers. This chapter summarizes a number of these papers, in some cases using the words of the authors, as a way of providing the reader with an overview of the extensive efforts undertaken as part of revising the GRE General Test.

Consideration of Various Test Designs

The format of the GRE General Test used prior to August 2011 was a computer adaptive test (CAT). In an adaptive test, the questions administered to test takers depend on their performance on previous questions in the test; subsequent questions that the test takers receive are those that are appropriate to their ability level. The goal of adaptive testing is to improve measurement precision by providing test takers with the most informative, appropriate questions. An artifact of this is that fewer questions are required to obtain a good estimate of test takers' ability levels, resulting in a shorter, but more precise, test.

The model used with the GRE General Test was adaptive at the question level. That is, test takers were routed to their next question based on their performance on the previous question. The introduction of the CAT version of the GRE General Test was innovative and took advantage of technology (that is, computer delivery). However, the CAT design did not allow some of the goals underlying the revision of the test to be attained. As a result, a different test design was needed.

Initially, a computer-delivered linear test (that is, test forms at a particular administration would contain the same set of questions), delivered on a limited number of fixed dates, was considered (see Liu, Golub-Smith, & Rotou, 2006). Between 2003 and 2006, a number of question tryouts, pilot tests, and field trials were run to determine the question types, time limits, and appropriate configurations for the Verbal Reasoning and Quantitative Reasoning measures of the revised test.[1] These studies provided detailed information about the functioning of the linear test. However, in 2006, it became apparent that a fixed administration model using a linear test also would not accommodate all of the original goals of the redesign. During the next year, various evaluations were done to examine alternatives to the linear model. In the end, it was decided that a multistage approach would be used with the Verbal Reasoning and Quantitative Reasoning measures of the revised test.[2]

The multistage adaptive test (MST) is adaptive at the stage (section), not question, level, and the determination of the next set of questions an examinee receives is based on performance on an entire preceding stage. The MST model allows for frequent test administrations while still providing score accuracy (Robin & Steffen, Chapter 3.3, this volume). A number of pilots and simulations were undertaken beginning in 2007 to determine the final number of stages, section configurations (number and types of questions), and timing for the MST GRE General Test.

Initial Concept: A Revised Linear Test

Verbal Pilots

The Verbal Reasoning measure of the GRE General Test measures test takers' ability to analyze and evaluate written material and to synthesize that information. It also measures test takers' ability to recognize relationships among words and concepts and among parts of sentences. One of the reasons for revisiting the Verbal Reasoning measure was the desire to remove those question types that did not reflect the skills needed to succeed in graduate school (i.e., the analogy and antonym question types). Analogies and antonyms rely heavily on test takers' vocabulary knowledge, are short, and are easily memorized question types. In May and June of 2003, seven potential new Verbal Reasoning question types were examined as part of a pilot study (Golub-Smith, 2003). These question types included (a) paired passages with a set of bridging questions, (b) missing sentence (in a passage), (c) extended text completions (with two or three blanks), (d) logical reasoning, (e) antonyms in context, (f) synonyms in context, and (g) paragraph reading (100-word paragraph followed by a single question). The goal of the pilot was to examine the statistical characteristics of the new question types, the completion time required for the questions, and whether differences in performance among subgroups would be increased. Results of the pilot provided support for further explorations: The questions were appropriately difficult and discriminated between high- and low-ability test takers; took considerably longer to answer than antonym and analogy questions; and did not exacerbate score differences between males and females and White, Black, Hispanic, and Asian test takers.

To evaluate these seven question types further, a factor analysis using a small number of the new and current (i.e., analogy, antonym, reading comprehension, and sentence completion) question types was run (Golub-Smith & Rotou, 2003). The data came from the spring[3] 2003 pilot described above. The goal of the analysis was to determine if the new question types were measuring the same verbal construct as the current question types. Factor analyses were run for each section and on various combinations of the sections.

Golub-Smith and Rotou (2003) found that, overall, two factors were observed for the analyses using the combined sections. Factor I appeared to be a global verbal factor, which included understanding text and, to a lesser extent, vocabulary knowledge. Factor II seemed to be related to a complex interaction with question difficulty and location. The analyses on the sections containing current questions yielded results similar to the analyses of the combined sections, with two factors being observed. However, only one factor emerged when the sections composed of new question types were analyzed. This finding was not surprising, given that the sections were small and consisted of only a few questions. While there were several limitations to the study, the results still contributed to the redesign efforts. In particular, the results provided evidence that replacing the analogy and antonym questions with new question types did not appear to change the construct being measured.

As described in Golub-Smith, Robin, and Menson (2006), nine additional Verbal pilots were conducted between December 2003 and April 2004. The goals of the pilots included (a) examining the performance of the potential new question types and their feasibility for use, (b) refining the construction of the new question types, (c) examining possible test configurations, and (d) determining appropriate timing for the new questions.

Based on the results of these pilots, a study was conducted in fall[4] 2004 using a prototype configuration for the revised linear test (Golub-Smith et al., 2006). Six Verbal question types were included: text completions one,[5] two, and three blanks; sentence equivalence;[6] logical reasoning; paragraph reading (120 words); short reading (150 words); and long reading (450 words).[7] The study included different full-length configurations of Verbal measure sections, designed to cover the full range of question types and various mixtures of passage-based and discrete questions.

The configurations were evaluated using specific criteria: reliability, distributional characteristics, reproducibility, impact on question production, timing, subgroup impact, domestic[8] versus Asian[9] test takers' performance differences, and construct validity. None of the configurations met all of the criteria. Thus, the configuration chosen to be included in the subsequent field test described below was a hybrid of several configurations.

A Verbal field test study was held between March and May 2005 (Golub-Smith et al., 2006). The field test had three purposes: (a) to evaluate the psychometric properties of the field test configuration, (b) to compare the field test form to the old Verbal measure, and (c) to examine timing issues. Two new forms and one old form of the test were used in the study. Students from 54 institutions were included in the field trial.

As described in Golub-Smith et al. (2006), results of the Verbal field test study provided support for the use of the new question types. In particular, the following were observed: (a) the new Verbal forms were more difficult than the old form; (b) as expected, the domestic group performed better than a small international group composed of test takers who were non–U.S. citizens attending schools in the United States or Canada; (c) internal consistency estimates of reliability for the new forms were acceptable; (d) standard errors of measurement for the new forms built to different specifications were reasonably similar to those of the old form; (e) correlations of the total scores between the old and new forms indicated a moderately high relationship between the two measures; (f) correlations between the discrete and passage-based questions indicated more structural cohesiveness for the new forms compared to the old form; (g) most participants had adequate time to complete the new forms; and (h) differences in subgroup performance on the field test forms were similar to those on the old form.

Quantitative Pilots

The Quantitative Reasoning measure of the GRE General Test measures test takers' ability to understand, interpret, and analyze quantitative information; to solve problems using mathematical models; and to apply basic mathematical skills and concepts. One of the goals in the revision of the Quantitative measure was to better align the skills assessed in the test with the skills needed to succeed in graduate school. As a result, potential new types of question formats were developed for the Quantitative measure. These formats allowed the assessment of quantitative reasoning skills in ways not possible using standard, single-selection multiple-choice questions. The new question types were designed to ask test takers to evaluate and determine the completeness of their responses. In addition, the proportion of real versus pure mathematics questions[10] was to be increased, the proportion of geometry questions decreased, and on-screen calculators introduced. The reader should also refer to Bridgeman, Cline, and Levin (2008; Chapter 1.5 in this volume) for a discussion on the impact of calculator availability on Quantitative questions.

Between 2004 and 2005, six pilot studies were conducted on the potential new Quantitative question types (Rotou, 2007a). Some of the issues addressed in the pilots included the comparability of the new question types with standard multiple-choice questions, the number and composition of data interpretation sets (a set is composed of questions that share the same stimulus), appropriate time limits for the new question formats, and possible configuration designs for the measure (e.g., total number of questions, number of new question types in a section). Each of these pilots provided specific information about potential changes to the Quantitative Reasoning measure of the GRE General Test.

The first Quantitative pilot study was conducted in April 2004 (Steffen & Rotou, 2004a). Four new question types were included in the study: (a) numeric entry (test takers calculate their answer and entered it using the keyboard), (b) multiple-selection multiple choice (test takers select one or more answer choices), (c) order match (test takers select a response that constructs a statement), and (d) table grid (test takers determine if a statement is true or false). The goal of the pilot was to examine the comparability of the new question types with the standard multiple-choice questions used in the current version of the GRE General Test. Test takers who had recently taken the GRE General Test were recruited to participate in the pilot. Sections containing the new question types were created and paired with five sections that included standard multiple-choice questions. The sections were designed so that each standard multiple-choice question had a corresponding new question type measuring the same reasoning skill in a paired section. Results indicated that the new format questions tended to be more difficult, more discriminating, and require more time than the standard multiple-choice questions.

In September 2004, a pilot was conducted to further examine the psychometric properties of the new questions, question timing, the impact of question position on question and section performance, and the number of questions that could be associated with a common stimulus in the data interpretation sets (Steffen & Rotou, 2004b). Some of the sections were the same as those delivered in the April 2004 pilot and consisted of a mix of standard multiple-choice and new question types. Other sections consisted of the same questions as in the first sections, but in different orders to examine question position effects. Other sections consisted of data interpretation sets with various numbers of questions (two, three, four, and five). All sections were administered in the research section of the operational GRE General Test.

As described in Steffen and Rotou (2004b), results indicated consistency in terms of question statistics (e.g., difficulty and discrimination) across the two pilots. In addition, question position did not appear to impact examinee performance on the question or the section. The length of the data interpretation sets had no effect on the question statistics and suggested that the number of questions associated with each set should range from four to five. In addition, differences in subgroup performance (male-female students; White-Black, White-Asian, and White-Hispanic students) were examined. Results indicated that the use of the new question types did not appear to increase the standardized differences between groups.

The data interpretation sets were further evaluated in another pilot administered in October 2004 (Rotou, 2004b). This pilot examined the number of data interpretation sets on the test and the composition[11] of the sets. In addition, start-up effects, defined as effects that occur when questions appearing as the first question in a set of questions require more time to complete than similar, subsequent questions, were examined. Test takers who had recently taken the GRE General Test volunteered to participate in the pilot. Results indicated that the number and composition of the sets had no impact on participant performance or section reliability. In addition, no start-up effects were apparent.

In December 2004, a pilot was conducted to collect additional information about the psychometric properties and timing of the new question types (Rotou, 2004a). Test sections consisting of a mix of new question types and standard multiple-choice questions were administered in the research section of the operational GRE General Test. Results were consistent with the previous pilots and indicated that the new question types had higher discrimination levels and required more time compared to the standard multiple-choice questions.

A final pilot study was conducted in January 2005 to evaluate possible configuration designs for the revised test (Rotou & Liu, 2005). The study examined the proportion of real versus pure questions, total number of questions, and the number of new question types in a section. A number of pilot sections were created and administered in the research section of the operational GRE General Test using different time limits. Results indicated that the configuration of the section (total number of questions, number of new question types, and proportion of context-based questions) had no significant impact on performance. This result was seen for both domestic and international test takers. Section configuration also did not seem to have an impact on section time, although international test takers tended to take more time than domestic test takers. As expected, those sections containing more questions displayed higher reliability levels.

Based on the results of the earlier pilots, a configuration study was conducted in May 2005 (Rotou, Liu, & Sclan, 2006).The study examined possible configuration designs with the goal of determining the best configuration and statistical specifications for the linear test. Four new question types were included in the study: (a) numeric entry, (b) multiple-selection multiple choice, (c) order match, and (d) table grid.

Three different configuration designs were used. The total number of questions and the number of new question types varied across the configurations. The first configuration included only standard multiple-choice questions but allowed the use of a calculator. The other two configurations included new question types. In order to calibrate all sections concurrently, some question overlap was used. The pilot sections were delivered in the research section in the operational GRE General Test. Since only 40 minutes are allocated to the research section, it was not possible to administer an entire full-length configuration to each examinee. However, even though each examinee only took a half-length form, a statistical method (i.e., item response theory) was used to estimate the properties of a full-length test.

As summarized in Rotou et al. (2006), results indicated that the amount of time spent on the section was similar across all configurations. About 50% of the domestic test takers who indicated English is their first language completed the section in about 31–35 minutes, 75% completed it in 37–40 minutes, and 90% completed it in about 40 minutes. International test takers spent more time completing the section than did the domestic test takers: 50% completed the section in 35–38 minutes, while 75% completed the section using the maximum amount of time. Examinee performance, based on percentage correct, was similar for the sections containing the new question types. International test takers performed better than the domestic test takers on all sections. Finally, standardized differences between male and female test takers were similar to those seen with operational scores. Results for the comparison between Black and White test takers, however, indicated that the standardized differences for the pilot sections were somewhat *smaller* than those seen with operational scores.

Based on the results of this study, it was proposed that the configuration used with the revised linear test consist of the one with the least number of questions. In order to ensure that there is meaningful information at the top end of the scale, it was also recommended that the statistical specifications be made slightly more difficult than those used in the configuration study.

Combined Verbal and Quantitative Field Trial

A large field trial for the revised linear GRE General Test combining the Verbal Reasoning, Quantitative Reasoning, and Analytical Writing measures was conducted between October 2005 and February 2006. The goals of the field trial included evaluating the measurement characteristics of the revised linear test, determining the adequacy of the statistical specifications for the revised test, and confirming the timing and section configurations. Golub-Smith (2007) detailed the results of the field trial for the Verbal measure, and Rotou (2007b), the results for the Quantitative measure. A brief summary is presented below.

Participants in the field trial included test takers who had recently taken or were planning to take the GRE General Test. Participants were paid for their time and were given the chance to win one of 10 awards of $250; this was done to ensure that participants were appropriately motivated to perform their best on the field trial test. Additional screening analyses were done after the field test was completed to ensure that the final sample consisted of only participants who took the test seriously. The final sample consisted of approximately 3,190 participants. The participants used in the study adequately represented the 2005 GRE General Test test-taking population. However, a comparison of means and standard deviations with the operational scores of the study participants indicated that they were, on average, a more able group than the full GRE General Test test-taking population.

Two forms were administered at 43 domestic and six international test centers. The two forms were created as parallel forms; they shared a set of common questions that allowed performance from the different forms to be linked to each other. Four versions of each of the two forms were created, resulting in eight different test versions. The versions differed in terms of the order in which the Verbal and Quantitative measures were given (i.e., whether Verbal or

Quantitative came first and whether two sections of the same measure were delivered sequentially or alternated with sections of the other measure). Two Analytical Writing prompts, one issue and one argument, were always given prior to the first Verbal or Quantitative measure. The readers should see Robin and Zhao (Chapter 1.8, this volume) for a discussion of the configuration study for Analytical Writing.

Results of the field trial, described by Golub-Smith (2007) and Rotou (2007b) include the following:

- Analyses of the raw scores for each section found no significant differences in the total score across the forms. In addition, no significant differences were seen in performance based on the order of the Verbal and Quantitative measures.
- A review of the question-level statistics indicated that both the Verbal and Quantitative field trial forms appeared to be easier than would have been expected based on pretest statistics. This may have been due to the field test group being more motivated than the group used to obtain the pretest statistics.
- Overall standard errors of measurement were comparable across the domestic and international groups. In addition, the correlations between the Verbal and Quantitative measures were similar for the domestic and international groups. Reliability estimates for the field trial forms were acceptable for both the domestic and international groups.
- Mixed results were found for the timing analyses. As indicated by Golub-Smith (2007), very few domestic participants spent the entire allotted 40 minutes on each of the Verbal measure sections, and 80% or more reached the last question in all but one section. However, as she pointed out,

  > The use of [this] criterion is only meaningful if it is based on test takers who spend the total allotted time in a section. If an examinee does not reach the end of the test but spends less than 40 minutes in a section, one can assume factors other than speededness were the cause, for example, fatigue or lack of motivation. (Golub-Smith, 2007, p. 11)

- Rotou (2007b) indicated that timing results for the Quantitative measure sections showed that these sections were somewhat speeded. The percentage of domestic participants who spent the entire time on a Quantitative measure section ranged from 24% to 47%, and between 69% and 83% reached the last question. Based on these data, it was decided to reduce the number of questions in the revised Quantitative measure.

Overall, the results of the field trial indicated that the measurement properties of the field test forms were acceptable and allowed the statistical specifications for the revised linear test to be finalized.

Rethinking the Concept: A Multistage Test

The decision to move to an MST for the Verbal and Quantitative measures required that additional studies be completed. While the Analytical Writing measure did not change in that test takers would still respond to two different essays, there were changes in the prompts themselves. Essay variants were created by using a given prompt (parent) as the basis for one or more different variants that require the examinee to respond to different writing tasks for the same stimulus. The reader should refer to Bridgeman, Trapani, and Bivens-Tatum (2011; Chapter 1.10 this volume) for a detailed discussion of the comparability of essay variants.

The earlier pilots and field trials conducted on the linear version of the test provided foundational information as to the functioning of the new question types, data regarding timing issues and potential section arrangements, and insight into subgroup performance. While it was desirable for the testing time to remain similar to that used with the CAT version, analyses indicated that the MST needed to be longer than the CAT in order to maintain adequate reliability and measurement precision.

Therefore, the best structure for the MST had to be determined. Decisions related to the appropriate overall test length, the optimal number of stages, the optimal number of questions and time limit for each stage, and the final test specifications (i.e., content mix as well as the psychometric specifications) needed to be made.

As a first step, a series of simulation analyses were run, examining possible configurations for the MST (Lazer, 2008). Configurations containing different numbers of stages (e.g., 1-2, 1-2-3) were examined with a goal of selecting the simplest and most effective design that would meet the required test specifications. Total test length (e.g., 35, 40, or 45 questions) and number of questions per stage were evaluated. For example, for a 40-question test containing two stages, the first stage might contain 10 questions followed by a 30-question second stage or 15 questions followed by 25 questions or 20 questions each stage and so forth. For a 40-question test containing three stages, the first stage might contain 10 questions, the second 10 questions, the third 20 questions; or 15 questions followed by 15 questions followed by 10 questions. In addition, various psychometric indicators were examined: the distribution of the discrimination indices for the questions (e.g., uniform across all stages, maximum information provided in first stages, or maximum information provided in later stages); the range of question difficulty by stage; and routing thresholds (i.e., level of performance required to route test takers to the next stage).

Results of these simulations indicated that 40 questions for both the Verbal and Quantitative measures were appropriate (Lazer, 2008). The results also indicated that a simple

MST model was the most efficient because it provided routing accuracy as well as providing test takers with the appropriate difficulty level of questions to ensure measurement precision.

During spring 2009, pilots were conducted using the research section of the operational GRE General Test (Liu & Robin, 2009; Zhao & Robin, 2009a). The goals of the pilots were to evaluate test length and timing for the MST, evaluate different question configurations, and, as possible, evaluate subgroup impact. Multiple MST sections were created for the Verbal and Quantitative measures, reflecting various combinations of MST stage, level of difficulty, section timing, and number and types of questions. Each examinee who voluntarily responded to the research section received only one MST section; some sections were deliberately administered to more test takers than others. The number of test takers included in the analyses ranged from 149 to 899, depending upon the section.

As indicated in Liu and Robin (2009) and Zhao and Robin (2009a), results for the Verbal and Quantitative measures were similar. No significant differences were seen between Verbal configurations, and the 20-question sections appeared to work best for Quantitative. Most test takers answered all of the questions in the research section, and very few spent the total allotted time, regardless of the number and types of questions or level of difficulty. Subgroup comparisons indicated that male test takers tended to outperform female test takers.

To examine the composition of the data interpretation questions further, an additional pilot was conducted in summer[12] 2009 (Zhao & Robin, 2009b). The study was designed to understand the impact on test performance if one of the data interpretation set questions was replaced with a discrete data question. Again, the pilot was conducted using the research section of the operational GRE General Test. Multiple versions of the Quantitative MST were developed; each examinee who voluntarily responded to the research section took only one version. About 9,600 test takers were included in the analysis. Results indicated that, in general, replacing one of the data interpretation set questions with a discrete question did not influence examinee performance. In addition, the inclusion of the discrete question appeared to reduce the time requirements slightly for two thirds of the MST versions. The final conclusion was that replacement of a data interpretation set question with a comparable discrete question was an acceptable option.

Conclusion: The GRE revised General Test

Based on the results of a decade of studies, the GRE revised General Test was launched in fall 2011. The test is administered using an Internet-based testing platform in testing centers around the globe, ensuring accessibility and convenience for the maximum number of test takers. The structure of the test includes two 30-minute Verbal Reasoning measure sections containing 20 questions each, two 35-minute Quantitative Reasoning measure sections containing 20 questions each, and the Analytical Writing measure containing two essays. The Verbal measure includes four new question types (text completion [with one, two, or three blanks], sentence equivalence, select-in-passage, and multiple-selection multiple choice), as well as standard multiple-choice questions. The Quantitative measure includes two new question types (multiple-selection multiple choice and numeric entry), as well as quantitative comparison and standard multiple-choice questions.

The revised test provides many advantages to test takers—such as the ability to review and change answers, the opportunity to skip a question and revisit it later, and an on-screen calculator—as well as providing enhanced measurement precision (Robin & Steffen, Chapter 3.3, this volume). Ultimately, the goals set forth by the GRE Board when approving the exploration of revisions to the test were met.

References

Bridgeman, B., Cline, F., & Levin, J. (2008). *Effects of calculator availability on GRE Quantitative questions* (Research Report No. RR-08-31). Princeton, NJ: Educational Testing Service.

Bridgeman, B., Trapani, C., & Bivens-Tatum, J. (2011). Comparability of essay question variants. *Assessing Writing, 16*, 237–255.

Golub-Smith, M. (2003). *Report on the results of the GRE Verbal pilot.* Unpublished manuscript, Educational Testing Service, Princeton, NJ.

Golub-Smith, M. (2007). *Documentation of the results from the revised GRE combined field test Verbal measure.* Unpublished manuscript, Educational Testing Service, Princeton, NJ.

Golub-Smith, M., Robin, F., & Menson, R. (2006, April). *The development of a revised Verbal measure for the GRE General Test.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Golub-Smith, M., & Rotou, O. (2003). *A factor analysis of new and current GRE Verbal item types.* Unpublished manuscript, Educational Testing Service, Princeton, NJ.

Lazer, S. (2008, June). *GRE redesign test design update.* Presentation made at the GRE Board meeting, Seattle, WA.

Liu, J., & Robin, F. (2009). *March/April field test analyses summaries—Verbal.* Unpublished manuscript, Educational Testing Service, Princeton, NJ.

Liu, M., Golub-Smith, M., & Rotou, O. (2006, April). *An overview of the context and issues in the development of the revised GRE General Test.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Rotou, O. (2004a). *December quantitative research pilot: Psychometric properties and timing information of the novel response item formats.* Unpublished manuscript, Educational Testing Service, Princeton, NJ.

Rotou, O. (2004b). *Quantitative rapid pilot two: The structure of data interpretation sets.* Unpublished manuscript, Educational Testing Service, Princeton, NJ.

Rotou, O. (2007a). *Development work for the GRE Quantitative measure.* Unpublished manuscript, Educational Testing Service, Princeton, NJ.

Rotou, O. (2007b). *Documentation of the results from the rGRE combined field test Quantitative measure.* Unpublished manuscript, Educational Testing Service, Princeton, NJ.

Rotou, O., & Liu, M. (2005). *January configuration study for the Quantitative measure.* Unpublished manuscript, Educational Testing Service, Princeton, NJ.

Rotou, O., Liu, M., & Sclan, A. (2006, April). *A configuration study for the Quantitative measure of the new GRE.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Steffen, M., & Rotou, O. (2004a). *Quantitative rapid pilot one: Psychometric properties and timing information of the novel response item formats.* Unpublished manuscript, Educational Testing Service, Princeton, NJ.

Steffen, M., & Rotou, O. (2004b). *September quantitative research pilot: Impact of item sequence on performance.* Unpublished manuscript, Educational Testing Service, Princeton, NJ.

Zhao, J., & Robin, F. (2009a). *Summary for the March/April 2009 package field test data for the GRE Quantitative measure.* Unpublished manuscript, Educational Testing Service, Princeton, NJ.

Zhao, J., & Robin, F. (2009b). *Summary of the GRE Quantitative July/August 2009 package field test results.* Unpublished manuscript, Educational Testing Service, Princeton, NJ.

Notes

[1] See Robin and Zhao (Chapter 1.8, this volume) for information on the configuration study for Analytical Writing measure.

[2] The Analytical Writing measure is not adaptive. Examinees respond to two different essays during the test administration.

[3] *Spring* refers to data collected sometime during January through July.

[4] *Fall* refers to data collected sometime during August through December.

[5] The one-blank text completion question was a reformatted version of the previous sentence completion question type.

[6] The sentence equivalence questions evolved from the vocabulary (synonyms) in context question type.

[7] Short reading and long reading were question types used on the CAT version of the test.

[8] *Domestic* refers to test takers who indicated they are U.S. citizens and took the test in a test center in the United States or a U.S. territory.

[9] *Asian* refers to test takers who indicated they are citizens of Taiwan, Korea, Hong Kong, or China.

[10] Real mathematics questions reflect a real-world task or scenario-based problem, while pure mathematics questions deal with abstract concepts.

[11] The composition of the data interpretation set refers to the number and type (e.g., new question types, standard multiple choice) of questions associated with a particular set.

[12] *Summer* refers to data collected sometime during July and August.

1.3 Supporting Efficient, Evidence-Centered Question Development for the *GRE*® Verbal Measure[1]

Kathleen Sheehan, Irene Kostin, and Yoko Futagi

New test delivery technologies, such as Internet-based testing, have created a demand for higher capacity question-writing techniques that are (a) grounded in a credible theory of domain proficiency and (b) aligned with targeted difficulty specifications. This paper describes a set of automated text analysis tools designed to help test developers more efficiently achieve these goals. The tools are applied to the problem of generating a new type of Verbal Reasoning questions called the paragraph reading (PR) question. This new question type was developed for use on the *GRE*® revised General Test. It consists of a short passage, typically between 90 and 130 words, followed by two, three, or four questions designed to elicit evidence about an examinee's ability to understand and critique complex verbal arguments such as those that are typically presented in scholarly articles targeted at professional researchers. This new question type was developed at ETS as part of an ongoing effort to enhance the validity, security, and efficiency of question development procedures for the GRE General Test.

Two different approaches for enhancing the efficiency of the PR question development process are considered in this paper. The first approach (Study 1) focuses on the passage development side of the question writing task; the second approach (Study 2) focuses on the question development side of that task.

Study 1

The approach in Study 1 builds on previous research documented in Sheehan, Kostin, Futagi, Hemat, and Zuckerman (2006) and Passonneau, Hemat, Plante, and Sheehan (2002). This research was designed to capitalize on the fact that, unlike some testing programs that employ stimulus passages written from scratch, all of the passages appearing on the GRE Verbal measure have been adapted from previously published source texts extracted from scholarly journals or magazines. Consequently, in both Sheehan et al. (2006) and Passonneau et al. (2002), the problem of helping question writers develop new passages more efficiently is viewed as a problem in automated text categorization. These latter two studies documented the development and validation of an automated text analysis system designed to help test developers find needed stimulus materials more quickly. The resulting system, called SourceFinder, includes three main components: (a) a database of candidate source documents downloaded from appropriately targeted online journals and magazines, (b) a source evaluation module that assigns a vector of acceptability probabilities to each document in the database, and (c) a capability for efficiently searching the database so that users (i.e., question writers) can restrict their attention to only those documents that have been rated as having a relatively high probability of being acceptable for use in the particular source-finding assignment at hand. The SourceFinder database currently includes more than 90,000 documents downloaded from over 60 different journals and magazines designated as potentially appropriate for use in developing new passages and questions for the GRE revised General Test. Estimates of the acceptability status of each document, relative to a specified number of potential passage development assignments, are stored along with each document. These estimates enable question writers to limit their search to only those documents that have been rated as having a relatively high probability of being acceptable for use in satisfying the particular passage development assignment at hand.

Since PR passages are developed from paragraphs, as opposed to entire documents, however, the goal of this study was to assign a PR-specific acceptability rating to each paragraph in the database. Test developers could then use these new estimates, in combination with SourceFinder’s existing search capability, to restrict their attention to only those paragraphs that had been rated as having a relatively high probability of being acceptable for use in developing a new PR passage.

Method

A sample of 114 paragraphs from the SourceFinder database was selected and rated for acceptability by two GRE test developers. Raters were asked to provide two types of ratings: (a) a quantitative estimate of each paragraph’s acceptability status expressed on a scale of 1 (*definitely reject*), 2 (*probably reject*), 3 (*uncertain*), 4 (*probably accept*)*,* and 5 (*definitely accept*), and (b) a brief, written description of the aspects of text variation considered during the evaluation process. Next, hypotheses were generated about the aspects of text variation that might account for the observed similarities in the comments provided for similarly rated paragraphs. Then, natural language processing tools were developed to automatically extract candidate explanatory features. Finally, statistical models were developed to generate predictions of text acceptability that closely reflected the ratings provided by the test developers. These models were then validated on a sample of 1,000 paragraphs that were not used in the model development but that had been evaluated on the 5-point acceptability scale as part of the operational development work by test developers.

Results

The validation results confirmed that the proposed filtering technique can help question writers increase the percentage of acceptable stimulus paragraphs located per unit time interval from the current level of about 10% to nearly 30%.

Conclusion

Because the process of locating acceptable source material is one of the most time-consuming parts of the question development process, methods developed in this study should translate directly into efficiency gains. Indeed, the algorithms implemented to achieve this increase have already been incorporated into the operational SourceFinder system and are available to GRE question writers.

Study 2

Techniques for facilitating question development efficiency have been discussed by Sheehan and colleagues (Sheehan, 1997, 2003; Sheehan et al., 2006; Sheehan & Mislevy, 1990). This research demonstrated that question writers can work more efficiently by generating new question that conform to prespecified task models designed to provide unambiguous evidence about examinees' mastery status on targeted proficiencies. Study 2 focused on additional tools to help question writers write questions that are optimally configured to provide unambiguous evidence on the examinees' mastery status.

Method

Information on 125 PR questions that had been reviewed and pretested was assembled. The information included question difficulty (proportion of examinees getting the right answer), question discrimination (how well the question separates students with strong skills from those with weaker skills), question type classifications (inference, primary purpose, rhetorical purpose, and vocabulary in context), and question format classifications (multiple choice, highlight sentence, and select all correct options). Next, hypotheses were developed that focused on text characteristics that may either facilitate or impede an examinee's ability to develop a mental representation that is sufficiently rich to distinguish among the various options presented with an question. A tree-based regression approach (Brieman, Friedman, Olshen, & Stone, 1984) was used to confirm the hypotheses about critical task features and associated skills that were developed. The proposed models were then validated by considering the percentage of difficulty variance accounted for by the specified question classifications (i.e., how effective the model is in identifying the proportion of examinees who will get the question right).

Results

Difficulty variance accounted for by the model ranged from slightly more than 30% for questions designed to test vocabulary skills to slightly more than 40% for questions designed to test additional verbal reasoning skills, such as generating near and far inferences and understanding complex oppositional reasoning. Thus, the model should be useful for helping test developers write questions at the required difficulty levels.

Conclusion

From the test development perspective, the results are useful for facilitating targeted, evidence-centered question generation. The GRE question writers should be able to use the models to generate new questions that provide more precise evidence about the targeted skills and that, as a result, are more likely to scale at targeted difficulty levels. The detailed information about required skills developed in this study also may be used to describe critical construct elements in ways that may be more illuminating to students, admissions officers, and other GRE stakeholders. Gitomer and Bennett (2002) referred to this as unmasking the construct and argued that test designers have an obligation to present such information to test users. The student, evidence, and task models developed in this study provide a straightforward approach for satisfying that obligation.

References

Brieman, L., Friedman, J. H., Olshen, R., & Stone, C. J. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth.

Gitomer, D. H., & Bennett, R. E. (2002). *Unmasking constructs through new technology, measurement theory, and cognitive science* (Research Memorandum No. RM-02-01). Princeton, NJ: Educational Testing Service.

Passonneau, R., Hemat, L., Plante, J., & Sheehan, K. (2002). *Electronic sources as input to GRE reading comprehension item development: SourceFinder prototype evaluation* (Research Report No. RR-02-12). Princeton, NJ: Educational Testing Service.

Sheehan, K. M. (1997). A tree-based approach to proficiency scaling and diagnostic assessment. *Journal of Educational Measurement, 34*, 333–352.

Sheehan, K. M. (2003). Tree-based regression: A new tool for understanding cognitive skill requirements. In H. F. O'Neil & R. S. Perez (Eds.), *Technology applications in education: A learning view* (pp. 222–227). Mahwah, NJ: Erlbaum.

Sheehan, K. M., Kostin, I., Futagi, Y., Hemat, R., & Zuckerman, D. (2006). *Inside SourceFinder: Predicting the acceptability status of candidate reading comprehension source documents* (Research Report No. RR-06-24). Princeton, NJ: Educational Testing Service.

Sheehan, K. M., & Mislevy, R. (1990). Integrating cognitive and psychometric models to measure document literacy. *Journal of Educational Measurement, 27*, 1–18.

Notes

[1] Based on *Supporting Efficient, Evidence-Centered Item Development for the GRE® Verbal Measure* (GRE Board Research Report No. 03-14), by K. M. Sheehan, I. Kostin, and Y. Futagi, 2007, Princeton, NJ: Educational Testing Service.

1.4 Transfer Between Variants of Quantitative Questions [1]

Mary Morley, Brent Bridgeman, and René Lawless

Using test development software, ETS is able to create close variants of *GRE®* Quantitative questions with the intention that they will be psychometrically and otherwise exchangeable and equivalent with the base questions (Bejar et al., 2002). Using base questions as models, this tool allows for many immediately usable variant questions to be generated in an efficient manner. Ideally, it is possible to use different close variants with different test takers and effectively lengthen the shelf life of a question model beyond that of a single question, which, for security purposes, may only be used in an operational test a limited number of times.

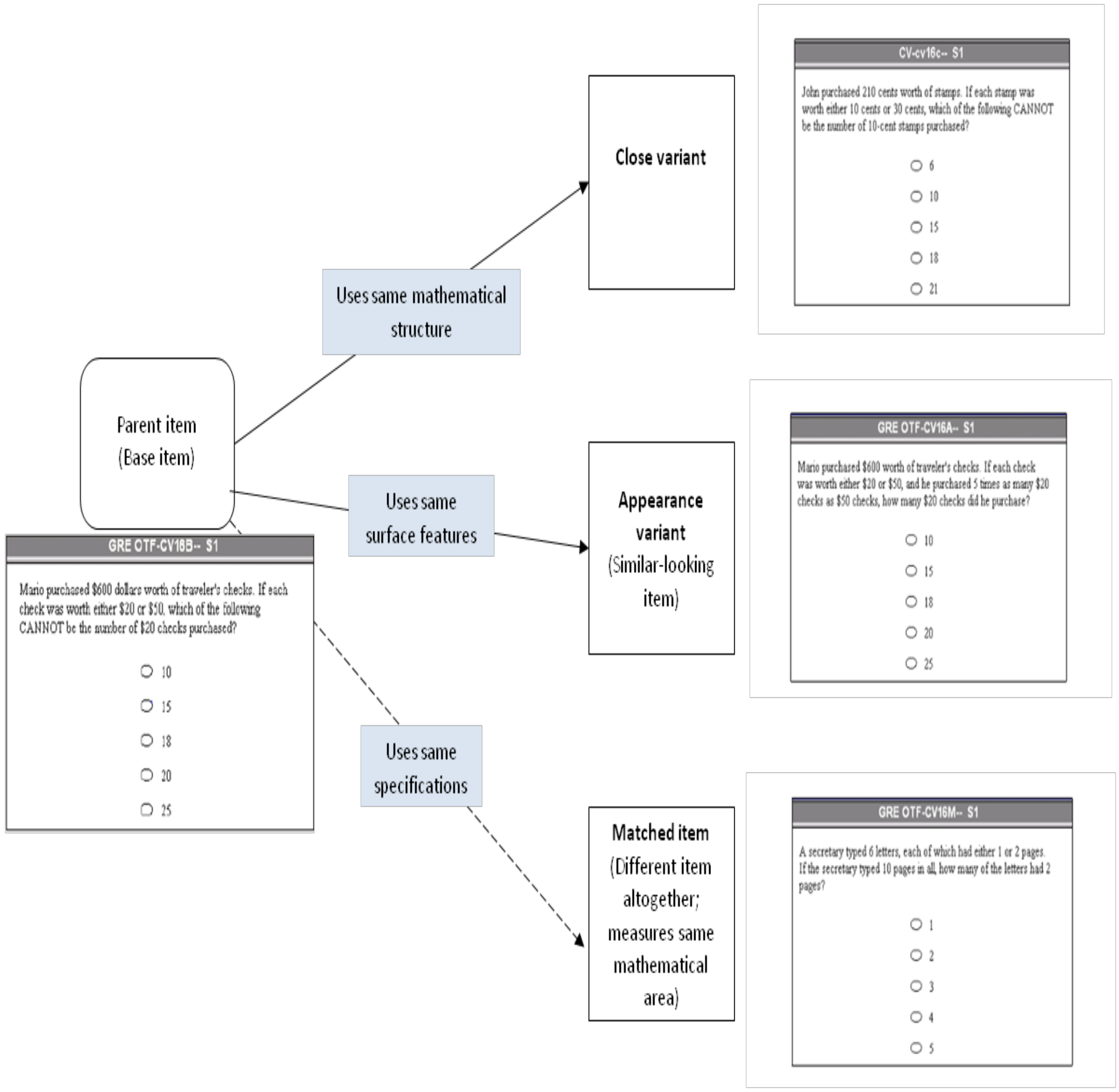
At the time of this study, ETS had instituted a number of policies to minimize the potential threat that question variants might have to test security and score validity. These policies assumed that test takers *would* be able to transfer solution strategies across variants and thereby obtain a solution inappropriately. In other words, there was concern that test takers could memorize rules that are only useful for answering questions belonging to a narrow class and these rules may lead to the right answer for the wrong reasons (e.g., if the question involves jelly beans, just add the two numbers given in the problem to get the answer).

The purpose of this study was to investigate the extent to which test takers were able to transfer solution strategies from one question to a variant question. In order to do this, the authors isolated three types of question variations possible for a question model: (a) matched questions, which were similar in the sense that they address the same content area (e.g., algebra), were about the same level of difficulty, had the same formatting (e.g., multiple choice), and had the same context (e.g., word problem vs. pure math problem); (b) close variants, which were essentially the same question mathematically, with altered surface features (e.g., names, numbers, and contexts); and (c) appearance variants, which were similar in surface features only (e.g., names, figures, etc.) and required different mathematical operations to solve. Examples of a base question and each variation type appear in Figure 1.4.1.

Novick (1988) demonstrated that students with a deeper understanding of the mathematics behind a presented problem are able to recognize the structural features and are better equipped to correctly solve the problem. Conversely, novices are reliant on the surface features of the new problem, as they are more salient. This information has several implications for this study and helped to define these research questions:

- Will transfer occur between close variants?
- Will the presence of appearance variants influence student performances on close variants in the same test?

- Is transfer related or associated with any student characteristics (e.g., ability level, ethnicity) or question characteristics (e.g., question format)?


Parent item
(Base item)
Uses same mathematical structure
Close variant
Uses same surface features
Appearance variant
(Similar-looking item)
Uses same specifications
Matched item
(Different item altogether; measures same mathematical area)

GRE OTF-CV16B-- S1

Mario purchased $600 dollars worth of traveler's checks. If each check was worth either $20 or $50, which of the following CANNOT be the number of $20 checks purchased?

○ 10
○ 15
○ 18
○ 20
○ 25

CV-cv16c-- S1

John purchased 210 cents worth of stamps. If each stamp was worth either 10 cents or 30 cents, which of the following CANNOT be the number of 10-cent stamps purchased?

○ 6
○ 10
○ 15
○ 18
○ 21

GRE OTF-CV16A-- S1

Mario purchased $600 worth of traveler's checks. If each check was worth either $20 or $50, and he purchased 5 times as many $20 checks as $50 checks, how many $20 checks did he purchase?

○ 10
○ 15
○ 18
○ 20
○ 25

GRE OTF-CV16M-- S1

A secretary typed 6 letters, each of which had either 1 or 2 pages. If the secretary typed 10 pages in all, how many of the letters had 2 pages?

○ 1
○ 2
○ 3
○ 4
○ 5

Figure 1.4.1. Examples of a base question and its close, appearance, and matched variant questions.

Method

Data were collected from 406 undergraduate college students who had not previously taken the GRE General Test. Testing was conducted in four sites: East Lansing, New Orleans, Philadelphia, and Princeton. The sample contained 64% female and 72% White students. Six pretest forms and one posttest form, each consisting of 27 questions, were developed for a computer-based test format. Each form was designed to span the levels of difficulty typically found on paper-and-pencil GRE test forms. GRE mathematics test developers reviewed all of the close variants and appearance variants for content and correspondence to their respective base questions. The experimental manipulation was accomplished by randomly assigning participants to different pretests; all completed the same posttest.

The posttest consisted of retired GRE questions that approximately met the specifications for an actual GRE quantitative test. The 27 questions selected for the posttest served as base questions from which the variants in the pretest were developed. The researchers divided these questions into three sets of nine questions each, and then put together six different pretest forms each containing a different configuration of the three sets (see Table 1.4.1). Each set of questions contained only one type of variant (close, matched, or appearance). As an example of form configuration, Form 2 was composed of matched, close, and matched sets, respectively, while Form 5 contained matched, close, and appearance sets, respectively.[2] Participants were randomly assigned to one of the six forms. This elaborate design was necessary in order to isolate the transfer effects of each variant type. At the beginning of the pretest, participants' computers displayed general directions, along with a message making them aware that some of the problems in the second test resembled problems in the first test, either visually or mathematically. Upon pretest completion, participants were then administered the posttest.

Table 1.4.1

Pretest Form Configuration

Form	Set A	Set B	Set C
1	Close	Matched	Matched
2	Matched	Close	Matched
3	Matched	Matched	Close
4	Close	Appearance	Matched
5	Matched	Close	Appearance
6	Appearance	Matched	Close

Results

The posttest number correct was analyzed separately for each question set. Results showed that appearance variants were always more difficult than close variants, and they were generally more difficult than matched questions. Close variants were generally easier than matched questions. Having previously seen a question with the same mathematical structure appears to enhance performance, but having seen a question that appears to be similar, but that actually has a different underlying mathematical structure, degrades performance. This finding was confirmed when the average scores on the forms containing appearance variants (Forms 4–6), were compared against those that did not (Forms 1–3; Table 1.4.1). Forms that contained appearance variants had lower average scores, indicating that these questions increased the difficulty for the participants.

The researchers also analyzed the difficulty of individual questions in order to see if some types were more difficult than others. Across all types, appearance variants were clearly the most difficult. In addition, close variants were easier than matched questions for four questions in Set A and for six questions in each of Sets B and C. Finally, analyses revealed that none of the investigated test-taker characteristics (i.e., gender, race/ethnicity, college major, and ability) had an impact on whether or not a test taker was able to transfer solution strategies between question variants.

Conclusion

On the tests assembled for this study, participants performed better on close variants, indicating that they transferred solution strategies from their pretest experience; however, appearance variants seemed to interfere with this transfer. This result demonstrates that the presence of appearance variants in tests that also contain close variants can cause interference with test-taker performance. This discovery suggests that tests can be designed to capitalize on the extent to which test takers set up test-taking schemas. By administering tests containing both of these types of variants, constructs can be better measured and test takers' use of inappropriate short-cut transfer strategies may be hindered.

An economical approach to test development might involve first writing question models that produce close variants, paired with writing question models that generate appearance variants of the first model. In essence, the two models would produce appearance variants of each other. This approach could be used to hinder test takers' use of inappropriate transfer strategies. In addition, by producing question models in this fashion, question shelf life may be extended. However, it should be assumed that, if a question modeling approach is adopted, test preparation schools may alter their curricula to include instruction in question modeling. These schools could not only teach question models, but could also make students aware of the existence of appearance variants and help them discriminate between appearance variants and close variants. This, in turn, could force coaching schools to focus more on teaching the mathematics underlying the questions, leading to legitimate improvements in student performance.

References

Bejar, I. I., Lawless, R. R., Morley, M. E., Wagner, M. E., Bennett, R. E., & Revuelta, J. (2002). *A feasibility study of on-the-fly question generation in adaptive testing* (GRE Board Professional Report No. 02-23). Princeton, NJ: Educational Testing Service.

Novick, L. R. (1988). Analogical transfer, problem similarity, and expertise. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 510–520.

Notes

[1] Based on *Transfer Between Variants of Quantitative Questions* (GRE Board Research Report No. 00-06R), by M. Morley, B. Bridgeman, and R. Lawless, 2004, Princeton, NJ: Educational Testing Service.

[2] Appearance variant questions were only displayed in half of the test forms.

1.5 Effects of Calculator Availability on *GRE*® Quantitative Questions[1]

Brent Bridgeman, Frederick Cline, and Jutta Levin

Professional standards for assessing quantitative reasoning skills suggest that calculators be provided to examinees (National Council of Teachers of Mathematics [NCTM], 2000). As part of the redesign of the *GRE*® General Test, the viability of providing an on-screen calculator was explored. Because the quantitative portion of the GRE General Test emphasizes reasoning skills, not computational skills, complex calculations are not required. Nevertheless, providing access to a calculator helps assure that trivial computational errors do not interfere with assessing the intended reasoning construct. To estimate the likely effects on question difficulty when calculators became available on the GRE Quantitative Reasoning measure, a special study was conducted.

The purpose of this study was two-fold: (a) to identify if an adjustment factor was needed for questions that were pretested without a calculator and (b) to evaluate the effects of calculator availability on overall scores. In particular, the effect of calculator availability by different subpopulations was of interest.

Procedure

Six 28-question linear tests were assembled, for a total of 168 questions. These forms reflected the content changes that were being considered for the GRE revised General Test. Specifically, the test forms had an increased emphasis on questions classified as *real* (i.e., word problems) in contrast to *pure* (i.e., numbers with minimal words); they placed less emphasis on geometry questions and a slightly higher proportion of data-interpretation questions than did the previous version of the GRE General Test. Several new formats for quantitative questions were also being considered as part of the test redesign; however, these question formats were not included in this study because they were developed and tried out in a calculator-available mode.

Each question was reviewed and categorized in two ways. First, test development experts rated each of the 168 questions for their calculator sensitivity (negative effect if a calculator was used [i.e., a lower score for examinees who had access to a calculator than for those with no access], positive effect if a calculator was used [i.e., a higher score for examinees who had access to a calculator than for those with no access], or no effect). Three experienced test development experts performed the ratings, and a consensus rating was reached after discussion. Second, each question was classified as having pure or real mathematical content.

Each of the six tests was administered with and without a calculator, resulting in 12 groups of examinees (six for the calculator condition and six for the no-calculator condition). Examinees were randomly assigned to one of the groups. For students in the calculator condition, an on-screen, four-function plus square root key calculator was made available and could be turned on for any question.

At the end of the regular GRE General Test administered in fall[2] 2003, examinees viewed a screen that invited them to volunteer to participate in a research project. In order to motivate examinees to stay motivated on the research questions, they were told, “It is important for our research that you try to do your best on this section” and that a $250 award would be given “to those 100 test takers who score the highest on questions in the research section relative to how well they did on the preceding scored sections.” The final sample included 13,159 participants, or about 1,100 examinees in each of the 12 groups.

Results

Question-Level Impact

Two question types were examined in the study. *Standard multiple-choice* questions contain five options from which an examinee chooses the correct answer. *Quantitative comparison* (QC) questions require the examinee to compare a quantity in one column with that of a quantity in a second column. In this case, the examinee determines which quantity is larger, whether they are equal, or if there is insufficient information to decide.

Results indicated that there was only a minimal calculator effect on most questions in that a greater percentage of examinees did not get the questions correct when using a calculator compared to those examinees not using a calculator. The percentage correct for each question across calculator and no-calculator conditions was virtually identical (within 2 percentage points) for 109 of the 168 questions. Only 15 of the questions showed differences of more than 5 percentage points.

The QC questions are designed to be answered quickly, with relatively little calculation needed. Thus, a calculator would generally not be expected to be very useful on this type of question. This seemed to be the case, since only two of the 78 QC questions showed a calculator effect of more than 5 percentage points.

For the multiple-choice questions, only 13 of the 90 questions showed a calculator effect of more than 5 percentage points.

Results indicated that questions identified by test development experts as likely to show a positive calculator effect tended to be somewhat easier when the calculator was available. This effect seemed to be the greatest for middle-difficulty questions. Because most examinees get easy questions correct even without a calculator, the advantage of having a calculator for these questions is trivial. Very difficult questions on the GRE General Test are typically conceptually difficult, not computationally difficult, so a calculator is of little benefit on these questions as well. But even in the middle-difficulty range, most questions showed little or no calculator effect.

Some differences were seen across pure and real questions, but this was affected by whether the question was considered to be calculator sensitive. Across question types, questions rated as not likely to show calculator effects did show minimal average differences in the percentage of examinees answering correctly regardless of whether they were classified as pure or real. Questions rated as likely to show some calculator effects typically showed larger average differences, about 3 or 4 percentage points, between the calculator and no-calculator conditions. The QC questions classified as real were an exception, showing a small average difference (0.25 percentage points) even when rated as likely to be sensitive to calculator use.

Effects on Total Score

Given that only 15 of the 168 questions showed calculator effects of more than 5 percentage points, the effect on examinees' total scores would be expected to be modest. Differences between total scores under the two conditions, by subgroup, are shown in Table 1.5.1.

An analysis of variance was run, with the operational quantitative score, calculator availability, gender, and ethnic/race group as independent variables. Results indicated a small but statistically significant main effect for calculator availability but no significant interactions. These results suggest that introduction of a calculator on the GRE revised General Test should not have much impact on gender or ethnic/race score differences.

Table 1.5.1

Average Total Scores Under Calculator and No-Calculator Conditions

	Average score (*SD*)		Sample size	
Ethnic/race and gender group	No calculator	Calculator	No calculator	Calculator
White/male	648 (121)	656 (120)	1,214	1,348
White/female	571 (126)	582 (128)	2,226	2,451
Asian American/male	692 (106)	707 (97)	218	213
Asian American/female	642 (123)	641 (116)	233	265
African American/male	524 (153)	541 (143)	104	91
African American/female	453 (130)	460 (125)	266	244
Hispanic/male	590 (135)	612 (131)	107	91
Hispanic/female	518 (131)	517 (117)	158	181
Other/male	637 (123)	654 (116)	174	183
Other/female	570 (137)	582 (134)	157	162
Total/male	642 (128)	654 (123)	1,817	1,926
Total/female	563 (133)	574 (132)	3,040	3,303

Note. Numbers in parentheses are standard deviations (*SD*s).

Differences in Question Difficulty by Calculator Use

The above analyses compared the question difficulties for the groups of students who did or did not have access to a calculator during the tests. Another way of examining calculator impact is to determine if there are differences in question difficulties for students who had access to a calculator and used it compared to students who had access but chose not to use it on a particular question. (*Used* will be shorthand for indicating that the examinee switched on the calculator for a particular question; it is possible that for some questions the calculator was turned on but not actually used.)

Results indicated that calculator use was relatively rare. For 86 of the 168 questions, fewer than 20% of the examinees who had a calculator available actually used the calculator. The calculator was used by more than one half of the examinees on only 20 questions; the question with the highest calculator use still had only 61% of the examinees choosing to use the calculator.

Differences in percentage correct were sometimes quite substantial between students who chose to use or not use the available calculator. The largest difference was seen for a question on which 36% of the examinees chose to use the calculator; the percentage correct for these examinees was 71% compared to 27% for the examinees who chose not to use the calculator. These results must be interpreted cautiously because the students choosing to use the calculator also had higher quantitative scores on the operational test than those students who did not use the calculator (657 compared to 567).

There were a few questions in which the average operational scores were higher in the group choosing not to use the calculator, but there was still an apparent advantage to using the calculator. For example, for a question on which 53% of examinees chose to use the calculator, the average quantitative score was 593 for those who used the calculator and 611 for those who did not use the calculator. The percentage correct was 84% for those who used the calculator but only 66% for those who did not.

Calculator Effects on Question Time

For each question, the average time to complete the question was computed separately for both calculator conditions. Question times appeared to be faster for examinees who had access to the calculator, especially on those questions that had been rated as calculator sensitive by the test development experts. In particular, for those 20 questions on which more than one half of the examinees used the calculator, there appeared to be a time advantage to using the calculator.

Conclusion

For most of the GRE Quantitative Reasoning questions studied, the effect of having access to a calculator was relatively small. Although a few exceptions were found, test development experts were fairly accurate in identifying which questions were likely to be calculator sensitive. For calculator sensitive questions, the effect was about 4 percentage points higher than the existing difficulty estimate. Real QC questions were an exception in that calculator use did not seem to impact their difficulty level, even when identified by test development experts as calculator sensitive.

The substantial effects noted for examinees who chose to use the calculator compared to those who chose not to use it when available are open to different interpretations and do not necessarily reflect a true calculator effect. Nevertheless, they suggest that continued monitoring is desirable as examinees become more familiar with the way to use the calculator most effectively.

Any time differences related to calculator use should not be of great concern. The time limits for the GRE revised General Test are being set based on field trials that include access to a calculator, so time differences will be taken into account (see Wendler, Chapter 1.2, this volume).

This study indicated that calculator benefits appear to be relatively constant across gender and ethnic/race groups, with no significant interactions of gender or ethnicity/race. As test preparation materials are developed for the GRE revised General Test, information that demonstrates the most effective ways to use a calculator should be included so it is available to all students.

References

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics.* Reston, VA: Author.

Notes

[1] Based on *Understanding the Impact of Calculator Availability on GRE® Quantitative Questions* (GRE Board Research Report No. 03-09), by B. Bridgeman, F. Cline, and J. Levin, 2008, Princeton, NJ: Educational Testing Service.

[2] *Fall* refers to data collected sometime during August through December.

1.6 Calculator Use on the *GRE*® revised General Test Quantitative Reasoning Measure[1]

Yigal Attali

The use of calculators in mathematics assessment and instruction is now commonplace, and particularly so in U.S. national standardized assessments such as the *SAT*®, ACT, and NAEP tests. The utility of calculators to enhance students' understanding and use of numbers and operations has been affirmed by relevant stakeholders and researchers alike (Ellington, 2003; National Council of Teachers of Mathematics, 1980). Although previous research findings (Bridgeman, Harvey, & Braswell, 1995) indicated that quantitative questions generally become easier for test takers to solve when calculators are available, this was found to be only marginally true in preliminary calculator research for the *GRE*® revised General Test (Bridgeman, Cline, & Levin, 2008; Chapter 1.5, this volume). On average, calculator availability increased the percent of test takers answering a question correctly by 1%.

It is important to make the distinction between calculator availability and calculator use. Research on calculator usage by SAT test takers revealed that female, White, and Asian test takers were more likely to make use of a calculator (Scheuneman, Camara, Cascallar, Wendler, & Lawrence, 2002). Further, test takers who reported using the calculator on one third to one half of the questions appeared to perform better than those who used it more or less often. However, the SAT uses a paper-based format and test takers bring their own calculator to the test. Therefore, the results of the study may not generalize completely to the GRE revised General Test, which provides an on-screen calculator. In addition, since the GRE revised General Test is administered on computer, it is possible to capture highly accurate calculator usage information.

The purpose of the current study was to explore calculator usage on the GRE revised General Test by test-taker group and question type. It examined the relationship of test-taker characteristics (ability level, gender, and race/ethnicity) and question characteristics (difficulty, typical response time, question type, and content) with calculator use. The study also explored whether response accuracy was related to calculator use across questions.

Procedure

Demographic and calculator usage data were collected for a random sample of 1,000 domestic[2] test takers from each of 20 different Quantitative Reasoning (Quantitative) sections. Data was drawn from sections administered in the fall[3] of 2012, a full year after the GRE revised General Test was launched. This helped ensure that test takers were aware of the availability of the online calculator. The calculator provided on the test is a basic four-function calculator with parentheses, square-root, and memory buttons.

Quantitative score, gender, and race/ethnicity were collected for each test taker. For each question to which a test taker responded, response accuracy (correct/incorrect), response time, and calculator usage (used/did not use) information was also collected. Questions were classified by question type (single-selection multiple-choice, multiple-selection multiple-choice, quantitative comparisons, or numeric entry) and question content classifications (real vs. pure; [4] arithmetic vs. algebra).

Results

Analyses by Test-Taker Characteristics

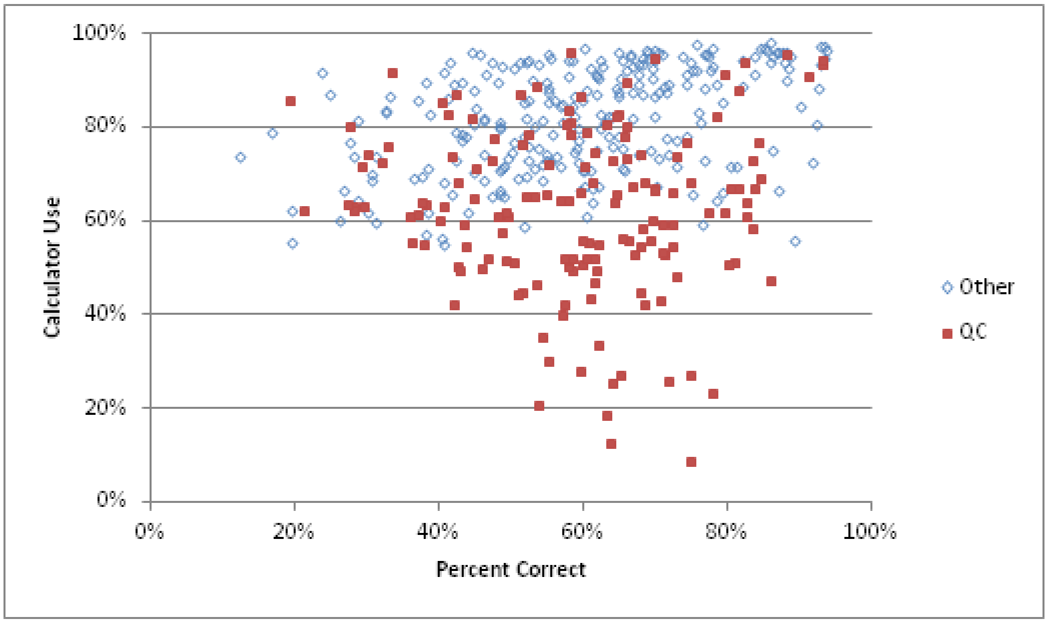
Analyses revealed that overall calculator usage was very common, with the calculator being used in around 75% of question responses. The correlation between calculator use and Quantitative score was fairly low (*r* = .11). However, the relation was not monotonic: Test takers with the lowest 10% to 20% and highest 5% of Quantitative scores showed the lowest amount of calculator use.

In addition, consistent with previous research, differences in calculator use across some gender and race/ethnicity groups were noted. Significant differences were found between men and women, with women displaying higher calculator use than men. Significant differences were also found for some race/ethnic groups, with White and Asian test takers showing higher calculator use than Black and Hispanic test takers.

Analyses by Question Characteristics

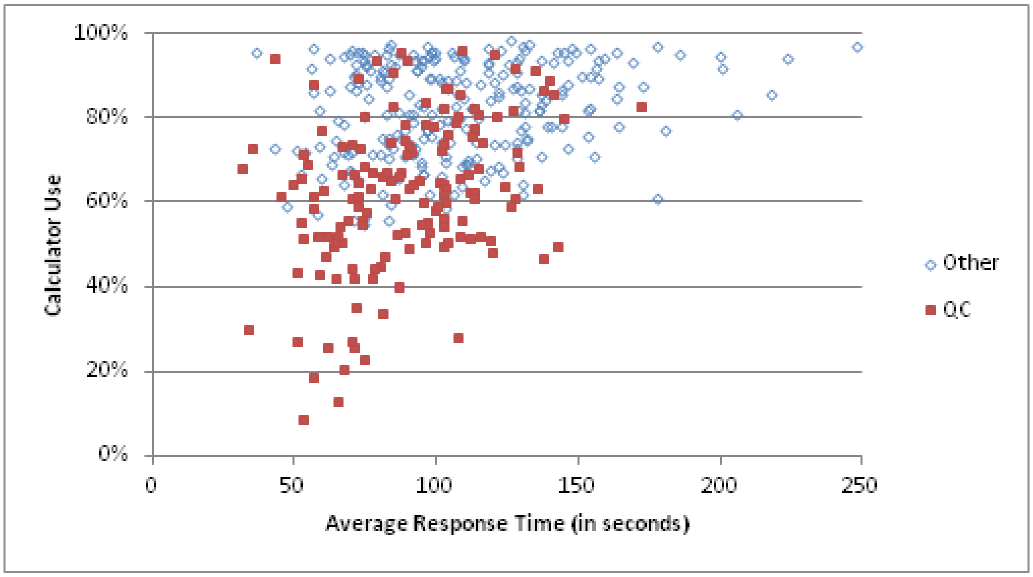
Calculator usage was also explored in relation to various question characteristics. The relationship between calculator use and question difficulty, defined as the percent of test takers who respond correctly to the question, was examined. Figure 1.6.1 shows the percent of test takers using a calculator by question difficulty level. The calculator was used less on quantitative comparisons questions. However, since quantitative comparisons questions are designed to require minimal computation, this result is not surprising.

The quantitative comparisons question type also did not display a relationship between calculator use and question difficulty (*r* = .01). Calculator usage was higher for the other question types (single-selection multiple choice, multiple-section multiple choice, and numeric entry) with a moderate relationship (*r* = .42) between calculator use and question difficulty, indicating that the easier a question was, the more likely it was that test takers used a calculator on the question. Question-level analyses also revealed a moderate relationship (*r* = .36) between average response time and calculator use. Figure 1.6.2 displays the relationship between response time (in seconds) and calculator use. Again, quantitative comparisons questions acted differently than single-selection multiple-choice, multiple-section multiple-choice, and numeric entry question types. On average, quantitative questions had shorter response times. Again, since quantitative comparisons questions are designed to elicit relatively swift responses, this result is not surprising.



Note. Other refers to single-selection multiple-choice, multiple-selection multiple-choice, or numeric entry questions. *QC* refers to quantitative comparisons questions.

Figure 1.6.1. Item difficulty and calculator use.



Note. Other refers to single-selection multiple-choice, multiple-selection multiple-choice, or numeric entry questions. *QC* refers to quantitative comparisons questions.

Figure 1.6.2. Average response time and calculator use.

Analyses were also conducted to determine if there was a difference in calculator use across question content classifications. Results revealed that test takers used the calculator less for questions classified as pure than for those classified as real and less for questions classified as algebra than for those classified as arithmetic. In both cases, however, the differences were not large. These results align with expectations, as questions classified as arithmetic or real generally require more computations than their counterparts. In addition, questions classified as pure often contain abstract ideas, and those classified as real often reflect real-world tasks or scenario-based problems.

Finally, the relationship between calculator use and correctness of response was evaluated. In general, a positive association was seen between calculator usage and responding correctly to a question. This indicated that, for most questions, test takers who used the calculator were more likely to answer correctly than test takers with the same Quantitative score who did not use the calculator. This result was stronger for easier questions and for questions where a higher percentage of test takers used the calculator. There are several possible explanations for this finding, including that test takers who do not understand the question or the computations needed to find the correct answer will tend not to use the calculator, or that test takers who do not use the calculator make more computational mistakes, but the true underlying mechanism(s) cannot be determined by this study.

Conclusion

The decision was made to make calculators available on the Quantitative Reasoning measure of the GRE based on preliminary research (Bridgeman et al., 2008), as well as other factors. In theory, the availability of a calculator should increase the utility of the Quantitative measure by eliminating the errors that can occur in hand computations. However, it also introduces other factors, such as test takers knowing when and how to use the calculator to their advantage. This study documented operational calculator usage patterns, further contributing to our understanding of the impact of calculator availability on the behavior and scores of test takers.

This study indicated that the calculator is used by most students and slightly more by test takers who are female, White, or Asian. The high frequency with which test takers used the calculator suggests that they may be using it to verify even simple calculations. It is also possible that many of the questions require advanced computations that can be greatly expedited by using a calculator (very few questions showed calculator usage below 50%).

Analyses revealed interesting findings in terms of test-taker ability (as measured by their Quantitative score) and calculator usage. The highest ability (top 5%) test takers may feel more confident in their computation abilities and, thus, use the calculator less. In contrast, lower ability (lowest 10%–20%) test takers may not be confident enough to know how or when to use the calculator.

Analyses by question difficulty level showed that calculator usage was higher on easier questions for single-selection multiple-choice, multiple-section multiple-choice, and numeric entry question types. This may be due to a couple of different factors. More difficult questions may be more conceptual than computational or fewer test takers may be comfortable with going through the necessary computations on the calculator. Further, questions with higher calculator usage also required more time to answer. This makes intuitive sense, as questions that are more computational in nature may require more time to answer.

Finally, analyses indicated that for most questions, but especially for easier questions, test takers who used the calculator are more likely to answer correctly than test takers with the same Quantitative score who do not use the calculator. The findings can provide input during the development of test preparatory materials to ensure that all test takers are aware of the most effective ways to use the calculator on the GRE revised General test.

References

Bridgeman, B., Cline, F., & Levin, J. (2008). *Effects of calculator availability on GRE quantitative questions* (Research Report No. RR-08-31). Princeton, NJ: Educational Testing Service.

Bridgeman, B., Harvey, A., & Braswell, J. (1995). Effects of calculator use on scores on a test of mathematical reasoning. *Journal of Educational Measurement, 32*, 323–340.

Ellington, A. J. (2003). A meta-analysis of the effects of calculators on students' achievement and attitude levels in precollege mathematics classes. *Journal for Research in Mathematics Education, 34*(5), 433–463.

National Council of Teachers of Mathematics. (1980). *An agenda for action: Recommendations for school mathematics in the 1980s.* Reston, VA: Author.

Scheuneman, J. D., Camara, W. J., Cascallar, A. S., Wendler, C., & Lawrence, I. (2002). Calculator access, use, and type in relation to performance on the SAT I: Reasoning test in mathematics. *Applied Measurement in Education, 15*, 95–112.

Notes

[1] Based on *Calculator Use on the GRE revised General Test Quantitative Reasoning Measure*, by Y. Attali, unpublished manuscript, Princeton, NJ: Educational Testing Service.

[2] *Domestic* refers to test takers who indicated they are U.S. citizens and took the test in a test center in the United States or a U.S. territory.

[3] *Fall* indicates that data were gathered sometime between August and December.

[4] Real mathematics questions reflect a real-world task or scenario-based problem, while pure mathematics questions deal with abstract concepts.

1.7 Identifying the Writing Tasks Important for Academic Success at the Undergraduate and Graduate Levels[1]

Michael Rosenfeld, Rosalea Courtney, and Mary Fowles

The developmental process used to create the *GRE*® Analytical Writing measure[2] included a number of steps, including feedback from focus groups of graduate faculty, input and guidance from several committees and technical advisory panels, and a series of formal research studies (Powers, Burstein, Chodorow, Fowles, & Kukich, 2000; Powers & Fowles, 1997, 2000; Powers, Fowles, & Welsh, 1999; Schaeffer, Briel, & Fowles, 2001). This study is part of the ongoing effort by the GRE program to gather validity information on the Analytical Writing measure and focuses on evidence based on test content. The findings of this study also provide foundational support for the continuation of the Analytical Writing measure in the GRE revised General Test.

One purpose of this study was to verify that the skills measured by the Analytical Writing measure are considered relevant for entry-level graduate students at both the master's and doctoral levels. A second purpose was to gather data to determine if the scores for the Analytical Writing measure are appropriate for use as an outcomes measure for upper-division undergraduate students. This summary, however, focuses on only the first purpose: to determine the alignment of the skills measured by the Analytical Writing measure with those writing tasks deemed important for academic success at the graduate level.

Procedure

Creating the Writing Tasks

A survey consisting of writing tasks statements was developed. Because the survey was administered to faculty members across a range of subject areas, it was important that the statements be written in language that would be clear and understandable to nonwriting specialists. All drafts of the survey were reviewed by a number of groups and individuals, such as ETS test development experts and scientists, an external five-person advisory committee composed of writing experts, faculty members who were part of the ETS Visiting Minority Faculty program, and a number of faculty who represented a range of disciplines and taught both undergraduate and graduate courses.

The initial draft of the writing survey was developed based on a review of literature associated with writing. Fifty task statements were included in the survey. Based on feedback from the various groups, the survey was revised, resulting in a final survey composed of 39 task statements, three importance rating scales (geared to the level of student that the faculty member taught), and a background information section.

The survey was sent to 1,512 faculty members (792 undergraduate and 720 graduate) across six disciplines: English, education, psychology, natural sciences, physical sciences, and engineering. A total of 33 schools were involved in the study. Coordinators at each college or university were responsible for distributing the surveys to the faculty.

Confirming the Link Between Analytical Writing Skills and the Writing Tasks Statements

Each of the GRE scoring guides used for the tasks in the Analytical Writing measure (analyze an issue and analyze an argument) have six levels. Each level describes the skills that are typically demonstrated in essays at each score level. For this study, the two scoring guides were merged into a single document consisting of nine skills. Overlapping skills appeared only once, but distinctly different skills remained as separate entities. Five ETS writing assessment specialists rated each of the nine skills in terms of its importance for competent performance on each of the 39 task statements. The ratings were conducted independently.

Results

Writing Tasks Analyses

Analyses were designed to identify the writing tasks statements that were judged by faculty to be important for competent academic performance within and across subject areas at the three educational levels (undergraduate, master's, and doctorate). Averages and standard deviations were computed for each task statement at each of the three educational levels. In addition, correlation coefficients were computed to evaluate the profile of task ratings within and across the three levels of education.

Of the 33 schools that agreed to participate, 30 returned surveys, resulting in a 91% institution participation rate. A total of 861 completed surveys were received. The institutions represented schools from four geographic regions, as well as a mixture of public and private schools and level of degrees offered (doctoral/research, master's, and baccalaureate only); two were Historically Black Colleges and Universities and three were Hispanic Serving Institutions.

Faculty responding to the surveys spanned all six subject areas and represented teaching at all three educational levels. When asked how important higher-level writing skills were in course assignments, the mean rating across all respondents was *very important*.

For those faculty that taught master's level students, mean ratings of each task ranged from *moderately important* to *very important* for entering master's level students to be able to perform competently. The average of the importance ratings for each of the 39 task statements were correlated with each other across six subject areas. For five of the six areas, the correlations were very similar. English was the one area that demonstrated a different profile of ratings, most likely due to three task statements that were primarily geared toward English.

For faculty who taught doctoral level students, mean ratings of each task ranged from *moderately important* to *extremely important* for entering doctoral level students. As seen with the master's level data, the correlations of the average of the ratings for the 39 task statements across the six subject areas were similar for five of the areas. Once again, English demonstrated a different profile of ratings.

Linking Analytic Writing Skills and Writing Tasks

A total of 29 writing tasks were judged to be important by college faculty in each of the six subject areas at both the master's and doctoral levels. These tasks were considered to be the core of important writing tasks at both graduate levels. A mean rating of *very important* was used as the minimum criterion for establishing a link between each writing task and the GRE scoring rubrics. Results indicated that all of the skills in the scoring rubrics were judged to be important for successfully performing one or more of the core tasks.

Conclusion

This study provides additional evidence of the content relevance of the GRE Analytical Writing measure. It does so by defining the domain of writing tasks seen as important for competent performance across a range of academic areas at the two graduate levels and by demonstrating the linkage between these writing tasks and the writing skills assessed in the GRE scoring rubrics. The results allow institutions to understand entering graduate students' abilities to perform important writing tasks required for graduate study. This study also provided fundamental information on core writing tasks and helped define the content measured by the writing prompts included on the GRE revised General Test.

References

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2000). *Comparing the validity of automated and human essay scoring* (Research Report No. RR-00-14). Princeton, NJ: Educational Testing Service.

Powers, D. E., & Fowles, M. E. (1997). *Correlates of satisfaction with graduate school applicants' performance on the GRE Writing measure* (Research Report No. RR-96-24). Princeton, NJ: Educational Testing Service.

Powers, D. E., & Fowles, M. E. (2000). *Likely impact of the GRE Writing assessment on graduate admissions decisions* (Research Report No. RR-00-16). Princeton, NJ: Educational Testing Service.

Powers, D. E., Fowles, M. E., & Welsh, C. K. (1999). *Further validation of a writing assessment for graduate admissions* (Research Report No. RR-99-18). Princeton, NJ: Educational Testing Service.

Schaeffer, G. A., Briel, J. B., & Fowles, M. E. (2001). *Psychometric evaluation of the new GRE writing assessment* (GRE Board Research Report No. 96-11). Princeton, NJ: Educational Testing Service.

Notes

[1] Based on *Identifying the Writing Tasks Important for Academic Success at the Undergraduate and Graduate Levels* (GRE Board Research Report No. 00-04), by M. Rosenfeld, R. Courtney, and M. Fowles, 2004, Princeton, NJ: Educational Testing Service.

[2] The Analytical Writing measure became a part of the GRE General Test in 2002.

1.8 Timing of the Analytical Writing Measure of the *GRE*® revised General Test

Frédéric Robin and J. Charles Zhao

The Analytical Writing measure has been part of the *GRE*® General Test along with the Verbal Reasoning (Verbal), and Quantitative Reasoning (Quantitative) measures since 2002. Until the launch of the GRE revised General Test in August 2011, Analytical Writing consisted of two types of writing tasks: [1] a 45-minute analysis of an issue (issue) and a 30-minute analysis of an argument (argument; Educational Testing Service [ETS], 2010). While planning for the launch of the revised test, the GRE program decided to revise the Analytical Writing measure in order to better allocate the testing time among the three measures and to discourage the use of memorized or formulaic text. The proposed changes were to reduce the total Analytical Writing time from 75 to 60 minutes, no longer offer test takers with a choice between two issue tasks, and require more focused responses (ETS, 2013a). As the revised tasks were developed, a preliminary study was conducted to help choose the most appropriate Analytical Writing timing configuration.

In this chapter, we first summarize the results of a study that led to the choice of the revised Analytical Writing timing configuration eventually implemented (Zhao, Zhu, Guo, Zeller, & Bannerjee, 2006). We then provide a comparison of the psychometric properties of the Analytical Writing measure before and after the launch of the revised test and show the extent to which the continuity of the measure has been maintained.

Timing Study

The choice of the 30-minute issue and 45-minute argument timing configuration used when the Analytical Writing measure was first implemented was supported by a study of the effects of applying different time limits to the proposed GRE writing test conducted by Powers and Fowles (1996). In that study it was found that:

> Examinees who described themselves as slow writers/test takers did not benefit any more (or any less) from generous time limits than did their quicker counterparts. In addition, there was no detectable effect of different time limits on the meaning of essay scores, as suggested by their relationship to several non testing indicators of writing ability. (p. 433)

These findings suggest that a further reduction of the time limits for the issue and argument tasks, desirable in order to avoid a likely increase in the total revised GRE testing time, might be possible.

Therefore, the timing configuration study (Zhao et al., 2006) discussed in this chapter was designed to assess the potential of three alternative 60-minute Analytical Writing timing configurations in preserving the psychometric properties of essay scores. The properties assessed included the following:

1. Test performance
2. Speededness[2]
3. Consistency of tryout and operational scores
4. Correlation between Analytical Writing and Verbal scores
5. Test takers' evaluation

Method

Three possible issue and argument timing configurations each resulting in a total of 60 minutes were investigated: 30/30 (reducing only the issue time limit), 35/25 (reducing the time for the issue task more than that for the argument task), and 40/20 (reducing the time for the argument task more than that for the issue task). In September 2004, 1,183 paid volunteers who had previously taken the GRE General Test and therefore had known operational Analytical Writing measure scores, were administered an Internet-based writing tryout. The tryout test consisted of one enhanced (requiring more focused writing) issue task and one enhanced argument task. To form groups of equal ability, study participants were randomly assigned to one of the three timing conditions. Fourteen experienced essay readers were recruited for the study. They were divided into three groups, each responsible for evaluating essay responses under one timing condition. Before starting their scoring activities, readers were trained on the use of a modified scoring rubric specially prepared for the study.

Each essay was scored by at least two readers; a third reader was called on when the first two readers' scores differed by more than 1 point. The final score for the essay was computed as the average of the two closest scores. In order to monitor the scoring consistency across timing conditions, about 10% of the essays randomly selected from each timing condition were seeded into other timing conditions for a second reading.

In addition to taking an Analytical Writing tryout test, the study participants were asked to respond to a 10-question exit questionnaire designed to elicit their qualitative feedback on the appropriateness of the timing configurations and the quality of the testing experience.

Results

After screening out test records not having (a) responses to both tasks or (b) operational Analytical Writing and Verbal measure scores, the analysis sample included 574 participants. Of these participants, about three quarters were U.S. test takers and one quarter international test takers from China, India, Korea, Taiwan, and Japan, a proportion relatively similar to that of the operational test-taker population. The study participants' tryout test results and their responses to the exit questionnaire were evaluated across the three timing conditions and compared with their operational test results. Additional results from 758,447 operational tests delivered to U.S. and international test takers from October 2002 to September 2004 (2002–2004 operational group) are also provided for comparison. (Data collected from test takers from China, Hong Kong, Korea, and Taiwan were collected through a different delivery network and did not contain timing information.)

Test performance. As indicated in Table 1.8.1, the study participants were spread about equally across the three tryout timing conditions and had very similar operational reported score means and standard deviations. This confirmed the effectiveness of the study's assignment to timing conditions in producing equivalent groups. The study participants were generally more able than typical GRE test takers, as comparisons of their mean operational Analytical Writing measure scores (4.58–4.60) to that of the 2002–2004 operational group (4.24) indicate. However, participants performed much lower under any of the tryout conditions than they did operationally, even when the issue and argument timing was the same or nearly the same (see the noted values in Table 1.8.1). This strongly suggests that the lack of motivation on the part of the study participants had a significant influence on their performance. Despite this challenge, which can be expected to happen with volunteers, valuable information was gathered from the pattern of results.

Table 1.8.1

Tryout and Operational Score Means (*M*) and Standard Deviations (*SD*)

Timing condition (sample size)	Issue		Argument		Analytical Writing	
	M	*SD*	*M*	*SD*	*M*	*SD*
Study tryout						
30/30 (*N* = 204)	3.63	1.02	3.61 [a]	0.91	3.74	0.87
35/25 (*N* = 203)	3.58	1.13	3.39	1.09	3.59	0.99
40/20 (*N* = 167)	3.42 [a]	0.90	3.30	0.85	3.47	0.80
Study operational						
45/30 (*N* = 204)	4.50	0.94	4.44	1.00	4.60	0.87
45/30 (*N* = 203)	4.48	0.93	4.43	0.95	4.58	0.85
45/30 (*N* = 167)	4.46 [a]	0.84	4.48 [a]	0.88	4.58	0.74
2002–2004 operational						
45/30 (*N* = 758,447)	4.23	0.97	4.01	1.09	4.24	0.94

[a] Even under the same or nearly the same timing conditions, participants performed much lower under the tryout conditions than they did on the actual operational test.

In particular, with the issue task, it appeared that mean performance actually decreased from 3.63 to 3.42 when given more time (30, 35, and 40 minutes). For the argument task, mean performance improved from 3.30 to 3.61 with more time, but in that case, timing conditions went from being clearly too limited to more reasonable (20, 25, and 30 minutes). The highest tryout Analytical Writing mean score (3.74) was obtained with the 30/30 timing condition. The 30/30 timing condition also led to the smallest discrepancies between the mean and standard deviation of the tryout and operational Analytical Writing scores (3.74 vs. 4.58 or 4.60, and 0.87 vs. 0.74, 0.85, or 0.87). Finally, the spread of scores (indicated by *SD*s) did not appear to be affected by any of the timing conditions.

Speededness. Table 1.8.2 shows the extent to which test takers used the time allocated to them. Regardless of the testing condition, we see that 50% or more of the test takers finished their essay before time ran out (P50 values are less than 100% of allocated time). Furthermore, the whole time usage patterns appear to be nearly the same across the tryout timing conditions, for both issue and argument tasks. These results suggest that while many test takers could have used more time, they were able to adapt to the various timing conditions and manage their time accordingly. Taken together with the above, this indicates that performance may not decrease as the issue time limit is decreased, and these results suggest that the 30/30 timing condition may not affect test speededness in any significant way.

Consistency of tryout and operational scores. The Spearman’s rank-order correlations between the study tryout and operational Analytical Writing measure scores served as a measure of the extent to which each of the alternative 60-minute timing conditions produces similar rankings of the test takers when compared with those obtained under the 75-minute operational timing condition. As Table 1.8.3 shows, the 30/30 and 35/25 conditions resulted in similar Analytical Writing score consistency (0.62 and 0.63). The 40/20 condition resulted in much lower Analytical Writing score consistency (0.53), mostly because of the argument task: 20 minutes is clearly not enough time for the argument task.

Correlation between Analytical Writing and Verbal scores. Because of the nature of the Analytical Writing and Verbal measures, some degree of correlation between them exists. As shown in Table 1.8.4, over a very large population sample using 2002–2004 operational data, it reaches 0.60. By comparison, the smaller and less representative study samples had smaller and decreasing Analytical Writing/Verbal correlations, from about 0.55 operationally to 0.45, 0.48, and 0.54 under the tryout conditions. Correlations at the level of the task scores were relatively unstable; as a result, no distinctive patterns among the timing conditions were noticeable.

Feedback from test takers. Table 1.8.5 summarizes the study participants’ answers to the exit questionnaire inquiring about their experience of the tryout test. The participants generally rated the argument task as *interesting or very interesting* or *indicated [their] writing ability fairly well or very well* much more often than they did the issue task (60% vs. 40% of answers, respectively). They were also more likely to indicate that they had made as much effort as they did operationally with the argument task than they did with the issue task (65% to 55% of answers, respectively). In terms of timing, relatively high proportions of participants across all the issue and argument timing conditions indicated having *just enough or more than enough time to finish*.

Table 1.8.2

Study Tryout and Operational Time Usage

Writing task	Timing condition (sample size)	Time usage [a]			
		P25	*P50*	*P75*	*P95*
Study tryout					
Issue	30 (*N* = 204)	63	87	100	100
Issue	35 (*N* = 203)	63	91	100	100
Issue	40 (*N* = 167)	58	80	100	100
Argument	30 (*N* = 204)	53	77	97	100
Argument	25 (*N* = 203)	64	88	100	100
Argument	20 (*N* = 167)	65	85	100	100
Study operational [b]					
Issue	45 (*N* = 162)	89	98	100	100
Issue	45 (*N* = 150)	91	98	100	100
Issue	45 (*N* = 126)	91	98	100	100
Argument	30 (*N* = 162)	80	90	97	100
Argument	30 (*N* = 150)	83	90	97	100
Argument	30 (*N* = 126)	80	93	97	100
2002–2004 operational					
Issue	45 (*N* = 758,447)	89	98	100	100
Argument	30 (*N* = 758,447)	83	93	100	100

[a] As the percentage of the time allocated, test takers used to finish their essay; reported at the 25th (P25), 50th (P50), 75th (P75), and 95th (P95) percentiles of the observed test time distribution. [b] Lower sample sizes, as operational timing was not available for some of the participants.

Table 1.8.3

Study Tryout and Operational Score Rank-order Correlations

Timing condition (sample size)	Issue	Argument	Analytical Writing
30/30 (*N* = 204)	0.57	0.46	0.62
35/25 (*N* = 203)	0.57	0.53	0.63
40/20 (*N* = 167)	0.53	0.37	0.53

Table 1.8.4

Correlations Between Analytical Writing Tasks, Analytical Writing, and Verbal Scores

	Pearson correlation			
Timing condition (Sample size)	Issue/Argument	Issue/Verbal	Argument/Verbal	Analytical Writing/Verbal
Study tryout				
30/30 (*N* = 204)	0.58	0.42	0.36	0.45
35/25 (*N* = 203)	0.59	0.52	0.46	0.54
40/20 (*N* = 167)	0.55	0.50	0.37	0.48
Study operational				
45/30 (*N* = 204)	0.57	0.44	0.53	0.54
45/30 (*N* = 203)	0.59	0.52	0.46	0.55
45/30 (*N* = 167)	0.52	0.54	0.40	0.54
2002–2004 operational				
45/30 (*N* = 758,447)	0.62	0.54	0.56	0.60

The essay readers expressed their preference for the 30/30 timing condition. As they put it, under this timing condition, the test takers did not seem to add the extra *padding* in their issue essays sometimes seen in responses under longer timing conditions.

Table 1.8.5

Percentage of Participant Responses to Tryout Exit Questionnaire by Task and Timing Condition

	Issue				Argument			
Exit questionnaire	30/30	35/25	40/20	All	30/30	35/25	40/20	All
Interesting, or very interesting, writing task (TASK)	39	42	38	40	69	61	58	63
Indicating writing ability fairly well or very well (ABILITY)	45	44	44	44	70	60	56	62
About the same effort had the test counted for admission (EFFORT)	54	55	55	55	67	61	65	64
Just enough, or more than enough, time to finish (TIME)	69	78	88	78	92	82	70	81

Conclusion

Like most studies relying on volunteer participation and tryout data, this study had important limitations: (a) the relatively low level of effort participants invested in the tryout, (b) the use of only one set of issue and argument tasks, and (c) the revised test's early stage of development. Nevertheless, it was felt that the results provided sufficient evidence to support the selection of the 30/30 Analytical Writing timing condition for further development and eventually for operational implementation.

Transition to the GRE revised General Test: Monitoring Analytical Writing Testing Outcomes

Before the GRE revised General Test was launched in August 2011, many more Analytical Writing tasks were developed and tried and new assembly, delivery, and scoring processes were implemented to ensure that the goals of the revised GRE program would be met (Briel & Michel, Chapter 1.1, this volume; Robin & Kim, Chapter 2.3, this volume). Since the launch of the GRE revised General Test, testing outcomes have been closely monitored, and now more than two years' worth of operational data have been collected. In this chapter, we summarize the main Analytical Writing results obtained before and after the launch of the revised test and show the extent to which the continuity of the Analytical Writing measure has been maintained. As before, a summary of test speededness, performance, and correlations between Analytical Writing and Verbal measures are presented.

Analyses were conducted on the GRE 2009 and revised GRE 2012 domestic[3] samples, which represent a large majority (about three quarters) of the total test-taker population.[4] As the results summarized in Table 1.8.6 show, some changes in average performance between the GRE and the revised GRE are noticeable. Overall, the Analytical Writing scores decreased from 3.97 to 3.83 (0.14) points. In 2012, more test takers than in the past were making use of all the time allocated, with both issue and argument tasks. But the Verbal/Analytical Writing and argument/Analytical Writing correlations did not change. Also noticeable is the higher correlation between issue and argument tasks, which indicates that the revised measure is a slightly more cohesive measure of analytical writing ability.

Conclusion

A possible explanation for these results is that by using sets of directions that require more focused writing, the issue and argument tasks have increased somewhat in difficulty and have become more similar. On the whole, it appears that the psychometric properties of the revised Analytical Writing measure are very similar to those of the original Analytical Writing measure.

Table 1.8.6

Comparison of Major Outcomes Between the GRE and the revised GRE for the Domestic Sample

	2009 (*N* = 384,986; 30/45 timing)			2012 (*N* = 351,142; 30/30 timing)		
Outcome	Issue	Argument	Analytical Writing	Issue	Argument	Analytical Writing
Performance						
Mean	3.98	3.80	3.97	3.68	3.66	3.83
SD	0.8	1.0	0.8	0.8	0.9	0.7
Time use						
*P*25	89	83		93	87	
*P*50	98	93		100	100	
*P*75	100	100		100	100	
Correlation						
Issue		0.54	0.84		0.61	0.84
Argument			0.89			0.87
Verbal			0.55			0.56

References

Educational Testing Service. (2010). *Analytical writing*. Princeton, NJ: Author.

Educational Testing Service. (2013a). *Introduction to the Analytical Writing measure*. Princeton, NJ: Author.

Educational Testing Service. (2013b). *GRE guide to the use of scores*. Princeton, NJ: Author.

Powers, D., & Fowles, M. (1996). Effects of applying different time limits to a proposed GRE writing test. *Journal of Educational Measurement*, *33*, 433–452.

Zhao, J. C., Zhu, R., Guo, F., Zeller, K. E., & Bannerjee, B. (2006, April). *Timing configuration and psychometric properties of the redesigned Analytical Writing measure of the GRE General Test*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Notes

[1] In the GRE revised General Test, the Analytical Writing measure consists of two 30-minute tasks (an issue statement or a brief argument passage) and a set of writing instructions (ETS, 2013a).

[2] *Speededness* indicates the extent to which test takers' performance on a test may be affected by the time limit. Some degree of speededness may or may not contribute to the construct measured by the test. In order to evaluate Analytical Writing speededness, we rely on both the proportion of test takers who are finishing their task ahead of the time limit and the impact (or lack of impact) that reducing or increasing the time allowed may have on Analytical Writing scores.

[3] Domestic samples include U.S. citizens testing in test centers in the United States or a U.S. territory.

[4] Similar analyses of data collected from international, gender, and ethnicity subgroups to assess subgroups differences and the fairness of the GRE revised Analytical Writing measure are beyond the scope of this chapter.

1.9 Psychometric Evaluation of the New *GRE*® Writing Measure [1]

Gary Schaeffer, Jacqueline Briel, and Mary Fowles

This study collected initial psychometric information about the new *GRE*® General Test Analytical Writing measure in order to guide its final design, delivery, and scoring process before its introduction in fall [2] 1999. Two prompt types were evaluated: analyze an issue (issue) and analyze an argument (argument). The issue task requires the examinee to think critically about a general topic and respond using a set of specific instructions. The argument task requires the examinee to discuss the logical soundness of the author's case according to specific instructions.

The study investigated four aspects of the Analytical Writing measure:

1. Prompt difficulty. Similar prompt types are expected to be at about the same level of difficulty. Thus, examinees' scores on various prompts were examined to determine if they were representative of the scores that would have been obtained on any other prompt of the same type.
2. Order effects. Alternative study designs were examined to determine whether scores on prompts were affected by the order in which the two prompt types were administered.
3. Score distributions. Subgroup (i.e., race/ethnicity and gender) performances were examined for each of the two prompt types across the study designs. If either type resulted in much larger than expected mean score differences between the subgroups, serious consideration would be given as to whether that prompt type could be used in the operational assessment.
4. Relationships between prompt scores. The magnitude of the relationship between issue and argument scores was examined to determine whether two writing scores or a single combined score would be reported.

Procedure

More than 2,300 students, recruited nationally from 26 participating U.S. colleges and universities, took part in the study. Participants were college undergraduates planning to take the operational GRE General Test within a year of the study. About 58% of participants were female, 20% African American, 20% Asian, and 48% White.

Data were collected from September through December 1997. A total of 40 prompts (20 issue and 20 argument) were used. Each participant wrote two essays in response to two prompts. Each pair of prompts was administered to the same number of participants, and each prompt was administered in eight positions (four times in the first position and four times in the second position). Participants were randomly assigned to one of four conditions (study designs):

- 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.

Eighteen readers scored the essays. All were college faculty with expertise in writing, and all had completed GRE General Test reader training. Each essay was evaluated independently by two readers, who assigned a holistic score on a 6-point scale between 6 (*highest*) and 1 (*lowest*). If the two readers' ratings were identical or adjacent, the scores were averaged to compute the final score. If the two readers' scores differed by more than 1 point, a third reader was used.

Results

Prompt Difficulty

No apparent relationship was found between prompt difficulty and prompt classification. Although sample sizes were 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.

Order Effects

Results suggested that an order effect was present. 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. The correlation between prompt types was also higher in the issue-argument order, and the estimated reliability was somewhat higher in this order.

Score Distributions

Score distributions were examined by gender and race/ethnicity (i.e., African American, Asian, Hispanic, and White) groups. Overall, more participants received higher scores on the issue prompt compared to the argument prompt regardless of the order of presentation. In addition, more Asian and African American participants received high scores (i.e., an average score greater than or equal to 4) in the issue-argument order than in the argument-issue order. Comparison of the score distributions for these two groups across the two prompt orders confirmed there was a significant difference.

Standardized scores were computed to compare performances for gender and race/ethnicity subgroups across the four study conditions. Women scored higher than men across all prompt types and study conditions, but only one condition (i.e., issue-issue) was statistically significant.

Differences between African American and White participants and between Hispanic participants and White participants were smaller than the differences found between these groups on other GRE General Test measures (Graduate Record Examinations Board, 1998). Further, the magnitude of the differences observed in the current study were similar to those found in similar tests, such as the analytical writing measure of the Graduate Management Admissions Test (Breland, Bridgeman, & Fowles, 1999).

Relationship Between Prompt Scores

Several analyses were conducted to evaluate the relationship between the different prompt scores. 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.

Reliability analyses were also conducted to assess the consistency of test scores given that participants had to respond to different prompts. If the assessment is sufficiently reliable, it should not matter which particular set of prompts an individual received. Results indicated that the magnitude of the observed correlations between the two prompts across the four conditions of prompt administration, which ranged from .51 to .62, suggested that, for the group of examinees as a whole, the two prompt types measured relatively similar writing constructs. In addition, the patterns of the correlations were generally similar for the subgroups. However, correlations among issue prompts (.52 to .67) were found to be considerably higher than correlations among argument prompts (.36 to .56) across all subgroups. Finally, it was found that the reliability was higher in the issue-argument order than in the argument-issue order (0.70 vs. 0.63).

In addition, if the assessment is sufficiently reliable, very similar score distributions should be seen regardless of which particular prompts are given. Results indicated that, for the total group and most subgroups, about 84% of participants had score differences of less than or equal to 1 point for two issue prompts, and about 82% had score differences of less than or equal to 1 point for two argument prompts; about 98% of the participants in these two conditions had differences of less than or equal to 2 points. Thus, for the majority of participants, there were only minor differences in their scores on two prompts of the same type.

Conclusion

The results of this study guided the decisions made about the design and scoring of the operational Analytical Writing measure. Based on these results, the following decisions were made:

- The issue and argument writing tasks appear to assess relatively similar constructs, supporting the decision to include both types of prompts in the operational assessment and to report a single score based on examinees' average performance on the two prompts.
- Within each task type, most of the prompts were comparable in difficulty, and no important subgroup interactions with prompt classifications were detected. In addition, the assessment was found to be sufficiently reliable. This supports selecting prompts for individual examinees at random from a large pool.
- Analyses indicated some small advantage to administering the issue prompt first and the argument prompt second. Based on this finding, it was decided to administer the prompts in the issue-argument order on the operational test.

References

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

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

Notes

[1] Based on *Psychometric Evaluation of the New GRE® Writing Assessment* (GRE Board Professional Report No. 96-11P), by G. A. Schaeffer, J. B. Briel, and M. E. Fowles, 2001, Princeton, NJ: Educational Testing Service.

[2] *Fall* in this context refers to tests that were taken beginning in August.

1.10 Comparability of Essay Question Variants[1]

Brent Bridgeman, Catherine Trapani, and Jennifer Bivens-Tatum

Ensuring accurate, valid scores for all test takers is a critical issue for testing programs. Evidence has shown that some test takers memorize substantial chunks of existing well-written texts and include this material in their own essays on tests assessing analytical writing ability. In order to discourage this practice and maintain the validity of the inferences that can be made from test scores, the potential use of *essay variants* on the *GRE*® revised General Test Analytical Writing measure was studied.

Essay variants are created from the same prompt: A given prompt (parent) may be used as the basis for one or more different variants that specify different writing tasks in response to the same stimulus. This results in prompts that require responses that are much more closely tied to the specific content of the essay questions. Because the use of variants makes the writing task less predictable, it should reduce the use of prememorized material.

Thus, for example, one test taker may be shown a prompt that presents an argument and a recommended course of action and be asked to do the following: *Write a response in which you discuss what questions would need to be addressed to decide whether the recommendation is likely to have the predicted result. Be sure to explain how the answers to the questions would help to evaluate the recommendation.* Another test taker would see the same prompt, but be asked to do the following: *Write a response in which you examine the unstated assumptions of the argument above. Be sure to explain (a) how the argument depends on those assumptions and (b) what the implications are if the assumptions prove unwarranted.*

Because there are no equating procedures for writing prompts or variants, fairness considerations require that the prompts and variants be of comparable difficulty and yield comparable score distributions. The idea of generating a variety of different questions with comparable difficulty levels from a single parent has been studied in other contexts (Bejar, 1993; Bejar & Braun, 1999; Embretson, 1998; Hively, Patterson, & Page, 1968; Morley, Lawless, & Bridgeman, 2005) but not with use in a high-stakes writing test. This study examined (a) the comparability of score distributions (averages and dispersion) across prompts and variants, (b) differential difficulty of variant types across gender and race/ethnicity subgroups and for test takers whose best language is not English, and (c) consistency of reader reliability across prompts and variants.

Procedure

The GRE revised General Test consists of two separately timed analytical writing tasks: analyze an issue and analyze an argument. The issue task requires the test taker to think critically about a general topic and respond using a set of specific instructions. The argument task requires the test taker to discuss the logical soundness of the author's case according to specific instructions. Six argument variant types and six issue variant types were evaluated in the study.

In the last section of the GRE General Tests administered in winter[2] 2009, test takers viewed a screen inviting them to volunteer to participate in a research project. Participants were randomly assigned to one of the prompt/variant combinations. Not every possible variant type could be generated from all prompts (parent), but examples for every variant type were generated from at least two parents. As a result, sample sizes were somewhat larger for issue essays than argument essays because there were more prompt/variant combinations available: 10,827 issue essays and 7,573 argument essays. Essays were evaluated on a 6-point rating scale in an online scoring environment. Two independent raters rated each essay; raters had been trained on the scoring rubrics for the new variant types. The two ratings were averaged and adjudication rules per GRE program policy were applied.[3]

Results

Comparability of Score Distributions

Averages and score distributions (the number of test takers at each score level) were calculated for each variant type. As indicated in Table 1.10.1, differences across variant types were also small. The grand average across all argument variant types was 2.95, with the *evaluate a recommendation/predicted result* variant type having the largest discrepancy from the grand average (0.13). The largest difference on the 1 (*fundamentally deficient)* to 6 (*outstanding*) rating scale was 0.20 (or a standardized difference of 0.24). Results indicated that average differences between argument prompts within variant type were generally quite small. An exception was the two prompts in the prediction variant type with a difference of 0.25. Similar results were seen for the issue prompts. The grand average of 2.93 for the issue variant types was very close to the grand average of 2.95 for the argument variant types. Most issue variants had averages that fell between 2.80 and 3.05.

Distributions across the different variant types were comparable. The modal score for every variant type was 3. The most notable feature of these distributions is the very low frequency of scores at the higher end (i.e., 5 and 6) of the scale. Several factors may contribute to these low frequencies: (a) the no-stakes nature of the test may have produced unmotivated test takers; (b) the prompt types are unfamiliar to test takers, and, therefore, there was no opportunity to practice on the new prompt types; and (c) the prompt types are unfamiliar to the raters who may have initially scored them very severely.

Table 1.10.1

Types of Variants, Number of Essays, and Average Scores for Each of the Writing Tasks

Analyze an argument	Number of essays	Average scores	Analyze an issue	Number of essays	Average scores
Alternate explanations	1,488	2.89	Claim with reason	1,482	2.83
Evaluate a recommendation	2,210	2.97	Generalization	1,256	2.99
Evaluate a recommendation/ predicted result	566	3.08	Position with counterarguments	1,454	3.05
Evaluate a prediction	766	2.88	Recommendation	2,228	2.98
Specific evidence	1,483	2.88	Recommended policy position	2,277	2.97
Unstated assumptions	1,060	3.01	Two competing positions	2,130	2.89
Total	7,573	2.95	Total	10,827	2.95

Differential Difficulty Across Subgroups

Analysis of covariance (ANCOVA), with the operational writing score as the covariate, was used to evaluate potential interactions with background variables (gender, race/ethnicity, and language background). These background variables were obtained from a background questionnaire that was voluntarily completed at the time test takers registered for the test.

There were no significant interactions of variant type with either gender or race/ethnicity for the argument variant types. Similarly, in the ANCOVA comparing test takers who indicated English as their best language with those for whom English was not their best language, there was a significant main effect for variant type, but the interaction with variant type was not significant.

As with the argument variant types, there was no significant interaction of variant type with either gender or race/ethnicity for the issue variant types, suggesting that no variant type favored any gender or race/ethnicity group. The results for the comparison of test takers who noted English as their best language and test takers who noted that English was not their best language mirrored the results for the argument variant types. For the issue variants, no variant type was differentially difficult for test takers who noted that English was not their best language.

Consistency of Rater Reliability

Rater reliability, using quadratic weighted Kappa and percent exact agreement, was evaluated separately for each parent/variant combination. Although some variation occurred among the different variant types, no variant type appeared to be more or less reliable than any other type. This result was true for both the argument and issue variant types.

Conclusion

The use of variants offers a practical approach in the need to produce a large number of comparable essay questions for a high-stakes test. This study indicated that differences between variant types in terms of averages, distributions, and rater reliabilities were small enough to support the use of variants in the analytical writing tasks for the GRE revised General Test. No variant type was differentially difficult for any of the subgroups examined, which suggests that the use of variants should not impact subgroup differences. This approach seems to be a win-win situation, in that it both enhances validity by reducing the impact of prememorized material and reduces test creation costs.

The study also indicates that the use of variants on the GRE revised General Test should not introduce any more variability than has been observed using parent prompts only. Nevertheless, some variants do create greater variability and, thus, the particular variant used at a given testing session would not be a matter of total indifference to the test taker. As a result, a pairing approach that matches relatively easy issue prompts/variants with relatively difficult argument prompts/variants (and vice versa) is being evaluated as a further revision to the test.

References

Bejar, I. I. (1993). A generative approach to psychological and educational measurement. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (1st ed., pp. 323–357). Hillsdale, NJ: Erlbaum.

Bejar, I. I., & Braun, H. (1999). *Architectural simulations: From research to implementation, Final report to the National Council of Architectural Registration Boards* (Research Memorandum No. RM-99-02). Princeton, NJ: Educational Testing Service.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tasks: Application to abstract reasoning. *Psychological Methods*, *3*, 380–396.

Hively, W., Patterson, H. L., & Page, S. H. (1968). A “universe defined” system of arithmetic achievement tests. *Journal of Educational Measurement*, *5*, 275–290.

Morely, M., Lawless, R., & Bridgeman, B. (2005). Transfer between variants of mathematics test questions. In J. Mestre (Ed.), *Transfer of learning from a modern multidisciplinary perspective* (pp. 313–335). Greenwich, CT: Information Age Publishing.

Notes

[1] Based on “Comparability of Essay Question Variants,” by B. Bridgeman, C. Trapani, and J. Bivens-Tatum, 2011, *Assessing Writing*, *16*(4), pp. 237–255.

[2] *Winter* in this context refers to January and February.

[3] Normal adjudication rules are that ratings from the two raters that differ by more than a point go to a third reader, and the two closest scores are averaged.

Section 2: Creating and Maintaining the Score Scales

A major consideration as part of the revision to the *GRE*® General Test was the score scales that would be used with the revised test. While it was recognized that changes to the scales would impact score users and test takers alike, compelling reasons required that the scales for the Verbal Reasoning and Quantitative Reasoning measures be changed. Two of the chapters in this section detail the considerations used in the decision to change the Verbal and Quantitative scales and provide information on the method used to define the scales. Information on the scale for the Analytical Writing measure is also provided. While the reporting scale for Analytical Writing remained the same, it was important to ensure that the meaning of the scale was stable, despite some alterations to the prompts and section timing. Two chapters in this section are devoted to Analytical Writing issues.

- Chapter 2.1 provides a brief history of the GRE score scales and sets a context for the decision to change the scales for the Verbal and Quantitative measures. It provides an overview of the properties of a useful score scale and how these properties were used to define the new Verbal and Quantitative score scales: (a) the reference group scores should be centered near the scale's midpoint, (b) the score distribution should be unimodal (i.e., have one distinct peak), (c) the score distribution should be nearly symmetric (i.e., have the same shape on both sides of the midpoint of the scale), (d) the score distribution should follow a commonly recognized form (e.g., a bell-shaped distribution), (e) the scores' working range should go beyond the reported range, (f) the number of points on the scale should not be greater than the number of possible raw score points, and (g) the scale should be evaluated periodically and repaired if necessary.

- Chapter 2.2 presents the process by which the new score scales for the Verbal and Quantitative measures were defined. The chapter outlines the challenges that were faced, such as using actual test takers instead of participants in a field trial to create the new scales, resolving demographic shifts related to the growing population of test takers from outside of the United States, and aligning the Verbal and Quantitative scales. It describes the goals of the rescaling, the scaling procedures that were used, and how well the resulting scales met the goals of the rescaling. While this chapter covers several fairly technical methods, it is still written to be informative to those with a nontechnical background.

- Chapter 2.3 focuses on the Analytical Writing scale, providing overviews of the scoring process and the monitoring of the scale to ensure that reported scores are fair and accurate for all test takers. The chapter describes how Analytical Writing is assembled and delivered to test takers. Two ratings are produced for each essay: one using the *e-rater®* scoring engine automated scoring software and one using a human rater. Ratings range from 0 (*bottom*) to 6 (*top*). Comprehensive postadministration analyses are conducted monthly to evaluate the quality of the ratings, determine the measurement characteristics of the essay prompts, and monitor performance for the total group of test takers and major subgroups (by region, gender, and race/ethnicity). In addition, yearly monitoring analyses are performed that provide empirical evidence for scoring stability and test fairness. Comparisons of the correlations between(a) the Verbal/Quantitative and Analytical Writing scores and (b) performance of total, regional, and race/ethnicity groups show the continuity of the scale for the previous and the revised version of the Analytical Writing measure.

- Chapter 2.4 describes an evaluation of the ETS automated scoring engine, e-rater, as a way to ensure stability of the scale used with the Analytical Writing measure. Analyses examined the agreement of e-rater with human scores in terms of percent agreement, correlation, and mean differences and the relationship of external variables to the scores produced by e-rater. The study also attempted to determine if changes in agreements between human and e-rater scores provided a plausible method for monitoring changes in human scoring over time. Results indicated that exact and adjacent rates of agreement between human raters and e-raters were acceptable and on par with previous research. In addition, results indicated that monitoring discrepancies between scores generated by human raters and e-rater over time helps to assure consistency in the meaning of Analytical Writing scores.

2.1 Considerations in Choosing a Reporting Scale for the *GRE®* revised General Test[1]

Marna Golub-Smith and Cathy Wendler

In order to be able to interpret test performance and make meaningful distinctions among individuals, test scores need to be reported on a scale with some predefined units. The choice of reporting scale[2] is a critical and fundamental need for a testing program (Wendler & Walker, 2006). This chapter describes some of the important issues leading to the decision to redefine the scales for the Verbal Reasoning and Quantitative Reasoning measures of the GRE® revised General Test.

Brief History of the GRE Score Scales

The first *GRE* consisted of eight tests (called the Profile Tests) in specific subject areas administered in the fall of 1937. However, the score scale most familiar to the graduate community was not defined until 15 years later. The original scale was created in 1941 using a group of first-year graduate men attending four Eastern universities. The 200 to 800 scale used for the GRE General Test until the introduction of the GRE revised General Test in 2011 was established in a special study conducted in the spring of 1952 (Schultz & Angoff, 1956). The rationale for changing the scale at the time was that the groups being tested were more heterogeneous and generally lower in ability than the 1941 group and that the tests had undergone revisions to content and scope. In the 1952 study, 2,095 college seniors representing 11 colleges took the GRE Aptitude and Advanced Tests[3] in their major area of study. It was believed that the test performance of these students would provide a reasonable estimate of the ability of other college students in the United States. The raw scores obtained by this group were placed on a scale with a maximum permitted range of 200 to 900, by setting the raw score mean equal to a scale score of 500 and the standard deviation to 100. Thus, for this 1952 group of examinees, the mean scale score and standard deviation were identical for both the Verbal and Quantitative measures (Briel, O’Neill, & Scheuneman, 1993).

Raw scores were originally corrected for guessing by subtracting a fraction of the wrong responses from the total number of right responses. Unanswered questions received a zero weight. In October 1981, the method for calculating raw scores was changed to a summation of the total number of right answers. In addition, the scale’s top score was truncated to 800 as a result of the change to rights scoring. While a slight shift in the scale occurred with this change, the continued use of the (truncated) scale was justified based on the work of Angoff and Schrader (1981).

Since 1952, the group of students who take the GRE General Test has changed considerably. Examinees are much more diverse in terms of ability level, gender, and ethnicity/race than the group used to set the scales in 1952. In addition, examinees now come from many countries; slightly more than 30% of test takers are from countries outside of the United States. The test itself also continued to change after 1952 and included revisions to question types, content, test length, timing, scoring, and mode of delivery (from paper based to computer delivered).

One of the results of the shift in population and changes to the GRE General Test was that the means for both the Verbal and Quantitative measures shifted from what was perceived as the midpoint of the scale (i.e., 500). Thus, in 2002, 50 years after the scale was defined, examinees scoring 500 might erroneously conclude that their scores were the same—"about average"—on both measures, when in fact their verbal skills were above average and their quantitative skills below average. In addition, as the population of examinees from outside of the United States grew, the number of examinees achieving the top scale scores between 760 and 800 increased. While this is not surprising given that the majority of international students major in the sciences and, therefore, have strong quantitative skills, it nevertheless exemplified the differences between current examinees and the original reference group used to set the scale (Golub-Smith, 2005).

Properties of a Useful Score Scale

Since the scale score is what is reported to examinees and institutions, it forms the framework by which scores are interpreted. Therefore, the scale should be aligned with the intended use of the scores in order to facilitate meaningful score interpretation and minimize misinterpretations. The particular scale used also has implications for test specifications, equating, and test reliability and validity (Dorans, 2002; Petersen, Kolen, & Hoover, 1989; Wendler & Walker, 2006).

For a test such as the GRE General Test, where performance over a broad spectrum of ability impacts admissions and other decisions, it is important that the scale facilitates meaningful score interpretation across most of the score range. Dorans (2002) described a set of properties that a scale should have if it is aligned with the intended uses of the scores. Dorans, Yu, and Guo (2006) provided a way to evaluate the extent to which a scale is aligned with its intended uses. The seven properties the scale should possess are as follows (Dorans, 2002, p. 60):

- The scores of the reference group used to define the scale should be *centered* near the midpoint of the scale. The average score (mean or median) in the reference group should be on or near the middle of the scale.
- The distribution of the aligned scores for the scale-defining reference group should be *unimodal*, and that mode should be near the midpoint of the scale.
- The distribution should be nearly *symmetric* about the average score.

- The shape of the distribution should follow a commonly *recognized form*, such as the bell-shaped normal curve.[4]
- The *working range* of scores should extend enough beyond the *reported range* of scores to permit shifts in population away from the scale midpoint without missing the endpoints of the scale.
- The number of scale units should not exceed the number of raw score points, which is usually a simple function of the number of questions.
- A score scale should be viewed as an infrastructure that is likely to require repair.

The Reference Group Scores Should Be Centered Near the Scale's Midpoint

The *reference group* is the population used to define the scale and is generally the group for whom the test is designed. The *testing population* is the group who actually takes the test. Ideally, the testing population and reference group are identical. However, over time some general mismatch with those for whom the test is designed and those who show up to take it can occur. In terms of defining a new scale, the average score (mean or median) in the reference group should be ideally on or near the midpoint of the scale, and this should reflect the distribution of scores for the testing population.

The Distribution of Scores Should Be Unimodal

The distribution of the scores for the reference group on the scale should have one distinct peak (mode), and that mode should be near the midpoint of the scale.

The Distribution of Scores Should Be Nearly Symmetric

The distribution of scale scores should have the same shape on both sides of the average score at the midpoint of the scale. This helps establish interpretability for scores both above and below the center of the scale.

The Distribution Shape Should Follow a Commonly Recognized Form

Dorans (2002) originally recommended the normal, bell-shaped distribution because of its symmetry, its single mode, and its correspondence to intuitions about distributions of general proficiencies.

The Scores' Working Range Should Extend Beyond the Reported Scores' Range

This ensures use of the full score scale range without introducing unwanted distortions that may be caused by future shifts in the reference population ability distribution. It also allows the average score of the reference group to move away from the scale midpoint without stressing the endpoints of the scale. If, for example, the highest raw score falls short of the maximum reported score, scores at the top end of the scale may need to be forced up to the maximum reported score using a method that may not produce equivalent scores across different versions of the test.

There Should Not Be a Greater Number of Scale Points Than Raw Score Points

The number of raw score points is usually a function of the number of questions that exist on the test (i.e., 1 point is given for each correct answer, and an examinee's raw score is the total number of questions correctly answered). A fundamental requirement for a useful scale is that there is at least one question for each point on the reported scale. Too many score points on the scale compared to the number of raw score points may suggest more precision than is justified and lead to improper differentiation of examinees.

The Scale Should Be Periodically Evaluated and Repaired

Repair to the scale should be considered if the average score of the testing population moves significantly away from the midpoint of the scale, if the distribution moves sufficiently away from one of the endpoints of the scale to jeopardize the integrity of the scale at that endpoint, when the original reference group has changed to the point where it is no longer appropriate, or when substantial content revisions change the meaning of the existing scale.

Conclusion: Rescaling Considerations

Based on the Dorans (2002) definition of the properties of a useful scale, the 200 to 800 scale used in the past for the GRE General Test met the criteria for rescaling and warranted repair efforts. That is, the average (mean) scores for the testing populations had shifted away from the midpoint of the scale, the demographics of the original reference group bore little resemblance to the current testing population, and a number of content and scoring changes were made to the test. Given this context, the *Standards for Educational and Psychological Testing* and the *ETS Standards for Quality and Fairness* advise against using the same scale metric for reporting scores from a revised test (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999; Educational Testing Service, 2002). As a result, new scales, ranging from 130 to 170, were introduced for the Verbal and Quantitative measures with the release of the GRE revised General Test in August 2011.

In addition to reflecting professional standards, introducing new scales allowed the number of scale points to better align with the configurations for the revised test. The procedures used to define the new scales (Golub-Smith & Moses, Chapter 2.2, this volume) were based on the performance of a large operational cohort of GRE examinees to minimize the risk that any ceiling or floor effects would be introduced as a result of the choice of scaling population. The scaling realigned the Verbal and Quantitative scores, so that the mean scores and standard deviations for distributions of both measures would be more closely matched. In addition, the use of a new 130 to 170 metric avoided confusion with the 200 to 800 metric and other widely used score scales for this population. Finally, in order to minimize possible confusion in the user community by introducing the new metric, an extensive communication effort with score users was undertaken.

References

Angoff, W. H., & Schrader, W. B. (1981). *A study of alternative methods for equating rights scores to formula scores* (Research Report No. RR-81-08). Princeton, NJ: Educational Testing Service.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Briel, J. B., O'Neill, K. A., & Scheuneman, J. D. (Eds.). (1993). *GRE® technical manual*. Princeton, NJ: Educational Testing Service.

Dorans, N. J. (2002). Recentering and realigning the SAT score distributions: How and why. *Journal of Educational Measurement, 39,* 59–84.

Dorans, N. J., Yu, L., & Guo, F. (2006). *Evaluating scale fit for broad-range admissions tests: An illustration using GRE data* (Research Memorandum No. RM-06-04). Princeton, NJ: Educational Testing Service.

Educational Testing Service. (2002). *ETS standards for quality and fairness*. Princeton, NJ: Author.

Golub-Smith, M. (2005). *Scales for the new GRE.* Unpublished manuscript, Educational Testing Service, Princeton, NJ.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd. ed., pp. 221–262). New York, NY: Macmillan.

Schultz, M. K., & Angoff, W. H. (1956). The development of new scales for the aptitude and advanced tests of the Graduate Record Examinations. *Journal of Educational Psychology, 47,* 285–294.

Wendler, C. LW., & Walker, M. E. (2006). Designing multiple test forms for large-scale tests. In S. Downing & T. Haladyna (Eds.), *Handbook of test development* (pp. 445–467). Hillsdale, NJ: Erlbaum.

Notes

[1] Based on *Determining an Appropriate Scale for the GRE General Test: Considerations Going Forward*, by C. Wendler and M. Golub-Smith, April 2007, paper presented at the annual meeting of the American Psychological Association, San Francisco, CA.

[2] The *raw score* scale is the most simple scale and generally involves adding up the number of correct answers for a test taker. Raw scores, however, are limited in their generalizability because they are specific to the particular version of the test and cannot be compared across test versions because the test forms differ somewhat in difficulty. Therefore, test scores are reported on a *score reporting* scale (usually referred to as the score scale for a test). The raw scores are linearly or nonlinearly transformed to the score scale using a number of different statistical techniques. The score scale is often established by taking the raw score distribution of a particular group of examinees (referred to as the *reference group*) and placing the distribution on the reporting scale by setting the mean and standard deviation to specified values. On this type of scale, the score indicates an examinee's relative standing in the reference group.

[3] In 1952, the GRE General Test was called the GRE Aptitude Test; the GRE Subject Tests were called the GRE Advanced Tests.

[4] Although not exactly normal, the GRE scale score distributions had features similar to those of normal distributions.

2.2 How the Scales for the *GRE®* revised General Test Were Defined

Marna Golub-Smith and Tim Moses

A score scale provides the framework by which test users interpret performance and make decisions. Unlike physical properties such as temperature and length, in the measurement of cognitive abilities there is no observable relationship between the underlying ability and the test score. Therefore, the scale needs to be defined in such a way as to facilitate interpretations and decisions based on the test scores that are appropriate for the assumed abilities. One way to do this is to define the characteristics of a scale in terms of a distribution of scores for a representative group of test takers, or *reference group*. For the resulting scale to remain meaningful, this reference group must be reflective of the intended testing population, as well as subsequent test takers.

In August 2011, with the launch of the *GRE®* revised General Test, the scales for the Verbal Reasoning and Quantitative Reasoning measures were redefined. This revision of the GRE General Test's existing scales was only the second such revision since the program was established. In describing the rationale for the first rescaling in 1952, Schultz and Angoff (1956) explained how the population using the test at the time had changed in such a manner that the scale had "little inherent normative value" (p. 285). As was described previously (Golub-Smith & Wendler, Chapter 2.1, this volume), since 1952, there have been many changes to the content, scoring, and delivery of the GRE General Test, as well as demographic shifts in the populations taking the test. The introduction of the GRE revised General Test brought additional changes, including new question types and new testing tools (the ability to review answers and the use of a calculator), for the Verbal Reasoning and Quantitative Reasoning measures. All of these cumulative changes over time impacted the interpretations of the scores to the point where a redefinition of these scales was warranted. The actual procedures used to define the new scales for the GRE General Test are discussed here.

Challenges and Solutions for the Rescaling

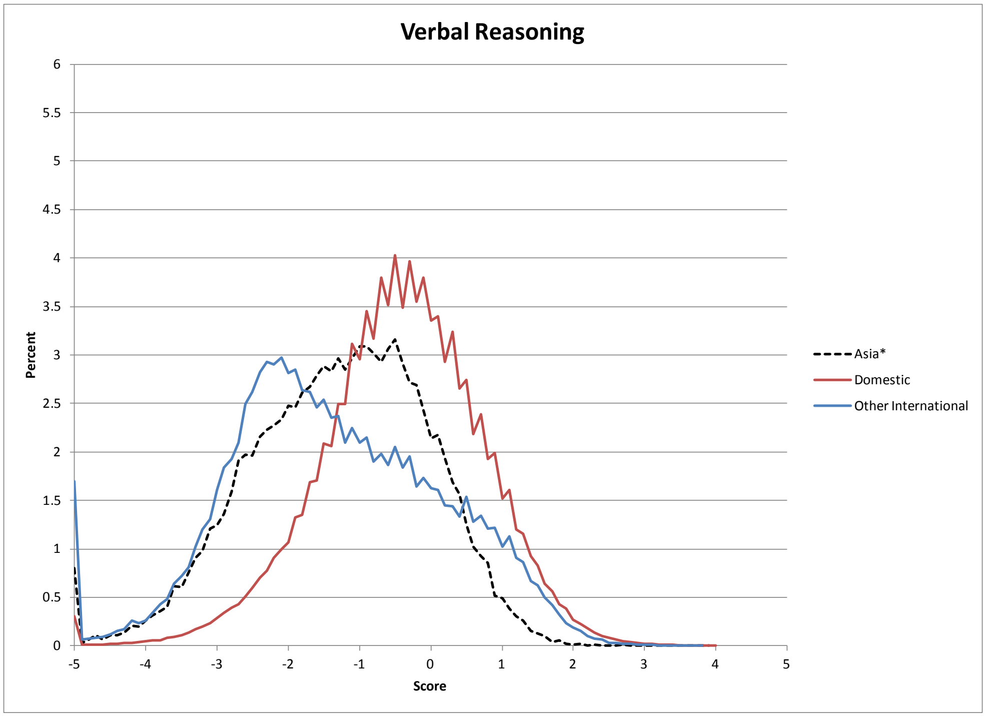
There were several challenges in rescaling the GRE revised General Test. GRE statisticians wanted the scales to be defined on a reference group that was representative of the GRE testing population, and they wanted the reference group test performance to reflect its actual ability. That eliminated the possibility of scaling using a *field trial* group because volunteer field trial groups tend to be less motivated than actual examinees and result in biased estimates of proficiency (see Wise, 2007; Wise & DeMars, 2005). In order to have enough time to perform the scaling, the GRE program had to delay the reporting of scores for the scaling group. From previous experience introducing new tests, there was a concern that examinees might decide to wait awhile before registering to take the new test, especially since they would not receive their scores for 8 to 12 weeks after testing. Therefore, to encourage examinees to test early, an incentive 50% discount was offered for those testing in August and September. The final scaling reference population was composed of those examinees who tested between August 1 and October 2, 2011.

The GRE population demographic shifts that prompted the rescaling also posed some unique challenges. Unlike 60 years ago, when the 200 to 800 scale was defined, the total GRE population is a heterogeneous mixture of three distinct subpopulations: domestic examinees (consisting of U.S. citizens who tested at a test center in the United States or U.S. territory), examinees from Asia (consisting of examinees from China, Taiwan, Hong Kong, and Korea), and other international (consisting of non–U.S. citizens who were not from China, Taiwan, Hong Kong, or Korea) examinees. Slightly more than 30% of examinees between 2007 and 2010 took the GRE General Test in test centers outside the United States. The patterns of performance on the Verbal Reasoning and Quantitative Reasoning measures are very different for these subpopulations. International examinees, especially those from Asia, tend to major in engineering and the sciences and outperform domestic examinees on the Quantitative Reasoning measure. The opposite is true for the Verbal Reasoning measure.

Figures 2.2.1 and 2.2.2 illustrate this difference in the performance of the three subgroups on the two measures, based on their performance on the GRE General Test during the 2008–2009 testing year. The figures plot the score distributions of Verbal Reasoning and Quantitative Reasoning scores. These scores are expressed on a scale that is derived from item response theory (IRT; Lord, 1980), the mathematical model the GRE General Test scores have been based on since the mid-1990s. This scale has a mean of 0 and a standard deviation of 1.

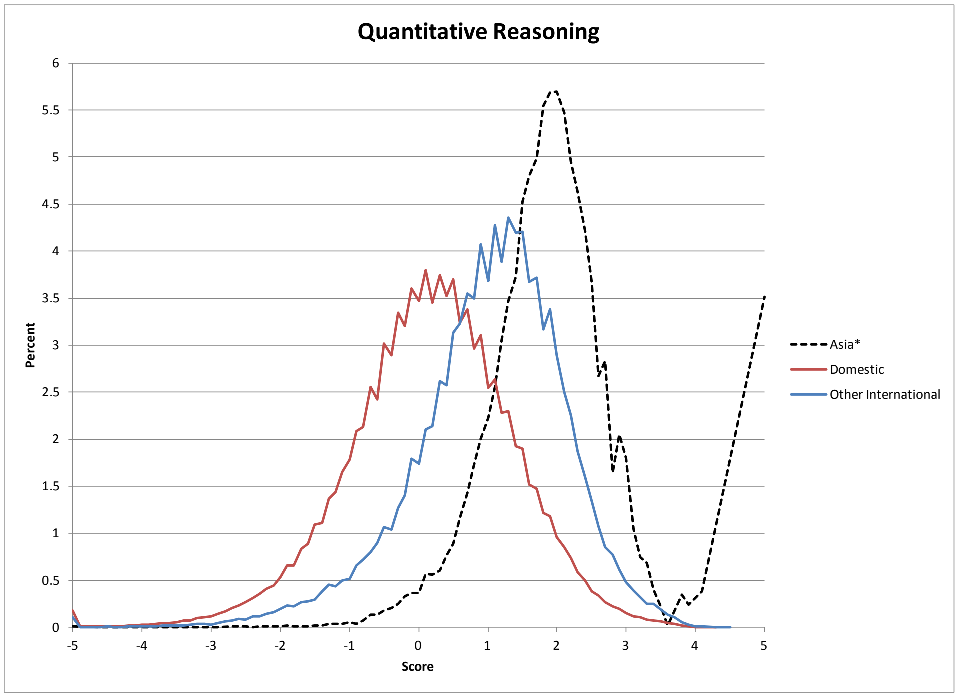
Since one of the purposes for the scaling was to better align the Verbal Reasoning and Quantitative Reasoning scales, it was important to include all populations in the scaling of the two measures. However, from a series of simulations of the GRE revised General Test, we knew that the traditional scaling methods (Kolen & Brennan, 2004), such as those based on normalizing the observed distributions and then defining the scale by setting the average and standard deviation, would not work very well. In fact, it would lead to large gaps in the upper part of the scale.

The solution to this problem came by way of a procedure that not only set the average and standard deviation of the scale, but also set the degree of skewness[1] and kurtosis[2] (Moses & Golub-Smith, 2011). While it is generally believed that measures of cognitive ability are relatively symmetric[3] across large populations, we did not observe that for the GRE General Test. By definition, the population choosing to apply to graduate school is a narrow self-selected group. So this ability to adjust the scale characteristics would help us meet our goals.



Note. Asia consists of examinees from China, Taiwan, Hong Kong, and Korea.

Figure 2.2.1. A distribution of the Verbal Reasoning scores for the 2008–2009 norm group.



Note. Asia consists of examinees from China, Taiwan, Hong Kong, and Korea.

Figure 2.2.2. A distribution of the Quantitative Reasoning scores for the 2008–2009 norm group.

Defining the Goals of the Rescaling

Given that we were planning to use an empirically defined scaling procedure that set the first four moments[4] of the distribution, rather than a theoretical one (e.g., normalization[5]), it was important to articulate some goals to use as criteria for selecting a reasonable scaling solution. Prior to the launch of the GRE revised General Test, seven scaling goals were defined:

1. The range of the Verbal Reasoning and Quantitative Reasoning scales would be 130–170, in 1-point increments.
2. The total number of score gaps, especially at the top of the scale, would be minimized.
3. The pile-up of scores at the top of the scale for the Quantitative Reasoning measure (Figure 2.2.2) would be reduced, but there would be enough density at the top so that if the difficulty of the tests would change over time, gaps would not be introduced at the top.
4. The distribution of scale scores for both Verbal Reasoning and Quantitative Reasoning measures would have an average of 150 and similar standard deviations for the entire group that is tested during the first year of the GRE revised General Test.
5. The scale transformation would facilitate the interpretation of the concordance relationship between the old and new scales.[6]
6. The scale score distributions would not deviate too far from symmetry. Verbal Reasoning and Quantitative Reasoning measures would have score distributions with somewhat similar shapes. These intended distributional characteristics describe the entire group that would be tested during the first year.
7. Conditional standard errors of measurement[7] would be similar across the score scale.

These goals represented the *ideal*. The program recognized that all the goals could probably not be entirely met simultaneously. In the end, priorities would have to be set.

Scaling Procedures

The Verbal Reasoning and Quantitative Reasoning measures on the GRE revised General Test are multistage adaptive tests (MST; Robin & Steffen, Chapter 3.3, this volume). Each test consists of different paths that share a common routing section. Examinees receive a routing section and then a second-stage section. Performance on the routing section determines which second-stage sections an examinee receives. In total, an examinee is administered 40 operational questions, 20 in each section. Multiple versions of every section are administered daily.

The test score on the MST is the sum of the number of correct answers on both sections. This raw number correct score is then transformed into an equated number correct score (ENR) on a 50-question reference test to control for the differences in the difficulties of each MST. This is accomplished by using IRT true score equating. (For a description of this type of equating, refer to Holland & Dorans, 2006, or Kolen & Brennan, 2004.) Rounded ENR scores range from 0 to 50. It is this distribution of rounded ENR scores on which the scaling transformations were based.

As mentioned above, the scaling reference population for the Verbal Reasoning and Quantitative Reasoning measures consisted of all examinees with reportable Verbal Reasoning and Quantitative Reasoning scores who tested between August 1 and October 2, 2011. The size of the scaling population was 146,504 examinees. Table 2.2.1 provides a distribution of the demographics that defined this scaling population compared with the demographics of the 2007–2010 3-year-norm group. The notable difference between the scaling population and the norm group was that the former was composed of slightly more domestic examinees and seniors. In terms of ethnicity and gender, the distributions were quite similar. With regard to major field and graduate objective, the large amount of missing data for the scaling population makes it hard to compare. Just why the scaling population had so much demographic data missing is unknown.

The ENR distribution used to set the scale was a composite distribution. It consisted of a weighted sum of three smoothed distributions: domestic, Asia, and other international. The individual weights were set so that the resulting distributions would reflect the 2007–2010 norm group distributions with an overall percentage weighting of 65, 7, and 27 for the groups, respectively.[8] The smoothing method was based on using loglinear models to fit the distributions' first seven moments (Holland & Thayer, 2000).

The scaling procedure involved finding a set of parameters in a polynomial function of the ENR scores that would result in a distribution with a prespecified scale average, standard deviation, skewness, and kurtosis (Moses & Golub-Smith, 2011). The exact specification of what these values should be to produce a viable scale were empirically determined through multiple iterations using a range of values that had been proposed from previous simulation studies. The evaluation was based on meeting the overall scaling goals listed above. These goals were operationalized and prioritized so that potential scaling solutions produced from the multiple iterations were only considered when they met the following four criteria:

- The percentage of examinees with a scale score of 170 was between 0.5% and 4.0% for the Quantitative Reasoning measure and between 0.1% and 4.0% for the Verbal Reasoning measure.

Table 2.2.1

A Comparison of the Demographics of the Scaling Population Compared to the 2007–2010 Norm Group

	Scaling group		Norm group	
Category	*N*	%	*N*	%
Total	146,504	100	1,653,273	100
Nationality				
Domestic	105,832	72	1,082,622	65
Asia [b]	8,535	6	118,370	7
International	32,137	22	452,281	27
Gender				
Male	55,060	38	670,914	41
Female	78,204	53	911,625	55
Missing	13,240	9	70,734	4
Ethnicity [c]				
White	73,521	74	796,558	75
Asian	6,556	7	63,149	6
Black	8,287	8	97,522	9
Hispanic	7,031	7	68,583	6
Other	4,563	5	42,174	4
Broad undergraduate major field				
Engineering	13,306	9	154,677	9
Business	3,704	3	57,475	3
Education	5,444	4	73,319	4
Humanities & arts	13,374	9	181,043	11
Life sciences	24,816	17	304,477	18
Missing	35,550	24	278,893	17
Other fields	12,171	8	168,925	10
Physical sciences	12,559	9	143,366	9
Social sciences	24,144	16	289,255	18
Undecided	1,436	1	1,843	[a]
Educational status				
Senior	58,564	40	524,455	32
Graduate student	13,145	9	150,849	9
Junior	4,579	3	60,433	4
Missing	279	[a]	116,053	7
Other	9,764	7	134,901	8
Sophomore	704	[a]	9,888	1
College graduate	42,720	29	460,537	28
Master’s degree	16,749	11	196,157	12
Graduate objective				
Doctorate	45,650	31	578,834	35
Intermediate	983	1	16,057	1
Master’s degree	55,520	38	821,353	50
Missing	42,255	29	201,250	12
Nondegree	200	[a]	5,467	[a]
None planned	304	[a]	6,985	[a]
Postdoctoral	1,592	1	23,327	1

[a] Indicates a percentage less than 0.5. [b] Asia consists of examinees from China, Taiwan, Hong Kong, and Korea.
[c] Ethnicity is only collected for examinees who state they are U.S. citizens.

- The top scale score of 170 was possible for every panel (the unique collection of routing and second-stage sections).
- There was no more than one score gap for scale scores between 168 and 170 in every panel’s observed score distribution.
- The scale scores were reasonably *lined up* with the 200–800 scale.[9]

While multiple solutions survived the initial screening, the scaling solutions actually chosen were the ones that were most successful at satisfying the scaling goals.

An evaluation of predicted old GRE 200–800 scores for the scaling population led to the assumption that this group was more able, both in the Quantitative Reasoning and Verbal Reasoning scores, than the entire GRE testing population. The goal of the scaling was to have an overall average for the population of 150. However, given that the scaling group seemed to be more able than the overall population, GRE statisticians determined that the average for this group should be set to a value above 150; otherwise, the full-year average might be lower than 150. To estimate what this value might be, the performance of the examinees taking the GRE General Test in the period from August to October 2009 was compared to the entire testing year 2009–2010. The standardized differences between these groups were 0.1 for Verbal Reasoning and 0.2 for Quantitative Reasoning. Based on these differences and assumed scale standard deviations of 8.75, it was estimated that the averages for Verbal Reasoning and Quantitative Reasoning should be set at 151 and 152, respectively, to ultimately have an average of 150 at the end of the year.

Conclusion: End-of-Year Scale Characteristics and Results

The 2011–2012 testing year ended in June 2012, 11 months after the GRE revised General Test was introduced. One question that can begin to be answered now is how successful was the GRE General Test in meeting its overall scaling goals, especially related to setting the scale average to produce an overall average of 150 for the total year (Scaling Goal #4). The norms for the 11 months ending June 2012 indicate an observed average of 150.8 and a standard deviation of 8.5 for Verbal Reasoning and 151.3 and 8.7 for Quantitative Reasoning. While the new scale scores’ averages and distributions are slightly higher than might have been desired, the results should be interpreted in terms of being obtained in an abbreviated and atypical year. Many more examinees took the GRE General Test between May and July 2011 than in that time period in previous years. With regard to the averages and distributions, the results may not settle until there is a larger and more stable 3-year-norm group.

With regard to the reduction in the pile-up of scores at the top of the Quantitative Reasoning scale (Scaling Goal #3), results indicate success. Compared to the 4% of the population in 2007–2010 that obtained the top score of 800, only 1.3% obtained the top score of 170 on the GRE revised General Test. Figure 2.2.3 provides a plot of the end-of-year Verbal Reasoning and Quantitative Reasoning distributions. The other scaling goals were also regarded as basically met in that the new scale scores had very few gaps in their score ranges, and they lined up reasonably well with the 200 to 800 scale in terms of their averages and the conversions of the highest, middle, and lowest scores.

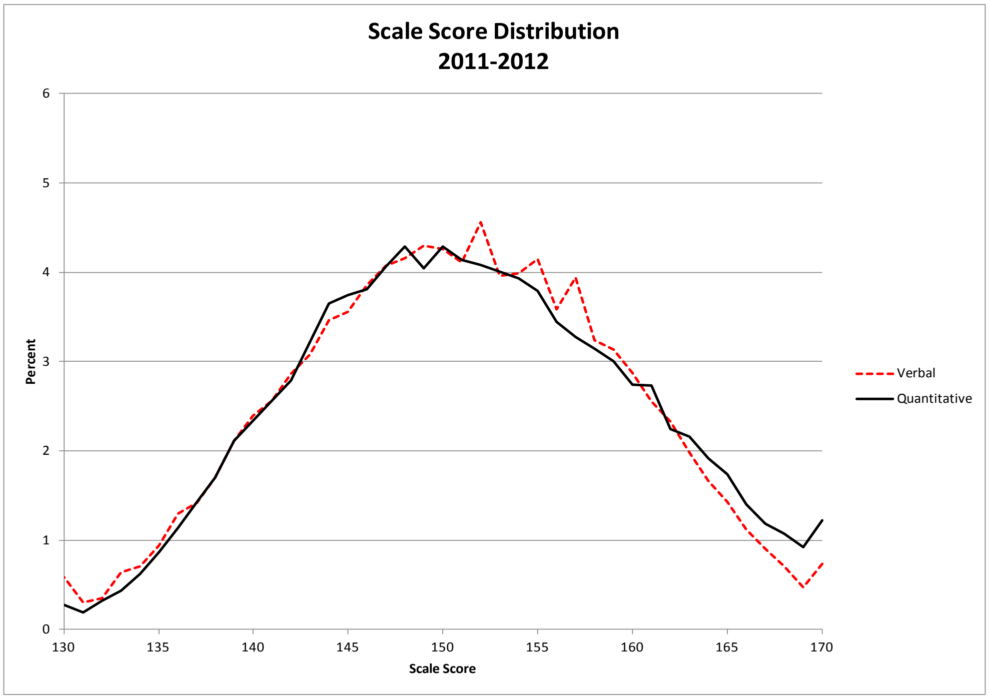


Figure 2.2.3. A distribution of GRE revised General Test scale scores for the 2011–2012 year.

References

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 189–220). Westport, CT: American Council on Education & Praeger.

Holland, P. W., & Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics*, *25*, 133–183.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling and linking: Methods and practices* (2nd ed.). New York, NY: Springer-Verlag.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Moses, T., & Golub-Smith, M. (2011). *A scaling method that produces scale score distributions with specific skewness and kurtosis* (Research Memorandum No. RM-11-04). Princeton, NJ: Educational Testing Service.

Schultz, M. K., & Angoff, W. H. (1956). The development of new scales for the aptitude and advanced tests of the Graduate Record Examinations. *Journal of Educational Psychology*, *47*, 285–295.

Wise, S. L. (2007, October). Examinee effort and test score validity. Paper presented at the meeting of the Northeastern Educational Research Association, Rocky Hill, CT.

Wise, S. L., & DeMars, C. E. (2005). Examinee motivation in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, *10*, 1–18.

Notes

[1] Skewness is a measure of the symmetry of a frequency distribution.

[2] Kurtosis is a measure of how peaked or flat a frequency distribution is. Relative to a normal distribution with a kurtosis value of 0, kurtosis values greater than 0 reflect peaked or leptokurtic distributions, and kurtosis values less than 0 reflect flat or platykurtic distributions.

[3] A symmetric distribution is one where the observations equidistantly above and below the center have the same frequency, resulting in a shape that is the same both above and below the center. In a symmetric distribution, the average is equal to the median, and both define the center of the distribution.

[4] Moments of distributions are characteristics that describe the distribution's shape. The first moment is the average, the second moment is related to the variance and the standard deviation, the third moment is related to skewness, and the fourth moment is related to kurtosis.

[5] Normalization is a scaling technique whereby an observed frequency distribution is transformed to approximate the characteristics of a normal distribution, a theoretical bell-shaped symmetric (skewness = 0) probability distribution.

[6] The concordance between the two scales was developed to help score users transition to the new scale. The concordance was estimated in a special study by matching scores that corresponded to the same percentile rank for a group of GRE examinees who took both tests. For each score on the old GRE scale, the concordance provided an approximate equivalent on the new scale.

[7] The conditional standard error of measurement is an index that provides a measure of the error in an examinee's score at a given point on the score scale.

[8] Actual percentages were unrounded and summed to 100.

[9] This criterion evaluated the impact of the scale transformation on the concordance of the highest, middle, and lowest scores.

2.3 Evaluating and Maintaining the Psychometric and Scoring Characteristics of the Revised Analytical Writing Measure

Frédéric Robin and Sooyeon Kim

The Analytical Writing measure is part of the *GRE*® revised General Test launched in 2011. The measure assesses the same critical thinking and analytical writing skills as the previous GRE Analytical Writing version that had been in use since 2002. It consists of two separately timed analytical writing tasks: analyze an issue (issue) and analyze an argument (argument; Educational Testing Service [ETS], 2013a). The test taker is allowed 30 minutes for each writing task. The Analytical Writing tasks assess test takers' ability to understand, analyze, and evaluate arguments according to specific instructions and to convey their evaluation clearly in their writing. The two tasks are complementary in that one requires test takers to construct their own argument by taking a position and providing evidence supporting their views on an issue and the other requires test takers to evaluate someone else's argument by assessing its claims and evaluating the evidence it provides. Each task is accompanied by a specific instruction.

Compared to the modifications made to the Verbal Reasoning and Quantitative Reasoning measures of the revised test, changes to the Analytical Writing measure were minimal. The current version of the Analytical Writing measure differs from the previous version in two aspects: (a) reduced testing time for the issue task (45 to 30 minutes) and (b) varied sets of instructions for issue and argument tasks (Robin & Zhao, Chapter 1.8, this volume; Zhao, Zhu, Guo, Zeller, & Bannerjee, 2006). The goal of utilizing varied sets of instructions is to reduce the predictability of the writing task in order to diminish the probability of test takers employing memorized text in their essay.[1] Given these changes, extensive revisions in the issue and argument rating rubrics were made and raters were retrained. However, there was no change in the construct measured, and the raters were still instructed to rate each task holistically using the same 0 to 6 point score scale. Therefore the expectation that the scale of the revised version would be comparable to the scale of the previous one, leading to the same interpretation of the Analytical Writing scores (ETS, 2013a) needed to be confirmed.

Revisions were also made to the assembly process that generates the pairs of issue and argument tasks delivered to the test takers. Besides adapting to the new delivery design and delivery infrastructure (Robin & Steffen, Chapter 3.3, this volume), these revisions focused on maintaining test security[2] and on ensuring that all the tests delivered have the same level of difficulty and, therefore, provide the same opportunity to test takers to demonstrate their writing ability.

In this chapter, we briefly describe test form assembly to help the reader understand the extent to which it contributes to enhancing test security and fairness. We then describe the scoring process and the analyses that are conducted on a regular basis to monitor and evaluate the scoring of the responses. Finally, using the test data collected before and after the launch of the revised test, we provide empirical information that confirms the expectation that the scores on the revised version would be comparable to scores on the previous version.

Test Assembly and Delivery

As part of the GRE General Test, the Analytical Writing measure is delivered to test takers worldwide. In this context, one concern is to ensure that test takers cannot predict the test they will be assigned to. This is addressed by assembling large numbers of tests, essentially by random draws from the relatively large operational pool of disclosed issue and argument topics (ETS, 2013b). Then, under the operational testing setting, a particular form is selected from the large assembly batch at the time of delivery. Another concern is that, although prompts are developed to be equivalent in difficulty and are scored using the same task-specific rating rubrics, tryout and operational data have shown that variations in prompt difficulty,[3] while not large enough to require adjustment through statistical equating,[4] are not negligible (see Schaeffer, Briel, & Fowles, 2001). Thus, to enhance fairness, descriptive statistics characterizing prompt difficulty derived from yearly test data are now used in the test assembly process to further reduce variations in test difficulty. Because of the added test difficulty constraint, for example, the pairing of relatively easy (hard) issue and argument prompts will be rejected. More specifically, the assembly process is designed to reject any pairs for which their averaged means (i.e., difficulty) differ from the grand mean of all available pairs by more than 0.1—well below the standard error of measurement of 0.4 (ETS, 2013c, Table 5).

Scoring Process, Monitoring, and Major Outcomes

Unlike the Verbal Reasoning and Quantitative Reasoning measures, which are objectively scored, the scoring of the Analytical Writing measure relies on ratings of expert judges. As a result, the reliability and accuracy of the Analytical Writing scores depend on the reliability and accuracy of the ratings, as well as on the measurement properties of the prompts themselves. Therefore, the scoring process was developed to maximize the reliability of the ratings and to minimize any idiosyncratic or systematic rater effects, such as severity or leniency. In this section, we describe the scoring process and its monitoring, and we summarize major outcomes of the Analytical Writing measure scores collected during the first year of the GRE revised General Test.

For each issue or argument task, at least two 0 to 6 ratings are produced: one by a trained human rater and one by the *e-rater*® scoring engine (Williamson, Xi, & Breyer, 2012; also see Section 4 chapters, this volume). Human raters provide integer ratings on the 0 to 6 scale, whereas e-rater provides ratings on a continuous scale ranging from 0 to 6. If the ratings closely agree, then the final essay score is the human rater's rating. Otherwise, one more human rater is called on to provide the second rating and the final score is the average of the two human ratings. The final Analytical Writing score is the average of the issue and argument scores, rounded to one half point.[5] Because each rater scores either issue or argument, not both, at least two independent human ratings are used in producing the final Analytical Writing reported score.

All the raters are thoroughly trained in the use of the rating rubrics in order to enhance the reliability of the ratings and, therefore, the reliability of the Analytical Writing scores. Additionally, daily rating sessions are organized for each task so that each rater can concentrate on either issue or argument. Raters are required to pass a certification test, consisting of a number of benchmark essays for which consensus ratings exist, prior to starting the reading of operational responses. Finally, an ongoing monitoring of their ratings is conducted by expert raters who are also available to provide support and feedback as needed.

Comprehensive postadministration analyses are conducted on a monthly basis to (a) evaluate the quality of the ratings, (b) assess the measurement characteristics of prompts delivered, and (c) monitor test takers' performance over time. These evaluations rely on the following types of statistics compiled for the total test taker group, as well as for major subgroups (by region, gender, and race/ethnicity):

- Rating agreement per prompt, indicated by the percentage of same ratings (no difference), the percentage of adjacent ratings (1 point difference), and the percentage of discrepant ratings (more than 1 point difference)
- Rating consistency per prompt, indicated by the correlation between the two ratings, and the weighted kappa coefficient (Fleiss, Cohen, & Everitt, 1969)
- Rating distributions of first and second ratings and their differences per prompt
- Descriptive statistics per prompt; descriptive statistics of the issue, argument, and final Analytical Writing scores
- Correlation between scores on issue and argument, and the Analytical Writing measure reliability[6]
- Correlations between the Analytical Writing scores and the GRE Verbal and Quantitative scores

After a period of transition from the previous version of the Analytical Writing measure, the evaluation results have shown stable measurement outcomes meeting the desired levels of reliability and accuracy for the scoring process and for the reported scores. Based on the data of the GRE revised General Test collected from October 2011 to April 2012, the *GRE Guide to the Use of Scores* (ETS, 2013c, Table 5), which is updated annually, provided a summary of the main measurement outcomes, indicating that:

> The reliability of the Analytical Writing measure is estimated at 0.79 . . . . Overall, the two ratings used in each essay score are in agreement about 66 percent of the time; they differ by one score point about 33 percent of the time; and they differ by two or more score points about one percent of the time. (p. 19)

Table 5 in the *Guide* also reported that the standard error of measurement (SEM) of individual scores (defined as the "Score range in which [a test taker's] true score probably lies," ETS, 2013c, p. 18) was 0.4 and that the SEM of score differences was 0.5. Group level outcomes summarized over the 2012 calendar year are presented later in this chapter.

Monitoring Potential Scoring Drift and Ensuring Fair Measurement

Despite a testing program's best efforts to ensure reliable and accurate measurement, some scoring drift can occur over time. For example, changes in the pool of raters, as some raters retire and new ones join, may result in variations in the scoring trend, or the seasonal variations in the test takers' performance levels may influence raters' ratings in some ways. Statistical analyses have been implemented for the Analytical Writing measure to monitor scoring trends and further to detect any suspicious scoring patterns. Such analyses are particularly helpful in evaluating the need for additional rater training and in evaluating its effectiveness. Yearly monitoring analyses provide much stronger empirical evidence for scoring stability and test fairness than do monthly monitoring analyses.

Trend scoring is a method for monitoring the quality of human scoring over time and for controlling for systematic changes in the score distribution (trend equating; see Kim, Walker, & McHale, 2010; Lane, 2010; Tate, 2000). Specifically, a set of previously scored essays is seeded into operational essays being scored. As with validity responses,[7] raters cannot detect which of the essays they score are trend essays and which ones are operational since the prompts represented in the trend set are also represented in the current session. Thus, any changes in the trend set score distributions can be assumed to be due to the raters.

Trend scoring was conducted for the Analytical Writing measure in spring[8] and summer[9] 2012 to assess any change in raters' scoring behaviors (e.g., scoring stringency or leniency). About 1,300 papers (including both task responses) were randomly selected from the operational administrations in fall[10] 2011. These papers were rescored twice, by the same pool of raters, in 5-month intervals through trend scoring procedures. The score distribution of the 1,308 trend papers rescored in spring 2012 was directly compared to that of the same set of trend papers rescored in summer 2012. Further, the summer 2012 trend scores were equated to the spring 2012 trend scores using single-group equipercentile equating as a method to determine the extent to which scores may have been affected.

Table 2.3.1 shows the percentage and cumulative percentage distributions of the first trend scores and the second trend scores for the 1,308 examinees. Means and standard deviations of the scores are also presented at the bottom of the table. Figure 2.3.1 displays a graphical comparison of the cumulative distributions of the two sets of trend scores. As shown, the two distributions were nearly identical across the full range of the score scale. The equating results also indicated that the scoring standards, as intended, were properly used to score the papers across administrations. Both the cumulative distributions of the scores and equating results provided evidence of raters' scoring consistency over time.

Table 2.3.1

Percentages, Cumulative Percentages, and Descriptive Statistics
of the Analytical Writing's Trend Scores (*N* = 1,308)

	March 2012 Analytical Writing score		August 2012 Analytical Writing score	
Writing score	Percentage	Cumulative percentage	Percentage	Cumulative percentage
6.0	1.07	100.00	1.30	100.00
5.5	2.22	98.93	2.83	98.70
5.0	6.19	96.71	6.04	95.87
4.5	11.93	90.52	12.31	89.83
4.0	20.80	78.59	20.49	77.52
3.5	18.96	57.80	19.11	57.03
3.0	20.95	38.84	20.72	37.92
2.5	9.02	17.89	8.26	17.20
2.0	5.05	8.87	4.97	8.94
1.5	2.52	3.82	2.83	3.98
1.0	1.15	1.30	1.07	1.15
0.5	0.15	0.15	0.08	0.08
0.0	0.00	0.00	0.00	0.00
Mean	3.53	.	3.56	.
SD	0.95	.	0.97	.

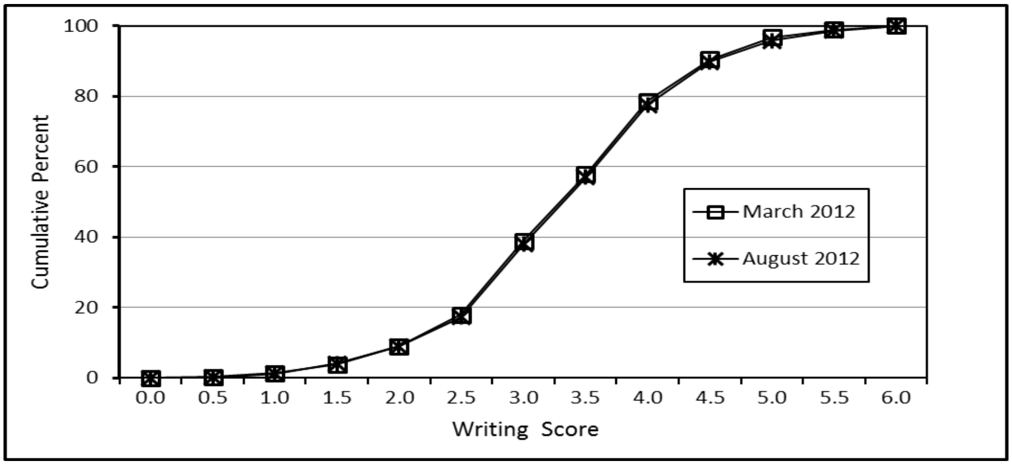


Figure 2.3.1. Cumulative percentages of the trend scores (*N* = 1,308).

Any change caused by a scoring shift could result in unfair scores to the test takers. Because of this, it is expected that trend scoring will continue to be used periodically to monitor raters' scoring behaviors (e.g., leniency or stringency on scoring, score rubric usage) and further to detect any potential scoring shift over time.

Comparability of the Old and Revised Test Scores

As mentioned previously, expectations were that the relatively minor changes made to the Analytical Writing measure would not change the meaning of the scores and that scores obtained before and after the launch of the revised test would still be comparable. Without administering both versions of the test for some period of time and, thus, without the availability of randomly equivalent group data, it is not possible to confirm these expectations directly. As an alternative, the correlations between (a) the Verbal/Quantitative scores and Analytical Writing scores and (b) the performance of total, regional, and race/ethnic groups were compared across two calendar years (2009 and 2012). The choice of these two years avoided the 2010–2011 transition period during which test-taker demographics and motivations for taking or retaking the test may have been different than seen in other testing years.

Table 2.3.2 presents descriptive statistics of the Analytical Writing total scores for the group taking the GRE revised General Test group (revised GRE group) and for the group taking the previous version of the GRE (old GRE group). The revised GRE group included all test takers who took the test from January to December in 2012; the old GRE group included all test takers who took the test from January to December in 2009. Descriptive statistics were also calculated separately for each major subgroup classified by the test takers' geographic regions, gender, and race/ethnicity.

On average, test takers' performance was very similar, without indicating any abrupt changes before and after the launch. The standardized mean differences (SMD) between the mean scores of the reference groups and focal groups, divided by the pooled standard deviation, showed similar patterns and magnitudes before and after the launch across subgroups. Overall, the magnitude of the SMDs increased slightly after the launch, possibly as a result of the enhanced writing directions and some changes in the composition of the test-taker population, particularly the Asian (i.e., test takers from China, Hong Kong, Korea, and Taiwan) and the Indian and Japanese subgroups, whose proportions and SMDs increased the most.

Correlations of the Analytical Writing scores with the Verbal/Quantitative scores may be considered as an indirect (and weak) method of assessing the continuity of the Analytical Writing scores—in particular, considering the significant changes made to the Verbal and Quantitative measures and scales (Robin & Steffen, Chapter 3.3, this volume; Wendler, Chapter 1.2, this volume). Nevertheless, expectations are that similar results should be obtained. Table 2.3.3 presents correlations of the Analytical Writing final score with the GRE General Test scores, separately by major subgroup, for each of the revised and old GRE groups.

Table 2.3.2

Descriptive Statistics of GRE Analytical Writing Measure

	Revised GRE (January–December 2012)					Old GRE (January–December 2009)				
Group	*N*	%	*M*	*SD*	*SMD*	*N*	%	*M*	*SD*	*SMD*
Total	560,543	100.0	3.6	0.8	.	579,156	100.0	3.7	0.9	.
Nationality										
Domestic (ref)	351,142	62.6	3.8	0.7	.	390,132	67.4	3.9	0.9	.
Asian[a]	65,736	11.7	2.9	0.5	1.26	41,752	7.2	3.2	0.5	0.91
India and Japan	54,121	9.7	2.9	0.7	1.20	46,276	8.0	3.0	0.7	1.06
Other international	89,544	16.0	3.3	0.8	0.75	100,996	17.4	3.3	1.0	0.70
Gender										
Male (ref)	234,338	41.8	3.4	0.9	.	237,033	40.9	3.6	0.9	.
Female	288,212	51.4	3.6	0.8	-0.22	322,329	55.7	3.8	0.9	-0.15
Missing	37,993	6.8	3.7	0.8	-0.26	19,794	3.4	3.1	0.8	0.53
Ethnicity										
White (ref)	246,115	70.1	3.9	0.7	.	286,284	73.4	4.0	0.8	.
Asian	21,457	6.1	3.9	0.8	0.07	23,913	6.1	4.0	0.8	0.04
African American	28,397	8.1	3.3	0.8	0.83	34,786	8.9	3.4	0.8	0.75
Hispanic	24,781	7.1	3.6	0.7	0.41	25,055	6.4	3.7	0.9	0.38
Others	30,392	8.7	3.8	0.8	0.23	20,094	5.1	4.0	0.9	0.01

Note. M = Mean; *SD* = Standard deviation; *SMD* indicates standardized mean difference from a particular reference group in each category. Race/ethnicity is only collected for test takers who state they are U.S. citizens.

[a] Asian group includes test takers from China, Hong Kong, Korea, and Taiwan.

Although the magnitude of correlations with the Verbal/Quantitative scores changed after the revision, the patterns of correlations were fairly comparable despite the revision across the major subgroups. As expected, the Analytical Writing scores are correlated more strongly with the Verbal scores than with the Quantitative scores, due to their similarity in measured constructs.

Table 2.3.3

Correlations of GRE Analytical Writing Scores With GRE General Test Scores

	Revised GRE (January–December 2012)		Old GRE (January–December 2009)	
Group	Quantitative/ Analytical Writing	Verbal/ Analytical Writing	Quantitative/ Analytical Writing	Verbal/ Analytical Writing
Total	.13	.67	.15	.59
Domestic	.45	.56	.40	.51
Asian[a]	.34	.57	.28	.52
India and Japan	.56	.71	.40	.61
Other international	.31	.71	.17	.65

[a] Asian group includes test takers from China, Hong Kong, Korea, and Taiwan.

Score reliability was compared before and after the launch as a more direct indicator to assess the continuity of the Analytical Writing scores. The score reliability based on the entire 2012 group of test takers was .82, whereas the score reliability based on all 2009 test takers was .77. This is a noticeable increase that may have resulted for one or both of the following reasons: (a) more specific instructions were provided for each prompt (Bridgeman, Trapani, & Bivens-Tatum, 2011; Briel & Michel, Chapter 1.1, this volume) or (b) raters received extensive retraining, ongoing monitoring, and feedback for several months after the launch. As shown in Table 5 of the *GRE Guide* (ETS, 2013c), the reliability of the Analytical Writing scores, based on the performance of all test takers from August 2011 to April 2012, is 0.79. This estimate was obtained after the launch but just before the operational retraining for the revised test and was only slightly higher than the reliability of the old Analytical Writing measure.

Summary

In this chapter, we briefly described the processes and analyses the GRE program uses to assemble and evaluate the Analytical Writing measure, to evaluate the stability of the rating process, and to monitor the Analytical Writing testing outcomes. We then summarized the key results obtained a little more than 1 year after the launch of the redesigned tests and then showed the extent to which the major program goals, in terms of rating and measurement accuracy and in terms of score equivalence across time and across major test taker groups, have been achieved. Finally, we provided empirical comparisons across the previous and revised version of the Analytical Writing measure that showed the continuity of the assessment and the validity of the score comparisons that have been and most likely will be made.

Conclusion

As new tests are delivered and new prompts are added to the disclosed operational Analytical Writing measure prompt pool, the test assembly process and the testing outcomes will continue to be closely monitored to ensure the continued quality of the Analytical Writing measure.

References

Bridgeman, B., Trapani, C., & Bivens-Tatum, J. (2011). Comparability of essay question variants. *Assessing Writing, 16*(4), 237–255.

Educational Testing Service. (2013a). *Introduction to the Analytical Writing Measure*. Princeton, NJ: Author.

Educational Testing Service. (2013b). *GRE published topic pools for the Analytical Writing Measure*. Princeton, NJ: Author.

Educational Testing Service. (2013c). *GRE guide to the use of scores*. Princeton, NJ: Author.

Fleiss, J. L., Cohen, J., & Everitt, B. S. (1969). Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin, 72,* 323–327.

Kim, S., Walker, M. E., & McHale, F. (2010). Investigating the effectiveness of equating designs for constructed response tests in large scale assessment. *Journal of Educational Measurement, 47,* 186–201.

Lane, S. (2010). *Performance assessment: The state of the art* (SCOPE Student Performance Assessment Series). Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

Schaeffer, G. A., Briel, J. B., & Fowles, M. E. (2001). *Psychometric evaluation of the New GRE Writing Assessment* (GRE Board Professional Report No. 96-11P). Princeton, NJ: Educational Testing Service.

Tate, R. L. (2000). Performance of a proposed method for the linking of mixed format tests with constructed-response and multiple-choice items. *Journal of Educational Measurement*, *36*, 336–346.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31*(1), 2–13.

Zhao, J. C., Zhu, R., Guo, F., Zeller, K. E., & Bannerjee, B. (2006, April). *Timing configuration and psychometric properties of the redesigned Analytical Writing measure of the GRE*

General Test. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Notes

[1] Writing test developers developed an approach to prompt writing that can generate several different instructions by specifying different writing tasks in response to the same stimulus. Thus, for example, one test taker may be shown a prompt and asked to:

> Write a response in which you discuss what questions would need to be addressed to decide whether the recommendation is likely to have the predicted result. Be sure to explain how the answers to the questions would help to evaluate the recommendation.

Another test taker would see the same prompt, but be asked to:

> Write a response in which you examine the unstated assumptions of the argument above. Be sure to explain how the argument depends on those assumptions and what the implications are if the assumptions prove unwarranted.

[2] That is, making it impossible for test takers to predict which prompt among the disclosed pool they will be assigned.

[3] Prompt difficulty is estimated from data collected from test takers who are U.S. citizens and test in a test center in the United States or a U.S. territory.

[4] Test equating is a statistical method that makes scores from different test forms interchangeable by adjusting for differences in difficulty among forms.

[5] In rare cases, when the two ratings are *discrepant* (i.e., far apart), an additional rater is also called on. If the third rating is midway between the first two human ratings, the task score is the average of the first two ratings; otherwise, it is the average of the third rating and whichever of the first two ratings is closer to the third rating.

[6] The Spearman-Brown formula was applied to the correlation between scores on the two writing tasks to estimate the Analytical Writing reliability (i.e., split-half reliability). The result will be an underestimate of the alternate-forms reliability.

[7] Validity responses are exemplars having established scores that are mixed in with the operational responses during operational scoring. They are used to assist scoring leaders in monitoring the accuracy of scoring. Raters are not aware of which responses are validity responses.

[8] *Spring* refers to data collected sometime during January through July.

[9] *Summer* refers to data collected sometime during May through August.

[10] *Fall* refers to data collected sometime during August through December.

2.4 Using Automated Scoring as a Trend Score: The Implications of Score Separation Over Time [1]

Catherine Trapani, Brent Bridgeman, and F. Jay Breyer

By November 2010, Educational Testing Service's automated scoring software for essays*,* the
e-rater® scoring engine, was implemented as a confirmatory[2] score for the Analytical Writing measure on the *GRE*® General Test and as a contributory[3] score for the writing measure on the *TOEFL*® test. The governing boards of both the GRE General Test and the TOEFL requested that initial implementation of e-rater be monitored closely to see if e-rater and the writing scores were behaving as anticipated during the implementation research. Initial analyses found a larger than expected difference between the first human score and the e-rater score, the automated scoring engine used with both GRE and TOEFL, for Analytical Writing on both the analyze an issue (issue) and analyze an argument (argument) tasks and for the TOEFL independent prompt. The full study reports on the results of the follow-up analyses of
e-rater performance for both Analytical Writing and TOEFL, but only the Analytical Writing results are summarized here.

The research questions were as follows:

1. Are agreements with human scores consistent with implementation research in terms of percent agreement, correlation, and average differences between human scores and e-rater scores?
2. Are associations with external variables consistent with prior results?
3. Are changes in agreements between human and automated scores a plausible method for monitoring changes in human raters over time?

Three assumptions must hold in order to expect comparable results from one sample to the next:

- The general ability levels of examinees must be constant.
- The nature of writing submissions must be constant.
- Human rating standards must be constant.

Method

The sample consisted of all electronic essays written by examinees on the Analytical Writing measure that tested from October 2008 through June 2009 (approximately 356,000 essays). Descriptive statistics; agreement rates between human and automated ratings; and correlations among the various ratings, question scores, and section scores, as well as other operational sections of Analytical Writing, were obtained.

Results

Analyses of performance over time confirmed that, as with many educational assessments, the ability of GRE General Test candidates varies with seasons; higher ability examinees tend to take the test in the fall,[4] and lower ability examinees tend to test in the spring.[5]

The study also looked at the eight feature scores used to generate automated essay scores: (a) development, (b) organization, (c) grammar, (d) usage, (e) mechanics, (f) style, (g) word length, and (h) lexical complexity. An additional variable, number of words per essay, was also studied. This variable is not used explicitly in the scoring model, but it is very highly correlated with the sum of the development and organization features. Averages were gathered for each of these features within each cohort group using commonly administered prompts. To examine the question of whether the nature of the written submissions has changed, these averages were compared between the operational 2008–2009 data and the model-build evaluation set from 2006–2007. The standardized average difference (i.e., the difference in the averages expressed in standard deviation units) was calculated, and a threshold of 0.10 (standard deviations) was used to flag change. Based on the observed differences, the number of words, mechanics, and average word length have all changed over time for issue essay responses; for argument responses, usage and average word length have changed over time. The reason for these changes is unclear.

We compared agreement between human and e-rater scores over time. Table 2.4.1 displays the results of this comparison. The percent agreement between human and e-rater scores, and the correlation of human and e-rater scores, slipped slightly over time but was still well within acceptable standards. The standardized average difference between human and automated scores was essentially 0 in the implementation study, but was 0.20 in the operational results. Fortunately, the confirmatory approach ensures that the only consequence of lower agreement between a human and e-rater is the need for additional second human ratings. As another positive aspect, the observed correlations between e-rater scores and scores on other operational GRE sections were on par with the correlations between human raters and other operational GRE scores. Given the overall successful performance of e-rater relative to humans, some differences in the nature of writing submissions, and the confirmation of the assumption of little to no change in examinee ability, more research was done to see if human scoring had changed over time.

Human Trend Scoring Study

A possible explanation for the growing discrepancy between human and e-rater was that human raters were becoming more severe over time. This possibility was explored by having current human raters assign scores to the same essays that were scored years earlier (with no knowledge of the score previously assigned). In 2009, three issue prompts and three argument prompts of 500 essays each from the 2006–2007 e-rater implementation cohort were rated by certified human raters. Complete evaluation of human–human and human–e-rater results were obtained for both scoring periods for comparison. The average difference in human scores over this two year period was approximately 0.23 lower for identical papers. It was concluded that human scoring had changed in significant ways over the intervening years (Rick Morgan, personal communication). Because the size of this difference (0.23) mirrored the discrepancy between humans and e-rater in the latter years (0.20), it appears that tracking this discrepancy over time is a useful first step in monitoring the consistency of human scores over time.

Table 2.4.1

Human and e-rater Agreement Statistics Over Time for GRE Analytical Writing Measure

Agreement statistic	Implementation study	Operational results	Difference
Issue			
Exact agreement	59%	57%	-2%
Exact + adjacent agreement	98%	97%	-1%
Correlation	0.79	0.75	-0.04
Standardized difference	-0.02	0.22	0.20 [a]
Argument			
Exact agreement	53%	52%	-1%
Exact + adjacent agreement	96%	95%	-1%
Correlation	0.78	0.75	-0.03
Standardized difference	-0.01	0.21	0.20 [a]

[a] Exceeds the threshold.

Conclusion

In this study, exact and adjacent rates of agreement between human raters and e-rater, as well as the correlations among various scaled operational scores, are on par with values expected as a result of prior research. However, the growing difference in average performance by humans compared to e-rater led to the suspicion that human standards may have been changing over time. This suspicion was confirmed in the human trend score study. Monitoring of e-rater and human discrepancies over time, combined with ongoing rater training efforts, helps to assure consistency in the meaning of scores over time.

Notes

[1] Based on *Using Automated Scoring as a Trend Score: The Implications of Score Separation Over Time*, by C. Trapani, B. Bridgeman, and F. J. Breyer, April 2011, paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

[2] In the confirmatory model, e-rater is used as a check on the human score. If the two ratings are in agreement, the human score is used as the question score. If not in agreement, the e-rater score is discarded and additional human rating(s) are obtained.

[3] In the contributory model, the e-rater rating is used in place of a second human score and is subject to typical adjudication policies. Typically, the question score will be the average of a human and an automated score.

[4] *Fall* in this context refers to a time period ranging from August through December.

[5] *Spring* in this context refers to a time period ranging from March through July.

Section 3: Test Design and Delivery

The revision of the *GRE*® General Test allowed the exploration of various test designs for the Verbal Reasoning and Quantitative Reasoning measures. The test design used with the Verbal and Quantitative measures sections of the GRE had been adaptive at the question level; that is, test takers were routed to their next question based on their performance on the previous question. However, this design did not allow for some of the goals underlying the revision of the test to be reached. As a result, different test designs were considered and evaluated. Following these evaluations, it was determined that a multistage adaptive test (MST) model would best fit the needs of the revised test. The MST design is adaptive at the stage (section), not question, level, and the determination of the next set of questions an examinee receives is based on performance on an entire preceding stage. Chapters in this section describe the efforts related to the decision to use the MST model with the revised test.

- Chapter 3.1 presents a summary of the limitations and advantages of various paper-based (PBT) and computer-based (CBT) testing models. This summary was the result of an extensive review of research literature and was written as a way to assist practitioners who were contemplating various test designs. Three basic advantages for moving to a CBT model are identified: (a) the ability to measure constructs or skills that cannot be fully measured by PBT, (b) improving the precision and efficiency of the measurement process, and (c) increased convenience for the test taker and the test administrator. Descriptions of the various administrative models available for a CBT (fixed form, random form, multistage form, question-adaptive, and computerized classification) and a comparison along the important test properties of efficiency, security, requirements for question development, complexity, and cost are given. The information in this chapter provided guidance as to the test design that would best suit the needs of the GRE revised General Test.
- Chapter 3.2 describes a study that examined the comparability of paper-based and computer-based question presentation formats of the GRE revised General Test. The goal of this study was to identify challenges that test takers might encounter when test content developed for a particular test design (such as computer-based) was delivered using a different design (such as paper-based). The data were gathered using think-aloud methodology (also called a cognitive lab), which requires test takers to vocalize their thoughts while responding to test questions. All participants were nonnative English-speaking adults with various experiences with taking tests on the computer. Two question types were examined in the study: verbal text completions and quantitative numeric entry. In addition to providing

feedback on particular questions, a short survey designed to identify any difficulties participants encountered on the test was given. Results indicated that overall performance was not noticeably affected by the format in which questions were presented. However, some question formats were problematic, and some further test modifications were made. Overall, this study provided support of the comparability of the paper-based and computer-based formats for the text completion and numeric entry question types.

- Chapter 3.3 outlines the major goals for the test revision and describes the efforts undertaken to explore alternative designs and scoring models for the GRE revised General Test. These goals included the following: (a) support the rotation of test content to make it unpredictable to the test taker, (b) enhance the level of measurement for the increasingly diverse group of test takers, (c) provide support for the revision of the score scales for Verbal Reasoning and Quantitative Reasoning, (d) maintain a testing time of 4 hours or less, and (e) offer a more test taker–friendly experience. The chapter reviews the test specifications at a high level and discusses the alternate test designs and scoring models that were considered for the GRE revised General Test. It also provides information on the MST design that was chosen and the results of the evaluation of the functioning of the design a year after its implementation. This evaluation shows that the quality of the test has been maintained and, in some cases, enhanced. While this chapter discusses several complex, technical methods, it is still written to be informative to those with a nontechnical background.
- Chapter 3.4 reports on a study related to further understanding the impact that the MST design could have on the scoring and equating of the GRE revised General Test. In particular, the impact of context effects was examined. *Context* refers to the position in which a question appears in a test, as well as the content, format, and specific features of the other questions that surround it. Changes in question position or the surrounding questions, as seen in an MST, could inadvertently modify the characteristics (such as difficulty level) of the test question. Results indicated that questions appeared to be easier when they appeared earlier in the test and more difficult when they appeared later in the test, and Quantitative Reasoning questions seemed to be more subject to change in difficulty level than Verbal Reasoning questions. Trying out (pretesting) questions in random locations throughout the test seems to reduce position effects. Results also indicated that the MST tests were more speeded than linear tests, especially for Quantitative Reasoning questions. Results of this study provided guidance on the final design of the test. For example, questions were pretested in random locations throughout the test, and additional field trials were held to determine the appropriate test

configuration for the Quantitative Reasoning measure (see Golub-Smith & Wendler, Chapter 2.1, this volume). As with Chapter 3.3, this chapter is fairly technical in nature but written to be informative to those with a nontechnical background.

3.1 Practical Considerations in Computer-Based Testing [1]

Tim Davey

As part of the redesign of the *GRE*® General Test, ETS conducted a thorough review of the research literature on the limitations and advantages of paper-based and computer-based testing (CBT). This chapter presents a summary of the findings of this review in order to assist practitioners who are considering whether to move to CBT and deciding which computer-based test might best suit their needs.

The Advantages of Computer-Based Testing

Three basic advantages for moving to CBT were identified in the literature review. First, computer-based tests can measure constructs or skills that cannot be fully or appropriately captured by paper-based tests (Bennett, 2002; Parshall, Harmes, Davey, & Pashley, 2010). Standardized tests often are criticized as artificial and abstract, measuring performance in ways divorced from real-world behaviors. At least some of this criticism is due to the constraints that paper-based administration imposes upon test developers. Paper testing is restricted to displaying static text and graphics, offers no real means of interacting with the examinee, and limits the ways in which examinees can respond. Computers can free test developers from these restrictions and interact dynamically with examinees, accept responses through a variety of modes, and even score those responses automatically. Therefore, CBT can be a richer, more realistic experience that allows more direct measurement of the traits in question.

However, the test developer should adopt innovation only when necessary to best measure a construct and resist its temptations when conventional question types will suffice. As is the case with all matters related to test development, the decision process must start with a comprehensive analysis of the construct being measured and how evidence of a student's standing on it is best collected. If this analysis uncovers gaps between what a test should be measuring and what it could measure (if released from the constraints imposed by paper-and-pencil tests), question types that appear to best address these gaps can then be identified or designed and successive rounds of pilot testing can be conducted to inform design revisions (Harmes & Parshall, 2010).

Second, CBT can improve the precision and efficiency of the measurement process (Parshall, Spray, Kalohn, & Davey, 2001; van der Linden & Glas, 2000; Wainer, 1990). This is the result of the computer's capacity to be *adaptive*, that is, to interact with and tailor itself to the student being tested. As an adaptive test proceeds, answers to earlier questions determine which questions are asked later. Therefore, the test progressively changes as the student's performance level is gradually revealed. As a result, adaptive CBT can be more *efficient* than conventional tests that present the same questions to every student. That is, an adaptive test can match the precision of a conventional test while containing fewer questions or match the length of a conventional test but return more precise measurement, particularly of students at either extreme of the performance continuum.

A third advantage of CBT is increased convenience. A computer-based test can eliminate the need for someone to distribute and collect testing materials and keep track of time. The computer can collect identification data, orient the student to the testing process, and administer and time the test. Computer-based tests can also provide an immediate score report at the conclusion of the test. At the classroom level, this might enable a teacher to change quickly the instructional approach taken with a particular concept. At the school or district level, immediate information might allow tactical shifts in the instruction process. Scores generated by CBT can be entered automatically into classroom, school, district, or state level databases to allow various reports to be easily produced that summarize and track the performance of individual students and defined groups.

Considerations in Designing a Computer-Based Test

The review of the literature identified five test properties that need to be considered during the design process. Because different CBT models posses these properties to varying degrees, the designer’s task is to identify the model that best matches the specific test purposes or objectives.

Measurement Efficiency

Test reliability and test length are strongly related; reliable tests tend to be longer, and shorter tests tend to be less reliable. In choosing the most appropriate design for a test, developers must balance reliability and length. The CBT models differ considerably in the ratio of reliability to length, that is, in their *measurement efficiency*. A more efficient test is one that offers more measurement precision per question or unit time.

Test Security

Although computer-based tests are subject to many of the same security concerns that afflict conventional tests, they appear to be less vulnerable to students copying from one another. Unlike answer sheets that sit open on desktops and are accessible to prying eyes, answers usually appear on the screen of a computer-based test only fleetingly before being replaced by the next question. Designs for CBT that vary the questions administered or the order of their administration across students are even better in this regard. Questions stored in encrypted files on a computer are also much better protected prior to administration than a box of booklets placed in a desk drawer or closet.

Question Development Requirements

Designs for CBT differ considerably in the number of questions that need to be developed to properly support administration. Designs for CBT that present each student with only a portion of the available questions (e.g., adaptive tests that select questions from an available question pool) may require that substantially more questions be developed. The stakes attached to testing also play a role in determining development requirements. For security reasons, consequential, high-stakes tests, such as admissions or exit tests, require that questions be replaced more frequently than do tests used for formative purposes. Finally, the desire to change the nature of measurement through the use of innovative CBT questions may be much more difficult and expensive to develop than text-based multiple-choice questions.

Design Complexity

Computer-based tests differ widely in terms of the complexity of their administrative model, scoring methodologies, and the statistical mechanisms required to ensure comparability across examinees and across time. Generally, adaptive CBT models, where each student may be administered a unique combination of questions, are more complex to administer and score than models where all test takers will be administered the same questions. Similarly, replacing or changing a question pool requires more sophisticated statistical mechanisms than replacing one test form with another. All things equal, simpler test designs are preferable to more complex designs. Simpler designs are more robust and more resistant to unanticipated problems and less expensive to develop and maintain. Complex designs can be more efficient and more secure but are likely to impose higher question development requirements and maintenance costs.

Computer-Based Testing Administrative Models

The test administration model controls the questions with which a student is presented and the order in which they are presented. Because it strongly impacts all of the test properties discussed above, choosing which model is appropriate for a given test situation is crucial.

Five distinct test administration models are described below. The last three are adaptive because the testing process can change in response to each student's performance. The description of each model is summarized in Table 3.1.1, which highlights strengths and weaknesses relative to the five test properties.

Fixed Form

The simplest type of CBT essentially replicates the administration model as well as the construction and scoring methods of conventional paper tests. Each student is presented with the same set of questions, either in the same or in a randomly scrambled order.

Random Form

Each test taker is presented with a subset of questions drawn from a larger pool of questions. Rules for drawing from the question pool are imposed to ensure that the different forms drawn for different students each measure the same content and are parallel in difficulty and reliability. Scores are usually computed using item response theory (IRT) methods.

Multistage Test Form

The simplest of the adaptive administration models, the multistage test (MST) form, begins by presenting each student with a first-stage, or *routing test*, whose questions broadly sample the content domain, focusing on questions of middle difficulty. After the student completes the routing test, a score is calculated. Students who performed well are assigned a second-stage test composed mainly of more difficult questions; students who struggled are administered an easier second-stage test. Upon completion of the second-stage test, the test ends and a final score is produced that aggregates performance across both the routing and second stages of the test. More elaborate branching designs also are possible, with additional decisions and a third or fourth stage following the second.

Question-Adaptive Model

The question-adaptive model computes a score following each question and, based on the student's performance level, makes a decision as to what question to present next. Questions are selected from a pool based on the performance level a student has demonstrated on questions administered earlier in the test. Question selection is usually designed to meet three goals: (a) maximize test efficiency by measuring students as precisely as possible with as few questions as possible; (b) construct for each student a test that is properly balanced in terms of question substance or content; and (c) protect individual questions from either overuse, which can threaten test security as they become known to students, or underuse, which can waste resources (Davey & Nering, 2002; Mills & Steffen, 2000).

Computerized Classification Test

Rather than assigning each student a precise numeric score, computerized classification test (CCT) attempts to classify students into groups, such as pass/fail or basic/proficient. Students assigned to the same classification group are considered as having performed equivalently on the test. Questions are selected from a question pool that best targets the classification threshold or cut-point most crucial to a given student's classification. For example, consider a scenario in which students are to be classified as basic, proficient, or advanced. For a student who is performing well on early questions, the critical threshold would be that which divides proficient from advanced. This threshold would then be targeted by questions chosen in the latter part of the test, until a classification can be reliably made. Conversely, the last questions for a struggling student would focus on the threshold between basic and proficient.

Table 3.1.1

Rating Computer-Based Testing Models on Each of the Five Important Test Properties

Testing model	Efficiency	Security	Question development requirements	Complexity	Cost
Fixed form	Low	Low	Low	Low	Low
Random form	Low	Medium	Medium	Medium	Medium
MST	High	Medium	Medium	Low	Medium
Question-adaptive	High	Medium/high	High	High	High
CCT	Very high	Medium/high	High	High	High

Note. MST = multistage test; CCT = computerized classification test.

Conclusion

This chapter attempts to convey that no single administration model is ideal for all tests and under all circumstances. Instead, the most appropriate model depends upon the nature and unique characteristics of a test. The question types needed to test the construct, the stakes attached to the test scores, characteristics of the test taking population, and the subjective values of both the test's owner and score users must all be considered when choosing the administration model to be used.

References

Bennett, R. E. (2002). Inexorable and inevitable: The continuing story of technology and assessment. *Journal of Technology, Learning, and Assessment*, *1*(1). Retrieved from http://ejournals.bc.edu/ojs/index.php/jtla/article/view/1667/1513

Davey, T., & Nering, M. (2002). Controlling question exposure and maintaining question security. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 165–191). Mahwah, NJ: Erlbaum.

Harmes, J. C., & Parshall, C. G. (2010, April). *A model for planning, designing, and developing innovative questions.* Paper presented at the annual meeting of the National Council on Measurement in Education, Denver, CO.

Mills, C. N., & Steffen, M. (2000). The GRE computer adaptive test: Operational issues. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 75–99). Norwell, MA: Kluwer.

Parshall, C. G., Harmes, J. C., Davey, T., & Pashley, P. J. (2010). Innovative question types for computerized testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 215–230). New York, NY: Springer.

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. C. (2001). *Practical considerations in computer-based testing.* New York, NY: Springer.

van der Linden, W. J., & Glas, C. G. (2000). *Computerized adaptive testing: Theory and practice.* Norwell, MA: Kluwer.

Wainer, H. (Ed.). (1990). *Computerized adaptive testing: A primer.* Hillsdale, NJ: Erlbaum.

Notes

[1] Based on *Practical Considerations in Computer-Based Testing,* by T. Davey, 2011, Princeton, NJ: Educational Testing Service, retrieved from http://www.ets.org/Media/Research/pdf/CBT-2011.pdf

3.2 Examining the Comparability of Paper-Based and Computer-Based Administrations of Novel Question Types: Verbal Text Completion and Quantitative Numeric Entry Questions[1]

Elizabeth Stone, Teresa King, and Cara Cahalan Laitusis

This study examined the comparability of paper- and computer-based item (that is, question) presentation formats of the *GRE*® revised General Test to identify challenges that test takers encounter when test content developed for computer administration is delivered in a paper-based format. Although the GRE revised General Test was designed to be delivered on a computer, it will be delivered in a paper-based format to some test takers in countries outside of the United States, depending on their location, and some test takers who test in a test center in the United States or a U.S. territory, primarily test takers who take the test with accommodations. Since scores from paper- and computer-based delivery will be reported on the same scale, it is important to identify and mitigate, as much as possible, any comparability issues between the two testing formats.

Research completed to date has primarily examined the comparability of paper- and computer-administered single-selection multiple-choice test questions (see Gallagher, Bridgeman, & Cahalan, 2002; Schaeffer et al., 1998). Gallagher et al. (2002) found that gender and ethnic differences were similar across paper-and-pencil and computer formats. Gallagher et al. examined average performance levels across paper-based and computer formats for the GRE General Test but with a very small increase in performance on the computer-based tests when compared to the paper-and-pencil tests for some racial and ethnic groups. For instance, male test takers were found to achieve slightly higher scores than female test takers on the computer-based tests for some of the assessments, including the GRE General Test. However, documentation is lacking for many of the new question types proposed for the GRE revised General Test.

To examine whether differences in question type or test format may result in errors unrelated to the construct being measured, the study used *think-aloud methodology* (also known as cognitive interviews or cognitive labs), which requires examinees to vocalize their thought processes while answering test questions. The study participants' verbalizations provided researchers with clues as to how the participant was approaching and processing the academic task and could identify obstacles that participants faced when completing the test questions that were not related to what the test was intended to measure.

Think-aloud methods have been used to assist in the development of psychometric instruments by alerting researchers to how questions may be interpreted by test takers in an operational setting (Campbell, 1999; Desimone & Le Floch, 2004; Paulsen & Levine, 1999). The results of think-aloud studies provide test developers with important feedback about whether students followed the logic the developers expected (Paulsen & Levine, 1999). Other studies have used this methodology to gain understanding after the fact about why particular questions behaved as they did during field testing (Johnstone, Altman, & Thurlow, 2006). All of this information can be used to revise test questions to strengthen the reliability of test scores and the validity of inferences based on those scores (Desimone & Le Floch, 2004; Paulsen & Levine, 1999). The examination of response processes is also a key source of validity evidence suggested by the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). Although the responses to the specific questions were not the focus of the study, the think-aloud approach was used to try to identify possible obstacles that participants faced, with a focus on difficulties due to question type or test format. The think-aloud approach was supplemented with a posttest survey that asked explicitly about the question types and test formats.

Method

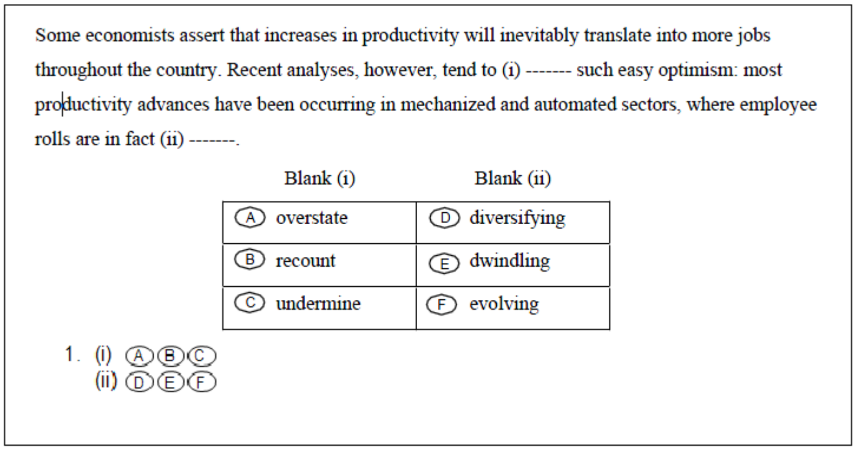
Study Sample

The sample consisted of 25 nonnative English-speaking adults. Fourteen did not have any experience taking a test on the computer. Sixteen were from regions of the world in which the paper format will more likely be administered.

Test Questions

As a first step in this study, ETS researchers, test developers, and psychometricians identified the new question types that appear to be dissimilar (in terms of presentation or response) between the paper- and computer-based formats. From the question types identified as potentially problematic, the text completion and numeric entry question types were selected for this study because they appear to present the most significant challenges for paper-based test takers. Figures 3.2.1 and 3.2.2 show examples of these question types.

For each question type, 15 test questions were grouped into three blocks of five questions each. The question blocks were administered in three different formats: a computer-based format, a paper-based format in which answers were written on the test booklet, and a paper-based format in which answers were recorded on a paper answer sheet. Participants were randomly assigned to one of four experimental groups for each content area. Each group was administered a form consisting of the same three question blocks administered under the three different delivery formats (see Table 3.2.1).

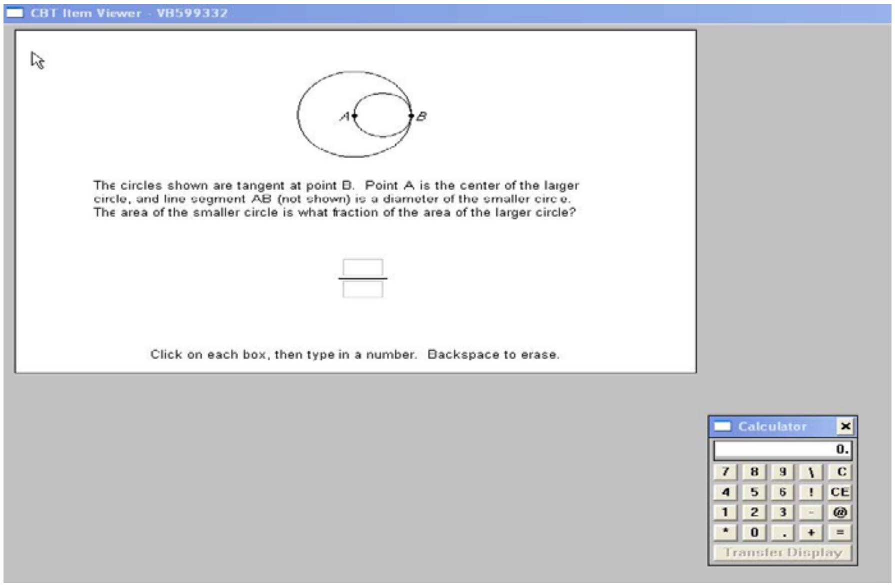


Some economists assert that increases in productivity will inevitably translate into more jobs throughout the country. Recent analyses, however, tend to (i) ------- such easy optimism: most productivity advances have been occurring in mechanized and automated sectors, where employee rolls are in fact (ii) -------.

Blank (i)	Blank (ii)
Ⓐ overstate	Ⓓ diversifying
Ⓑ recount	Ⓔ dwindling
Ⓒ undermine	Ⓕ evolving

1. (i) ⒶⒷⒸ
(ii) ⒹⒺⒻ

Figure 3.2.1. Two-blank verbal text completion question. Top (sentences) of the figure shown in paper-based format; bottom of the figure in computer-based format.



The circles shown are tangent at point B. Point A is the center of the larger circle, and line segment AB (not shown) is a diameter of the smaller circ e. The area of the smaller circle is what fraction of the area of the larger circle?

Click on each box, then type in a number. Backspace to erase.

Figure 3.2.2. Numeric entry question shown in computer-based format. This figure also displays the on-screen calculator.

Table 3.2.1

Design for Both Studies (Verbal and Quantitative, With Number of Participants Assigned to Each Intact Three-Condition Form)

	Condition 1		Condition 2		Condition 3	
	Question Block A		Question Block B		Question Block C	
Group	Format	Question	Format	Question	Format	Question
1 (6,6)	Paper AS	1-5	Paper TB	6-10	Computer	11-15
2 (6,6)	Paper TB	1-5	Paper AS	6-10	Computer	11-15
3 (6,7)	Computer	1-5	Paper AS	6-10	Paper TB	11-15
4 (6,5)	Computer	1-5	Paper TB	6-10	Paper AS	11-15

Note. AS = answers made on answer sheet; TB = answers made directly on test booklet.

Cognitive Interviews

Each think-aloud session lasted 1 to 1.5 hours and was administered one on one. The protocol used for the cognitive labs included a short introduction to the new question type (e.g., text completion) and provided guidance about how to think aloud. The researcher administering the protocol kept notes on the think-aloud activities. To capture an initial measure of each participant's English proficiency, participants were instructed to read the question aloud and to begin immediately thinking aloud through the process of arriving at the answer. This allowed the interviewer to evaluate at a basic level each participant's ability to understand each question and to monitor any mistakes involving interpretation of words.

Posttest Survey

Upon completing the think-aloud portion, a short survey was read aloud to the participant. The survey questions were designed to identify any difficulties the participant encountered and to elicit participants' suggestions for revising the format of the questions. Participants were asked explicitly to compare formats in the posttest survey.

Results

Overall performance on the test questions was not noticeably affected by the format in which the questions were presented. The cognitive lab results appeared to indicate that, aside from a few individual issues, none of the participants had consistent trouble demonstrating what they knew when presented with the test questions in the computer-based format or either of the paper-based formats. The majority of observations and responses provided evidence to support comparability of the paper and computer formats for both the quantitative numeric entry and the verbal text completion questions.

Nearly half of all participants indicated that they did not have a response format preference for either the verbal or the quantitative test. Among the participants who had a format preference, the two formats that were most often chosen were the paper test that allowed test takers to respond directly in the test booklet (32% verbal and 20% quantitative) and the computer test (28% verbal and 20% quantitative).

The majority of participants felt comfortable taking the computer-based test. On the posttest survey, 60% of respondents reported being extremely comfortable and 28% reported being somewhat comfortable with the computer test. No participants indicated that they felt extremely uncomfortable. (Possible responses included *Extremely comfortable*, *Somewhat comfortable*, *Neither comfortable nor uncomfortable*, *Somewhat uncomfortable*, and *Extremely uncomfortable*.)

Participants appeared to need more prior direction on both the verbal text completion and the quantitative numeric entry questions, both of which are innovative and novel question types. Because the population taking the operational test will include test takers whose first language is not English, the researchers concluded that further clarification of the directions describing how to select answers was warranted.

Participants reported that the layout of options in the test versus entry blanks on the answer sheet caused some confusion. A final issue involved possible confusion that may occur when transferring the vertical options to the answer sheet, which displays the answer spaces in a horizontal format (see Figure 3.2.1 for an example of different orientation of the options in the test book versus the answer sheet). This apparent confusion was observed by the researcher administering the protocol. Several participants appeared to be carefully entering their responses and going back to make sure that they had entered their responses correctly.

Conclusion

While the cognitive lab results focused on in this paper were those that were considered of possible interest in evaluating comparability of format, overall performance was not noticeably affected by the format in which the questions were presented. Nevertheless, in order to ensure that the GRE revised General Test reflected the needs of all test takers, some of the instances in which question format was problematic for some test takers were revised based on the results of this study.

For example, for the paper-based test, the answer sheet is now embedded in the test booklet (as opposed to a separate sheet of paper). This alteration also eliminated formatting difficulties with the multiple-blank verbal questions, as test takers no longer have to transcribe their answers onto a separate answer sheet that may not have response options listed in the same orientation. Overall, while there were several features that may be of interest to investigate further, the majority of observations and responses provided evidence to support comparability of the paper and computer formats for the quantitative numeric entry and verbal text completion question types.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Campbell, J. R. (1999). *Cognitive processes elicited by multiple-choice and constructed-response questions on an assessment of reading comprehension* (Unpublished doctoral dissertation). Temple University, Philadelphia, PA.

Desimone, L. M., & Le Floch, K. C. (2004). Are we asking the right questions? Using cognitive interviews to improve surveys in educational research. *Educational Evaluation and Policy Analysis*, *26*, 1–22.

Gallagher, A., Bridgeman, B., & Cahalan, C. (2002). The effect of computer-based tests on racial/ethnic and gender groups. *Journal of Educational Measurement*, *39*, 133–147.

Johnstone, C. J., Altman, J., & Thurlow, M. (2006). *A state guide to the development of universally designed assessments.* Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Paulsen, C. A., & Levine, R. (1999, April). *The applicability of the cognitive laboratory method to the development of achievement test items.* Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Schaeffer, G. A., Bridgeman, B., Golub-Smith, M. L., Lewis, C., Potenza, M. T., & Steffen, M. (1998). *Comparability of paper-and-pencil and computer adaptive test scores on the GRE General Test* (Research Report No. RR-98-38). Princeton, NJ: Educational Testing Service.

Notes

[1] Based on *Examining the Comparability of Paper-Based and Computer-Based Administrations of Novel Item Types: Verbal Text Completion and Quantitative Numeric Entry Items* (Research Memorandum No. RM-11-03), by E. Stone, T. King, and C. Cahalan Laitusis, 2011, Princeton, NJ: Educational Testing Service.

3.3 Test Design for the *GRE*® revised General Test

Frédéric Robin and Manfred Steffen

The multistage adaptive test (MST) design for the *GRE*® revised General Test was developed to improve upon its computer adaptive test (CAT) predecessor, which had been in operation since the early 1990s. When it was launched in 1993, the GRE CAT brought great flexibility to test takers who then could take a shorter test on demand. However, since the early 2000s, the increased availability and power of Internet communication has created new test security challenges. At the same time, as previously described in this compendium, there was a growing desire to revise the content of the Verbal Reasoning and Quantitative Reasoning measures[1] and to realign their associated reporting scales (Briel & Michel, Chapter 1.1, this volume; Golub-Smith & Moses, Chapter 2.2, this volume; Golub-Smith & Wendler, Chapter 2.1, this volume; Wendler, Chapter 1.2, this volume). Hence, it was concluded that the development of a new Internet-based test delivery infrastructure and a new test design would be necessary to address potential security challenges and to effectively support the extensive revisions of the test content and scales.

The major goals for the new test design included the following:

- Support the rotation of content to make it unpredictable to the test taker, which is necessary to ensure test security in today's world where the Internet allows information to be quickly communicated to a large audience
- Enhance the level of measurement for an increasingly diverse testing population
- Support the revision of the scales
- Maintain the overall testing time to 4 hours or less
- Offer a more flexible test-taking experience by allowing candidates to move forward and backward, skip and revisit questions, and change their answers within each test section

These goals, as well as significant changes to a large proportion of the questions (Educational Testing Service [ETS], 2013b), led to revisions of the test specifications that define the content representation, measurement accuracy, and test security required for each test form.

In this chapter, we briefly outline the test specifications used to guide the development and implementation of the new test design for the GRE revised General Test. We then review the rationales and the studies that led us to the choices of MST design and number-correct scoring for the GRE General Test. Next, we describe in some details the development process and the characteristics of the Verbal Reasoning and Quantitative Reasoning MSTs obtained. Finally, we present a summary of the major measurement outcomes and show the extent to which each one of the many versions of the test produced and delivered over a 1-year period meet the program's measurement specifications.

Test Specifications Considerations

Test specifications define the mix of content, the measurement accuracy, and the level of security each test form must provide to satisfy a testing program's goals. They provide the criteria by which alternative test designs can be evaluated. Once the testing program is operational, they are used to monitor outcomes, detect potential issues that may develop and need to be addressed, and ensure that the desired goals continue to be achieved over time (Davey & Pitoniak, 2006). Test specifications for tests delivered on demand are typically organized into three main areas: content, measurement, and test security (Davey & Pitoniak, 2006; van der Linden & Glas, 2010).

For the GRE revised General Test, the *content specifications* prescribe the appropriate number of questions from specific primary and secondary content domains, and question type classifications (ETS, 2013c). Content specifications also prescribe that questions that are too closely related or provide clues to one another should not appear together in the same test.

The GRE revised General Test *measurement specifications* prescribe that every version of the test should be

- free from potentially significant gender or ethnicity bias,[2]
- highly reliable over the full range of ability to be measured,[3]
- accurately[4] scored, and
- delivered under nonspeeded timing conditions.[5]

Exploration of Alternative Test Designs

Different designs were evaluated as part of the development of the GRE revised General Test. A priori, the computerized linear test (CLT), section level adaptive or MST, and CAT designs all have features that could make them a good fit for the GRE revised General Test. All of these designs make use of computers for delivery and allow for the use of the new question types developed for the revised test (Wendler, Chapter 1.2, this volume). CLT would afford test takers the most flexibility in working through their tests and would be the simplest design to implement. However, relatively long and time-consuming tests would be required for achieving the desired level of measurement. CAT, on the other hand, would minimize the required test length by taking advantage of question level adaptation. However, in practice, CAT requires test takers to respond to questions sequentially without any possibility to review and revise answers. Furthermore, with the increased exposure constraints necessary for test security, ensuring that every CAT test measured every test taker equally well proved to be not only increasingly challenging but also costly in terms of test development (Mills & Steffen, 2000; Stocking & Lewis, 2000).

An MST design is also adaptive. However, it takes a middle-of-the-road approach between nonadaptive linear and question level adaptive designs, as it adapts at the end of each test section rather than after each question response (Zenisky, Hambleton, & Luecht, 2010). For example, a simple two-stage, three-level MST design using a panel[6] composed of four half-length test sections might be considered. In this case, the first section to be administered (routing section)[7] is made up of mostly average difficulty questions, and the other three sections are made up of mostly easy, middle difficulty, or difficult questions, respectively. Such a design, noted as MST13 and illustrated in Figure 3.3.1, is simple enough that many minimally overlapping or nonoverlapping panels can be effectively assembled and reassembled in large batches ahead of delivery. At delivery time, a panel is then selected and testing proceeds.

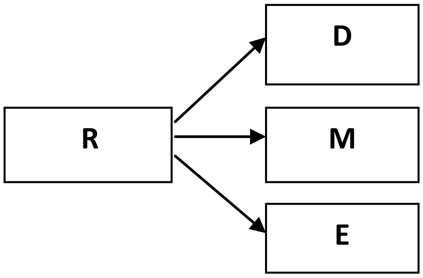


Figure 3.3.1. Schematic representation of a prototype GRE MST panel. As operational testing starts, the test taker is assigned to the routing section (R). After the routing section is completed, the second section is selected from the easy (E), middle (M), and difficult (D) sections, based on the test taker's performance on the routing section.

From a practical point of view, such a design has at least three important advantages. First, each panel can be fully evaluated before it is made available for delivery. If any of the sections or test forms the panel can produce fails to meet specifications, the panel is rejected and reassembled until all the specifications are met. Second, the problem of maintaining test security can be handled in a straightforward manner, at least by comparison with CAT. Depending on the size of the question bank, the limits imposed on question reuse, and the maximum allowable question overlap across forms, new panels may be produced and rotated very quickly. This way, test takers cannot predict the questions or group of questions they will be assigned. Third, within each section, test takers can be allowed to move through the test as they wish, revisit questions, and change their answers.

Given the potential of MST for the GRE revised General Test, research was conducted to develop the simplest and most effective MST design that would meet the GRE test specifications described above. Following previous research (Hendrickson, 2007; Zenisky et al., 2010), several design features were considered, including

- scoring method,
- number of sections (or stages),
- relative section length across stages,
- average question discrimination,
- extent to which sections differ in average difficulty, and
- range of question difficulty within section.

Exploration of Alternative Scoring Models

In order to narrow down the number of experimental conditions to investigate, the issue of scoring was considered first. Scoring is an important part of test design. Its role is to convert the pattern or the sum of question scores resulting from the responses provided by test takers on a specific test form to a score on the test's reporting scale. In doing so, scoring must adjust for the specific characteristics of each test form—in particular, in the case of adaptive testing in which the difficulty of forms can vary greatly as questions or sections are tailored to the test-taker performance. When testing is done only a few times a year, a traditional equating approach can be employed to account for each form's characteristics (Kolen & Brennan, 2004). However, with frequent or continuous test delivery, the use of item response theory (IRT) is generally required in order to produce comparable scores across the many forms delivered.

Two alternative IRT models were considered: the three-parameter model, which accounts for question difficulty, discrimination, and guessing (which may occur with questions that require test takers to select among a limited number of answer choices), and the two-parameter model, which accounts only for difficulty and discrimination. Analyses of both Verbal Reasoning and Quantitative Reasoning tryout data indicated that the IRT two-parameter model would provide a good fit to the data. This result was anticipated because the revised test greatly reduced the use of simple multiple-choice questions and, consequently, significantly reduced the possibility of successfully guessing answers. With fewer parameters to estimate, the two-parameter model requires smaller sample sizes than the three-parameter model previously used with CAT. Therefore, the two-parameter IRT model was chosen for the GRE revised General Test.

Two alternative classes of IRT scoring methods were considered: methods that make use of the information contained in the pattern of correct and incorrect responses (pattern scoring) and methods that make use of the total number of questions correctly answered[8] (number-correct scoring). The tradeoff between these approaches is that pattern scoring generally provides more reliable measurement (Hambleton, Swaminathan, & Rogers, 1991; Thissen & Wainer, 2001), while number-correct scoring provides measurement that tends to be more robust to suboptimal test adaptation or misleading responses caused by factors unrelated to the ability being measured, such as misunderstanding the directions, anxiety, and poor time management (Meijer & Nering, 1997; Stocking, 1996; Stocking, Steffen, & Eignor, 2002).

Evaluation of the GRE Multistage Test Design

A series of pilot and simulation studies was conducted with prototypes built using questions representative of the content of the GRE revised General Test and representative of the expected timing statistics and IRT parameters (Liu & Robin, 2009; Robin, Steffen, & Bontya, 2009; Zhao & Robin, 2009). These studies demonstrated that, for both the Verbal Reasoning and Quantitative Reasoning measures, the desired level of reliability and efficiency (that is, the testing time remained under 4 hours) could be obtained with the MST design illustrated in Figure 3.3.1. The results also showed that the use of more complex MST designs and the use of pattern scoring would not increase the reliability of the scores substantially enough to offset the additional test development challenges and constraints on the test-taking experience that would be created. Therefore, the choice of GRE MST design was narrowed down to the following features:

- Two-stage and three-level panels composed of 20-question sections
- IRT number-correct scoring using the two-parameter model
- Routing threshold set so that, roughly, one third of the test takers are assigned to easy, medium, and difficult second-stage sections, and the ability of each of these groups is well matched to the difficulty of the second-stage sections they are assigned to

Typical scoring and measurement results obtained with such a design are provided in Figures 3.3.2 and 3.3.3. Figure 3.3.2 shows all the possible reported scores (rounded scaled scores) that can be obtained from the panel's easy, medium, and difficult tests' number-correct scores. As expected, scoring takes into account the specific characteristics of each one of the three test forms that may be assigned from the panel. For example, the reader can see that the same number correct of 20 would be scored as 146, 149, or 153, depending on the difficulty of the form assigned. As expected, there are restrictions to the range of scores that can be obtained from the easy, medium, and difficult tests, as the assignment to these tests is restricted to only low, medium, and high performance in the routing section, respectively. The number correct to reported scores conversions shown also illustrate the extent to which the desired robustness of the scoring process is achieved. As can be seen, the impact of, for example, mistakenly responding on (any) one or two questions which may otherwise have been answered correctly is never more than 1 or 2 scaled score points.

The scoring conversion plots also highlight potential issues that needed to be paid attention to as the test design and the assembly specifications[9] were finalized. One such issue is scoring gaps, an example of which can be seen with the difficult form in Figure 3.3.2. In that case, the number-correct scores of 38 and 39 correspond to reported scaled scores of 167 and 169, skipping 168. Because this issue is closely related to scaling, the assembly blueprints were developed concurrently with the revised scales to make sure that no operational panels would have more than one gap toward the top of the scale—criteria set as part of the test specifications and the scaling goals (Golub-Smith & Moses, Chapter 2.2, this volume). Another issue, highlighted by the significant overlap in scoring between the easy, medium, and difficult forms, is the uncertainty associated with the routing decision. This is to be expected, since the routing decision is made with only partial information (responses to the routing section). As a result, some test takers whose true ability is close to one of the two routing thresholds will be assigned to the easy rather than the more appropriate middle second section (or vice versa), or assigned to the middle rather than the better suited difficult section (or vice versa). This issue, discussed in further detail below, was taken into account in the development of the assembly specifications and the choice of the ability thresholds used to make routing decisions.

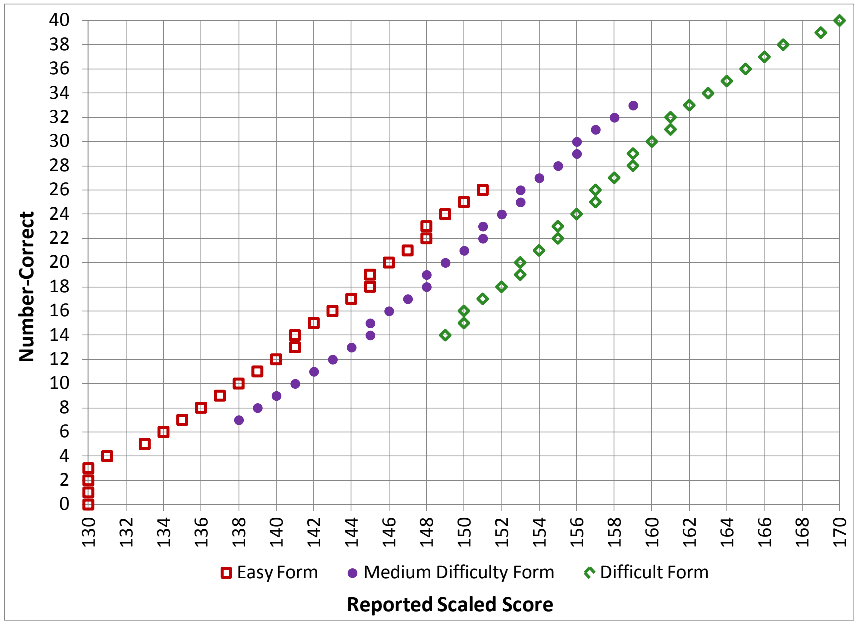


Figure 3.3.2. Number correct to reported scale score conversions for an MST13 panel. Note that because the reported scores are rounded to the nearest integer, more than one (up to two, except at the bottom of the scale) number-correct scores convert to the same reported score.

Figure 3.3.3 shows the measurement outcomes associated with each of the three forms test takers may be assigned to for the MST panel used in Figure 3.3.1. It shows that, unless ability is below 138, even in the least probable suboptimal routing situations in which low or high ability test takers are assigned to the medium form instead of the easy or difficult forms (dashed vertical lines), the standard error of measurement and scoring error value remain acceptable with values close to 3.0 and plus or minus 0.3, respectively.

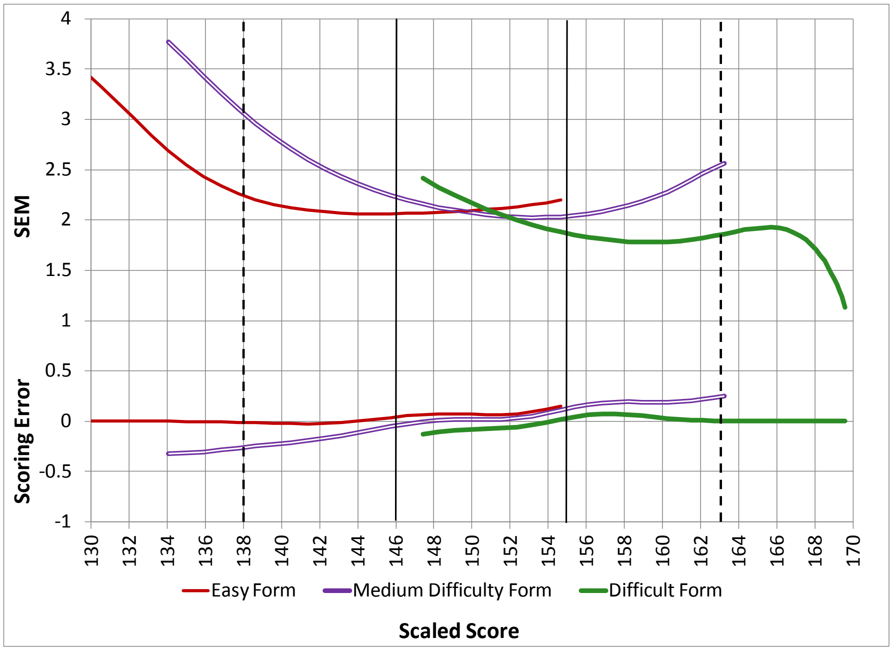


Figure 3.3.3. Measurement outcomes for the same MST panel as in Figure 3.3.1. The routing thresholds have been set at estimated scale values of 146 and 155 (solid vertical lines) so that approximately 30%, 40%, and 30% of the norm group would be assigned to the easy, medium, and difficult forms, respectively. The top solid lines indicate the estimated standard error of measurement (SEM) for each form. The bottom solid lines indicate the estimated scoring accuracy for each form—that is, the differences between the average score one would be expected to obtain by being assigned to a specific path and the average score one would be expected to obtain by following the MST assignment.

When the test design was finalized, the operational assembly specifications and processes were implemented. In particular, effective specifications imposing limits on question exposure, the amount of overlap between sections and panels, and the estimated time most test takers will need to complete each section were set.

Monitoring the Tests

More than 1 year after the launch of the GRE revised General Test in August 2011, large numbers of MST panels have been produced and delivered, and extensive monitoring and quality control analyses have been conducted. In this section, we summarize the main results obtained and show the extent to which the ongoing test assembly process is able to consistently meet the desired test specifications and, therefore, to provide reliable, accurate, and fair opportunities for all test takers to demonstrate their ability. For that purpose, we will concentrate on the characteristics of the hundreds of Verbal Reasoning and Quantitative Reasoning MST panels delivered in 2012. While very similar measurement outcomes were obtained with earlier panels, this time period is most representative of the current MST development processes. It includes adjustments to the initial assembly configuration implemented to optimize the routing rates and the reuse of questions as the new scales and the new operational question bank were established.

Unbiased

The first of the test measurement specifications listed earlier is that tests should be free from significant gender or ethnicity bias (Holland & Wainer, 1993). This specification is essentially handled as part of the question development process in which the questions identified as potentially biased are screened out of the question bank available for operational test assembly (Robin, Chapter 6.6, this volume).

Reliable and Accurate

Next, it is essential to make sure that all the MST test forms provide the level of reliability and scoring accuracy required. Figure 3.3.4 shows the full range of standard errors of measurement and scoring accuracy over all of the hundreds of Verbal Reasoning and Quantitative Reasoning panels delivered in 2012.[10] These results are consistent with the values published in the *GRE Guide to the Use of Scores* (ETS, 2013a, Table 6A). Most importantly, these results show that the differences between the least and most reliable and accurate of the MST panels delivered are quite small and that the lowest quality panel delivered still does meet the required level of reliability and scoring accuracy.

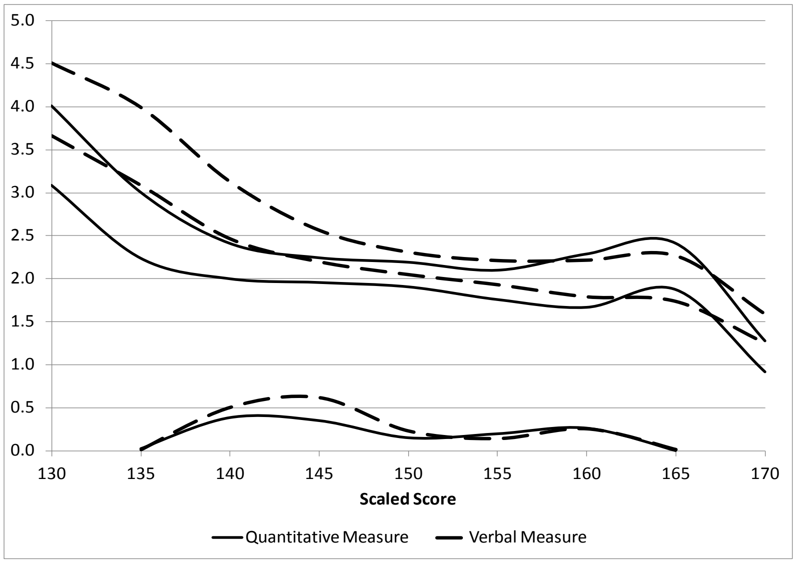


Figure 3.3.4. Verbal (dashed lines) and Quantitative (solid lines) minimum and maximum standard error of measurement (top lines) and maximum absolute scoring (in)accuracy (bottom lines) over all operational MST panels delivered in 2012.

Nonspeeded

Finally, it is specified that no test section should be delivered under speeded conditions. As part of the assembly process, the time most test takers will need to complete each section is estimated based on the question timing statistics collected though pretesting by means of unidentified test sections delivered along with the operational sections (ETS, 2013b). Drafted sections that have an estimated time greater than their empirically determined threshold value are discarded. The empirical thresholds for routing the easy, medium, and difficult sections were initially set based on tryout data and then adjusted as operational data were collected, so that 90% or more of the test takers would be expected to answer 80% or more of the questions in each section (16 or more questions out of 20). Also, it was expected that, given nonspeeded sections, test takers would be able to revisit questions and not be forced to engage in rapid response behavior as time may be running out.

The post administration analyses conducted after the operational administration on a regular basis included the percentage of MST panels flagged for missing the 90% rule just described, the average number of questions that test takers revisit one or more time, and the number of questions the 90th percentile of test takers respond to in less than 10 seconds (an indication of rapid response behavior consistent with guessing). As indicated in Table 3.3.1, most test takers had enough time to revisit more than four questions and very few of them appear to have engaged in rapid response behavior with more than one question. Therefore, we believe that all the Verbal measure sections were indeed delivered under nonspeeded conditions. With the

Quantitative Reasoning measure, 13% of the routing sections were flagged with low-ability test takers. However, as with Verbal, most test takers had enough time to revisit more than five questions and very few of them appear to have engaged in rapid response behavior.

Table 3.3.1

Summary of Speededness Evaluations for all MST Panels Delivered in 2012

Outcome	First-stage section			Second-stage section		
	Test takers' ability					
	Low	Medium	High	Low	Medium	High
Verbal						
Percentage of sections flagged	2	0	0	0	0	0
Average questions revisited	5	7	10	6	8	10
Average rapid responses	0.5	0.3	0.1	0.5	0.2	0.2
90th percentile of rapid responses	1	0	0	0	0	0
Quantitative						
Percentage of sections flagged	13	8	0	1	2	2
Average questions revisited	5	6	7	5	6	6
Average rapid responses	0.4	0.1	0.0	0.2	0.1	0.1
90th percentile of rapid responses	2	1	0	1	1	1

Note. Routing for the first stage section was easy; the routing for the second-stage section was medium or difficult.

Conclusion

In this chapter we described the main features of the multistage adaptive test design implemented with the Verbal Reasoning and Quantitative Reasoning measures in the GRE revised General Test. We showed how the technical choices made support the critical goals set for the revised test, such as enhancing the program's level of measurement for an increasingly diverse testing population and offering a more flexible, more easily understandable, and more secure test-taking experience.

Little more than a year after the successful launch of the new test, large amounts of empirical information have already been evaluated. Some questions have been retired and new ones have been added to the growing operational question bank. The quality of the revised test is being maintained, and continuous improvements of the operational processes will continue to be implemented.

References

Davey, T., & Pitoniak, M. (2006). Designing computerized adaptive tests. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 543–573). Mahwah, NJ: Erlbaum.

Educational Testing Service. (2013a). *GRE guide to the use of scores*. Princeton, NJ: Author.

Educational Testing Service. (2013b). *GRE revised General Test content and structure*. Princeton, NJ: Author.

Educational Testing Service. (2013c). *Introduction to the Analytical Writing measure*. Princeton, NJ: Author.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Hendrickson, A. (2007). An NCME instructional module on multi-stage testing. *Educational Measurement: Issues and Practice, 26*(2), 44–52.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning.* Hillsdale, NJ: Erlbaum

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer-Verlag.

Liu, J., & Robin, F. (2009). *Summary of GRE Verbal March/April 2009 field test results.* Unpublished manuscript, Educational Testing Service, Princeton, NJ.

Meijer, R. R., & Nering, M. L. (1997). Trait level estimation for nonfitting response vectors. *Applied Psychological Measurement, 21,* 321–336.

Mills, C. N., & Steffen, M. (2000). The GRE computerized adaptive test: Operational issues. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 27–52). Boston, MA: Kluwer.

Robin, F., Steffen, M., & Bontya, A. (2009). *Evaluation of alternative MST designs for GRE.* Unpublished manuscript, Educational Testing Service, Princeton, NJ.

Stocking, M. L. (1996). An alternative method for scoring adaptive tests. *Journal of Educational and Behavioral Statistics*, *21,* 365–389.

Stocking, M. L., & Lewis, C. (2000). Methods of controlling the exposure of items in CAT. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 163–182). Boston, MA: Kluwer.

Stocking, M. L., Steffen, M., & Eignor, D. R. (2002). *An exploration of potentially problematic adaptive tests* (Research Report No. RR-02-05). Princeton, NJ: Educational Testing Service.

Thissen, D., & Wainer, H. (2001). *Test scoring*. Mahwah, NJ: Erlbaum.

van der Linden, W. J., & Glas, C. A. W. (2010). *Computerized adaptive testing: Theory and practice*. Boston, MA: Kluwer.

Zenisky, A., Hambleton, R. K., & Luecht, R. M. (2010). Multi-stage testing: Issues, designs, and research. In W. J. van der Linden & C. A. W. Glas, (Eds.), *Elements of computerized adaptive testing* (pp. 355–372). New York, NY: Springer.

Zhao, J., & Robin, F. (2009). *Summary of GRE Quantitative March/April 2009 field test results.* Unpublished manuscript, Educational Testing Service, Princeton, NJ.

Notes

[1] The Analytical Writing measure is also part of the GRE General Test. However, as the measure includes two writing prompts delivered in a traditional linear fashion, it is not discussed in this chapter.

[2] In addition to extensive fairness reviews performed by expert test developers, pretest data are collected and questions identified as differentially functioning across gender and ethnicity subgroups are screened out of the operational question bank from which MST assembly is conducted (Holland & Wainer, 1993).

[3] Reliability and standard error of measurement values indicate the level of uncertainty associated with the reported scores. They are reported in the *GRE Guide to the Use of Scores* (ETS, 2013a, Tables 5 and 6).

[4] In addition to providing reliable scores, a test form should provide scores that can be accurately compared with the ones obtained from any other form (ETS, 2013a, Tables 5 and 6; Kolen & Brennan, 2004).

[5] Nonspeeded means that the large majority of test takers should be able to complete each of the sections they are assigned to within the testing time limits.

[6] A panel, or an MST panel, refers to the collection of sections available at testing time for test assembly and delivery.

[7] In this case, the first section is referred to as routing section because it is used to route (assign) the test-taker toward one of the more or less difficult sections available in the next testing stage.

[8] For example, consider a four-question test and two response patterns: 0101 and 1010, where 0 indicated an incorrect question response and 1 a correct response. While number-correct scoring will produce the same number-correct score of 2, pattern-scoring is likely to result in two different scores as question difficulty and discrimination may differ across questions.

[9] The assembly specifications are used to translate the test specifications into terms that the automated assembly software can process.

[10] In Figure 3.3.3, the standard errors of measurement and scoring accuracy statistics were estimated separately for each of the forms that an MST panel can produce. Here the same statistics are estimated and combined at the level of the entire panel, thus accounting for the probabilities that a test taker could be assigned to any of the forms the panel can produce.

3.4 Potential Impact of Context Effects on the Scoring and Equating of the Multistage *GRE*® revised General Test [1]

Tim Davey and Yi-Hsuan Lee

The decision to move the *GRE*® General Test from a computer adaptive test design, where subsequent questions examinees are given are dependent upon performance on earlier questions, to a multistage adaptive design, where examinees are routed to subsequent sections based on performance on a previous section, was made for theoretical and practical reasons. However, this change in test design brought with it a number of operational and psychometric challenges. One of these challenges, the impact of question-level context effects, was examined in this study.

Context refers to the position in which a question appears in a test, as well as the content, format, and specific features of the other questions that surround it. Prior research indicates that question position is a primary driver of context effects (Dorans & Lawrence, 1990; Haladyna, 1992; Harris, 1991). Changes in question position may inadvertently modify the characteristics (e.g., difficulty level) of an individual test question. Test forms used in high-stakes testing, such as the GRE revised General Test, are created using exacting specifications to ensure comparability in difficulty and reliability across different versions of the test. This comparability depends on questions performing in the same manner as when they were pretested. In addition, changes to the characteristics of a question may impact the equating process by which scores on different versions of the test are made comparable. The equating process also assumes that questions perform the same across all versions of the test. Changes in the characteristics of the questions could confound this process and result in concerns about the comparability of scores across alternate test versions.

In order to evaluate the impact of context effects, two research questions were examined:

1. Are question position effects evident in linear test forms administered to GRE examinees?
2. Are question position effects likely to pose a particular challenge to a multistage test?

Method

The operational GRE General Test contains a variable section that is generally used for pretesting new questions. Each examinee receives a single variable section that includes either verbal or quantitative questions. Data for this study were collected from variable sections administered as part of the operational test with a computer adaptive test design. This ensured that the data came from highly motivated examinees. However, since the variable section must look like an operational section, it did not allow the administration of some of the new question types being considered for use in the GRE revised General Test, nor did it allow some of the operational changes that were being made with the revised test (e.g., the use of a calculator, time limit or test length changes, allowing the examinee to review or revise previous answers). Therefore, while the results of the study provided guidance in making decisions about the GRE revised General Test, they have limitations in terms of their generalizability. Data were collected in three rounds.

In Round 1, sets of 28 quantitative and 30 verbal questions were arranged to form linear test sections; that is, unlike the operational GRE revised General Test, they were not adaptive. A total of 13 quantitative and seven verbal sections were created, each containing the same questions in different orders. Examinees took one section of either quantitative or verbal questions. The sections were administered in scrambled orders so that each question appeared with equal frequency in each of several general locations throughout the test. One of the orderings for both quantitative and verbal questions was designated as the fixed *base* ordering and was administered to a larger sample than were the other orderings. About 5,693 examinees in total were involved in this data collection occasion.

In Round 2, sets of 80 quantitative and 84 verbal questions were arranged to form linear test sections. The quantitative questions were divided into three sets of 28 questions each, with four of the 80 questions appearing in two different sets. The three sets were assembled into 39 sections, each containing the questions in different orders. The verbal questions were similarly assembled into 21 sections, each of which was ordered in distinct ways. Each examinee took one section of either quantitative or verbal questions. The sections were administered to 11,245 examinees and were randomly spiraled across examinees, with each question appearing in various locations throughout the test section with roughly equal frequency.

In Round 3, all questions administered in the first two data collections were combined to approximate a multistage test. Examinees began with a set (module) of eight moderately difficult questions and then, based on their performance, were routed to any of five second-stage modules. These modules ranged from quite easy to quite difficult. Again, contingent on performance, examinees were routed to the next set of modules. Each examinee thus took four different sets of modules, consisting of either quantitative or verbal questions. While this multistage design did not directly reflect that ultimately used in the GRE revised General Test, its impact on examinees was expected to be similar. A total of 2,947 examinees took a verbal multistage test and 2,916 took a quantitative multistage test.

Results

Operational GRE scores served as an independent measure of the ability level of the examinees. Results indicated that no significant differences in score distributions for verbal or quantitative were seen across the different groups of examinees that were administered each question ordering. This means that observed differences in performance across question orderings is likely to be the result of where the question appears in the test, rather than due to underlying group differences.

Analyses of position effects compared differences in difficulty level, indicated by proportion of correct responses or *p* value, for each question in each of the various positions in which it appeared. These *p* values were calculated in such a way that they were relative to a question's position on the test. Results indicated that eight of the 28 quantitative and four of the 30 verbal questions were significantly impacted by where they appeared on the test. Questions appeared to be easier when their position moved forward on the test and more difficult when their position moved backward on the test. Quantitative questions were more likely to demonstrate a change in difficulty level when their position was changed compared to verbal questions. For the verbal, questions based on passages (i.e., one passage, multiple questions) were more likely to change in difficulty when their position was changed than were discrete questions.

One strategy for reducing the impact of position effects might be to pretest questions in a variety of positions throughout a section rather than in a single, fixed position. This was evaluated by comparing the *p* values from the various scrambled orders to the *p* values from the base ordering. This comparison simulates pretesting questions in random positions throughout a test section and then administering them in fixed positions in operational test sections. Results indicated no differences in difficulty level between the random and fixed orderings. Thus, pretesting questions in random locations throughout the test appears to effectively diminish position effects.

To determine if position effects existed in the multistage tests, residuals, which were the differences between (empirical) *p* values and model predicted values, were calculated for each question relative to its position on the test and aggregated at the module level. Analyses of the aggregated residuals revealed that modules composed of easier questions were even easier than pretest estimates predicted, while modules composed of difficult questions were even more difficult than pretest estimates predicted. This effect seemed to grow more pronounced as the test progressed through the final two stages.

Position effects can be influenced by a number of test administration conditions, such as the time limits allocated to complete the test section, or speededness. Examinees may underestimate the extent to which a section is speeded and spend too much time on earlier questions, causing them to rush through questions near the end of the section. The result is that late-appearing questions look more difficult than expected. In addition, a test given under a specific time limit may be less speeded for examinees at higher ability levels compared to less able examinees. Speededness is also influenced by test difficulty. A time limit that accommodates a set of easy questions may be insufficient for a set of difficult questions. For a multistage test, this aspect is especially challenging because more able examinees generally receive sections containing more difficult sets of questions.

To determine the extent to which speededness affected the results of the position effects analyses, the completion rates for each test section were examined. Examinees were first divided into five ability levels based on their GRE scores, and completion rates were computed within each of the levels. For the verbal sections, completion rates increased modestly with increasing levels of examinee ability. However, the opposite was seen with the quantitative sections: The examinees at the lower ability levels were much more likely to complete the section than were the examinees at the higher ability levels. Thus, the multistage tests, especially those containing quantitative questions, were more speeded than the linear forms.

Conclusion

The results of this study provided guidance on the design of the GRE revised General Test. First, to ensure that position effects are mitigated, questions are pretested in random locations throughout the test. Second, the impact of speededness was a critical factor to consider in the design of the revised test because multistage tests, like all adaptive tests, are more subject to speededness than are linear forms of the same length with the same time limits. As a result, additional field studies were held to determine the appropriate test configuration for the Quantitative Reasoning measure as a way of minimizing the influence of speededness and context effects (Rotou, Liu, & Sclan, 2006).

References

Dorans, N. J., & Lawrence, I. M. (1990). Checking the statistical equivalence of nearly identical test editions. *Applied Measurement in Education*, *3*, 245–254.

Haladyna, T. (1992). Context-dependent item sets. *Educational Measurement: Issues and Practice, 11*(1), 21–25.

Harris, D. (1991, April). *Practical implications of the context effects resulting from the use of scrambled test forms.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Rotou, O., Liu, M., & Sclan, A. (2006, April). *A configuration study for the quantitative measure of the new GRE.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Notes

[1] Based on *Potential Impact of Context Effects on the Scoring and Equating of the Multistage GRE® revised General Test* (GRE Board Research Report No. 08-01), by T. Davey and Y.-H. Lee, 2011, Princeton, NJ: Educational Testing Service.

Section 4: Understanding Automated Scoring

It is critical that scores from a test are scored accurately and fairly for all test takers. Since the Analytical Writing measure was implemented on the *GRE*® General Test, specially trained expert human raters have been used to score the essays. Over the last several years, however, considerable strides have been made in the development of an automated essay scoring engine that models those processes used by trained essay readers. As a result of these efforts, the ETS-developed *e-rater*® scoring engine was put in place in 2008 to score the Analytical Writing measure along with human raters. The use of e-rater was introduced prior to the release of the GRE revised General Test. However, the research conducted to support the use of automated scoring for the Analytical Writing measure in the previous version of the GRE General Test also provides the foundation for the use of e-rater with the revised test. Chapters in this section describe many of the studies done that provide the evidence for the accuracy and fairness of an automated scoring engine.

- Chapter 4.1 provides an overview to the functioning of an automated scoring engine. It describes the pros and cons, challenges, and strengths of using automated scoring. It acknowledges the rich history of automated essay scoring (AES) systems dating back to pioneering work by Ellis Page in the 1960s but focuses primarily on the AES system developed at ETS , e-rater. The chapter also describes the process by which e-rater models are developed and evaluated.
- Chapter 4.2 reports on a study that evaluated different aspects of the validity of automated scoring. Although the agreement of machine and human scores is a component of the validity argument for automated scoring, evidence of relationships to other measures of writing ability is also critical. In this study, data were collected from approximately 1,700 prospective graduate students from 26 colleges and universities across the United States; half of the group wrote one issue essay and one argument essay, while the other half wrote either two issue or two argument essays. Indicators of writing skills, such as course writing samples and test takers' perceptions of their writing skill level, were compared to the scores on the issue and argument essays generated by e-rater and those generated by human raters. Ratings based on the combination of one human rater and e-rater correlated with the external criteria to almost the same extent as ratings based on two human raters.
- Chapter 4.3 describes a study that evaluated the extent to which e-rater could be fooled into assigning scores that were inappropriate compared to what a human rater would assign. In the study, a number of writing experts were invited to create

essays with the intention of tricking e-rater into awarding scores that were either higher or lower than deserved. For essays that were written to fool e-rater to assign a score that was too high, predictions of higher scores were borne out in 26 of 30 instances (87%), while predictions for essays that were written to fool e-rater to give a score that was too low, were accurate less often (10 of 24, or 42% of the time). Two essay features were especially important in e-rater overvaluing the essay: (a) repeating the exact same set of paragraphs a number of times and (b) repeating the same paragraph but rewording the first sentence slightly. As a result of this research, later versions of e-rater include flags to identify essays with this kind of repetition so that they go directly to human raters and not receive an e-rater score.

- Chapter 4.4 provides the results of a study that examined different scoring models that could be used with e-rater. Some automated essay scoring engines rely heavily on content features that are unique to each essay prompt, but because e-rater emphasizes form over content, it can score many topics using the same standards. Such a generic scoring approach allows the same scoring model to be used across prompts, and new prompts can be introduced without requiring any changes to the scoring engine. This study examined the functioning of a generic scoring model and compared it to approaches that were more dependent on the content in particular prompts. In terms of average scores and correlations with scores assigned by human raters, the scores from the generic approach were comparable to scores from the much more time- and labor-intensive prompt-specific approach.
- Chapter 4.5 describes efforts that examined whether e-rater scores could successfully replace one of the two human rater scores used for GRE essays. Various criteria were used as part of this evaluation, including association with human raters, examination of the kinds of skills assessed, degradation from the agreement level of human/human ratings, association with external variables (such as scores on other GRE test sections, grades in English classes, etc.), subgroup differences, and operational impact. On most dimensions evaluated, e-rater and human scores were quite comparable. An exception was the agreement between human and e-rater scores for examinees from China. Specifically, e-rater tended to give much higher scores (by more than half of a standard deviation) to examinees from China. Because this could introduce serious bias if e-rater and human scores were simply averaged, the report recommended the use of the check score model for the GRE Analytical Writing scores. With a check score model, the e-rater score is compared to the one assigned by a single human rater. If there is no discrepancy, the human score stands. If the scores are discrepant, a second human reader reads the essay and the scores of the first and second human are averaged (and if the first and second humans disagree by more than a point, an additional human score is

obtained). In this system, the essay score is always based solely on evaluations by human raters. The check score model is currently used for scoring GRE Analytical Writing measure essays.

- Chapter 4.6 reports on a study that investigated the use of e-rater with essay variants. These variants, which are created from a *parent* prompt, ask a focused question that addresses a specific aspect of the prompt and requires the test taker to respond to that aspect. Variants were developed to address the problem of memorized responses to essay prompts that did not make such specific demands and to help test developers enlarge the pool of essay topics. This chapter extends the findings from a study described in Chapter 1.10, which investigated the comparability of essay variants as scored by human raters to e-rater scores on essay variants. Findings indicated that e-rater could score all variant types and that no significant differences in performance were found across different variant types.
- Chapter 4.7 focuses on the quality control role that automated engines can play in the scoring of essays. It addresses both the why and the how of automated essay scoring. In particular, the use of the check score model with GRE essays is discussed.
- Chapter 4.8 describes a study that investigated differences in scores produced by humans and e-rater by gender, ethnicity, and country of origin. For most groups studied, the average scores produced by e-rater and human raters were almost identical. A notable exception was essays from mainland China that received much higher ratings from e-rater than from human raters. This finding was especially puzzling because such differences were not noted in other Asian countries, or even in Taiwan, which shares the same language as mainland China. This study provided additional support for the use of a check score model for the GRE Analytical Writing measure, rather than averaging the human and e-rater scores.
- Chapter 4.9 examined possible root causes for those discrepancies seen in Chapter 4.8 between scores generated by human raters and those generated by e-rater across various subgroups. The research suggested that e-rater is not severe enough on grammatical language errors (compared to humans), tends to overvalue long essays, and occasionally undervalues content.

4.1 Overview of Automated Scoring for the *GRE*® General Test

Chelsea Ezzo and Brent Bridgeman

Constructed response (CR) questions such as essays are increasingly popular as evaluations of ability (Aschbacher, 1991) but bring added complexity and subjectivity to the scoring process. CR scoring must be made as uniform and consistent as possible, especially for programs where scores must have the same meaning across administrations. Depending on the testing program, either a single rater or multiple raters may score the CRs. This requires training and certification of raters, a process that is time-consuming and costly.

In addition, although essays are thought to be a more exemplary and direct measurement of a person's writing skills, they are very time consuming, as well as labor intensive, for humans to score (Attali, Bridgeman, & Trapani, 2009). As a result, there have been considerable strides in the development of applied automated essay scoring (AES) that model processes utilized by trained essay readers.

Critics argue that AES engines focus on surface structure and linguistic features of essays, rather than analyzing deeper meaning such as strength of argument. Focus, persuasive techniques, and connection to the audience contribute much to the quality of an essay, and there is some concern that computers cannot differentiate between the average and exceptional in such stylistic areas to the degree that humans can (DeLoughry, 1995). Others express concerns that AES has no mechanism for recognizing unusual responses that may receive a much different score from a human than from AES (Cohen & Wollack, 2006), or that AES may not be appropriate for some examinees, such as English-language learners (Weigle, 2013), or they simply do not believe that the kinds of essays that AES engines can score have much value for revealing writing ability (Condon, 2013).

However, proponents of the use of AES argue that because text production and ability to address complex issues are related, AES engines and human raters agree highly (Deane, 2013). Further, AES engines appear to be as reliable as human raters (Attali & Burstein, 2006) and may add stability to the scoring process (Walker, 2007). In fact, the scores produced by the first automated engine developed in 1966 were deemed nearly indistinguishable from scores assigned by human raters (Page, 1966).

Current major rating systems include the Intelligent Essay Assessor (IEA; Landauer, Laham, & Foltz, 2003), IntelliMetric (Elliot, 2001), Project Essay Grade (PEG; Page, 1994), and *e-rater*® scoring engine (Attali & Burstein, 2006). The latter two of these systems rely heavily on various regression procedures to hone in on the most predictive features in each essay (Attali, Bridgeman, & Trapani, 2010). Advantages provided by all of these finely tuned programs include unwavering impartiality, consistency, objectivity, and reliability (Schwartz, 1998; Weinstein, 1998).

The e-rater engine, which is used to score the Analytical Writing measure of the *GRE*® General Test, aggregates sets of features deemed vital to high-quality writing, then uses regression analysis to weight each feature and generate a final score (Attali et al., 2009). Under an analytic scoring framework, e-rater's feature categories were mapped by Quinlan, Higgins, and Wolff (2009) to the six-trait scoring model (Culham, 2003) that focuses on the dimensions of ideas and content, organization, voice, word choice, sentence fluency, and conventions. These dimensions are each represented in e-rater's 11 score features—nine representing aspects of writing quality and two representing content. Grammar, usage, mechanics, and style features together identify more than 30 error types, including errors in subject-verb agreement, homophone errors, misspelling, and overuse of vocabulary. Two prompt-specific (PS) vocabulary usage features relate to content of vocabulary used in the essay. Other feature types include organization, development, lexical complexity, and correct use of collocations and prepositions. Each feature uses a set of automated criteria to evaluate an essay in relation to the prompt and other essays written to the same prompt.

Developing e-rater scoring models is typically a two-stage process: (a) model training/building and (b) model evaluation (Ramineni, Trapani, Williamson, Davey, & Bridgeman, 2012). This is a fully automated process, given a random sample of population-representative training essays in the model building set. Prior to model building, each essay set is subjected to a number of advisory flags or filters, which have been established to indicate when an essay is inappropriate for the model build process. The filters eliminate essays that, among other things, poorly develop key concepts, are inappropriate in length, and are irrelevant to the prompt topic.

The e-rater engine focuses on three key, nonstylistic features whose strong presence has been identified by human raters as vital to a superior essay: structure, organization, and content (Powers, Burstein, Chodorow, Fowles, & Kukich, 2001). During the programming process, an advisory flag system is employed by which specific elements that render an essay inappropriate for automated scoring are identified (Ramineni et al., 2012). This process ensures that the essays used to program e-rater are adequate examples of the three key features of strong writing. Typically, automated scoring models must be calibrated for a specific writing prompt (Landauer et al., 2003; Page, 1994; Rudner, Garcia, & Welch, 2006). The e-rater engine, however, uses a validated technique that allows for a more generic approach (Attali et al., 2010; Ramineni, Williamson, & Weng, 2011). By emphasizing form over content, both in the programming process and operationally, it enhances score validity by standardizing the evaluated features across prompts, much as a human rater would.

A great deal of the research on AES examines the degree to which computer-produced scores agree with those ascribed by a human rater (Powers, Burstein, Chodorow, Fowles, & Kukich, 2000). To this end, sophisticated programs have been tested with many different types of assessments, such as the Graduate Management Admission Test (GMAT), the *PRAXIS*® assessments, and the TWE® (Burstein et al., 1998; Kaplan et al., 1998; Peterson, 1997). Generally speaking, a comparable level of agreement is consistently demonstrated with correlations ranging from .60 to .96. When reflecting on these results, it is important to keep in mind that there is no foundational evidence that demonstrates the superiority of the human assessment of writing (Bennett & Bejar, 1997). After all, the phrase *prone to human error* is prevalent for a reason.

The current version of e-rater demonstrates a high level of agreement with human raters (Attali et al., 2009; Attali & Burstein, 2006; Ramineni et al., 2012) with the greatest reliability coming from the use of e-rater in conjunction with a human rater (Bridgeman, 2004). Although it was shown that an early version of e-rater could be fooled into assigning higher scores than merited, most of the process mechanisms that caused such discrepancy have subsequently been modified or eliminated. While there have still been some observed differences in agreement in certain demographic and language subgroups (Bridgeman, Trapani, & Attali, 2012; Burstein & Chodorow, 1999; Ramineni et al., 2011), these differences have resulted in the modification of policies regarding the operational use of e-rater in both the GRE and *TOEFL*® tests (Ramineni et al., 2011). The use of e-rater provides a rapid, efficient, and reliable method for scoring Analytical Writing measure essays. However, AES is not the sole determinant of an examinee's score. For the Analytical Writing measure, e-rater is used as a check to the human rater score, triggering an additional reading by a human rater whenever the human score and the rounded AES score do not agree exactly (Monaghan & Bridgeman, 2005). In addition, the e-rater engine is continuously being improved. Research on new features and modified weights that enhance the measurement of essay quality is ongoing.

References

Aschbacher, P. R. (1991). Performance assessment: State activity, interest, and concerns. *Applied Measurement in Education, 4,* 275–288.

Attali, Y., Bridgeman, B., & Trapani, C. (2009). *E-rater performance for GRE essay variants.* Unpublished manuscript, Educational Testing Service, Princeton, NJ.

Attali, Y., Bridgeman, B., & Trapani, C. (2010). Performance of a generic approach in automated essay scoring. *The Journal of Technology, Learning, and Assessment, 10*(3).

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V.2. *Journal of Technology, Learning, and Assessment, 4*(3).

Bennett, R., & Bejar, I. (1997). *Validity and automated scoring: It's not only the scoring* (Research Report No. RR-97-13). Princeton, NJ: Educational Testing Service.

Bridgeman, B. (2004, December). *E-rater as a quality control on human scorers*. Presentation in the ETS Research Colloquium Series, Princeton, NJ.

Bridgeman, B., Trapani, C., & Attali, Y. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education, 25,* 27–40.

Burstein, J., Braden-Harder, L., Chodorow, M., Hua, S., Kaplan, B., Kukich, K., . . . Wolff, S. (1998). *Computer analysis of essay content for automated score prediction: A prototype* *automated scoring system for GMAT analytical writing assessment essays* (Research Report No. RR-98-15). Princeton, NJ: Educational Testing Service.

Burstein, J., & Chodorow, M. (1999, June). Automated essay scoring for nonnative English speakers. In *Computer-mediated language assessment and evaluation of natural language processing*. Joint symposium of the Association of Computational Linguistics and the International Association of Language Learning Technologies, College Park, Maryland.

Cohen, A. S., & Wollack, J. A. (2006). Test administration, security, scoring, and reporting. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 356–386). Westport, CT: American Council on Education and Praeger.

Condon, W. (2013). Large-scale assessment, locally-developed measures, and automated scoring of essays: Fishing for red herrings? *Assessing Writing*, 18, 100–108.

Culham, R. (2003). *6+1 traits of writing: The complete guide grades 3 and up*. New York, NY: Scholastic.

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing, 18*, 7–24.

DeLoughry, T. J. (1995, October 20). Duke professor pushes concept of grading essays by computer. *Chronicle of Higher Education,* pp. A24–25.

Elliot, S. M. (2001, April). *IntelliMetric: From here to validity*. Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Kaplan, R. M., Wolff, S., Burstein, J. C., Lu, C., Rock, D., & Kaplan, B. (1998). *Scoring essays automatically using surface features* (GRE Board Research Report No. 94-21). Princeton, NJ: Educational Testing Service.

Landauer, T. K., Laham, D., & Foltz, P. W. (2003). Automated scoring and annotation of essays with the Intelligent Essay Assessor. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 87–112). Mahwah, NJ: Lawrence Erlbaum.

Monaghan, W., & Bridgeman, B. (2005, April). *E-rater as a quality control on human scores* (R&D Connections No. 2). Princeton, NJ: Educational Testing Service.

Page, E. B. (1966). The imminence of grading essays by computer. *Phi Delta Kappan, 48*, 238–243.

Page, E. B. (1994). Computer grading of student prose, using modern concepts and software. *Journal of Experimental Education, 62*, 127–142.

Peterson, N. S. (1997, March). *Automated scoring of writing essays: Can such scores be valid?* Paper presented at the annual meeting of the National Council on Education, Chicago, IL.

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2000). *Comparing the validity of automated and human essay scoring* (GRE Board Research Report No. 98-08aR). Princeton, NJ: Educational Testing Service.

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2001). *Stumping e-rater: Challenging the validity of automated essay scoring* (Research Report No. RR-01-03).Princeton, NJ: Educational Testing Service.

Quinlan, T., Higgins, D., & Wolff, S. *Evaluating the construct-coverage of the e-rater scoring engine.* (Research Report No RR-09-01). Princeton, NJ: Educational Testing Service.

Ramineni, C., Trapani, C. S., Williamson, D. M., Davey, T., & Bridgeman, B. (2012). *Evaluation of e-rater for the GRE issue and argument prompts* (Research Report No. RR-12-02). Princeton, NJ: Educational Testing Service.

Ramineni, C., Williamson, D., & Weng, V. (2011, April). *Understanding mean score differences between e-rater and humans for demographic-based groups in GRE.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Rudner, L. M., Garcia, V., & Welch, C. (2006). An evaluation of the IntelliMetric^{SM} essay scoring system. *Journal of Technology, Learning, and Assessment, 4*(4).

Schwartz, A. E. (1998, April 26). Graded by machine. *The Washington Post,* p. C07.

Walker, M. E. (2007, April). *Criteria to consider when reporting constructed response scores*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Weigle, S. C. (2013). English language learners and automated scoring of essays: Critical considerations. *Assessing Writing, 18*, 85–99.

Weinstein, B. (1998, June 21). Software designed to grade essays. *Boston Globe Online* [Online serial].

4.2 Comparing the Validity of Automated and Human Essay Scoring[1]

Donald Powers, Jill Burstein, Martin Chodorow, Mary Fowles, and Karen Kukich

Prior research (e.g., Burstein et al., 1998; Burstein & Chodorow, 1999; Kaplan et al., 1998; Page & Petersen, 1995; Petersen, 1997) focused on the degree of agreement between scores generated using automated scoring methods and human essay scores and showed that automated scoring methods produce scores that agree strongly, albeit not perfectly, with those assigned by human raters. In some instances, the agreement between automated scoring and human raters was stronger than the agreement between pairs of human raters. In addition, limited evidence had been found of the relationship between automated scores and other measures of writing ability, such as scores from other sections of a test (e.g., reading, mathematics, or writing based on editing and error recognition tasks on the *PRAXIS®* assessments [Petersen, 1997]). The correlations between automated scores and these criteria were similar to those between human scores and the criteria, which suggest that the machine and human raters may be measuring similar constructs.

It is clear that there is some evidence, albeit limited, comparing the relationships of human and automated scores to other, independent indicators of writing skill. As one automated scoring engine, the *e-rater®* scoring engine, was being considered for use in combination with human-generated scores on the *GRE®* General Test, the aim of the study reported here was to generate more such evidence.

Procedure

This study examined relationships of each of two sets of GRE writing measure scores—those given by human raters and those generated by e-rater—to several independent, nontest indicators of writing skill, such as course-related writing samples and examinee perceptions of their success with writing. Data were collected from approximately 1,700 prospective graduate students from 26 colleges and universities across the United States. Each participant composed two GRE essays at a test center; half of the group wrote one analyze an issue (issue) essay and one analyze an argument (argument) essay, while the other half wrote either two issue or two argument essays. In addition, participants provided several nontest indicators of their writing skills (e.g., two samples of writing prepared for undergraduate course assignments—one of typical quality, and one of somewhat lower quality; self-evaluation of writing ability relative to their peers; self-reported grades in undergraduate courses requiring considerable writing; self-reported accomplishments in writing, such as publishing a letter to the editor or writing a paper for a professional meeting; and self-reported success with various kinds of writing, such as research papers or persuasive writing, and with various processes of writing, such as organizing, drafting, and editing or revising).

Results

Correlations between individual human raters were .85 for issue essays and .83 for argument essays, and agreement between human raters was very high (99% exact agreement or one point score differences). Agreement between e-rater and human raters was also very high (93% exact agreement or one point score difference). Performance on the GRE Analytical Writing measure correlated significantly but modestly with each of the nontest indicators. Correlations ranged from .06–.09 (with accomplishments) to .24–.38 (with scores on writing samples scored by GRE raters). The patterns were similar for human and e-rater scores. In other words, nontest indicators that related most strongly to human scores also related most strongly to e-rater scores. One reason for the similar pattern for the relationship of human and e-rater scores to nontest indicators was the greater concentration of e-rater scores—relative to human scores—around the average. With the e-rater model in use at the time of this study, e-rater was less likely than human raters to assign very high or very low scores, so there was less variation in e-rater scores than in human scores. In general, combining e-rater scores with either one of the human scores resulted in relatively small decreases in validity as compared to estimates based on two human raters.

Table 4.2.1 shows the correlations with the nontest indictors for essay scores based on human raters, e-rater, and combinations of human and e-rater scores. Correlations for the combination of one human rater and e-rater were only slightly lower than the correlations for two human raters (and for some criteria, were identical).

Table 4.2.1

Correlations of GRE Writing Measure Scores With Nontest Indicators of Writing Skill, for Human and Automated Scoring

Indicator	One human reader	Two human raters	e-rater	One human reader and e-rater	Two human raters and e-rater
Writing samples (raters' grades) [a]	.31	.38	.24	.33	.36
GPA in writing courses [b]	.29	.34	.27	.33	.34
Self-comparison with peers [c]	.24	.29	.17	.25	.27
Success with various kinds of writing [d]	.22	.26	.16	.23	.25
GPA overall	.18	.20	.17	.20	.21
Success with writing processes [e]	.17	.20	.13	.18	.19
GPA in major field	.12	.14	.13	.15	.15
Writing samples (professors' grades) [f]	.13	.16	.12	.15	.16
Accomplishments [g]	.06	.07	.09	.08	.08

[a] Two samples of undergraduate writing graded by trained essay raters. [b] GPA in undergraduate courses that required a "considerable" amount of writing. [c] Comparison of writing with peers in major field of study (well below average to well above average). [d] Reported success in college courses (not at all successful to extremely successful) with various kinds of writing (e.g., personal writing, persuasive writing, analysis/criticism, essay exams). [e] Reported success (not at all to extremely) with various writing processes (e.g., organizing ideas and revising). [f] Grades given by professors to the undergraduate writing samples evaluated in this study. [g] Reported accomplishments in writing (e.g., publishing a

letter to the editor, writing technical manuals or other instructional material, authoring or co-authoring an article published in a scholarly journal).

Conclusion

Significant but modest correlations were found between the nontest indicators and each of the two methods of scoring. Automated and human scores exhibited similar relations with the nontest indicators, which suggest that the two methods of scoring reflect similar aspects of writing proficiency. For a variety of reasons, the relationships between writing skill indicators and automated scores may be weaker than between the same indicators and human scores. E-rater may fail to focus as well as humans on features of writing reflected in the nontest indicators and tend to assign scores clustered in the middle of the score scale.

Some of the limitations of e-rater might be addressed by including additional features into the e-rater models: exploring the dimensionality of e-rater features to arrive at clusters of features (i.e., factors) that are more reliable or stable than individual features, employing more adequate data for specifying e-rater scoring models and improving e-rater's ability to discriminate (i.e., distinguish between ability levels). The authors also recommended that general e-rater models be used for different prompts in further studies, as this would mean that the same features are evaluated regardless of the prompt.

References

Burstein, J., Braden-Harder, L., Chodorow, M., Hua, S., Kaplan, B., Kukich, K., . . . Wolff, S. (1998). *Computer analysis of essay content for automated score prediction: A prototype automated scoring system for GMAT analytical writing assessment essays* (Research Report No. RR-98-15). Princeton, NJ: Educational Testing Service.

Burstein, J., & Chodorow, M. (1999, June). Automated essay scoring for nonnative English speakers. In *Computer-mediated language assessment and evaluation of natural language processing.* Joint symposium of the Association of Computational Linguistics and the International Association of Language Learning Technologies, College Park, MD.

Kaplan, R. M., Wolff, S., Burstein, J. C., Lu, C., Rock, D., & Kaplan, B. (1998). *Scoring essays automatically using surface features* (GRE Board Research Report No. 94-21). Princeton, NJ: Educational Testing Service.

Page, E. B., & Petersen, N. S. (1995). *The computer moves into essay grading: Updating the ancient test.* Phi Delta Kappan, 76, 561–565.

Petersen, N. S. (1997, March). *Automated scoring of writing essays: Can such scores be valid?* Paper presented at the annual meeting of the National Council on Education, Chicago, IL.

Notes

[1] Based on *Comparing the Validity of Automated and Human Essay Scoring* (GRE Board Research Report No. 98-08a), by D. E. Powers, J. C. Burstein, M. Chodorow, M. E. Fowles, and K. Kukich, 2000, Princeton, NJ: Educational Testing Service.

4.3 Stumping *e-rater*®: Challenging the Validity of Automated Essay Scoring[1]

Donald Powers, Jill Burstein, Martin Chodorow, Mary Fowles, and Karen Kukich

The use of an automated engine to score the *GRE*® Analytical Writing measure was introduced prior to the release of the GRE revised General Test. However, the research conducted to support the use of automated scoring for the Analytical Writing measure is relevant in that it provided the foundation on which automated scoring could be applied to the GRE revised General Test.

Early on, in order to gain the acceptance of automated scoring by the testing community and score users, researchers needed to better understand the various kinds of challenges that would be faced. The objective of this study was to evaluate the extent to which the *e-rater*® scoring engine, one of the automated scoring methods used by Educational Testing Service (ETS), may improperly reward or, conversely, unfairly penalize certain features of examinees' writing.

Some critics assert that computers, unlike human raters, are incapable of distinguishing exceptional, inspirational essays from those that, while technically correct, are clearly quite ordinary (DeLoughry, 1995). While computers may be able to analyze writing for the presence or absence of certain words or structures, they cannot understand or appreciate a writer's message in the same sense that human readers can. Thus, two scenarios are possible: Automated scoring methods may be influenced by extraneous features of examinees' writing and award higher than deserved scores, or automated scoring methods may fail to recognize features that are clearly relevant to good writing and therefore award lower than deserved scores.

In light of these possibilities, we believe that automated methods must undergo thorough evaluation by the critics and skeptical onlookers. This kind of scrutiny helps validate such methods because it reveals whether automated scoring is unduly susceptible to influences that are irrelevant to the construct of writing ability. This scrutiny is especially critical for a scoring program such as e-rater, which has since come to play a central role in the GRE General Test essay scoring process.

Procedure

The GRE Analytical Writing measure was designed to measure prospective graduate students' ability to perform various components of good writing (e.g., articulating complex ideas clearly and effectively; examining claims and their accompanying evidence; and controlling the elements of standard, written English). Both types of analytical writing prompts—analyze an issue (issue) task and analyze an argument (argument) task—were used in the study. The prompt for the issue task states an opinion on a topic of general interest and asks writers to address the topic, providing relevant reasons and examples to explain and support their views. The prompt for the argument task presents a written argument and requires test takers not to agree or disagree with the expressed position, but rather to evaluate and discuss its logical soundness. A total of four prompts—two argument and two issue—were used in the study.

A number of writing experts were invited to compose essays in response to the GRE prompts with the intention of tricking e-rater into awarding scores that were either higher or lower than deserved. Participants were asked to write two essays for each of the two prompts they received: one essay they thought would elicit a higher score than deserved and one they thought would receive a lower score than deserved. For each essay submitted, writers were asked to explain the reasons for their predictions. A description of the automated scoring method (i.e., e-rater) was provided to the participants, along with additional information on the GRE Analytical Writing measure. The description included e-rater's general approach, specific techniques used, and particular cue words on which it focused.

All essays were scored both by e-rater and two trained human readers using the same holistic GRE scoring criteria as used operationally. These readers had received training in the scoring criteria, were aware of the general goals of the study, but did not know the specific intentions of the writers. The discrepancy between the e-rater score and the average score assigned by the readers was computed for each essay.

Results

In total, 27 people wrote one or more essays. The writers consisted of undergraduate students, graduate students, professors in linguistics departments, writing assessment specialists from ETS, and a university language center coordinator.

The average scores assigned to the 63 essays by first and second readers, respectively, were 3.22 (standard deviation = 1.45) and 3.26 (standard deviation = 1.54) on the GRE scale of 6 (highest score that can be received) to 1 (lowest score that can be received). Essays that do not respond to the prompt, called *off-topic essays*, were assigned a 0.

The human readers agreed exactly with one another 52% of the time, while e-rater agreed exactly with human readers about 34% of the time. Readers agreed exactly with or within one point of one another 92% of the time, while e-rater agreed exactly with or within one point of readers approximately 65% of the time. The correlation between readers was .82, while the correlations between e-rater and individual first and second readers were .42 and .37, respectively. Thus, e-rater's agreement with human readers was less than the agreement between readers.

Of the total 63 essays submitted, 30 were predicted by the writers to receive a higher score from e-rater than they deserved and 24 were predicted to receive a lower score. (For the remaining nine essays, no prediction was made). In total, 36 of 54 predictions (67%) were in the correct direction. For the 26 essays that were correctly predicted to get a higher score from e-rater than from human readers, three were discrepant by three or more points, five by two to three points, 12 by one to one and one half points, and six by one-half of a point. For the 10 essays that were correctly predicted to receive a lower score from e-rater than from readers, three were discrepant by two points, three by one point, and four by one half of a point. Thus, predictions that e-rater would award a higher score than human readers were borne out in 26 of 30 instances (87%), while predictions that e-rater would award a lower score than human readers were accurate less often (10 of 24, or 42% of the time).

A number of strategies were created by the writers in an effort to trick e-rater into assigning a higher score. The strategies showing the greatest discrepancy between human reader and e-rater scores were (a) repeating the exact same set of paragraphs a number of times and (b) repeating the same paragraph but rewording the first sentence slightly, substituting a few key words, and reordering the subsequent sentences. Other strategies that succeeded in tricking e-rater included using faulty logic, rambling, missing the point, or progressing haphazardly, all the while using relevant content words, complex sentence structure, or other features valued by e-rater.

Conclusion

These findings point to certain aspects of writing that automated scoring engines may reward unduly or fail to notice or appreciate. The results of this study provided direction as to how to improve automated scoring methods such as e-rater. It helped clarify both the promise and the potential pitfalls of one specific system for automated essay scoring. Based on this and other research, e-rater has continued to be improved and enhanced.

Other research led to the development of techniques to detect essays that are unresponsive to essay prompts. Filters were devised for e-rater in order to identify off-topic essays. Some filters detect essays that have very little overlap with the lexical content of the prompt, and others spot writers' tendencies to repeat substantive words. In newer versions of e-rater, additional filters were added. Following a large research and development effort, e-rater was implemented to operationally score GRE essays in 2008.

References

DeLoughry, T. J. (1995, October 20). Duke professor pushes concept of grading essays by computer. *Chronicle of Higher Education,* pp. A24–25.

Notes

[1] Based on *Stumping e-rater: Challenging the Validity of Automated Essay Scoring* (GRE Board Professional Report No. 98-08bP), by D. E. Powers, J. C. Burstein, M. Chodorow, M. E. Fowles, and K. Kukich, 2001, Princeton, NJ: Educational Testing Service.

4.4 Performance of a Generic Approach in Automated Essay Scoring [1]

Yigal Attali, Brent Bridgeman, and Catherine Trapani

Typically, automated scoring systems are trained to score responses to specific essay *prompts* (the essay topics; e.g., Landauer, Laham, & Foltz, 2003). As a consequence, the scores will have different meanings across the prompts because different writing components are emphasized and different scoring rubrics may be established. The *e-rater*® scoring engine is different because its emphasis on form over content allows it to score many topics using the same standards (Attali & Burstein, 2006). This method, called a *generic scoring approach*, is more cost-effective and practical than prompt-specific scoring because e-rater does not need to be retrained for each newly developed prompt.

The current study compared generic and prompt-specific e-rater scores with human scores. The purpose was to investigate whether human scoring standards display enough variation across prompts that a generic scoring approach would be at a disadvantage in comparison to the more finely tuned prompt-specific approach. The study used scores produced for essays written to *GRE*® General Test and *TOEFL iBT*® prompts. Both present difficulties to a generic scoring approach, as the GRE prompts require critical analysis of a topic-specific prompt while the TOEFL iBT examinees do not speak English as their first language.

Method

E-rater's emphasis on form over content is evident based on the overall set of features it uses. Among others, these include measures of essay organization, style, grammar, mechanics/spelling, and vocabulary level. Only two features compare the content of essays to other essays written for the same prompt. They do this by analyzing the specific vocabulary used in comparison to the vocabulary used in essays getting high or low scores in the same prompt. All features were applied to up to 3,000 essays for each of the 113 GRE issue prompts and 139 GRE argument prompts. In addition, 205,566 *TOEFL*® essays across 26 prompts were included in these analyses. This analysis included only one of two types of prompts on the TOEFL iBT: the independent prompt.

Regression analysis was used to predict human scores based on the e-rater features. The generic scoring approach considered only the noncontent features. The prompt-specific approach additionally included the two vocabulary content features. In order to ensure comparable comparisons were made, all scores, both human and e-rater, were scaled such that they had the same standard deviations. Then, 500 essays from each GRE argument and issue prompt and 50% of the essays for TOEFL were placed into a training data set. The remainder of the essays was placed into a validation data set. In order to produce generic scores for a single prompt, a regression analysis was conducted on all essays of the training set, except for those written to that specific prompt. The results of the regression were implemented in the scoring of the validation set for that prompt. For prompt-specific scores, regression analysis was conducted on the training sample of a prompt. The results of this regression were then used to generate a score for the validation set of the prompt. The next section will discuss only the scores from the validation sets.

Results

Correlations were calculated between the scores of a first human rater and (a) a second human rater, (b) the e-rater score from the generic approach, (c) the score from the prompt-specific e-rater approach with content, and (d) the generic e-rater scores without prompt-specific content information. Higher correlations were found between the first human rater and the e-rater scores than between the two human raters (see Table 4.4.1). In addition, the correlation between the first human rater and the generic approach was no different from the correlations between the first human rater and the prompt-specific approaches. Next, the average scores assigned by humans were compared with the average scores assigned by e-rater. Differences were all quite small and were only very slightly higher for the generic approach.

Table 4.4.1

Level of Agreement Across Prompts of First Human Rater Scores
With Second Human Rater and e-rater Scores

Prompt	Second human rater	Generic approach	Prompt-specific approach without content	Prompt-specific approach with content
GRE argument (*N* = 139)				
Correlation	.79	.76	.76	.79
Average score differences	.02	.10	.03	.02
GRE issue (*N* = 113)				
Correlation	.74	.79	.79	.80
Average score differences	.02	.05	.03	.03
TOEFL independent (*N* = 26)				
Correlation	.70	.76	.76	.77
Average score differences	.02	.07	.01	.01

Finally, essay scores from the generic and prompt-specific approaches were correlated with other ability measures (i.e., the TOEFL reading, speaking, and listening scores [for TOEFL essays] and the GRE Verbal Reasoning scores [for GRE essays]).

Correlations were comparable for humans and both generic and prompt-specific e-rater approaches.

Conclusion

A generic approach to automated essay scoring allows a single model to be used across a variety of different prompts, rather than producing a unique model for each prompt. This approach is clearly more efficient and cost-effective, and this study demonstrates that. In terms of predicting human scores, little is lost by substituting generic for prompt-specific models. However, the findings of this study do not corroborate the further contribution of the prompt-specific approach beyond the generic approach. Indeed, they only revealed that the human rater scoring standards were consistent across prompts. As a consequence, prediction of scores at the prompt-specific level provides little to no advantage. Additionally, because the prompt-specific content had very little effect on the scores produced (see Table 4.4.1), it can be concluded that the human scores did not take content into important consideration in these types of writing assessments.

As a consequence of these findings, the GRE revised General Test makes use of the generic approach in the automated scoring of the issue and argument essays. In terms of the maintenance required by such a large-scale assessment, the generic approach helps to keep track of scores across time and raters. It also provides consistent, systematic results that avert the need for costly equating procedures. Thus, the implications of this study have already had a widespread impact on the meaningfulness and validity of automated essay scoring.

References

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V.2. *Journal of Technology, Learning, and Assessment, 4*(3).

Landauer, T. K., Laham, D., & Foltz, P. W. (2003). Automated scoring and annotation of essays with the Intelligent Essay Assessor. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 87–112). Mahwah, NJ: Erlbaum.

Notes

[1] Based on “Performance of a Generic Approach in Automated Essay Scoring,” by Y. Attali, B. Bridgeman, and C. Trapani, 2010, *The Journal of Technology, Learning, and Assessment*, *10*(3), pp. 1–6.

4.5 Evaluation of the *e-rater*® Scoring Engine for the *GRE*® Issue and Argument Prompts[1]

Chaitanya Ramineni, Catherine Trapani, David Williamson, Tim Davey, and Brent Bridgeman

The use of the *e-rater*® scoring engine in conjunction with human raters was implemented in 2008 for the *GRE*® General Test. Prior to e-rater's operational use, various automated scoring models were evaluated to determine their feasibility with both types of GRE Analytical Writing measure prompt types: analyze an issue (issue) and analyze an argument (argument). In particular, the current study investigated if e-rater scores could successfully replace one of the two human raters in operational scoring of the GRE General Test, thereby effectively reducing the program costs and ensuring fast and consistent score turnaround for the large number of GRE test takers.

Procedure

Multiple scoring models are available for e-rater. For purposes of this study, the following model types were evaluated:

- Prompt specific (PS). These are custom-built models for each prompt in the question pool. Their feature weights and intercept are customized for the human score distribution used to calibrate the prompt model. The intercept is used to set the average e-rater score equal to the average human score.
- Generic (G). Generic models are based on taking a group of related prompts, typically 10 or more, and calibrating a regression model across all prompts so that the resultant model is the best fit for predicting human scores for all the prompts, taken as a whole. As such, a common set of feature weights and a single intercept are used for all prompts regardless of the particular prompt in the set.
- Generic with prompt-specific intercept (GPSI). These models are produced by first creating a fully generic model, as described above, then adjusting the model for each prompt so that the intercept of the regression matches the human score average for the particular prompt.

After the automated (e-rater) scores for all essays are calculated, certain guidelines for performance are applied to an independent evaluation sample used to validate the scoring models. These guidelines provided the road map for the evaluation of various e-rater models for the GRE issue and argument prompts. The guidelines are as follows:

- Construct evaluation. First, it is critical to evaluate the fit between the goals and design of the assessment and the design of the automated scoring capability itself.

Among other steps, the construct of interest as reflected in the scoring rubric and the score reporting goals is compared with that represented by the capability.

- Association with human scores. The agreement of automated scores with human scores is typically evaluated on the basis of two statistics, the correlation between two raters, and quadratic weighted kappa,[2] among other criteria. A threshold of .7 (on a scale of 0.0 to 1.0) is recommended and used for both statistics.
- Degradation from human/human scores. The e-rater/human scoring agreement is recommended to not be more than 0.10 *lower*, in correlation and weighted kappa, than the human/human agreement. This standard prevents circumstances in which automated scoring may reach the aforementioned 0.70 threshold but still be notably deficient in comparison with human scoring.
- Standardized average score difference. The standardized average score difference between the human scores and the e-rater scores is recommended to not exceed 0.15. This standard ensures that the distribution of scores from automated scoring is centered on a point close to what is observed with human scoring in order to avoid problems with differential scaling.
- Association with external variables. Due to the imperfections of human scoring, it is important not only to investigate the consistency of automated scores with human scores but also to evaluate the patterns of relationships of automated scores, compared to their human counterparts, with external criteria. Scores on other GRE test sections and external criteria such as self-reported and academically relevant variables (e.g., grades in English class, academic majors) are some examples that are used for this purpose.
- Subgroup differences. In order to evaluate the fairness of automated scoring for subgroups of interest, two approaches are used. The first is extending the flagging criterion of standardized average score differences[3] with a more conservative threshold of 0.10 for all subgroups of interest in order to identify patterns of systematic differences in the distribution of scores between human and automated scoring. The second approach is to examine differences in the ability of automated scoring to predict a human rater score and an external variable of interest by subgroup.
- Operational impact analysis. Determining the predicted impact on the aggregate reported score for the writing section is the final criteria. This impact is evaluated by simulating the score that would result from substituting an automated score for a human score and determining the distribution of changes in reported scores that would result from such a policy.

The PS, G, and GPSI scoring models for issue and argument tasks were built and evaluated on GRE data drawn from test records obtained between September 2006 and September 2007. These data comprised more than 750,000 essay responses written to 113 issue prompts and 139 argument prompts. Along with the two human rater scores for each essay, several additional variables were included for analysis (e.g., examinee demographic variables and GRE Verbal Reasoning and Quantitative Reasoning measure scores). The G and GPSI models included all e-rater features (except for the two content features related to topic-specific vocabulary) in the final model build, while the PS models included the two content features.

Agreement statistics for automated scores with human scores were computed for all e-rater models. The agreement statistics were evaluated using the GRE data, and the model(s) determined to be best was subjected to the remaining evaluation criteria of association with external variables, subgroup differences, operational impact analysis, and agreement thresholds for adjudication.

Results

Based on the results of the analyses in the development stage, advisory flags were turned on during the development stage to filter the essay responses prior to building the e-rater models. These included advisory flags for excessive repetition, irrelevant use, inappropriate length, and excessive number of problems. The combination of these advisories resulted in successfully filtering out 96% of the responses on the issue prompt type that received a human score of 0. For responses on the argument prompt type, 93% of the responses that received a human score of 0 were successfully filtered out using this approach. The use of these rules resulted in only a very small number of cases being flagged and requiring a second human score (about 1% for issue prompts and about 3% for argument prompts).

Traditionally, the essay length limit for e-rater scoring had been set to 800 words. During the evaluation of the individual advisory flags, however, it was observed that a substantial number of essays were identified as too long. About 81% of issue and 92% of argument tasks were found to be in the range of 800 to 1,000 words; of these, 35% of issue and 74% of argument tasks received the highest score category of 6. As a result, the word limit for e-rater was increased to 1,000 words for GRE Analytical Writing measure tasks.

Evaluation of the differences in raw scores under human/e-rater scoring compared to human/human scoring was conducted. When aggregated over all prompts, scores from e-rater were highly similar to human scoring. Scores generated using all three scoring models (from G, PS, and GPSI models) met the correlation and weighted kappa criteria at the overall level for both issue and argument tasks, as well as at the prompt level for the issue task. However, 13 of the argument prompts under the G model and 11 of the argument prompts under the GPSI model failed to meet the recommended threshold.

At the overall level, the degradation threshold was met for both issue and argument tasks. In fact, the e-rater/human agreement reflected an *improvement* in agreement, higher on average by 0.06 for correlation, over human/human agreement for issue prompts. At the prompt level for the argument task, however, 17 prompts under the G model and nine prompts under the GPSI model showed degradation (i.e., they had correlations more than 0.10 *lower* than the human/human agreement). The standardized score differences at the overall level between e-rater and human scores were 0.01 on average for the issue prompts and 0.02 on average for the argument prompts, both well under the acceptable threshold of .15. At the prompt level, however, four prompts for issue tasks and 37 prompts for argument tasks exceeded the threshold under the G model.

Based on the results for the evaluation criteria at the aggregate and the prompt level for the three e-rater model types, GPSI and PS models were chosen as the best scoring models for issue and argument writing prompts, respectively. G models are, however, more preferable for the ease of implementation and maintenance, and therefore were the preferred model type for scoring GRE Analytical Writing measure essays.

Human scores and e-rater scores were correlated with external measures such as scores on other GRE General Test sections, undergraduate grade point average, and English as the best language. E-rater was determined to be appropriate for scoring because correlations between e-rater scores and these criteria were generally higher than for human scores and these criteria.

Analyses were then conducted estimating the degree to which e-rater and human scores differ across subgroups based on gender, ethnicity, and test center country, among others. In general, standardized average score differences of 0.05 or less are desirable for subgroups, and those between 0.05 and 0.10 may be considered acceptable. None of the subgroups revealed any substantial differences except for the country of China (differences as large as 0.60 for issue prompts with e-rater scores being higher than human scores) and African American test takers (difference as large as 0.18 for argument prompts with e-rater scores being lower than human scores). Although the allowable discrepancy threshold between the two human scores on a GRE writing task is one point, it was determined that a smaller threshold of half a point was the optimal level for discrepancy threshold between e-rater and human in operational practice, as it ensured no subgroups were flagged for differences in score.

Furthermore, due to the subgroup differences discovered using this model, it was decided that a more conservative approach that used the e-rater score as a check (or confirmatory) score would be used. Under this model, the e-rater score is used to check or confirm the human score. If the human and e-rater scores do not differ by more than 0.5 points, the human score constitutes the final score for the test taker on a given essay. Thus, test takers' scores are based on only one score that is determined by the human rater. If, however, the human and e-rater scores differ by more than 0.5 points, a second human rating is requested and the two human scores are averaged.

Compared to the previous writing score produced using two or more human ratings, the writing scores using the check score approach showed equal or slightly better association with scores on other GRE General Test sections, undergraduate and major grade point averages, and examinee English ability (see Table 4.5.1). There were no subgroup differences of concern under this model.

Table 4.5.1

Score Association With Other Measures Using Check Score Model for e-rater Scoring Engine

Score	GRE Verbal	GRE Quantitative	Undergraduate grade point average, overall	Undergraduate grade point average, major	English as the best language
New simulated writing score	0.62	0.19	0.22	0.21	0.26
Operational writing score	0.62	0.17	0.20	0.20	0.27

The anticipated number of second human ratings for scores based on all human scoring and scores based on one human with e-rater as check score were compared. Results showed that when using e-rater, roughly 41% cases for issue and 47% cases for argument tasks are expected to still require more than one human score. This finding suggests substantial savings in the operational costs associated with using a second human rater.

Conclusion

As part of ongoing efforts, it will be critical to monitor and evaluate e-rater performance in operation from time to time, owing to the changes in overall test format, test taker and human rater characteristics, and human scoring trends over time, as well as new feature developments and enhancements in the e-rater scoring engine. The introduction of automated scoring with the GRE General Test points to a major advancement in scoring a test on which high-stakes decisions are made.

Notes

[1] Based on *Evaluation of e-rater® for the GRE® Issue and Argument Prompts* (Research Report No. RR-12-02), by C. Ramineni, C. S. Trapani, D. M. Williamson, T. Davey, and B. Bridgeman, 2012, Princeton, NJ: Educational Testing Service.

[2] Quadratic weighted kappa is a statistic that measures the agreement between two raters. Weighted kappas generally range from 0 (random agreement between raters) to 1 (complete agreement between raters).

[3] Standardized average score differences are calculated by taking the differences in the average scores for subgroups and dividing the differences by the standard deviation.

4.6 *E-rater®* Performance on *GRE®* Essay Variants [1]

Yigal Attali, Brent Bridgeman, and Catherine Trapani

The purpose of this research was to conduct a preliminary evaluation of the performance of the *e-rater®* scoring engine on the new *GRE®* variant prompts. A variant prompt is created from a *parent* prompt (either analyze an argument [argument] or analyze an issue [issue]) and asks a focused question that addresses a specific aspect of the prompt. For example, one test taker may be shown a prompt that presents an argument and a recommended course of action and be asked to

> Write a response in which you discuss what questions would need to be addressed to decide whether the recommendation is likely to have the predicted result. Be sure to explain how the answers to the questions w*ould help to evaluate the recommendation.*

Another test taker would see the same prompt, but be asked to

> Write a response in which you examine the unstated assumptions of the argument above. Be sure to explain how the argument depends on those assumptions and what the implications are if the assumptions prove unwarranted.

These focused questions are intended to discourage test takers from writing just very general memorized responses.

Method

Test takers who took the operational test at computer testing centers in the winter [2] of 2009 were invited to participate in a research project immediately following completion of the regular GRE General Test. After screening out very short responses from test takers who did not appear to be making a sincere effort, 17,356 usable responses were obtained. Because more prompt/variant combinations were available for issue essays, sample sizes were somewhat larger for issue than for argument prompts/variants (6,708 argument and 10,648 issue; a list of variant examples can be found in the appendix). Participants were randomly assigned to one of the prompt/variant combinations. Essays were evaluated on 6-point rating scales in an online scoring environment by two independent raters who had been trained on the scoring rubrics for the new variant types.

For each task type, a generic e-rater scoring model was built based on all available essays. In each model, the e-rater scores were scaled based on the average of the mean (average) and standard deviations of H1 and H2 (first and second human scores).

Results

Table 4.6.1 presents agreement results for human and e-rater scores. The average difference between human raters and e-rater for each task type is 0 because each of the two generic models was trained on the entire sample of essays for the task type. The discrepancies between humans and e-rater in standard deviation units (*d*) by variant are generally very small (less than .07 except for Recommendation/Result [*Rec/Result*]). Note that the average human scores for the *Rec/Result* variant are the highest among all variants.

For argument, the correlations between two human scores (HH) and between H1 and e-rater scores (HE) are similar overall, with slight variations across variants. For issue, the HE correlation is .05 higher than HH, again with slight variations across variants. For argument, the HE and HH correlations are identical.

Table 4.6.1

Agreement of Human and e-rater Scores

	N	*M*		*SD*		*d*	Correlation	
Variant		H	E	H	E		H/H	H/E
Argument								
Alternative explanations	1,153	2.88	2.90	.87	.87	.02	.71	.74
Unstated assumptions	787	2.98	3.05	.96	.89	.07	.80	.74
Specific evidence	1,473	2.88	2.88	.89	.93	.00	.71	.74
Evaluate a prediction	562	2.91	2.92	.89	.96	.01	.72	.73
Evaluate a recommendation/predicted result	555	3.10	2.95	.88	.82	-.17	.70	.76
Evaluate a recommendation	2,178	2.99	2.99	.90	.90	.00	.75	.73
All	6,708	2.95	2.95	.90	.90	.00	.74	.74
Issue								
Claim with reason	1,463	2.84	2.80	.94	.91	-.04	.75	.81
Two competing positions	2,087	2.91	2.97	.89	.91	.06	.75	.82
Generalization	1,234	3.00	2.93	.97	.93	-.06	.79	.81
Recommended policy position	2,244	2.99	3.03	.92	.92	.05	.76	.84
Position with counterarguments	1,430	3.06	3.02	.91	.95	-.05	.77	.83
Recommendation	2,190	3.00	2.99	.96	.96	-.01	.78	.83
All	10,648	2.97	2.97	.93	.93	.00	.77	.82

Note. H = Human; E = e-rater.

Conclusion

E-rater performance on the new GRE prompts was evaluated on the basis of discrepancies between human and machine scores and agreement (correlations) between human and machine scores. Discrepancies and agreement were evaluated across task types (issue and argument), variant types, and prompts. E-rater performance on the new GRE prompts was comparable to performance on old argument and issue prompts, and no significant differences in performance were found across different variant types. The small discrepancies that were found between human and machine scores across variants and prompts are generally related to the level of human scores. When average human scores are higher, e-rater scores tend to be lower than human scores, and vice versa. This result has the effect of reducing differences across prompts and parents when both human and e-rater scores are used in combination.

The main limitations of this study are the lower stakes of the experimental section and the relatively small sample sizes that preclude fine-grained analyses. Additional research is ongoing now that the new GRE prompts/variants are operational and data on fully motivated samples are available.

Notes

[1] Based on *E-rater® Performance for GRE® Essays*, by Y. Attali, B. Bridgeman, and C. Trapani, 2010, unpublished manuscript, Princeton, NJ: Educational Testing Service.

[2] *Winter* in this context refers to a time period between October and December.

Appendix

Sample Argument Prompts and Variants

I. Alternate Explanations

Write a response in which you:

- discuss one or more alternative explanations that could rival the explanation offered above

  and

- indicate how your explanation(s) can plausibly account for the facts presented in the argument.

II. Evaluate a Recommendation

Write a response in which you discuss what questions would need to be answered in order to decide whether the recommendation and the argument on which it is based are well justified. Be sure to explain how the answers to these questions would help to evaluate the recommendation

III. Evaluate a Recommendation/Predicted Result

Write a response in which you discuss what questions would need to be addressed to decide whether the recommendation is likely to have the predicted result. Be sure to explain how the answers to the questions would help to evaluate the recommendation.

IV. Evaluate a Prediction

Write a response in which you discuss what questions would need to be answered in order to decide whether the prediction and the argument on which it is based are reasonable. Be sure to explain how the answers to these questions would help to evaluate the prediction.

V. Specific Evidence

Write a response in which you

- discuss what specific evidence is needed to evaluate the logical soundness of the argument above

  and

- explain how the evidence would weaken or strengthen the argument.

VI. Unstated Assumptions Sample Question

Write a response in which you examine the unstated assumptions of the argument above. Be sure to explain

- how the argument depends on those assumptions

  and

- what the implications are if the assumptions prove unwarranted.

Sample Issue Prompts and Variants

I. Claim With Reason

Write a response in which you

- discuss the extent to which you agree or disagree with the claim

  and

- explain how the given reason would affect your position on the claim.

II. Generalization

Discuss the extent to which you agree or disagree with the statement above, and explain your reasoning for the position you take. In developing and supporting your position, you should consider ways in which the statement might or might not hold true, and explain how those considerations shape your position.

III. Position With Counterarguments

Write a response in which you

- discuss the extent to which you agree or disagree with the claim

  and

- anticipate and address the most compelling reasons or examples that could be used to challenge your position.

IV. Recommendation

Discuss the extent to which you agree or disagree with the recommendation above and explain your reasoning for the position you take. In developing and supporting your position, describe specific circumstances in which adopting the recommendation would or would not be advantageous and explain how those samples shape your position.

V. Recommended Policy Position

Discuss your views on the policy above and explain your reasoning for the position you take. In developing and supporting your position, you should explain the possible consequences of implementing the policy.

VI. Two Competing Positions

Discuss which view more closely aligns with your own position and explain your reasoning for the position you take. In developing and supporting your position, you should explain what principles you used in choosing between the two views.

4.7 *E-rater®* as a Quality Control on Human Scores [1]

William Monaghan and Brent Bridgeman

Although essays are now quite common in high-stakes testing situations, they do present a practical problem for those in the testing industry—how to efficiently develop, administer, and score tests with essay sections. This paper focuses on the scoring of essays and the quality control role automated essay evaluation systems can play in the process.

Opponents of automated essay evaluation systems claim that computers lack the intrinsic human capacity to determine good writing from bad. However, testing organizations see such capabilities as being a necessity to efficiently score essay tests (Flam, 2004). A suitable compromise would be to have human readers score essays in tandem with an automated essay evaluation system, such as the ETS-developed the *e-rater®* scoring engine. This approach benefits those in the testing industry by creating less reliance on expensive human readings, and it lessens the concerns of critics, as human readers are an integral element in the system.

Why Automated Essay Scoring?

ETS made its mark by standardizing and then automating much of the testing process. This was done out of necessity, as much as for creating systems in which all test takers can demonstrate their proficiency in a common, fair way. When using essays for assessment purposes, however, ETS has found that having a single essay question or prompt and a single reader per essay does not produce reliable scores (Breland, Bridgeman, & Fowles, 1999). The remedy is to have test takers write at least two essays and to have at least two people read and rate each essay. Scoring costs for this method are substantial and include training and logistical support for each of the many raters necessary to complete this job. These costs are passed along to test takers as additions to their registration fees.

That is why ETS has invested in and developed automated essay evaluation capabilities such as e-rater. In the e-rater system, the computer is fed thousands of essays that human raters have scored. The essays range from those deemed to be high-quality responses to ones seen to be less than adequate. To score an essay, the system is set up to look for patterns that are evident in better essays. The system accomplishes this task in seconds. Studies show a high level of agreement between the sores human raters assign to an essay and what e-rater awards (Attali & Burstein, 2005).

Text Versus Context

Even with this high level of agreement and e-rater's apparent efficiency, a number of people still object to the idea of automated essay evaluation. They argue, and rightly so, that such systems can be fooled by clever nonsense or the inclusion of well-constructed sentences that together make no sense at all. This argument assumes that a human reader, who would detect such cases, is not in the scoring model at all. The opposite fear is to have brilliant writing constructed in such a nonconformist manner that the machine assigns a poor score. Again, a reader should be an effective guard against such a situation. Of course, students seeking instruction would have little to gain in using e-rater outside of its intended function.

Another worry is that the automated essay systems might be less valid for use in the scoring of essays written by English-language learners. Will a machine that is trained on the writing of native English speakers work in a situation where the majority of the testing population does not speak English as a first language? Will systems like e-rater have the same kind of validity in such instances?

Bridgeman (2004) said that a possible solution is to use e-rater to check the scores assigned by human raters. By having e-rater run in the background, the score e-rater provides can be compared to the one assigned by a single human rater. If there is no discrepancy, the scores stands. If the scores are discrepant, a second human reader receives the essay to see if a factor such as fatigue affected the score the first rater assigned or if the essay has elements that are unduly influencing the automated system. In this system, the essay score would always be based solely on human raters. The approach allows a testing organization to efficiently streamline the essay evaluation process while still providing valid score reporting.

Testing E-rater as Quality Control

To test his model, Bridgeman (2004) turned to the Analytical Writing measure of the *GRE®* General Test, which has each test taker write two essays: one on an issue prompt and the other on an argument prompt. A single score is reported for the Analytical Writing measure. Each essay received a score from two trained readers using a 6-point holistic scale. In holistic scoring, readers are trained to assign scores on the basis of the overall quality of an essay in response to the assigned task. If the two assigned scores differ by more than 1 point on the scale, the discrepancy is adjudicated by a third GRE reader. Otherwise, the scores from the two readings[2] of an essay are averaged. The final scores are based on the two essay scores that were averaged and then rounded up to the nearest half-point interval (e.g., 3.0, 3.5).

E-rater scoring models were developed for more than 100 prompts of each type (issue and argument). For the issue prompts, the e-rater scores agreed with the scores assigned by a human rater at the same rate that one human agreed with another. For the argument prompts, agreement of e-rater and human raters was slightly lower, but still quite high. The correlation between the scores assigned by two humans was .81, and the correlation of a human score and e-rater score was .76.

To evaluate the effectiveness of using e-rater as an additional score or as a check on the score from one human rater per prompt, Bridgeman (2004) studied 5,950 examinees who had taken the GRE Analytical Writing measure twice. He used the final score based on at least four human ratings (two for each prompt) from one administration as the criterion. This criterion was predicted from scores on a different administration that were based on two human ratings per prompt, one human per prompt, one human with the e-rater check procedure that required a second human rating, or one human plus e-rater. Results are summarized in Table 4.7.1.

Table 4.7.1

Agreement When Criterion Is Analytical Writing Total From a Different Administration

Readers per prompt	Within ½ point	Within 1 point
2 humans	76.6%	94.0%
1 human	72.9%	92.5%
Checked human	75.5%	93.9%
1 human + e-rater	77.7%	94.2%

The highest agreement, even higher than two human readers per prompt, was found when the score assigned from one human reader was combined with the e-rater score. But if test users are uncomfortable with having a score assigned by a machine being part of a person's score, the checked human approach results in agreement rates that are nearly as high.

Conclusion

Automated essay evaluation systems have a very high threshold to meet to gain people's full confidence as a valid scoring approach. This skepticism is healthy, and until these systems reach a level of sophistication to make such concerns unwarranted, educational measurement organizations should be judicious in the use of these systems, especially in assessments that help in making high-stakes decisions, such as those used in admissions. However, automated essay evaluation systems do have value if properly used. One such valid application, as this paper establishes, is as a quality control check on humans rating essay prompts. As a consequence, the GRE revised General Test now utilizes this method for essay scoring in the Analytical Writing measure.

ETS has explored and continues to explore other uses for e-rater as it works to perfect the system. Even this seemingly limited usage of this capability can reap rewards by making essay scoring more efficient and less costly.

References

Attali, Y., & Burstein, J. (2005). *Automated essay scoring with e-rater v.2.0* (Research Report No. RR-04-45). Princeton, NJ: Educational Testing Service.

Breland, H. M., Bridgeman, B., & Fowles, M. E. (1999). *Writing assessment in admission to higher education: Review and framework* (College Board Research Report No. 99-03). New York, NY: College Entrance Examination Board.

Bridgeman, B. (2004, December). *E-rater as a quality control on human scorers*. Presentation in the ETS Research Colloquium Series, Princeton, NJ.

Flam, F. (2004, August 30). An apple for the computer. *The Philadelphia Inquirer*, p. D-01.

Notes

[1] Based on *E-rater as a Quality Control on Human Scores* (R&D Connections No. 2), by W. Monaghan and B. Bridgeman, April 2005. Princeton, NJ: Educational Testing Service.

[2] If a third reader is used, the two closest scores are averaged to form the final score for the essay.

4.8 Comparison of Human and Machine Scoring of Essays: Differences by Gender, Ethnicity, and Country[1]

Brent Bridgeman, Catherine Trapani, and Yigal Attali

On average, essay scores generated by automated scoring machines (such as ETS's *e-rater®* scoring engine) are reasonably similar to those produced by a human scorer. However, it is of interest to investigate the extent to which automated scoring machines give systematically higher or lower scores than human raters to examinees from different counties, language groups, or ethnic minority groups. Previous differences have been observed between the e-rater and human scores for Arabic, Spanish, and Chinese native speakers (i.e., e-rater assigned higher scores than did humans to Chinese speakers, and humans assigned higher scores than did e-rater to Arabic and Spanish speakers [Burstein & Chodorow, 1999]). This study expands on these findings by using a much more recent version of e-rater and by examining human and e-rater scores for U.S. domestic subgroups.

Method

As part of the *GRE®* Analytical Writing measure, test takers are asked to write two essays: one in response to an analyze an issue prompt and one in response to an analyze an argument prompt. For this study, approximately 3,000 essays were randomly sampled from the 113 issue prompts and another 3,000 from the 139 argument prompts.

Results

For the issue prompt, across genders, ethnicities, and gender within ethnicity, the human rater score and the e-rater score always correlated more highly than did the two human rater scores. All scores from the first human rater had at least a .75 correlation with e-rater scores. These findings are striking because they indicate that it is better to use an e-rater score to predict a human score than it would be to use a score from another human. Additionally, although there are some differences in human and e-rater average scores for African American and American Indian males (humans assign slightly higher scores to these groups than e-rater does), the average scores for most gender and ethnic groups on the issue prompt are quite similar whether the score is from e-rater or a human rater.

Analysis of test takers from countries outside of the U.S. with the highest GRE volumes showed that, for most countries, there were only trivial differences between essay scores from humans and e-rater. There was one notable exception**:** e-rater tended to score essays written by test takers from mainland China much higher than did human raters (e-rater average score = 3.74; human rater average score = 3.29). In other Asian countries, the differences were much smaller, even in Taiwan, which shares the same language as mainland China. Thus, it would appear that the difference is cultural rather than linguistic. One possible explanation for this discrepancy is that the coaching schools in mainland China encourage the memorization of large chunks of generic, grammatically correct text. While human raters are trained to assign lower scores to off-topic text, e-rater does not have this level of discernment and, thus, might produce a higher score than is appropriate.

For the argument prompt, the correlations between human scores and e-rater scores were generally very comparable to the correlations between two human scores for any U.S. gender or ethnic group analyzed. An examination of score averages revealed that African American men and women tended to receive slightly higher scores from humans than they did from e-rater (3.3 vs. 3.1, respectively). Analysis of international test taker essays showed that those from mainland China received higher argument prompt scores from e-rater than they did from human raters, though to a lesser degree than was observed for the issue prompt. This supports the hypothesis that chunks of memorized, generic text, undetectable by e-rater, is a primary factor in the discrepancies, as it is likely more difficult to incorporate such text into an argument essay.

Conclusion

Although the differences between average human and e-rater scores for certain subgroups might seem disturbing at first, it is important to remember that their effects are greatly decreased in the current implementation of e-rater scores (i.e., for GRE, e-rater is used only as a check on the score assigned by a human rater). If exact agreement is not achieved between a human score and an e-rater score, the discrepancy is flagged and a second human rater score is obtained. E-rater scores and human rater scores are never averaged together, and all essay scores reported are based only on the judgments of human raters. Using this method with e-rater as *quality control*, it is possible to mediate the incongruities between e-rater and human scores for any given examinee.

In order to address the issue of providing memorized text in an essay response, ETS has also developed similarity detection software for use in GRE essay scoring (Educational Testing Service, 2009). This software enables ETS to identify automatically text that is highly similar to the text provided by a different examinee.

References

Burstein, J., & Chodorow, M. (1999). Automated essay scoring for nonnative English speakers. In M. Broman Olsen (Ed.), *Computer mediated language assessment and evaluation in natural language processing* (pp. 68–75). Morristown, NJ: Association for Computational Linguistics.

Educational Testing Service. (2009). *How the GRE® tests are scored.* Princeton, NJ: Author.

Notes

[1] Based on "Comparison of Human and Machine Scoring of Essays: Differences by Gender, Ethnicity, and Country," by B. Bridgeman, C. Trapani, and Y. Attali, 2012, *Applied Measurement in Education*, *25*, pp. 27–40.

4.9 Understanding Average Score Differences Between *e-rater*® and Humans for Demographic-Based Groups in the *GRE*® General Test[1]

Chaitanya Ramineni, David Williamson, and Vincent Weng

The standard for mimicking human scores has been met successfully overall for most e-rater evaluations; however, notable differences between machine and human scores have been observed in relation to test center country and ethnicity (Attali, 2008; Attali, Bridgeman, & Trapani, 2007; Bridgeman, Trapani, & Attali, 2012; Ramineni, Trapani, Williamson, Davey, & Bridgeman, 2012). Reasons for this may be that e-rater is placing a greater value on certain essay features than would human raters or that the human raters may be evaluating features not captured by e-rater currently. Regardless of the reason, if systematic in nature, the differences may have potential impact on fairness of scores for examinees. This impact is mitigated by the use of e-rater as check-score and the use of appropriate discrepancy limits between e-rater and human scores. However, there remains a need to understand the root causes of these demographically based score differences between e-rater and human scores to enhance the validity of the present automated scoring system.

This study made use of *GRE*® General Test data from 2009 and 2010. It used a combination of empirical methods and qualitative review of GRE essay data to develop and test hypotheses about the root cause of score discrepancies between e-rater and human raters. This study targeted three research questions:

1. Are there characteristics of writing (e-rater features) consistently associated with these demographically based differences?
2. Are these differences an artifact of the modeling procedures for e-rater?
3. Can we gain insight into the consistency and root causes of differences through expert review of discrepant cases?

Method

The data used for this study were collected between July 2009 and February 2010, spanning 101 analyze an issue (issue) prompts and 114 analyze an argument (argument) prompts. A thousand responses were sampled per prompt, and operational human and e-rater v10.1 scores for these responses were analyzed for the study. Since the e-rater models were implemented for both issue and argument writing tasks in 2007, only discrepant cases had double scores available. Therefore, new e-rater models were built to and evaluated against single human scores, and notable average score differences between e-rater and humans were observed under new evaluations for demographic groups (China,[2] Taiwan, African American), similar to previous evaluations.

This study first identified the subgroups of concern (average score differences between e-rater and humans beyond acceptable threshold level) for both issue and argument writing tasks. Next, it analyzed the e-rater feature scores for each of these subgroups and writing tasks to examine the differences in writing styles and characteristics for each group. Then it evaluated e-rater model performance under two different types of regression models, comparing one suggested by previous research compared to the model used currently, at the overall and the subgroup levels. Following that, for each of the subgroups of concern and writing task, a subset of maximally discrepant cases was identified and subjected to rescoring and review by expert human raters to gain deeper understanding and help formulate hypotheses about sources of discrepancies between e-rater and human scores.

Results

All the evaluation criteria and thresholds were sufficiently met at the overall score level. However, the analyses at the subgroup level revealed some standardized average score differences between e-rater and humans for some demographic subgroups of interest. As an example, Table 4.9.1 reports the observed differences for subgroups by ethnicity and test center country for the issue prompt. Other demographic characteristics, such as gender, undergraduate major, and English as best language, were examined as well, but revealed no formal concerns.

Table 4.9.1

E-rater Evaluation Results by Test Center Country and Ethnicity for Issue Prompts

		Human 1		e-rater		
Subgroup	*N*	Average	*SD*	Average	*SD*	Std diff (human minus e-rater)
Test center country						
China	4,005	2.96	0.58	3.41	0.76	0.68
India	7,887	2.99	0.78	3.06	0.87	0.09
Japan	303	3.18	0.76	3.16	0.89	-0.02
Korea	1,008	2.81	0.73	2.84	0.92	0.02
Taiwan	672	2.76	0.72	2.73	0.89	-0.06
Ethnicity						
White	56,058	4.00	0.75	3.97	0.78	-0.04
African American	6,263	3.53	0.77	3.46	0.88	-0.09
Hispanic	5,401	3.72	0.80	3.68	0.88	-0.05
Asian	8,746	3.50	0.88	3.59	0.93	0.09
American Indian	771	3.29	0.94	3.28	1.01	-0.03
Other ethnicity	4,977	3.62	0.93	3.63	0.96	0.01

Note. Std diff = difference between human and e-rater scores in standard deviation units. Differences of 0.10 or less in absolute value are very small.

Data for the following subgroups were sampled from the larger data set for the study: test takers in China for both issue and argument, and test takers in Taiwan, as well as African American test takers in the U.S., for argument. The average e-rater and human scores for the China subgroup on issue prompts were lower, but the average score difference between e-rater and human scores was greater than that for the overall. For the argument prompts, Taiwan received the lowest average e-rater and human scores; the average e-rater score for the African American group was lower than that for the China subgroup, but the average operational human score was slightly greater than that for China. The difference between the human and e-rater scores was positively large for China, implying higher scores by e-rater than by humans, while for the other two subgroups, Taiwan and African American, the difference was negative, implying lower scores by e-rater compared to humans for the essays from these subgroups.

To communicate the aforementioned differences at the e-rater feature level, their visual representations were created. Figure 4.9.1 is an example of one such bar graph, which visually presents the feature scores of the essays for test takers in China. Positive denotes the positive feature measuring collocation and preposition usage; word length is the average word length over the essay response; vocabulary denotes the sophistication of word choice; style, mechanics, usage, and grammar errors are typical measures of language control; development and organization are related to the structure of the text, where organization is a count of the number of discourse elements, and development is measured as the average length of discourse elements.

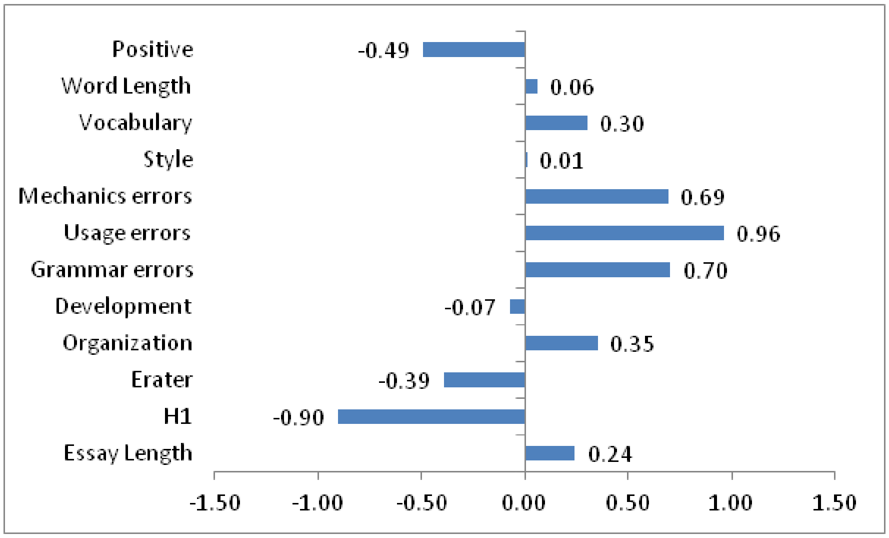


Figure 4.9.1. Plot of standardized feature scores, essay length, and human and e-rater scores for China issue prompt (*N* = 4,005).

On the issue task, the test takers from China wrote longer essays on average than the overall population and received lower human and e-rater scores on average than everyone. At the e-rater feature level, these examinees had more language errors (grammar, usage, mechanics) but obtained a higher organization score, and they scored higher for sophisticated word choice (vocabulary) than the overall population.

To further review the discrepancies between e-rater and human scores, a set of maximally discrepant cases across 10 prompts was identified for each subgroup of concern, and the cases were rescored by expert human raters. Five expert human raters rescored 313 argument responses for the three groups (China, Taiwan, and African American) and 90 issue responses for China.

Conclusion

Several interesting observations were made by raters during the qualitative review of these cases regarding the use of the scoring scale/rubric, the human rating process/procedures, and the e-rater scoring mechanism that help identify some differences between the objective e-rater and the holistic human scoring process guided by an analytic scoring scale. The human raters appear to be using conditional logic and a rule-based approach to their scoring, while e-rater uses linear weighting of all the features. Based on the results of these processes, it appears that e-rater is not severe enough on language errors, overvalues essay length, and occasionally undervalues content. Although e-rater continues to be successfully used in operational implementation under a check score model and, thus, is not detrimental to the reliability and validity of the writing scores, efforts are continuing to investigate and understand the root causes of the demographically based score differences between e-rater and human score.

Next steps include enhancing the e-rater construct coverage to score features such as argumentation and cohesion, exploring new methods to mitigate the undue influence of essay length, and conducting further research into the differences in the human and e-rater scoring process.

References

Attali, Y. (2008). *E-rater evaluation for TOEFL iBT Independent essays.* Unpublished manuscript, Educational Testing Service, Princeton, NJ.

Attali, Y., Bridgeman, B., & Trapani, C. (2007). *E-rater performance for GRE essays.* Unpublished manuscript, Educational Testing Service, Princeton, NJ.

Bridgeman, B., Trapani, C., & Attali, Y. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education, 25,* 27–40.

Ramineni, C., Trapani, C. S., Williamson, D. M., Davey, T., & Bridgeman, B. (2012). *Evaluation of e-rater® for the GRE® issue and argument prompts* (Research Report No. RR-12-02). Princeton, NJ: Educational Testing Service.

Notes

[1] Based on *Understanding Mean Score Differences Between E-rater and Humans for Demographic-Based Groups in the GRE*, by C. Ramineni, D. Williamson, and V. Weng, April 2011, paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

[2] China refers to mainland China and excludes Hong Kong and Taiwan.

Section 5: Validation Evidence

Providing evidence that scores support the claims made by a test is a critical component of test design. Validation of the intended use of *GRE*® General Test scores has been ongoing for decades and continues to be a priority for the GRE program. Chapters in this section provide foundational studies for the predictive validity of the test and long-term success in graduate school. While many of the studies used data from the older version of the GRE General Test, their results are still relevant for and applicable to the revised test because they show the value of a standardized measure of verbal and quantitative reasoning skills.

- Chapter 5.1 examines the validity of GRE scores using an approach that does not rely on the traditional reporting of correlations and regression equations. A correlation of .30 that, when squared, is said to *explain* less than 10% may not seem to be very useful, but focusing instead on the proportion of students who succeeded in graduate school at different levels of test performance tells a different story. Data from 145 graduate departments were analyzed, and within each department, students were divided into quartiles based on the GRE scores and on their graduate grades. In biology departments, for example, only 15% of the students in the bottom GRE quartile of their department were in the top grade quartile, while 43% of the students in top GRE quartile were in the top grade quartile. Other analyses examined the percentage of graduate students who excelled in their first year or two of studies (defined by a grade point average [GPA] of 3.8 or higher) for students at different levels of undergraduate grades and GRE scores.
- Chapter 5.2 provides results from a recent GRE validity study that used data from 10 public universities in the Florida state university system. Analyses examined how well GRE scores predict graduate GPA (for both master's and doctoral programs) beyond what undergraduate GPA can predict. Data were analyzed for seven program areas with a large number of students: (a) education, (b) engineering, (c) English language and literature/letters, (d) biological and biomedical sciences, (e) mathematics and statistics, (f) psychology, and (g) health professions and clinical sciences. Results were presented both as correlations between test scores and grades and as the percentage of students who were clearly struggling (C+ or below grade average) or clearly doing well (3.8 or higher grade average) at different GRE score levels. For example, for master's seekers in engineering departments, 40% of the students in the bottom Analytical Writing quartile in their department were clearly struggling, while only 25% in the top Analytical Writing quartile were struggling.

- Chapter 5.3 reports on a study that was conducted prior to the introduction of the Analytical Writing measure that examined the likely impact of Analytical Writing scores on graduate admission decisions. Graduate faculty members, representing 23 departments, each reviewed simulated admission folders that contained information on the applicant's GRE General Test scores, GRE Analytical Writing measure score, undergraduate GPA, and a simulated personal statement and recommendation from an undergraduate professor. The Analytical Writing score had a statistically significant, but small, impact on admissions decisions. Half of the participating faculty could review the actual essay and not just the score, but this had no significant impact on admissions decisions.
- Chapter 5.4 reports on a meta-analysis conducted by researchers at the University of Minnesota, which combined results from many different predictive validity studies into a single analysis. This study addressed some of the weaknesses of previous studies by looking across several graduate populations and academic areas, correcting for statistical artifacts present in some earlier studies, and testing the validity of multiple predictors and diverse criteria. In addition to predicting graduate GPA, the research showed that the GRE also predicts comprehensive examination scores, faculty ratings, publication citation counts, and, to a lesser extent, degree attainment.
- Chapter 5.5, with the same University of Minnesota researcher as in 5.4, extends those findings by including more recent studies in the meta-analysis and by presenting results separately for master's and doctoral programs. Results were consistent with those in the earlier research, and the GRE was shown to be an effective predictor of graduate outcomes for both master's and doctoral programs.
- Chapter 5.6 discusses a study that examined the use of GRE scores in predicting long-term success in graduate school, including cumulative GPA over 2 or more years and faculty ratings. This study also focused on fairness issues related to how well the GRE predicts outcomes for minorities and women. Women tended to receive slightly higher grades than predicted by the test, while African American and Asian American students performed slightly worse than predicted—but the differences were all very small. Hispanic students tended to do better than predicted in education but worse in English and chemistry; again, all differences were very small, suggesting that the GRE is a fair assessment across genders and ethnic groups.
- Chapter 5.7 examines the impact of disclosing GRE essay prompts to test takers prior to the test administration. Disclosing essay topics can help candidates prepare for the test by allowing them to practice on actual topics, and disclosure provides a

more level playing field for examinees who may not have access to the live topics that can be collected by coaching schools. Participants were sent 27, 54, or 108 essay prompts on which to practice. Although they were practicing for an actual operational test, most participants in the research study spent less than 1 hour practicing with the essays provided, and the number of prompts provided had a negligible effect on their study practices.

- Chapter 5.8 provides an overview of the scientific literature related to the role of noncognitive factors and other background variables on graduate school admissions and success. Noncognitive factors include, among others, creativity, emotional intelligence, self-efficacy, and motivation; background variables include mentoring and social support, prior accomplishments, financial support, ethnicity, race, and gender.

5.1 Understanding What the Numbers Mean: A Straightforward Approach to *GRE*® Predictive Validity[1]

Brent Bridgeman, Nancy Burton, and Frederick Cline

Numerous studies have been used to demonstrate the validity of the *GRE*® General Test for a variety of purposes (e.g., Burton & Wang, 2005; Kuncel, Hezlett, & Ones, 2001). While these studies provide a long tradition of foundational research on the GRE General Test, most of which should generalize to the GRE revised General Test as well, the results of such studies are difficult for lay audiences to interpret. The use of statistical concepts, such as multiple regression methods and variance, creates difficulty because the terminology has little intrinsic meaning outside of the social science realm. Additionally, evidence suggests that even trained social scientists may be severely underestimating the importance of seemingly insignificant correlations (Wainer & Robinson, 2003).

Bridgeman, Pollack, and Burton (2003) addressed these problems by presenting *SAT*® validity results in terms of the proportion of students who succeeded in college at different levels of performance (that is, different undergraduate grade point averages [UGPAs]). They concluded that if the most important outcome is the percentage of students who succeed in college, the substantial relationship between SAT scores and college performance is apparent. This study uses the same approach to examine the validity of GRE scores.

Method

Data were obtained from two sources: (a) departments that participated in the GRE Validity Study Service (VSS) between 1987 and 1991 and (b) a data set for a special validity study of graduate students who entered graduate school between the years 1995 to 1998 (Burton & Wang, 2005). These sources provided data on 3,303 students from 128 departments and 1,148 students from 17 departments, respectively, for a total of 4,451 students from 145 departments. Because of very small sample sizes in many departments, only biology, chemistry, education, English, and psychology departments were included. GRE Verbal Reasoning and Quantitative Reasoning scores, UGPA, and a weighted average of first-year graduate GPA (GGPA) were obtained for each student. Outcome measures included the student's graduate transcript, cumulative GGPA, academic milestones (such as graduation), and faculty ratings of the student's academic and professional skills.

Students in each department were divided into three categories (top quartile, middle half, and bottom quartile) based on their combined GRE scores. They were similarly divided into three categories based on their grades in the first year of graduate school. The within-institution analyses were then combined across all of the institutions in the study for departments of that type. Finally, the percentage of students in each category who obtained 3.8 or higher graduate average within each institution was calculated, and the results were then combined across institutions.

Results

Looking at the three categories of students grouped by GRE scores, it was seen that a large percentage of students who were in the top quartile of GRE scores also had first-year GGPAs that were in the top quartile of their class, and a large percentage of students who were in the bottom quartile of GRE scores had first-year GGPAs that were in the bottom quartile of their class. For example, Figure 5.1.1 shows that among students in the bottom quartile of GRE scores in a biology department, only 15% earned GGPAs in the top quartile. However, nearly three times as many students (43%) in the top quartile of GRE scores ended the year with GGPAs in the top quartile. These patterns were comparable across all departments.

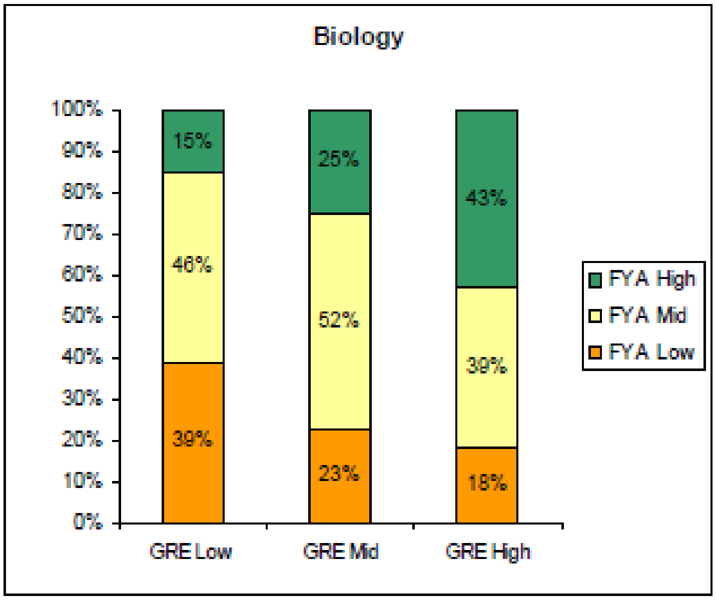


Figure 5.1.1. First-year grade point average (FYA) in the top quartile, middle half, and bottom quartile of a biology class. High and Low refer to the top and bottom quartiles for both GRE scores and FYA.

To further examine the data, a sample of students with GGPAs of 3.8 or higher was selected. It was found that students in the top quartile of GRE scores were much more likely to earn GGPAs of 3.8 or higher compared to students in the bottom quartile. This pattern was consistent across all departments. Similarly, students in the bottom quartile of GRE scores were much more likely to earn less than a B average compared to students in the top quartile of the within-department GRE score category. This pattern was also consistent across departments.

These results show the value of GRE scores in identifying successful and unsuccessful graduate students, but they do not address the question of whether the GRE General Test improves on what is already known from the UGPA. Further analyses show that even among students with comparable UGPAs, those with high GRE scores were more likely to earn high grade point averages (above 3.8 or perfect 4.0).

Students were divided into quartiles based on UGPA and separately into GRE quartiles, so that some students would be low quartile on UGPA but top quartile on GRE, and vice versa (see Figure 5.1.2). If the GRE General Test adds nothing to predicting GGPA, students in all combinations of GRE scores and UGPA (i.e., low UGPA, low GRE; low UGPA, high GRE; high UGPA, low GRE; and high UGPA, high GRE), should be equally likely to excel. In fact, results indicated that GRE scores do appear to matter. For example, in biology departments, not one student in the low UGPA, low GRE category completed the year with a 4.0. However, the rate of students earning a 4.0 jumped to 18% for students in the low UGPA, high GRE category. The impact of GRE scores across departments was a little more varied in this analysis than with the earlier analyses (i.e., in some departments the GRE General Test seemed to make a difference at one end of the scale but not at the other) but was consistent with the interpretation that the GRE General Test provides useful information concerning who will be highly successful even among students with similar undergraduate grades.

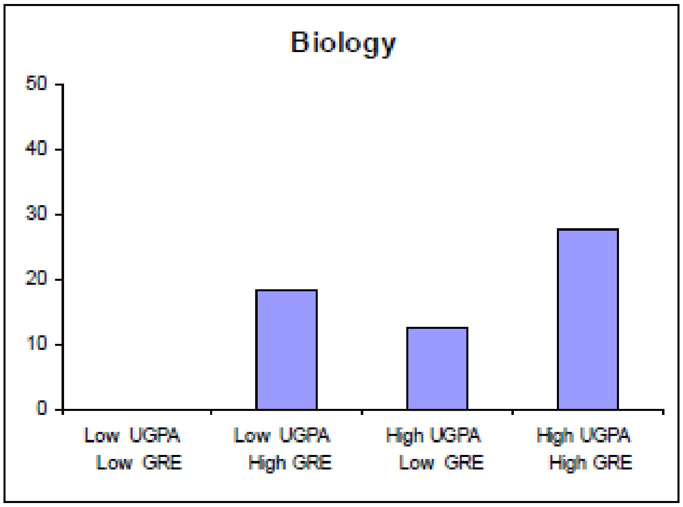


Figure 5.1.2. The vertical axis indicates the percentage of students in biology departments earning a 4.0 graduate GPA by undergraduate GPA (UGPA) and GRE high and low quartiles.

Conclusion

This study provides foundational support for the continued use of GRE scores as part of student selection into graduate schools. Correlations and regression coefficients are difficult to interpret, and seemingly low correlations that are said to *explain* less than 10% of the variance in graduate grades can mask large differences in terms of who is likely to be highly successful in graduate school. It is more helpful to compare academic success rates among students with high and low GRE scores. Although high GGPAs are not the only indicator of a successful student, it is nevertheless a significant academic accomplishment. It is therefore meaningful to observe that this accomplishment is much more likely among students with relatively high GRE scores. Although the focus of this study was directed toward GGPA, future analyses should focus on additional outcome variables, such as graduation rates.

References

Bridgeman, B., Pollack, J., & Burton, N. (2003). *Understanding what SAT I scores add to high school grades: A straightforward approach* (College Board Report No. 2004-4). New York, NY: College Entrance Examination Board.

Burton, N. W., & Wang, M. (2005). *Predicting long-term success in graduate school: A collaborative validity study* (GRE Board Research Report No. 99-14). Princeton, NJ: Educational Testing Service.

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2001). A comprehensive meta-analysis of the predictive validity of the Graduate Record Examinations: Implications for graduate student selection and performance. *Psychological Bulletin*, *127*, 162–181.

Wainer, H., & Robinson, D. H. (2003). Shaping up the practice of null hypothesis significance testing. *Educational Researcher*, *32*, 22–30.

Notes

[1] Based on *Understanding What the Numbers Mean: A Straightforward Approach to GRE Predictive Validity* (GRE Board Research Report No. 04-03), by B. Bridgeman, N. Burton, and F. Cline, 2008, Princeton, NJ: Educational Testing Service.

5.2 New Perspectives on the Validity of the *GRE*® General Test for Predicting Graduate School Grades[1]

David Klieger, Frederick Cline, Steven Holtzman, Jennifer Minsky, and Florian Lorenz

Making valid predictions of who will succeed in graduate school is extremely important. Although some errors in admissions and funding decisions are inevitable, it is crucial to minimize these mistakes by using sound tests. Graduate school attendance is a great emotional and/or financial investment for students and their families, graduate programs and departments, governmental and other funding sources, and taxpayers who subsidize higher education. For example, the annual cost of attending graduate school for master's degree students in 2007–2008 ranged from an average of $28,375 to $38,665, and the cost to educate doctoral students ranged from an average of $32,966 to $46,029 (Wendler et al., 2010).

Although previous research (Burton & Wang, 2005; Kuncel, Hezlett, & Ones, 2001; Kuncel, Wee, Serafin, & Hezlett, 2010; Powers, 2004) has already empirically demonstrated the predictive validity of the *GRE*® General Test, this study extends this research in two ways: (a) it uses different statistical methods to gain a greater perspective on the GRE General Test's validity for predicting overall graduate grade point average (GGPA), and (b) it investigates the utility of the GRE General Test to predict GGPA specifically for a diverse set of universities in a state university system.

The zero-order correlation is a statistic that gauges how accurately a test predicts an outcome, such as how well a score on a GRE measure (Verbal Reasoning, Quantitative Reasoning, or Analytical Writing) can tell us what overall GGPA an applicant would achieve if admitted to and enrolled in graduate school (see, e.g., Burton & Wang, 2005; Kuncel et al., 2001, 2010; Powers, 2004). Generally, a zero-order correlation with a value of 0.1 is considered small; 0.3, moderate; and 0.5 (or higher), large. However, the authors question the usefulness of this set of rules because reliance on them can blind people to the predictive value of a test. Furthermore, the separate GRE measures can be used together (and together with undergraduate grade point average [UGPA]) to make admissions and funding decisions. Also, admissions committees often use the GRE General Test and UGPA together in such a way that a high score on the GRE General Test can offset a low UGPA and vice versa (Powers, 2004; Walpole, Burton, Kanyi, & Jackenthal, 2002). Therefore, the authors look beyond the zero-order correlation to ask how well the GRE General Test predicts GGPA beyond what UGPA can predict.

Procedure

The Florida state university system (FUS)[2] provided the information used in this study. The FUS data are from 10 public universities within the same state, cover the academic years 2003–2004 through 2007–2008, and include information for any student who had either applied to or who was enrolled in any of those universities at any point during those years. The authors analyzed data for 21,127 students seeking master's degrees and 4,229 students seeking doctoral degrees (25,356 students in total).

Students were grouped together based on whether they were seeking a master's degree or doctorate and then based on their area of study. Analyses focused on seven program areas with a large number of students: (a) education, (b) engineering, (c) English language and literature/letters, (d) biological and biomedical sciences, (e) mathematics and statistics, (f) psychology, and (g) health professions and clinical sciences. The FUS provided GGPA for only those applicants who had actually been accepted and then enrolled, rather than for the entire applicant pool.

In addition to calculating zero-order correlations, we computed other statistics for measuring predictive validity. These include commonly used statistics of how much the GRE measures predict GGPA when used together with UGPA (*multiple correlations*) and how much the GRE predicts GGPA beyond what UGPA predicts (*incremental multiple correlations*). The authors also utilized *usefulness weights* (Budescu, 1993) and GRE *quartile comparisons* (Bridgeman, Burton, & Cline, 2009). Usefulness weights tell us how much of what the GRE General Test and UGPA together predict about GGPA is attributable to each particular measure of the GRE General Test and particularly to UGPA. The GRE quartile comparisons can tell us (a) how many times more likely an enrolled student was to achieve a GGPA of at least 3.8 (at least an A/A- average) if the student scored in the top 25% of the GRE General Test as opposed to the bottom 25%, and (b) how many times more likely an enrolled student was to have earned a graduate school grade of C+ or lower in at least one graduate school class if the student scored in the bottom 25% of the GRE General Test, as opposed to the top 25%.

Results

Analyses Based on Zero-Order Correlations

The first set of analyses focused on how well GRE scores predict GGPA based on zero-order correlations (i.e., the simple unadjusted correlation between the test and graduate grades). For each GRE measure at the master's level, virtually all of the GRE zero-order correlations are what Cohen (1988) classified as only small to medium in size (i.e., in the range of 0.1 to 0.3). Adjusted zero-order validity coefficients for doctoral programs are somewhat larger than those for master's level programs but, based on Cohen's rules of thumb, would still generally be considered small to moderate in size.

However, these traditional rules blind us to the predictive value of measures of the GRE General Test. For example, imagine a hypothetical situation in which an organization's current students (or employees) are randomly selected and half of them are satisfactory performers in school (or on the job). If the organization then decides to select the top 30% of new applicants after introducing a selection system with a validity coefficient of only 0.15, then 57% of those selected applicants will be successful students (or employees; Taylor & Russell, 1939). That 7% improvement in the success rate (i.e., 57% to 50%) seems especially beneficial when one considers that (a) over the course of even a small graduate program's history, it is admitting a large number of students, and (b) the costs of educating and training each graduate student in terms of time, money, and emotion can be very large.

Analyses Based on Measures Other Than Zero-Order Correlations

The remaining sets of analyses focused on how well GRE scores predict GGPA based on measures other than zero-order correlations. Multiple correlations revealed that use of the measures of the GRE General Test together in admissions and funding decisions, rather than alone, leads to larger correlations (i.e., often 0.3 and higher).

At both the master's and doctoral degree levels, incremental multiple correlations show that all GRE measures, together and individually, predict GGPA above and beyond what UGPA predicts, both overall and across reported program areas. All of these incremental multiple correlations are larger than 0.05 and, therefore, can provide substantial predictive value. Moreover, on average, across master's level program areas, GRE measures uniquely account for more than 40% of what the GRE General Test and UGPA collectively predict for GGPA when GRE measures and UGPA are used together to make an admissions or funding decision. On average, across doctoral level program areas, GRE measures account for more than half of what the GRE General Test and UGPA collectively predict for GGPA when GRE measures and UGPA are used together to make an admissions or funding decision (Table 5.2.1).

Except for doctoral programs in (a) English language and literature/letters and (b) mathematics and statistics, enrolled students who received scores in the lowest GRE quartile on any of the GRE measures were more likely than enrollees who received scores in the highest GRE quartile to achieve a grade of C+ or lower (see Table 5.2.2). As shown for all GRE measures, a higher percentage of students who had scored in the top GRE quartile achieved a GGPA of at least 3.8 than in the bottom quartile.

Table 5.2.1

Correlations Between Graduate Grade Point Average and GRE Sections Overall and for Seven Program Areas: Master's- and Doctorate-Seeking Students

	Program areas							
GRE measure	Overall	Education	Engineering	English lang. & lit./letters	Biological & biomed. sci.	Math. & stats.	Psychology	Health prof. & clin. sci.
Master's seekers								
Verbal	0.16	0.13	0.12	0.30	0.18	0.05	0.08	0.20
Quantitative	0.14	0.10	0.14	0.15	0.17	0.25	-0.02	0.18
Analytical Writing	0.19	0.15	0.15	0.33	0.19	0.08	0.31	0.18
Doctorate seekers								
Verbal	0.16	0.21	0.04	0.03	0.27	0.19	0.24	0.12
Quantitative	0.20	0.29	0.21	0.07	0.20	0.40	0.32	0.22
Analytical Writing	0.17	0.17	0.07	0.01	0.27	0.15	0.28	0.20

Table 5.2.2

GRE Quartile Comparisons: Master's Seekers

GRE measure		Program areas							
		Overall	Education	Engineering	English lang. & lit./letters	Biological & biomed. sci.	Math. & stats.	Psychology	Health prof. & clin. sci.
K (total # of universities contributing data)		10	10	7	9	9	8	9	10
N (total # of students contributing data)		21,127	4,649	1,481	552	445	230	461	3,772
		Probability of grade of C+ or lower							
Verbal	Low quartile	25%	17%	33%	12%	22%	45%	12%	30%
	High quartile	16%	12%	30%	6%	13%	31%	6%	17%
Quantitative	Low quartile	25%	17%	36%	10%	30%	49%	11%	28%
	High quartile	17%	14%	25%	4%	8%	31%	6%	19%
Analytical Writing	Low quartile	26%	18%	40%	11%	29%	47%	17%	30%
	High quartile	16%	11%	25%	4%	20%	38%	5%	18%
		Probability of cumulative GGPA ≥ 3.8							
Verbal	Low quartile	36%	54%	27%	49%	31%	14%	48%	33%
	High quartile	58%	76%	37%	74%	57%	21%	62%	57%
Quantitative	Low quartile	38%	55%	25%	55%	28%	7%	54%	36%
	High quartile	55%	71%	37%	77%	55%	29%	56%	55%
Analytical Writing	Low quartile	36%	55%	23%	42%	25%	18%	45%	36%
	High quartile	58%	76%	39%	75%	48%	23%	59%	54%

Note. GGPA = graduate grade point average.

Conclusion

Although past research establishing and confirming the generalizable validity of the GRE General Test is considerable (Burton & Wang, 2005; Kuncel et al., 2001, 2010; Powers, 2004), this study questions the traditional use of Cohen's (1988) rules to decide whether a zero-order correlation indicates that a test is useful. Admittedly, it is convenient to have an easy set of rules to determine the importance of a zero-order correlation, especially when the correlation lacks any larger context. However, the larger context does matter. Predictive validity information drives behaviors with major consequences. The authors contend that seemingly small correlations are still meaningful and significant in the sense that they should encourage behaviors that add value (e.g., persuade graduate programs to require and use the GRE General Test for making admissions and funding decisions).

References

Bridgeman, B., Burton, N., & Cline, F. (2009). A note on presenting what predictive validity numbers mean. *Applied Measurement in Education, 22,* 109–119.

Budescu, D. V. (1993). Dominance analysis: A new approach to the problem of relative importance of predictors in multiple regression. *Psychological Bulletin, 114,* 542–551.

Burton, N. W., & Wang, M. M. (2005). *Predicting long-term success in graduate school: A collaborative study* (GRE Board Report No. 99-14R). Princeton, NJ: Educational Testing Service.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2001). A comprehensive meta-analysis of the predictive validity of the Graduate Record Examinations: Implications for graduate student selection and performance. *Psychological Bulletin, 127,* 162–181.

Kuncel, N. R., Wee, S., Serafin, L., & Hezlett, S. A. (2010). The validity of the Graduate Record Examinations for master's and doctoral programs: A meta-analytic investigation. *Educational and Psychological Measurement, 70*, 340–352.

Powers, D. E. (2004). Validity of Graduate Record Examinations (GRE) General Test scores for admissions to colleges of veterinary medicine. *Journal of Applied Psychology, 89,* 208–219.

Taylor, H. C., & Russell, J. T. (1939). The relationship of validity coefficients to the practical validity of tests in selection. *Journal of Applied Psychology, 23*, 565–578.

Walpole, M., Burton, N. W., Kanyi, K., & Jackenthal, A. (2002). *Selecting successful graduate students: In- depth interviews with GRE users* (Research Report No. RR-02-08). Princeton, NJ: Educational Testing Service.

Wendler, C., Bridgeman, B., Cline, F., Millett, C., Rock, J., Bell, N., & McAllister, P. (2010). *The path forward: The future of graduate education in the United States*. Princeton, NJ: Educational Testing Service.

Notes

[1] Based on *New Perspectives on the Validity of the GRE® General Test for Predicting Graduate School Grades*, by D. Klieger, F. Cline, S. L. Holtzman, J. Minsky, and F. Lorenz, 2013, unpublished manuscript, Princeton, NJ: Educational Testing Service.

[2] Data analysis conducted by ETS staff. Raw data file provided by the Florida Board of Governors, State University System Office.

5.3 Likely Impact of the *GRE®* Writing Measure on Graduate Admission Decisions [1]

Donald Powers and Mary Fowles

The *GRE®* Board has a long-standing desire to better understand how its test offerings facilitate and influence graduate admissions decisions. The overarching aim of this study was to ascertain the likely role of the GRE Writing Assessment [2] in graduate admissions decisions. A secondary goal was to assess the influence of more traditional admissions criteria on graduate school admissions decisions.

The GRE Writing Assessment, which was introduced in the 1999–2000 testing year, requires GRE test takers to write two essays: One involves discussing an issue and the other involves analyzing an argument. On the basis of these two writing samples, a composite score is reported that reflects each writer's ability to analyze and discuss complex ideas in a clear, well-focused, coherent, and effective manner. During the development of the GRE Writing Assessment, the decision was made to not send test takers' actual essays along with their scores. This decision was based on the concern that, without proper training, graduate admissions staff may misuse or misinterpret the actual essays by, for example, focusing on extraneous features that are not integral to the construct of writing ability as defined by the GRE program. Because the prevalence of irrelevant writing features may vary by test-taker gender, ethnicity, or cultural background, sending the essays along with the scores might create the potential for unfairness to certain groups of test takers. However, because many experts in the field of writing assessment endorsed sending the essays along with the scores, the GRE Board left open the possibility of revising this policy. Therefore, a specific objective of the current study was to determine the probable effects of sending actual examinee essays to graduate institutions along with test scores and whether the presence of construct-irrelevant flaws in these essays might negatively influence admissions decisions.

Method

Twenty-three graduate faculty members, representing nine graduate psychology and 14 graduate history departments, each reviewed 27 simulated admission folders for a set of fictitious applicants. Each folder contained an application for graduate admission, including information on the applicant's GRE General Test scores, GRE Writing Assessment score, undergraduate GPA, and a simulated personal statement and recommendation from an undergraduate professor.

Half of the participating faculty in each department received the scores from the GRE Writing Assessment; the other half received scores and the actual essays on which the scores were based. To test whether extraneous flaws negatively influenced graduate faculty perceptions of applicants' writing skills, various types and numbers of construct-irrelevant flaws, judged by GRE Writing Assessment specialists to have no effect on the overall quality of any essay, were systematically introduced into half of these essays.

After becoming familiar with the GRE Writing Assessment (including information about scoring criteria and examples of essays at each score level), faculty participants were asked to judge the admissibility of each simulated applicant by indicating (a) the faculty member's own recommendation for admission (deny or admit) and (b) the faculty member's estimate of the likelihood that their program or department would admit the applicant. In making their judgments, participants were asked to review the applications twice, first considering only information about the applicants' writing skill as reflected by the GRE Writing Assessment scores for the score-only condition or by GRE Writing Assessment scores and essays for the scores-plus-essays condition. Then, on a second review, they were to consider all the available information for each applicant.

Faculty participants then were asked to indicate, on a 5-point scale ranging from 0 (*not considered at all*) to 5 (*extremely important*), the importance of each of the following factors in their admission decisions: GRE General Test scores, undergraduate grades, recommendations, and several traits listed on the simulated students' recommendation forms and personal statements.

Each admission recommendation (and each likelihood estimate) was treated as an independent observation so that 27 applicants reviewed by 23 faculty members resulted in 621 observations for admit/deny recommendations.

Hierarchical regression analyses were conducted on the observations to determine the contribution of the following factors in the faculty's decision making:

- GRE Writing Assessment scores
- GRE Writing Assessment essays in the admission folder
- The prevalence of construct-irrelevant errors in the essays
- A set of traditional preadmission measures (e.g., GRE General Test scores, undergraduate GPA in major and overall, rating of personal statement and recommendation, and selectivity of undergraduate school)

Results

Scores from the GRE Analytical Writing Assessment accounted for a statistically significant but small portion of the variation in faculty decisions above and beyond that explained by traditional preadmissions measures. The GRE writing scores explained 3%–4% of the variance in decisions made by history graduate faculty and 6%–10% of the variance in admissions decisions made by graduate psychology faculty. Thus, it is likely that GRE writing scores will play some role in graduate admissions decisions.

The availability of applicants' GRE essays had little additional influence on admissions decisions beyond that of the writing scores themselves. The presence of essays in folders was a significant influence in only one of the analyses, accounting for 2% of the variance in the estimates of likelihood for admission provided by psychology faculty and translating, on average, to a 7% *lower* likelihood of admission.

The presence of incidental writing errors in the test essays that were sent to admission departments had little effect on participants' judgments. There was little evidence that participants' judgments of the essays were influenced by relatively trivial, construct-irrelevant errors of the kind that GRE essay readers are trained to downplay when rating the overall quality of the essays, such as spelling or typing errors and minor grammatical errors or careless misstatements of fact.

When faculty participants based their judgments on all the information in the admission folders, the prevalence of errors in essays accounted for a small, statistically significant ($p < .05$) proportion of the variation in admit/deny recommendations but none of the variation in likelihood estimates. The effects are such that the presence of extra errors decreases the likelihood of admission somewhat, regardless of the quantity of added errors.

Faculty decisions were strongly related to applicants' standing on traditional preadmission measures (e.g., GRE General Test scores, undergraduate grades, faculty recommendations, and personal statements). In terms of the influence of each admissions criterion on faculty judgments about applicants, considerable variation occurred among participating faculty—often within the same department—with respect to the importance they attach to various kinds of preadmission information. However, the data indicate that GRE Verbal Reasoning scores appear to be an important factor in admissions decisions in both history and psychology departments. Lack of close correspondence between faculty-generated ratings of the importance of various admission criteria and the results of statistical analyses of their relative weights suggest that, to some degree, faculty perceptions of importance do not fully reflect the actual weights that the factors receive.

Conclusion

The results of this study contribute to the existing body of knowledge concerning the use of GRE test scores in graduate admissions. More broadly, the results shed light on the issue of whether the products or performances generated by test takers may contain useful information that is not captured solely in summary evaluations (i.e., test scores). Our findings indicated that, for one large-scale test, the GRE Writing Assessment, making applicants' responses available to admissions staff will probably not give rise to inappropriate judgments and misuse, at least with respect to the likelihood that faculty will focus on relatively irrelevant features of the applicants' writing.

Notes

[1] Based on *Likely Impact of the GRE® Writing Assessment on Graduate Admission Decisions* (GRE Board Research Report No. 97-06R), by D. E. Powers and M. E. Fowles, 2000, Princeton, NJ: Educational Testing Service.

[2] The GRE Writing Assessment became the Analytical Writing measure on the GRE General Test in 2002.

5.4 A Comprehensive Meta-Analysis of the Predictive Validity of the *GRE*®: Implications for Graduate Student Selection and Performance[1]

Nathan Kuncel, Sarah Hezlett, and Deniz Ones

The *GRE*® General Test has been a heavily weighted consideration in graduate school admission decisions in many departments for many decades. However, studies of the predictive validity of the test that have been done over the years have had widely varying results: Some found that the tests only weakly predicted graduate school success (e.g., Marston, 1971; Sternberg & Williams, 1997), while others found the tests to be strongly correlated with performance in graduate school (e.g., Broadus & Elmore, 1983; Sleeper, 1961). A meta-analysis is a way of combining results from many different studies into a single analysis. This allows for controlling many of the sources of error and uncertainty that can affect the results of individual studies.

This meta-analysis was designed to address the methodological weaknesses of previous studies in three major ways. First, in contrast to earlier studies that focused on a single population, academic discipline, or performance measure, this study looked at the predictive validity of the GRE General Test across multiple populations and academic areas using several criteria of graduate school success. Second, this study corrected for range restriction and criterion unreliability that existed in many of the studies used in meta-analysis; these statistical artifacts can reduce the magnitude of the correlation between the tests and graduate performance measures. Third, this study tested the validity of *multiple* predictors—scores from the GRE Verbal Reasoning and GRE Quantitative Reasoning measures, and undergraduate grade point average (UGPA), individually and in combination—and thus provides more accurate estimates of the validity of the GRE General Test than previous studies.

Method

Database of Articles Reviewed in the Meta-Analysis

The database consisted of articles and dissertations identified through PsychLIT (years 1887–1999), ERIC (years 1966–1999), and Dissertation Abstracts International (years 1861–1998), as well as all research reports published by ETS. The citation lists within these articles, dissertations, and reports were also examined for appropriate studies. The final database for the study included 1,753 independent samples and 6,589 correlations (i.e., relationships among eight criteria and five predictors) across 82,659 graduate students. Predictive validity studies of the following academic disciplines were included: humanities, social sciences, life sciences, and math-physical sciences. Because the data came from already published sources, only the pre-2011 version of the GRE General Test was analyzed. Results for the GRE revised General Test could differ slightly from these results, but because the foundational elements of verbal and quantitative reasoning skills remain largely the same, this study should still provide a good approximation of what would be found for the test.

Predictor Variables

Scores from two GRE General Test measures—Verbal Reasoning and Quantitative Reasoning—as well as scores from GRE Subject Tests (Subject Tests), and UGPA were used as predictor variables.

Graduate School Performance Measures

The criteria used to define graduate school success were first year graduate GPA (first-year GGPA), graduate GPA (GGPA), faculty ratings (ratings of students' research ability, professional work, potential, overall performance), comprehensive examination scores, time to degree, degree attainment, citation counts, and research productivity.

Results

Table 5.4.1 presents the findings of the predictive validity of the GRE General Test, GRE Subject Tests, and UGPA for the various criteria of graduate school success in the total sample of studies.

Table 5.4.1

Operational Validities of GRE Measures and Undergraduate Grade Point Average for the Total Sample

	Predictor			
Performance measure	Verbal	Quantitative	Subject test	UGPA
Graduate grade point average	.34	.32	.41	.30
First-year graduate grade point average	.34	.38	.45	.33
Comprehensive exam scores	.44	.26	.51	.12
Faculty ratings	.42	.47	.50	.35
Degree attainment	.18	.20	.39	.12
Time to complete	.28	-.12	.02	-.08
Research productivity	.09	.11	.21	na [a]
Publication citation count	.17	.23	.24	na [a]

Note. UGPA = undergraduate grade point average.

[a] *na* indicates that data were not available for these analyses.

Based on the moderately large predictor coefficients, the authors concluded that the Verbal Reasoning measure, Quantitative Reasoning measure, and Subject Tests were all valid predictors of GGPA; first-year GGPA; comprehensive examination scores; faculty ratings; citation counts; and, to a lesser extent, degree attainment. They also noted that while Verbal Reasoning, Quantitative Reasoning, and UGPA similarly predicted GGPA, first-year GGPA, comprehensive examination scores, and faculty ratings, UGPA did not predict degree attainment nearly as well as scores from the GRE General Test.

The GRE Subject Tests proved to be the strongest predictor of graduate school success for all success criteria (except time to completion). The authors suggest that the superior predictive power of the GRE Subject Tests, as compared to the scores from the GRE General Test, could be due to measuring interest in and motivation to master the field, factors that contributed to success in graduate school.

To test whether the predictive validity of the GRE General Test and the UGPA differed for different academic disciplines, separate meta-analyses were run on studies whose samples were drawn from the humanities, social sciences, life sciences, and math-physical sciences. The results of the predictive validities of Verbal Reasoning measure, Quantitative Reasoning measure, and UGPA for GGPA, first-year GGPA, and faculty ratings paralleled those of the overall sample, as did the finding of the stronger predictive validity of the GRE Subject Tests. (The authors note that findings from the subarea analyses were based on small sample sizes and thus should be interpreted with caution.)

Two separate meta-analyses were conducted on studies involving special populations: one on nonnative English speakers and the other on nontraditional students (defined as students older than 30 years). Both studies found the Verbal Reasoning and Quantitative Reasoning scores to be predictive of the GGPA and first-year GGPA for these populations.

A third meta-analysis on the effects of grade inflation (as measured by year of the study because of the belief that grade inflation has increased over time) on the predictive power of the GRE General Test found no relationship between year of study and observed validity.

Conclusion

Using multiple criteria and a wide range of academic disciplines, this study found significant correlations between GRE General Test scores and important criteria of graduate school success. Separate analyses of studies involving different discipline areas and student populations yielded results similar to those of the overall sample. The findings from this meta-analysis confirm that the GRE General Test is a valid predictor of success in graduate school. The lower validity found in some previous studies may have been due to statistical and methodological errors that were corrected in the current study.

References

Broadus, R. N., & Elmore, K. E. (1983). The comparative validities of undergraduate grade point average and of part scores on the Graduate Record Examinations in the prediction of two criterion measures in a graduate library school program. *Educational and Psychological Measurement, 43,* 543–546.

Marston, A. R. (1971). It is time to reconsider the Graduate Record Examinations. *American Psychologist, 26,* 653–655.

Sleeper, M. L. (1961). Relationship of score on the Graduate Record Examinations to grade point averages of graduate students in occupational therapy. *Educational and Psychological Measurement, 21,* 1039–1040.

Sternberg, R. J., & Williams, W. M. (1997). Does the Graduate Record Examinations predict meaningful success in the graduate training of psychologists? *American Psychologist, 52,* 630–641.

Notes

[1] Based on “A Comprehensive Meta-Analysis of the Predictive Validity of the Graduate Record Examinations: Implications for Graduate Student Selection and Performance,” by N. R. Kuncel, S. A. Hezlett, and D. S. Ones, 2001, *Psychological Bulletin*, *127*(1), pp. 162–181.

5.5 The Validity of the *GRE*® for Master's and Doctoral Programs: A Meta-Analytic Investigation [1]

Nathan Kuncel, Serena Wee, Lauren Serafin, and Sarah Hezlett

Although numerous studies over the past decade have confirmed the power of *GRE®* scores to predict various measures of graduate school performance, few studies have looked directly at whether the predictive power of the GRE General Test is different for master's and doctoral programs. If such differences were found, they would have important implications for how GRE General Test scores can be used in admission decisions for programs at each degree level and whether other application criteria should be given more weight. The purpose of this study was to conduct a meta-analytic investigation of the differential power of the GRE General Test to predict the performance of students enrolled in master's and doctoral programs. Although only studies using the previous version of the GRE General Test were included, because the foundational elements of verbal and quantitative reasoning skills remain largely the same, this study should still provide a good approximation of what would be found for the GRE revised General Test.

The study tested three hypotheses:

1. The GRE General Test will be a valid predictor of student performance in both master's and doctoral degree programs.
2. The predictive power of the GRE General Test is likely to differ by degree level, and the test may more strongly predict the performance of students in doctoral programs than master's programs. This hypothesis is based on findings from previous research that GRE General Test scores are moderated by two factors that differentiate master's and doctoral programs: course complexity (e.g., doctoral courses are typically more complex than master's courses) and the degree to which program activities are structured (e.g., doctoral programs are typically less structured than master's programs).
3. The predictive power of the GRE General Test may vary by degree level if disciplines that are more strongly correlated with GRE Verbal Reasoning scores (e.g., humanities) or GRE Quantitative Reasoning scores are more likely to have master's or doctoral programs.

Method

Nearly 100 studies of the predictive validity of the GRE General Test involving close to 10,000 students were included in the meta-analysis. A major source for these studies was the database used in a previous meta-analysis by Kuncel, Hezlett, and Ones (2001). In addition, studies were identified through new searches of ERIC, PsychINFO, and Dissertation Abstracts databases from the years 1999 to 2005. Because few studies exist that examine the validity of the more recently developed GRE Analytical Writing measure, scores on this measure were not included.

Three measures of graduate school performance were used in the meta-analysis: first-year grade point average (GPA), graduate GPA, and faculty ratings.

The study utilized a psychometric meta-analytical technique (see Hunter & Schmidt, 2004) whose properties allow for the reduction of possible biases arising from systematic differences in GRE General Test scores between successful applicants and the pool of all applicants (which includes those denied admission), differences in the variability of GRE General Test scores at different points in time, and possible measurement error arising from inconsistencies of the outcome measures (e.g., grades and faculty ratings). Outcomes for master's students in doctoral programs that require students to earn a master's degree as part of the program were not included in order to more clearly separate the two degree program levels.

Results

Support for the hypothesis that the predictive strength of the scores from the GRE General Test was likely to differ by degree program was mixed. Verbal Reasoning scores more strongly predicted graduate GPA for students in master's programs (.38) than doctoral programs (.27), a difference that was opposite to the hypothesized direction. On the other hand, the predictive validity of the Quantitative Reasoning scores for faculty ratings was stronger for students in doctoral programs (.30) than master's programs (.21), as hypothesized.

The validity of the Verbal Reasoning scores for faculty ratings did not differ by degree level. It is not clear why the two components of the test predicted the three indicators of graduate performance differently. Further studies are needed to identify variables that may influence the power of the GRE General Test to predict academic performance of students in different degree programs.

The predictive validity of Verbal Reasoning and Quantitative Reasoning scores for the three indicators of graduate school success in master's and doctoral programs ranged from .21 to .38. Averaged over the two test components and grade measures, the validity coefficient for the master's level was .30, while that for the doctoral was .27, a difference of only .03.

Conclusion

Overall, the GRE General Test proved to have predictive validity for both master's and doctoral level programs. The evidence of the predictive validity of the GRE General Test at both degree levels suggests that graduate programs can continue to use GRE scores as a tool in admission decision making. However, it is acknowledged that further studies are needed to identify variables that may influence the power of the GRE General Test to predict academic performance of students in different degree programs.

References

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Thousand Oaks, CA: Sage.

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2001). A comprehensive meta-analysis of the predictive validity of the Graduate Record Examinations: Implications for graduate student selection and performance. *Psychological Bulletin*, *127*, 162–181.

Notes

[1] Based on “The Validity of the Graduate Record Examinations for Master’s and Doctoral Programs: A Meta-Analytic Investigation,” by N. R. Kuncel, S. Wee, L. Serafin, and S. A. Hezlett, 2009, *Educational and Psychological Measurement*, *70*(2), pp. 340–352.

5.6 Predicting Long-Term Success in Graduate School: A Collaborative Validity Study[1]

Nancy Burton and Ming-mei Wang

In order for graduate schools to use test scores for selecting potential graduate students, it is critical that the appropriateness of such use be established. In addition to ensuring effectiveness of the admission process, the process itself needs to be fair for all prospective students, particularly for groups that are relatively new to graduate education or those that have been traditionally underrepresented in graduate school. While validity studies on the *GRE®* revised General Test are currently being conducted, the long tradition of research on the previous version of the test creates a foundation on which the GRE revised General Test is based.

The current study presents information on the predictive validity data for the GRE General Test by combining results from collaborating institutions and departments. While the data used here are from the version of the test used prior to August 2011, the findings are still relevant to the GRE revised General Test in that the underlying constructs being measured by both test versions are highly similar.

Understanding the relationship of GRE scores to first-year graduate grades is important. However, it has been commented that first-year grades do not represent the most important goals of graduate school (Sternberg & Williams, 1997; Yee, 2003). Kuncel, Hezlett, and Ones (2001) also summarized evidence that shows that GRE scores and undergraduate grades predict a number of long-term outcomes of graduate school. As a result, this study collected information on a broader definition of success in graduate school than previously investigated. This expanded outcome information is important because it allows score users to evaluate admission measures against a variety of goals considered important for graduate education.

Procedure

The institutions that participated in the study cover a wide breadth of the graduate community, representing a variety of disciplines and missions: regional professionally oriented master’s degree programs, programs primarily focused on teaching, and research programs that recruit nationally and internationally for top doctoral students. A total of 21 departments in biology, chemistry, education, English, and psychology from seven different graduate institutions participated. Data on 1,700 students who entered a master’s or a doctoral degree program in 1995–1996, 1996–1997, or 1997–1998 were obtained from the various departments.

The measures most commonly used for admission decisions acted as *predictors* in the study: GRE Verbal Reasoning and Quantitative Reasoning scores and undergraduate grade point average (UGPA). *Outcome measures* for this study were developed based on research literature and on interviews with GRE users about their most important goals for graduate students (Walpole, Burton, Kanyi, & Jackenthal, 2002). Data on cumulative graduate grade point average (GGPA) and faculty ratings were collected. Two faculty members familiar with a student rated the student on three dimensions: (a) professional knowledge, ability to apply that knowledge, and ability to learn independently (mastery of the discipline); (b) judgment in choosing professional issues and creativity and persistence in solving the issues (professional productivity); and (c) ability to communicate what was learned (communications skills). Background information on each student was also collected: gender, race/ethnicity, citizenship status, degree level (master's versus doctorate), and test mode (computer-based version vs. paper-based version). This summary presents only those findings by gender and race/ethnicity, not by degree level, citizenship, or test mode.

Only students with complete data on all three predictor variables and the outcome measure were included in the analysis. The minimum sample size for analysis was defined as nine students with complete data within each department. All possible combinations of the three predictors were used to compute prediction equations. Because the same set of students was used to compute each equation, the results from different equations are comparable. Measures used in admissions are restricted in range; that is, students with low UGPAs or low GRE scores are admitted less frequently than those with higher GPAs or scores. As a result, UGPA will look like a poorer predictor of graduate school outcomes than it really is. Therefore, a technique that corrects for the restriction in range was used to correct for this problem (Ramist, Lewis, & McCamley, 1990; Ramist, Lewis, & McCamley-Jenkins, 1994).

Results

Overall Analyses

The first set of analyses focused on how well GRE scores and UGPA predicted the following long-term graduate school outcomes: cumulative GGPA and mastery of the discipline, professional productivity, and communication skills as measured by faculty ratings. Analyses yielded the following results:

- When all three predictors (GRE Verbal Reasoning scores, GRE Quantitative Reasoning scores, and UGPA) are combined, the corrected correlations for faculty ratings are large (.5 or higher; Cohen, 1977).
- When all three predictors are combined, the corrected correlation for cumulative GGPA rounds to .5.
- Correlations for the two GRE scores combined are nearly as high as those for all three predictors combined. The UGPA does contribute to the prediction of all outcomes, but its greatest influence is on the prediction of cumulative GGPA.

- The correlation of UGPA alone is .32, so the GRE scores contribute .17 to the full correlation of .49.

Discipline Comparisons

Correlations were also computed by discipline. While the strength of various predictors on the four outcome measures varied to some degree across the disciplines, overall, the patterns of correlations for biology and chemistry departments were similar to the above results, while education and English departments have a slightly different pattern from the two science disciplines. For example, cumulative GGPA is predicted moderately well in education departments, while it is predicted strongly in both science disciplines. Communication skills are moderately predicted for chemistry students, but strongly predicted for education students.

Questions regarding the fairness of UGPA and GRE scores for various subgroups of the graduate school population were addressed next. Specifically, it was examined whether graduate school outcomes are equally predicted across different groups and whether predicted outcomes are systematically lower or higher than the actual outcomes for those groups. Separate equations were computed for the two groups being compared. A single equation was also generated by department to examine the differences between GGPAs that are predicted by each department's equation and the actual grades attained by students in that department. These differences were then investigated by subgroup to look for any systematic overprediction or underprediction for each subgroup. Because the number of students in a given subgroup was small, only the most frequently available outcome data, cumulative GGPA, was analyzed.

Overall, results support the appropriateness of using GRE scores and UGPA to predict academic success for the subgroups studied. Within a department, correlation coefficients are of comparable size. Over the departments studied, overpredictions or underpredictions tend to be small.

Gender Comparisons

There are small average differences between men and women, mostly in the expected direction. Men's grades are overpredicted in all disciplines but English. The amount of underprediction for women is very small; overall, women's grades are underpredicted by 1/100th of a grade point. The largest average underprediction occurs in chemistry departments, where women's cumulative average GPA is underpredicted by 6 1/100th of a grade point.

Ethnic Group Comparisons

Only the large education departments in three participating universities had the minimum required samples of both minority and White students. In education departments, graduate grades tend to be slightly overpredicted for African American and Asian students and slightly underpredicted for Hispanic and White students. Grades of African American students are consistently overpredicted, except in biology. The tendency to overpredict African American students' grades is also observed for undergraduates (Bowen & Bok, 1998; Jencks & Phillips, 1998; Ramist, Lewis, McCamley-Jenkins, 1994). Hispanic students' grades, underpredicted in education, are overpredicted in English and chemistry.

Conclusion

The results indicate that the combination of GRE scores and UGPA strongly predicts cumulative GGPA and faculty ratings. These results hold across various disciplines and subgroups.

This study shows that cumulative graduate grades, not just first-year grades, can be predicted using GRE scores and UGPA. In addition, key professional skills of graduate students, including their mastery of the discipline, their potential for professional productivity, and their ability to communicate what they know, are predicted using GRE scores and UGPA. This study provides foundational support for the continued use of GRE scores as part of graduate admissions.

References

Bowen, W. G., & Bok, D. (1998). *The shape of the river.* Princeton, NJ: Princeton University Press.

Cohen, J. (1977). *Statistical power analyses for the behavioral sciences* (rev. ed.). New York, NY: Academic Press.

Jencks, C., & Phillips, M. (Eds.). (1998). *The Black-White score gap.* Washington, DC: Brookings Institution Press.

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2001). A comprehensive meta-analysis of the predictive validity of the Graduate Record Examinations: Implications for graduate student selection and performance. *Psychological Bulletin*, *127*, 162–181.

Ramist, L., Lewis, C., & McCamley, L. (1990). Implications of using freshman GPA as the criterion for the predictive validity of the SAT. In W. W. Willingham, C. Lewis, R. Morgan, & L. Ramist (Eds.), *Predicting college grades: An analysis of trends over two decades* (pp. 253–258). Princeton, NJ: Educational Testing Service.

Ramist, L., Lewis, C., & McCamley-Jenkins, L. (1994). *Student group differences in predicting college grades: Sex, language, and ethnic group* (Research Report No. 93-1). New York, NY: College Board.

Sternberg, R. J., & Williams, W. W. (1997). Does the Graduate Record Examinations predict meaningful success in the graduate training of psychology? A case study. *American Psychologist, 52,* 630–641.

Walpole, M. B., Burton, N. W., Kanyi, K., & Jackenthal, A. (2002). *Selecting successful graduate students: In-depth interviews with GRE users* (GRE Board Research Report No. 99-11R). Princeton, NJ: Educational Testing Service.

Yee, C. (2003, September 25). Committee calls for new GRE changes. *The Chronicle Online.*

Notes

[1] Based on *Predicting Long-Term Success in Graduate School: A Collaborative Validity Study* (GRE Board Research Report No. 99-14R), by N. W. Burton and M.-m. Wang, 2005, Princeton, NJ: Educational Testing Service.

5.7 Effects of Pre-Examination Disclosure of Essay Prompts for the *GRE®* Analytical Writing Measure[1]

Donald Powers

The Analytical Writing measure of the *GRE*® General Test consists of two writing tasks: One requires examinees to present their perspectives on an issue and the other requires them to analyze an argument. These two tasks are designed to assess the ability to (a) discuss and critique an argument, (b) articulate and support complex ideas, and (c) sustain a focused and coherent discussion. The purpose of the current study was to document how examinees prepare for the GRE Analytical Writing measure and examine how prepublishing prompts used on the test impacts test preparation behavior, test performance, test validity, and examinee perceptions of the value of prompt prepublication.

Test score validity depends not only on the questions that comprise a test, but also on what happens before a test is administered, in particular, how examinees prepare for the examination. In the interest of minimizing any validity-compromising effects due to insufficient familiarity with a test, many test makers now provide a variety of materials designed to help test takers become familiar with the tests they take. One testing practice that has been instituted relatively recently to help examinees prepare for tests of writing skill (including the GRE Analytical Writing measure) is to prepublish the entire pool of essay prompts from which prompts are selected for each test administration.

Essay prompts are provided as part of test preparation for multiple reasons. One motive for prepublishing essay prompts is fairness: to ensure that all essay prompts are equally available to every examinee, not just the few who may obtain access using unethical means. A potential negative side effect of prepublication, however, is that some examinees may attempt to memorize exemplary essays and simply *regurgitate* these essays when testing. To minimize this prospect, some testing programs release relatively large numbers of prompts in hopes that a sufficiently large pool will discourage undesirable test-taking behavior.

On the positive side, prepublication of a smaller, reasonably manageable pool of prompts has the potential for increasing the validity of writing test scores by providing additional time for planning—a phase of composing that most writing experts view as integral to the writing process. Greater opportunity for planning prior to taking the test may allow examinees to devote less time to formulating and organizing their ideas and more time to translating and communicating them during the test. If prepublication helps examinees become more familiar with potential test topics, a writing test may be seen as more authentic (e.g., less a reflection of the ability to write extemporaneously and more of an indication of the kind of planful writing in most academic settings).

With few exceptions, the research that has been done on the subject has focused almost exclusively on the impact of releasing test questions *after* a test is administered

(Lockheed, Holland, & Nemceff, 1982; Stricker, 1984). However, findings from studies that did investigate the effects of predisclosure were inconsistent. Hale, Angelis, and Thibodeau (1983) found that students taking the *TOEFL*® examination performed better on disclosed multiple-choice test questions than on undisclosed ones. In another study that disclosed essay prompts for a beginning teacher certification test (Powers, Fowles, & Farnum, 1993), only a small difference was found between students' performance on disclosed and previously unseen topics, and no detectable effect of disclosure on test validity was present, as evidenced by correlation of essay scores with several other indicators of writing proficiency. A third study by Powers and Fowles (1998) revealed a negligible effect on students who saw essay topics prior to being tested, although the majority of examinees reported spending time thinking about the prompts they had received, and a small percentage engaged in more time-consuming preparation. None of these studies were carried out in the high-stakes situation of an actual test, thus limiting their generalizability.

Procedure

Sample

Approximately 2,000 individuals who were registered to take the GRE General Test were sent subsets of essay prompts and strongly encouraged to think about the prompts, develop outlines, and compose first drafts. (A variation of this *encouragement design* had been used successfully in previous studies of test preparation for the GRE General Test. See, for example, Powers & Swinton, 1984.) Because the study sought to examine whether the number of prompts that examinees received prior to taking the test affected their preparation for and performance on the actual test, different subgroups of examinees were sent 27, 54, or 108 essay prompts. Of the examinees who were contacted, a total of 199 responded to the request to provide information about their test preparation activities for the Analytical Writing measure. This sample was slightly more able than the general GRE examinee population; analyses showed this sample to have slightly outperformed, on all three measures of the test, a reference group of GRE examinees who took the exam during the same time interval.

Data Collection

The following data were collected on each study participant after he or she was tested:

- Scores on the two components of the Analytical Writing measure generated through the regular operational test-scoring process
- Surveys completed immediately after testing that asked about the participants' preparation for the Analytical Writing measure, including whether they had spent

time on any of a variety of test preparation activities (e.g., reading sample essays) and, if so, approximately how much

- Grades on courses and assignments that required *considerable* writing or focused heavily on logic, reasoning, or critical thinking
- Two samples of their course-related writing. These samples were then evaluated by college and university faculty or experienced evaluators of writing who had been trained to apply a scoring guide.
- Self-assessments of the examinee's success with various kinds of writing tasks (e.g., persuasive, descriptive, analytical writing) and thinking skills that have been deemed by graduate faculty to be important for success in graduate education (Powers & Enright, 1987)

Test files for each study participant were searched to identify those examinees who actually wrote on a prompt that they had received prior to taking the test.

Results

Results indicated that the strategy most frequently used by study participants (82%) to prepare for the Analytical Writing measure was to think generally about the potential topics. Participants commonly spent less than 1 hour doing so. Slightly less than one half of the study participants wrote sample essays to prepare, and very few (4%) reported memorizing essays.

Regardless of how many prompts they had been given, study participants used, on average, about six or seven to prepare for the exam. A slight majority used one to five prompts in their preparation. Thus, it can probably be assumed that typical GRE examinees will employ only a small fraction (about 10%) of the pool of the provided prompts in their preparation for the test.

No significant differences in test performance were found between students who had prepared in some way compared to those students who had not prepared for the prompt on which they were eventually tested. Modest correlations, mainly in the .20s, were found between students' scores on the two components of the writing test and each of the nontest indicators of reasoning and writing ability, including self-estimates of ability, self-reported writing problems, and evaluations of writing samples conducted by trained evaluators. For example, the correlation of two student-provided course-related writing samples with the combined writing measure score was .36. The correlation of a self-report index of success with various kinds of writing in college with the combined writing measure score was .26.

When asked if making the GRE essay topics available ahead of time is a good testing policy, most study participants said either *definitely yes* (44%) or *probably yes* (36%). Few students indicated it was not a good testing policy, with 13% indicating *probably not* and 7%

definitely not. The most frequent comment from those who endorsed the practice suggested that prepublishing the topics helped to reduce pressure/anxiety by “eliminating one of the unknowns” and giving examinees an idea of what to expect.

Conclusion

Within the limits of the data collected, the study found no evidence that participants benefited from encountering a prompt for which they had prepared. The correlations between scores on the Analytical Writing measure and the nontest indicators of reasoning and writing ability add to the accumulation of evidence of the validity of scores from the Analytical Writing measure. The results also extend previous research on the Analytical Writing measure because they are based on fully operational administrations of the test, not on experimental research.

References

Hale, G. A., Angelis, P. J., & Thibodeau, L. A. (1983). Effects of test disclosure on performance on the Test of English as a Foreign Language. *Language Learning, 33,* 449–464.

Lockheed, M., Holland, P., & Nemceff, W. (1982). *Student characteristics and the use of the SAT test disclosure materials* (College Board Research Report No. 82-3). New York, NY: College Entrance Examination Board.

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

Powers, D. E., & Fowles, M. E. (1998). Effects of preexamination disclosure of essay topics. *Applied Measurement in Education, 11,* 139–157.

Powers, D. E., Fowles, M. E., & Farnum, M. (1993). Prepublishing the topics for a test of writing skills: A small-scale simulation. *Applied Measurement in Education, 6,* 119–135.

Powers, D. E., & Swinton, S. S. (1984). Effects of self-study for coachable test item types. *Journal of Educational Psychology, 76,* 266–278.

Stricker, L. J. (1984). Test disclosure and retest performance on the SAT. *Applied Psychological Measurement, 8,* 81–87.

Notes

[1] Based on *Effects of Pre-Examination Disclosure of Essay Prompts for the GRE Analytical Writing Assessment* (GRE Board Research Report No. 01-07R), by D. E. Powers, 2005, Princeton, NJ: Educational Testing Service.

5.8 The Role of Noncognitive Constructs and Other Background Variables in Graduate Education [1]

Patrick Kyllonen, Alyssa Walters, and James Kaufman

Many *GRE*® studies conducted in the past asked faculty members to describe the qualities that they see as important to success in graduate school. A consistent finding was that noncognitive factors, such as motivation, creativity, personality, interests, and attitudes, should be part of graduate admissions (Briel et al., 2000). Graduate faculty members view such factors as increasing both the fairness and the validity of the admissions process.

The *ETS*® *Personal Potential Index* (ETS® PPI) was released in 2009, prior to the introduction of the GRE revised General Test. The ETS PPI is an innovative, web-based tool that allows an instructor or supervisor to provide applicant-specific information about six key attributes that graduate deans and faculty have identified as essential for graduate study: knowledge and creativity, resilience, communication skills, planning and organization, teamwork, and ethics and integrity.

Although a new admissions tool, the ETS PPI is based on a number of years of research that identified these factors. This paper provides a review of the scientific literature related to noncognitive factors and variables that impact them, their importance, how they can be measured, and their practicality for use in admissions decisions. The findings of this review helped define those factors measured by the ETS PPI.

Brief Overview of the Literature

The scientific literature classifies graduate school outcomes into two main categories: (a) traditional measures, such as attrition, time-to-degree, and grade point average and (b) performance factors, such as domain proficiency, general proficiency, communication, effort, discipline, teamwork, leadership, and time management. Noncognitive predictor factors, that is, factors that predict graduate school outcomes, may be categorized into three main categories: (a) personality, such as extroversion and introversion; (b) quasi-cognitive, such as metacognition; and (c) attitudinal factors, such as motivation and self-efficacy. In addition, two broad categories of background variables have been shown to affect noncognitive factors: (a) environmental (e.g., mentor support) and (b) group factors (e.g., gender and ethnicity).

Noncognitive Factors

Personality

A consensus has emerged within the psychological field of five major personality factors: extroversion, emotional stability, agreeableness, conscientiousness, and openness. Other personality factors appearing in the literature can be seen as combinations or facets (subfactors) of these. Including personality factors as part of graduate admissions would broaden the student qualities considered and could help to increase student diversity. While faculty members say personality is important, most of these dimensions are measured through self-reports and, therefore, can be faked by students to present a more desirable picture of them. Techniques, such as adjusting scores for social desirability bias, using a forced-choice approach, warning examinees of the consequences of faking, and using subtle questions, can be used to control faking, but these techniques are not necessarily feasible approaches in large volume, high-stakes settings.

Two methods for evaluating personality could be introduced as part of graduate admissions. The first one is through the use of a performance (ability) test that measures personality that is immune to faking. For example, one new promising method is the conditional reasoning approach. In this approach, questions look like reasoning questions but contain more than one correct answer; the examinee's choice of the *correct* answer is thought to reflect personality. The second way to evaluate personality is through the use of others' rating of personality, such as advisors, professors, and other members of the college community who typically write letters of recommendation for students now. This method is the approach adopted by the ETS PPI.

Quasi-Cognitive

These are factors that fall somewhere between cognitive and noncognitive factors. They may be measured with performance or ability tests, but they also reflect affective qualities. We consider four such factors here: creativity, emotional intelligence, metacognition and confidence, and cognitive style.

Creativity. Faculty typically rank creativity high on the list of qualities believed to be important to success in graduate school. However, they disagree on what creativity is and how it should be measured. Self-report measures tend to be obvious and easily faked. Performance measures, such as fluency tests (e.g., "How many ways can you use a brick?"), do not necessarily predict important criteria independent of general cognitive ability. Creativity has been assessed by others' judgments as overall ratings of the person or as ratings of particular creative products (e.g., essays), but whether such ratings are independent of other quality judgments has not been established.

Emotional intelligence. Emotional intelligence has received considerable attention in the popular media and some attention in scientific literature. Measures fall into two categories: self-reports and performance tests. Self-reports yield scores that duplicate personality measures, but they have the same problems as personality measures in that examinees can fake their responses. Performance tests offer the promise of not being as easy to fake. However, they have not been studied sufficiently and little is known about their validity.

Metacognition and confidence. These measures assess whether examinees can accurately predict whether they know the correct answer to a test question. Considerable research suggests that examinees' ability to predict if they will answer correctly is independent of their general ability. However, it is not clear how this could be used in an applied admissions context.

Cognitive style. A good deal of research has been conducted on cognitive styles, such as field-dependence, but there have been problems in discriminant validity (that is, such measures tend to overlap with cognitive ability or personality). In addition, for self-reports of cognitive styles, the ability to fake responses probably precludes their use in admissions.

Attitudinal

Attitudinal factors (such as self-concept, self-efficacy, motivation, attributions, interests, and social attitudes) influence choice of activities, goals, strategies, effort, and persistence. Attitudes are often domain specific. For example, being motivated in one domain, such as academics, may be largely independent from being motivated in another, such as athletics. Attitudinal factors are thought to be particularly important in understanding students traditionally underrepresented in graduate school.

Self-concept and domain identification. This refers to the way people characteristically think about themselves in a domain. The focus here is on academic self-concept and identification with the academic domain. Self-concept is typically measured with self-assessments, but a measure such as The Implicit Association Test could possibly be used as a self-concept measure.

Self-efficacy. This widely researched construct refers to a person's belief in his or her ability to achieve success in a particular area. Self-efficacy is important in selecting goals, adopting strategies, task persistence, and the effort put into a task. High self-efficacy is associated with a wide range of positive academic outcomes. Self-efficacy can be influenced by numerous causes, such as domain mastery, faculty support, personality, and interests.

Motivation. The term motivation has been used in many different ways over the years. Much of the motivation research literature falls into other categories (e.g., self-efficacy), but a few concepts are uniquely associated with motivation. One is the distinction between extrinsic and intrinsic motivation, with intrinsic motivation widely believed to be related to higher outcomes; this belief is not supported by recent meta-analytic research. Another is the distinction between a performance-goal and a learning-goal orientation, with the success of the orientation depending on one's ability (that is, performance goals work well when one is proficient; otherwise, learning goals work better).

Interests. A dominant framework for studying interests is that of Holland's (1959, 1973). This framework classifies interests in six categories: realistic, artistic, investigative, social, enterprising, and conventional. Numerous studies have shown links between interests and academic outcomes. Using this framework, gender differences in types of interests have been identified (for example, men score higher in *realistic*; women score higher in *social* and *artistic*), which could account for the difference in the graduate fields that men and women pursue.

Attributions. Reasons for successes and failures are habitually attributed to others or ourselves and to forces under our control or not under our control. Adaptive attribution styles (e.g., attributing failures to ourselves, but also to forces we can control and that are changeable) leads to greater persistence and intensity in performing tasks and, therefore, may be of considerable value in understanding and predicting higher education achievement.

Social attitudes and values. Several suggestions address the structure of beliefs, attitudes, and values (such as individualism, equality, and religiosity). These have primarily been studied in cross-cultural contexts, comparing countries with each other, and relating these to national indicators, such as gross national product, literacy level, and so on. However, it may be that social attitudes and values are also useful to study at the individual level, as they may influence success in higher education.

Background Variables

Environmental Influences

Various environmental influences other than personality and motivation are believed to affect the noncognitive factors discussed above. These variables, in turn, may impact performance in graduate school.

Mentor and social support. Mentor and social group support affect persistence to degree as well as other less tangible outcomes, such emotional well-being. Informal contact with one's mentor and mentor qualities such as interest in the student, accessibility, integrity, reliability, and communication skills are important for various graduate school student outcomes.

Prejudice and institutional integration. Encountering negative stereotypes about the intelligence or abilities of one's gender, ethnic, or age group may have negative effects on attitudes and school and test performance. There have been numerous demonstrations in the laboratory of such effects but few in actual higher-education settings. One way prejudice may impair school performance is by inhibiting integration into the university environment, leading to increased attrition by particular groups of students.

Financial support. Financial support is believed to reduce attrition, but it is not equally available to all. There is controversy over whether fellowships or assistantships are more effective, with two major research studies coming to different conclusions.

Prior accomplishments. Prior accomplishments, as reflected in transcripts, resumes, and standardized surveys, predict graduate school accomplishments and various undergraduate criteria. Although not widely researched, self-assessed standardized accomplishments measures, particularly when verifiable, have proven useful and may warrant further evaluation.

Group Factors

Black, Hispanic, and female students have lower participation, doctoral candidacy, and graduation rates than White and male students. The factors that are responsible for this and what might be done about it are still debated.

Ethnicity and race. Several explanations have been invoked to account for lower standardized test scores as well as higher attrition and lower performance in school of Black and Hispanic students. One is stereotype threat, which may result in students subconsciously changing their academic performance given generalized beliefs about the abilities of particular ethnic and race groups. Another is systemic, which affects identity development. Academic preparation is lower for Black and Hispanic students, as indicated by the number of high school science and mathematics courses taken and the selectivity of colleges attended. To some degree, faculty support, financial aid, and institution policies may combat these effects, as might consideration of a broader range of factors in admissions decisions.

Gender. The gender gap is closing, but differences remain in science, mathematics, and engineering, and in standardized test scores. Some of these differences reflect different interests between genders. Differences in academic preparation are rapidly closing, particularly in high school. Some of the difficulties women experience in mathematics and science may be reduced by considering a broader range of factors in admissions, coupled with institutional and faculty support. But diminished peer respect for abilities of those in support and preferential programs has been noted.

Conclusion

The purpose of this review was to consider the role played by noncognitive factors in graduate school and how such factors could be used, particularly in graduate admissions. Based on the results of the review of scientific literature, noncognitive assessment as part of graduate admissions holds much promise. These findings suggest particular factors and dimensions, question types, and approaches for such an endeavor. The findings also provide support for the continued evaluation of graduate school applicants and provide the foundation for the continued use of the ETS PPI as part of graduate admissions.

References

Briel, J., Bejar, I., Chandler, M., Powell, G., Manning, K., Robinson, D., . . . Welsh, C. (2000). *GRE horizons planning initiative: A research project funded by the GRE Board Research Committee, the GRE Program, and ETS Research and Division.* Unpublished manuscript.

Holland, J. L. (1959). A theory of vocational choice. *Journal of Counseling Psychology*, *6*, 35–45.

Holland, J. L. (1973). *Making vocational choices.* Englewood Cliffs, NJ: Prentice Hall.

Notes

[1] Based on *The Role of Noncognitive Constructs and Other Background Variables in Graduate Education* (GRE Board Research Report No. 00-11), by P. C. Kyllonen, A. M. Walters, and J. C. Kaufman, 2011, Princeton, NJ: Educational Testing Service.

Section 6: Ensuring Fairness and Accessibility

Assessments should be designed, developed, and administered in ways that treat all test takers equally and fairly regardless of their personal characteristics. As part of the *GRE®* program's commitment to fairness and access for all test takers, the revision of the GRE General Test presented an opportunity to evaluate the impact of the new question types and test design on many groups of test takers and to gain an understanding of the perception of the role of the GRE in graduate admissions. Chapters in this section describe a number of studies that focus on various aspects of GRE fairness and access.

- Chapter 6.1 reports on a survey conducted to explore test takers' perceptions of the role of GRE scores in graduate admissions. The goals of the study included understanding how test takers perceive the GRE General Test both before and after they take it; how test takers prepare for the test; the perceived importance of GRE scores in the graduate admissions process; and the sources of information, both formal and informal, that test takers consult regarding the GRE. The study included 3,362 students who identified themselves as White, Black, Hispanic, or Asian; the focus of the study was on the perceptions of those students who are traditionally underrepresented in graduate education (Black and Hispanic). Results indicated differences between racial minority and White test takers' self-reports on their levels of preparedness and anxiety to be relatively small. White test takers overall have fewer negative feelings about the test compared to other groups. In addition, although most test takers view the GRE as at least somewhat important in graduate admissions, they perceive it as being far less important than other factors, such as undergraduate grades, letters of recommendation, and life/work experience.
- Chapter 6.2 details the field trial that examined performance on and reactions to the question types proposed for use on the GRE revised General Test for test takers with disabilities. Participants in the field trial represented six major disability groups (visual, hearing, physical, learning, ADHD, and psychiatric). The test format included five test sections, administered in the following order: Section 1, new Quantitative questions; Section 2, new Verbal questions; Section 3, new Verbal questions; Section 4, current Quantitative questions; and Section 5, current Verbal questions. Participants also responded to two surveys. A short background survey was administered prior to testing and a longer postadministration survey designed to capture test-taker reactions to the proposed new question types was administered following the test. Participants who took the test in an alternate format were also asked for feedback on the format used for the different question types. Results

indicated that test takers with disabilities took considerably longer to complete the questions compared to test takers without disabilities who participated in an earlier pilot using the same questions. Test takers with learning disabilities also tended to skip over questions with high reading loads. However, most test takers with learning disabilities, hearing impairments, and physical disabilities who only required extended time on the computer-based test completed the sections with the proposed new question types within the most frequently granted time extension (50% extended time, or time and a half). The exit survey provided additional information on the proposed questions types. Participants felt that some question types needed clearer directions, and there was concern about questions that required significant use of memory. Alternate format test takers, particularly audio users, indicated that some of the new question types presented a larger challenge to complete compared to the current question types. Comments from the survey were used to refine the presentation of test material and test directions in alternate formats.

- Chapter 6.3 describes the efforts to develop a computer-voiced version of the GRE revised General Test (vGRE) for use by test takers who are blind or have low vision. The vGRE is designed for complete eyes-free usability and incorporates several features to improve accessibility for test takers with low vision. For example, all question and directions text can be enlarged to nearly any degree, the foreground and background colors can be selected by the test taker, text can be highlighted when spoken, and answer choices are underlined and boldfaced for better visibility. Usability research included consultations with experts in accessibility and with a small number of test takers who were blind or had low vision and had taken or planned to take the GRE General Test. Changes to the first version of vGRE were made, when feasible, based on the experts' comments. This version of the test was then administered to four study participants. Results indicated that the vGRE still did not pronounce some words with sufficient clarity. In addition, some participants had difficulty using keystrokes that were different from those used in screen readers. As much as possible, these concerns were included in the second version of vGRE, and these efforts resulted in the release of the voiced version of the test close to the launch of the GRE revised General Test.
- Chapter 6.4 presents an evaluation of different techniques that help ensure that the writing prompts used on the Analytical Writing measure are fair for all test takers. Three differential item functioning (DIF) methods were used that matched test takers across different subgroups (such as gender, ethnicity/race) based on their ability level and then compared their performance on a specific prompt. In this study, essay prompts were examined to determine if they were differentially

different for (a) female test takers; (b) African American, Asian, and Hispanic test takers; and (c) test takers whose best language is a language other than English. Both issue and argument prompt types were analyzed. Results indicated that no DIF values were large enough to warrant removal of a specific prompt from the pool of prompts. In addition, the three methods evaluated were in agreement in terms of the extent to which they identified the prompts having large DIF values. Even though the study found a range of DIF values across prompts, no prompt appeared to have a substantially higher DIF value than the other prompts. However, the study found that some combinations of issue and argument prompts should be avoided operationally. Results provide support for continuing sensitivity reviews as part of the GRE revised General Test to ensure that new prompts are appropriate for all test takers.

- Chapter 6.5 describes a study that examined the impact of extra time on test-taker performance on the Verbal and Quantitative measures. GRE test takers were invited to take an additional section of the test as part of the study; participation was voluntary, but participants were offered an incentive to perform as well as they could on the additional section. Participants took either a Verbal or Quantitative measure and were randomly assigned to a group using standard testing time or a group using 1.5 times the standard testing time. Participants were also grouped into three ability levels based on their operational GRE score: low, medium, and high. Results indicated that the benefit from extra time for both the Verbal and Quantitative measures was only 7 points on average. The effects of extra time seemed to be relatively constant across all race/ethnic categories. Participants in the low-ability group seemed to benefit more from additional time compared to those in the other ability groups. These results pointed to the need for additional field trials to ensure that appropriate time limits were established for the revised test.

- Chapter 6.6 provides an overview to the definition of and processes related to fairness as used by the GRE program. It describes the various test-taker groups that are of particular interest regarding test fairness: (a) U.S. citizens testing in domestic test centers, a group that is further divided into gender (female and male) and race/ethnicity (Asian, Black, Hispanic, and White test takers) groups using self-reported information; (b) non–U.S. citizens testing in China, Hong Kong, Korea, or Taiwan; (c) non–U.S. citizens testing in India or Japan; and (d) non–U.S. citizens testing in any other international test center. Smaller groups, such as those taking the test on paper or receiving testing accommodations, are also considered. The chapter details the procedures used to satisfy three critical fairness goals: all tests are free of bias, all test takers are given comparable opportunity to demonstrate their ability, and group differences are not exacerbated. Summaries of the pilot and field trials that evaluated these aspects of fairness are provided. Finally, the chapter compares data on critical test outcomes related to fairness, such as standardized differences between the various groups, for the test before and after the launch of the GRE revised General test.

6.1 Test-Taker Perceptions of the Role of the *GRE*® General Test in Graduate Admissions: Preliminary Findings[1]

Frederick Cline and Donald Powers

Historically, both the *GRE*® Board and the GRE program have been concerned about the access of traditionally underrepresented minority students to graduate education. It has been hypothesized that certain steps in the graduate admissions process may be perceived as more of a barrier by minority students than majority students. In particular, minority students may view admissions tests like the GRE General Test and the role of GRE scores in the admissions process more negatively than majority students.

Unlike undergraduate admissions, where admissions decision making is a centralized process, graduate admissions tend to be more diffuse and often decentralized. Thus, it is not surprising that prospective graduate students (and possibly minority students in particular) might be unclear about how admissions decisions are made at the graduate level, and quite possibly, they have misconceptions of the role of specific factors in the process. This situation is understandable, as many factors influence prospective graduate students' choice of a graduate school, including the chances of gaining admission.

Powers and Lehman (1983) studied the perceptions of a representative sample of GRE test takers who were asked to indicate their views of the importance of eight widely considered factors in graduate admissions. Candidates' perceptions were compared for each of the factors and for subgroups of candidates determined by sex, race/ethnicity, age, and intended graduate major field. Analyses by subgroup revealed that Black candidates and White candidates exhibited quite different patterns of perceptions, especially in their judgments of the importance of GRE scores, which Black candidates viewed as being more influential than did White candidates.

Other testing programs have also undertaken basic studies of test-taker attitudes about their test. Stricker, Wilder, and Bridgeman (2002) asked Graduate Management Admission Test (GMAT) test takers four questions about validity and fairness, finding that test takers in general had slightly negative views about the validity of the test—more so for minority test takers—and slightly positive views that the test was unbiased—less so for minority students.

The GRE General Test is but one of many factors that may be involved in minority students' (a) decisions to pursue graduate study, (b) decisions about where and in what fields to apply, and (c) eventual admission to pursue such study. Nevertheless, the GRE General Test is often an important factor.

The impetus for the study described here was to investigate potential differences in perceptions among minority test takers that might negatively impact the participation of qualified minority students in graduate programs. The major objective of the study described here was to determine minority student perceptions of the GRE General Test and its overall role in graduate admissions, in order to ascertain the extent to which test takers may perceive the GRE General Test as a significant barrier, due, for example, to high graduate admissions standards, test difficulty, or the degree of preparation needed to attain acceptable scores. To accomplish this objective, the study authors administered a survey to a representative sample of recent GRE test takers in order to answer the following questions:

1. How do GRE test takers perceive the GRE General Test both before and after they take it? Do perceptions vary by race/ethnicity?
2. Are there differences in how racial/ethnic subgroups prepare for the GRE? What factors affect the (possibly differential) selection of test preparation materials by subgroups?
3. What are test takers' perceptions of the significance of the GRE General Test in the graduate admissions process? Do perceptions differ according to test taker subgroups?
4. What informal and formal sources of information do test takers consult regarding the GRE General Test?

Method

From December 2008 through June 2009, approximately 35,000 GRE General Test examinees were contacted via e-mail, given a website address, and asked to complete a web-based survey within 3 weeks *after* their scheduled test administration. All test takers who indicated minority racial status when registering for the test were contacted to complete the survey. A random sample of 20% of the White test takers was contacted to form the comparison group.

The final response rate was 9.5% and yielded 3,362 test takers. The sample included 1,546 White, 744 Black (African American), 563 Hispanic (232 Mexican or Mexican American, 91 Puerto Rican, and 240 Other Hispanic or Latino), and 399 Asian (Asian American) test takers. Female test takers accounted for two thirds of the overall sample and 82% of Black respondents. While the goal in sampling was not to replicate the overall GRE race/ethnicity mix but instead to oversample specific groups for comparison reasons, it should be noted that the proportion of females in the sample is higher than the overall GRE population. The sample was limited to U.S. citizens or permanent residents only, as information about race/ethnicity is not collected for international test takers.

Nearly all (87%) of the respondents indicated their primary reason for taking the test was admission to graduate or professional school. Most respondents (71%) were either in their final year of an undergraduate program or were unenrolled with a completed undergraduate degree; an additional 14% were unenrolled with a completed master's degree. About one third indicated that they were the first generation in their family to have attended college. Of the respondents, 56% were seeking a master's degree, while 42% planned to pursue a doctorate. Intended field varied by race/ethnicity: 17% of Black and 18% of Mexican test takers reported an intended field in the physical or life sciences and engineering, compared to 42% of Asian and 30% of White test takers. Other Hispanic/Latino respondents had the highest percentage with an intended major in the social sciences, arts, and humanities at 45%, compared to 38% for White and 30% for Asian test takers.

An obvious limitation of the sample chosen is that it does not reflect the views of students who may have considered graduate education but who did not take the GRE General Test.

Results

Overall, differences between racial minority and White test takers' self-reports on their levels of preparedness and anxiety were relatively small, although overall White test takers had less negative feelings toward the test than the other groups (see Table 6.1.1). The differences in perception may be driven by differences in expected or actual performance on the GRE General Test itself, rather than by race/ethnicity. For the majority of questions, as test scores decreased, negative feelings about the test and its role in admissions increased. Since minority test takers tend, on average, to score lower than White test takers, higher levels of negative feelings about the test and its role by minority test takers may be related more to expected test performance than ethnicity.

- On average, across race/ethnicity, test takers report preparing for about 23–24 hours for the test, although the largest percentage prepared for 10 hours or less. A small minority, however, reportedly devoted more than 100 hours. By far, the most frequent test preparation practice undertaken by the vast majority of test takers across race/ethnicity was to read GRE study guides. A majority also reported taking official (or unofficial) GRE practice tests.
- Although the hours spent preparing for the test increased test takers' sense of being ready, 40% to 50% still reported feeling at least *very anxious* before taking the test. White test takers were least likely to feel unprepared but did not have the lowest anxiety levels.
- Although a majority of test takers view the GRE General Test as at least somewhat important in graduate admissions, they perceive it as being far less important than a variety of other factors—personal statements, letters of recommendation, undergraduate grades, life or work experience, personal qualities, and match to graduate faculties' research interests—which were all regarded as about equally important. The GRE scores were identified as extremely important in the admissions process less than 10% of the time, while essays and letters of recommendation were

considered extremely important 40% of the time (see Table 6.1.2 for this information by race/ethnicity). The Verbal Reasoning measure was more often expected to be more difficult than the Quantitative Reasoning and Analytical Writing measures, particularly for Asian and Hispanic test takers, and was also the section most often identified as being more difficult than expected and less valid as a measure of their verbal skills. Test takers may not be aware of differences in distributions on the Verbal and Quantitative scales, and most test takers set target GRE scores the same for both sections. Indeed, this misalignment of the Verbal and Quantitative scores was an important consideration in establishing a new score scale for the GRE revised General Test.

- Overall, the Analytical Writing measure was not considered to be difficult or time-consuming compared to the other two measures, possibly due to test takers not being as concerned about its use in admissions decisions.

Table 6.1.1

Test-Taker Test Preparation, Test Anxiety, and Perceptions of Use in Admissions by Race/Ethnicity

Survey question	Asian, Asian American, or Pacific Islander	Black or African American	Mexican, Mexican American, or Chicano	Puerto Rican	Other Hispanic, Latino, or Latin American	White (non-Hispanic)
Degree of preparedness to take the GRE General Test						
Not at all/barely ready when first considering the test	65%	64%	65%	62%	65%	57%
Not at all/barely ready after test preparation	19%	23%	21%	22%	22%	12%
Feelings of anxiety						
Extremely/very anxious when first considering the test	43%	56%	54%	48%	47%	37%
Extremely/very anxious after test preparation	38%	50%	44%	39%	49%	40%

Table 6.1.2

Graduate Admissions Factors Perceived as Most Important by Race/Ethnicity

Factor	Asian, Asian American, or Pacific Islander	Black or African American	Mexican, Mexican American, or Chicano	Puerto Rican	Other Hispanic, Latino, or Latin American	White (non-Hispanic)
Essay/personal statement	24%	22%	27%	17%	22%	21%
Letters of recommendation	11%	12%	10%	15%	16%	14%
Life or work experience	21%	20%	21%	23%	14%	20%
Student's personal qualities	5%	6%	7%	8%	6%	7%
Undergraduate grades/GPA	27%	20%	22%	17%	22%	26%
Match with faculty research interests	6%	7%	6%	9%	8%	6%
GRE Analytical Writing measure	1%	4%	2%	3%	2%	1%
GRE Quantitative Reasoning measure	5%	4%	3%	4%	6%	2%
GRE Verbal Reasoning measure	0%	4%	2%	4%	4%	2%

Note. GPA = grade point average.

Conclusion

The results of this study may have ramifications for modifying the kind of information that is provided to GRE test takers about graduate admissions and about the GRE General Test in particular. Misconceptions about the influence of the GRE can be addressed in any future outreach efforts or marketing campaigns that the GRE Board and the GRE program may decide to undertake. The relatively low response rate also needs to be considered, as those who responded may represent different experiences and perceptions than is typical for all GRE test takers.

Future work on these data should include analysis within ethnicity by gender and major field and a more complete attempt to disentangle the confounding of race/ethnicity, major field, and test results on perceptions of the GRE General Test. More importantly, replicating the results on a sample not already predisposed to graduate study would be beneficial, as students who truly believe the GRE is a potential barrier to graduate school may not have attempted the GRE General Test and, therefore, were not available in this study. Such a study is currently underway.

References

Powers, D. E., & Lehman, J. (1983). GRE candidates' perception of the importance of graduate admissions factors. *Research in Higher Education, 19,* 231–249.

Stricker, L. J., Wilder, G. Z., & Bridgeman, B. (2002). *Test takers' attitudes and beliefs about the Graduate Management Admission Test* (Research Report No. RR-02-10). Princeton, NJ: Educational Testing Service.

Notes

[1] Based on *Test Taker Perceptions of the Role of the GRE General Test in Graduate Admissions: Preliminary Findings,* by F. Cline and D. Powers, 2013, unpublished manuscript, Princeton, NJ: Educational Testing Service.

6.2 Field Trial of Proposed *GRE®* Question Types for Test Takers With Disabilities [1]

Cara Cahalan Laitusis, Lois Frankel, Ruth Loew, Emily Midouhas, and Jennifer Minsky

When a testing program is undergoing a revision, it is important to conduct research on proposed new question types and potential accommodations, to understand the impact that the revised test will have on persons with disabilities, and to ensure that it complies with federal law. As part of the work undertaken to revise the *GRE®* General Test, data collections occurred between 2004 and 2005 that examined the functioning and adaptability of the proposed new Verbal Reasoning (Verbal) and Quantitative Reasoning (Quantitative) question types with examinees with disabilities.

Procedure

Sample

Verbal and Quantitative question types were field tested in September 2004 on a small sample of test takers with a variety of disabilities. A total of 70 test takers participated in the field test. In addition, question-level timing data were collected for the 39 test takers who took the field test in the computer-based test format and for the six test takers with visual impairments who took the field test in an alternate test format (recorded audio).

The test takers who participated in the field test represented six major disability subgroups (i.e., visual, hearing, physical, learning, ADHD, and psychiatric). Although test takers from each disability group were included, we attempted to achieve a sample that was heavily weighted toward test takers who require audio rendering of test questions (i.e., individuals with visual disabilities and reading-based learning disabilities) because it was thought that the new GRE General Test question types might be the most problematic for these individuals.

Although we achieved an adequate sample of audio test takers with low vision and learning disabilities during the first data collection in September 2004, only one test taker who was blind requested an audio format (human reader), and that was in combination with a Braille test form. To increase the feedback from audio-users who are blind, a second data collection involving six such test takers was conducted in February 2005.

Test Format

Testing materials included two test sections (one Verbal and one Quantitative) composed of GRE General Test questions (current) in use at the time the study was conducted and three test sections (two Verbal and one Quantitative) of proposed question types (new). The sections were administered in the following order: Section 1, new Quantitative questions;

Section 2, new Verbal questions; Section 3, new Verbal questions; Section 4, current Quantitative questions; and Section 5, current Verbal questions.

The test section containing current question types were used to establish the test takers' performance levels on the operational GRE General Test and to determine how accessible the proposed new question types were relative to the current question types. The test questions in the sections composed of new question types were the same as those administered during an earlier 2004 pilot (Wendler, Chapter 1.2, this volume), but the order and number of test questions were modified. All test takers received at least five examples of each new question type. Data were collected on test takers' question responses, the amount of time required to complete the question, and any difficulties the test taker encountered while responding to the question.

The new Verbal question types included in the field trial were text completions one, two, and three blanks;[2] sentence equivalence;[3] and paragraph reading (120 words). Four new Quantitative question types were also included: numeric entry (test takers calculated their answer and entered it using the keyboard), multiple-selection multiple choice (test takers select one or more answer choices), order match (test takers select a response that constructs a statement), and table grid (test takers determine if a statement is true or false).

All materials were converted into five accessible formats (script for a reader, large print, Braille, audiocassette/CD, and digital talking book with computer presentation of test questions), and all necessary testing accommodations (e.g., extra time, reader, scribe, paper test, audiocassette/CD, Braille, large print) were provided. Most test takers (39 of 70) took the field test as a computer-based linear test[4] with extended time limits. Test takers who required alternate format tests (31 of 70) took the field test in the required alternate format.

Surveys

Two surveys were also given to participants. A short background survey was administered prior to testing. In order to obtain more information about test-taker reactions to the proposed new question types, all examinees also completed a postadministration survey that included at least three questions about each question type (clarity of question instructions, clarity of question type, degree to which question type measures Verbal/Quantitative reasoning). Test takers who took the test in an alternate format were also asked questions about the question types and about the format in which the question types were administered (e.g., layout preferences and verbal memory load).

Results

Due to the small sample of test takers with disabilities and the variety of accommodations provided, significance testing was not done. Instead, more qualitative approaches were employed. Question-level timing data, performance data, and exit survey responses by type of accommodation and disability were examined. Results provide useful information for this group of test takers.

Timing Results

It is generally true that it takes test takers with disabilities longer to answer a test question than it takes test takers without disabilities. Question-level timing data were available for the test takers without disabilities who participated in an earlier pilot study. While the same questions were used in our field trial as were used in the earlier pilot, the order of presentation was different, so results are not completely comparable. However, comparisons across the two groups provide general information about the functioning of the new question types for test takers with disabilities.

We found that, for several of the proposed question types, the difference in time was considerably greater than it was for the test takers without disabilities who participated in the earlier pilot. The timing data also revealed that test takers with learning disabilities tended to skip over questions with high reading loads (particularly long reading comprehension sets), so the actual times to complete these questions are most likely underestimations of the time required during a high-stakes testing situation.

However, most test takers with learning disabilities, hearing impairments, and physical disabilities who took the computer-based test and whose only accommodation requirement was extended time still completed the sections containing the new question types within the most frequently granted time extension (50% extended time, or time and a half).

Question-level timing data were also collected for the six additional participants who were blind or had low-vision who were tested in February 2005. For this sample, the timing ratio was found to be much larger for the proposed question types than for the current question types. Whether this was related to the format that the test takers used or to the nature of their disability could not be determined.

Exit Survey Results

Exit survey analysis indicated that some question types, such as numeric entry and three-blank text completion, would benefit from clearer directions, particularly for individuals with visual and hearing impairments. From discussions with blind consultants and with experts in the field of blindness, along with our own experience in producing alternate test formats for test takers with visual disabilities, several concerns were identified regarding the accessibility of some of the new proposed question types for this population. The types of questions of greatest concern were those that require significant use of memory (multiple blanks to be filled in, multiple elements to be ordered, selecting or referring back to text in a passage, scratch work for math questions). It should be noted that the latter two concerns are also salient for test takers generally and are not exclusive to individuals with visual impairments.

Comments from our participants about the proposed Quantitative questions included confusion about questions that required decimal-entry or fractional-entry responses. For example, none of the six participants tested in February using audio delivery answered the fractional-entry question type correctly. For proposed Verbal questions, participants indicated that the two-blank text-completion questions were difficult to maintain in memory when the sentences were repeated once for each possible combination of responses (that is, sentences were read a total of 10 times) and less so when the options were presented in place. Overall, participants found the current question types more accessible than the proposed ones, most markedly the Quantitative questions. While it is acknowledged that these results were based on small samples, nevertheless, they provided guidance for the revised test.

Conclusion

For the majority of test takers who received only extra time accommodations, the new question types did not appear to present any major complications. Alternate format test takers, particularly audio users, however, found that completing some of the new question types presented a greater challenge than completing the current question types. Comments received from participants were used in refining the details of how the material in alternate formats would be presented in the GRE revised General Test, as well as changes to accompanying directions. In addition, some of the more problematic question types were made more accessible in a computer-voiced format, which retains some of the interactive character originally designed into the new question types, while making that interactivity more accessible (Frankel & Kirsh, Chapter 6.3, this volume). This field test effort reflects ETS's mission to provide appropriate, valid, and fair tests for all examinees. As much as possible, the results of this field trial were used to guide the final version of the GRE revised General Test.

Notes

[1] Based on *Field Trial of Proposed GRE Item Types for Test Takers With Disabilities,* by C. Laitusis, L. Frankel, R. Loew, E. Midouhas, and J. Minsky, (2005), unpublished manuscript, Princeton, NJ: Educational Testing Service.

[2] The one-blank text completion question was a reformatted version of the sentence completion question type used on the GRE General Test.

[3] The sentence equivalence questions evolved from the vocabulary (synonyms) in context question type.

[4] While at this time the operational GRE was adaptive, in this field trial, test takers received a linear nonadaptive version of the test. All test takers received the same set of questions.

6.3 Development of the Computer-Voiced *GRE*® revised General Test for Examinees Who Are Blind or Have Low Vision

Lois Frankel and Barbara Kirsh

In the late 1990s, prospective *GRE®* General Test examinees with visual disabilities began requesting a computer-based version that would *speak* the test questions and allow them to use audio commands for navigating the test. Because the GRE General Test was delivered on a computer-based platform at testing centers, individuals who were blind wanted to access the test in the same way as other students using the tools they used to access other material in their school and leisure experiences. Personal computers were becoming part of many people's lives, and software that spoke printed text to expand the options for individuals who were blind was easily purchased. In addition, with the passage of the Americans with Disabilities Act (ADA) of 1990[1] and Individuals with Disabilities Education Act (IDEA),[2] the concept of mainstreaming children who were blind or had low-vision (as with children with other disabilities) in regular school classrooms instead of educating them in separate classes or schools was endorsed. Fewer children who were blind were taught by teachers who knew Braille, or who could teach them Braille, and computer voicing began to take precedence over Braille.

With greater legal protection, individuals with disabilities were better supported in educational achievement and had a wider range of employment options. Educational Testing Service (ETS) had long provided test accommodations to individuals with visual and other types of disabilities; the most widely used accommodations for individuals who are blind or have low-vision were Braille, large print, audio cassette, and human readers, along with additional time. To better address the consequences of the passage of the ADA, the ETS Office of Disability Policy was established in 1997. Among its early actions were codifying the process of requesting accommodations, developing and publishing specific guidelines for documenting various disabilities, appointing an external panel of 30 experts in a wide array of disabilities to review accommodations requests under the direction of the ETS director of the Office of Disability Policy, and standardizing the process of deciding upon appropriate accommodations across ETS testing programs.

Focusing on the needs and requests of individuals with disabilities reflects ETS's mission of expanding opportunities to all examinees. Although many of the requests for accommodations emanated from individuals with *invisible* disabilities—learning disabilities and attention deficit hyperactivity disorder (ADHD)—accommodations from prospective examinees with blindness and low vision had the most impact on questions for which alternate test formats would be requested and made available.

Method

The *Voiced* GRE General Test

Creating the test. Early explorations as to how to create a computer-based version of the GRE General Test with *voice* capabilities began with working with an accessible-technology vendor to customize its product to create a stand-alone testing platform that allowed the test and navigation to be voiced and fully keyboard navigable. When this approach did not work, an attempt was made to integrate off-the-shelf text-to-speech software (which provided speech and keyboard navigation) into the standardized computer software and hardware that delivered the GRE General Test. Unfortunately, this second approach did not work either. Instead, as the requests for a voiced version of the GRE General Test increased and digital technology continued to expand with the popularization and decreasing cost of personal as well as educational use, another approach was adopted.

A large cross-functional ETS team responsible for determining business and technical requirements for the voiced version of the GRE General Test was created. While these requirements were being developed, a revised version of the GRE General Test was being developed as well. It was expected that the voiced version of the GRE General Test would be added to the available alternate formats (e.g., Braille, large print, recorded audio, and reader script) for the GRE revised General Test; thus, the team developing the voiced version of the GRE General Test closely followed the progress of the development of the GRE revised General Test. During the development of the GRE revised General Test, several changes in question types and test design were considered. The team working on the voiced version worked to incorporate into the business and technical requirements the most likely features of the revised version, including the ability to review and change responses. However, the initial development of a voiced version of the GRE General Test used the then-current version of GRE General Test and created a stand-alone product rather than incorporating voicing into ETS's server-based standard test delivery platform. This version was released in 2008.

Because the team continued to note the changes that would be needed in order to create a voiced version of the GRE revised General Test, the development of this version of the test was accomplished more quickly and efficiently than would have been possible without the history of developing the initial version. The computer-voiced GRE revised General Test (vGRE) was launched in 2012, within several months of the release of the GRE revised General Test.

Key features of the test. The vGRE is designed for complete eyes-free usability. Users navigate the test and test questions through keyboard shortcuts and a keyboard-accessible menu. All prompts, menus, directions, and test content are delivered via computerized audio along with resizable text and graphics. In most cases, the text also rewraps when resized. All of the test content and interface (including prompts, directions, warnings, confirmations, menu items, etc.) are self-voicing and provide audio guidance to the user. For ease of use, it was desirable to match keyboard commands as closely as possible to those in popular screen readers or to standard Windows keystrokes. This matching was done to the extent possible. However, a complete match to screen readers was not feasible. While the mapping of keystrokes to functionality is similar across popular screen readers, it is far from identical. Furthermore, technical issues prevented capturing certain keys used by many screen readers. For example, the most popular screen reader, JAWS, uses the control key by itself to force immediate silence, but the development platform used by vGRE does not support capturing the control key by itself. In addition, vGRE needed new functions specific to the testing environment, such as selecting the directions, question, or answer choices for playback, or performing functions specific to particular question types. Accordingly, where matching screen reader commands were not feasible and where new commands were needed, mnemonic commands (e.g., alt-N for the next question) were employed to promote usability.

Because the vGRE is intended for test takers with low vision as well as for those who are blind, several features were included to improve accessibility for these test takers. All questions and directions text can be enlarged to nearly any degree desired, and users can select the foreground and background colors they find most readable. To prevent the need for horizontal scrolling when text is enlarged, paragraph text automatically rewraps. Text can be highlighted when spoken, and when a test taker selects an answer choice, that choice, in addition to being spoken for blind users, is shown underlined and boldfaced for better visibility.

The vGRE package, which installs from a CD and runs stand-alone on Windows XP and Windows 7 (as opposed to running from a server), includes an interactive tutorial and context-sensitive voiced help and is accompanied by large print and Braille quick-reference guides that list the menu and keyboard commands for all test functions. All figures in the Quantitative Reasoning measure sections are fully described by the system (descriptions are displayed and spoken on demand) and presented as large as possible on screen. They are also provided in large print and Braille figure supplements. A complete practice test, including the tutorial, is provided to examinees in advance of testing so they can fully acquaint themselves with the interface.

Usability Research

The assessment specialists at ETS who worked on the development of vGRE are professionals with considerable experience developing tests and test preparation materials for people with blindness or low vision. However, that experience cannot substitute for feedback from assistive technology users. Conducting usability studies early in the development cycle, as well as later in the cycle, was important in order to identify any accessibility problems in the interface being developed—thus ensuring that the final product is accessible—as well as to avoid making costly missteps. To ensure that the feedback received from the usability studies was as relevant as possible to the needs of test takers who are blind or have low-vision, usability research was done with both professional accessibility consultants (who are themselves blind or have low vision) and blind or low-vision individuals who are representative of the population for whom the test is intended. One consultation was done early in the development process, and a second consultation plus a usability study were completed close to the release date of the vGRE.

Early consultation. In July 2011, consultations were held with one consultant who is blind and one who has low-vision, both of whom are experts in accessibility. To prepare for these consultations, a series of functionality prototypes and accessibility questions were developed. The consultants were provided with Braille and large print quick-reference guides to the keystrokes and guided through the prototypes. Then several questions about how certain functions might work were posed. For example, it was proposed that when an answer choice is selected, its display would change to bold and underlined and audio feedback would be given (e.g., “Choice C, vindicated, selected.”). Subsequently, moving the cursor to a choice that previously had been selected would result in a sound-cue preceding the speaking of the choice (e.g., “[sound] Choice C, vindicated.”). The consultants trying out these approaches in the prototype provided evaluative comments and suggestions.

The consultants’ comments were evaluated and changes were made where appropriate and feasible. Results showed that many of the proposed approaches were supported by the consultants. The changes included such things as correcting pronunciations of some prompts; adding keys for various functions, such as those to play the answer choices for a given blank in a multiple-blank text completion question (e.g., alt-1, alt-2, and alt-3 now play the choices for the first, second, or third blank); or requiring repetition of certain dangerous keyboard commands (like the one that deletes all responses to a question) as a way to confirm the intended action.

Prerelease consultation. In June 2012, when development was near completion, a two-part usability study was undertaken. One part involved the same two consultants who provided the early consultation, and the second involved four study participants who were blind or had low vision, had educational levels consistent with the GRE test-taking population, and had taken the GRE General Test or were contemplating taking it in the near future.

A study protocol was developed from the protocol used for the first version of the vGRE and updated to reflect changes made for the revised test. In addition, staff members from the ETS User Experience area were consulted to further refine the instrument and observation protocol. The instrument, administered by ETS User Experience staff members who were not involved in the development of the test and so could be considered neutral observers, took each participant through important portions of the vGRE and had the participant interact with the system in prescribed ways. Participants’ interactions with the system were observed, and follow-up questions were asked as needed, based on the protocol. In particular, participants were asked to “answer the test questions as if you were really taking a test” but also assured that their performances would not be scored and would not affect their scores on any future test they might take. They also were provided with figure supplements and quick-reference guides. In the practice test, participants worked with particular questions selected because those questions incorporated as much as possible of the functionality newly developed for the GRE revised General Test.

Results

Results indicated that, despite efforts in producing the text, there were still some words that vGRE did not pronounce sufficiently clearly. In addition, it was found that some users had difficulty using keystrokes that were different from those used in screen readers and that they felt the tutorial was too long. Some concerns mentioned by participants reflected insufficient attention on their part to the tutorial or quick-reference guide. Some of the problematic pronunciations could be and were corrected (but a few were not correctable), and the identified programming bugs were fixed, but not all suggested changes were feasible. For example, the system could not be made to resemble a JAWS screen reader in all respects, nor could the tutorial be shortened given the amount of material that needed to be covered. However, the tutorial is included in the practice material provided to prospective test takers in advance of testing, so they can use it to become familiar with the test interface prior to testing.

Conclusion

The development of the vGRE underscores ETS's mission to provide appropriate, valid, and fair tests for all examinees. The ability to release this version of the test so close to the launch of the GRE revised General Test was accomplished through careful planning and listening to and working with external accessibility experts. The vGRE adds to the list of accommodations available for examinees. As additional versions of vGRE are produced, further enhancements to the system and functionality will be made.

Notes

[1] Americans with Disabilities Act (ADA) of 1990, 42 U.S.C., 12101 *et seq.* (1990).

[2] Individuals with Disabilities Education Act (IDEA) of 1990, 20 U.S.C. 1400 *et seq.* (1990).

6.4 Ensuring the Fairness of *GRE*® Analytical Writing Measure Prompts: Assessing Differential Difficulty [1]

Markus Broer, Yong-Won Lee, Saba Rizavi, and Donald Powers

The *GRE*® program has historically expended considerable effort to ensure that the GRE General Test is fair and equitable and that differences between test takers are due to actual differences in ability rather than a product of unfairness or bias in the test. This concern was fundamental in the creation of the Analytical Writing measure, as well as in the entire effort to revise the GRE General Test. To minimize the likelihood of unfairness in the Analytical Writing measure, test developers craft writing prompts that are as equivalent as possible and function similarly for all test takers. In this way, any between-group difference in performance is due to construct-relevant factors rather than to influences that are irrelevant to the assessment of writing ability.

For multiple-choice questions, well-established methods exist for detecting questions that are differentially difficult for certain subgroups of test takers. The method used for the GRE General Test examines differential performance at the question level (e.g., differential item functioning [DIF]), which occurs when test takers of equal ability but with different group membership (e.g., gender, ethnicity) have unequal probabilities of success on a question (Angoff, 1993; Clauser & Mazor, 1998; Hambleton, Swaminathan, & Rogers, 1991). For multiple-choice measures, the total score on the test is commonly used to determine test takers' ability level.

However, no comparable procedures exist for determining when essay prompts are differentially difficult for subgroups of test takers. One reason for this is the lack of a reliable, internal criterion on which test takers can be matched with respect to the overall ability or skill being measured. This overall matching must be accomplished before between-group performance comparisons can be made on individual questions or, in the case of the GRE Analytical Writing measure, essay prompts. Because the Analytical Writing measure contains only two essay prompts, it is not possible to derive a comparable internal matching criterion, and alternative strategies (e.g., using an external matching criteria such as scores on multiple-choice tests measuring similar abilities) have proven to be less than optimal. A further complication is that essay responses are scored polytomously (that is, the test taker receives a score of 1 to 6), not simply as correct or incorrect, and DIF can occur in some or all score categories (Dorans & Schmitt, 1993; French & Miller, 1996).

The purpose of the current study was to evaluate the usefulness of several alternative DIF methods for detecting GRE essay prompts that are differentially difficult for (a) female test takers; (b) African American, Asian, and Hispanic test takers; and (c) test takers whose best language is a language other than English. Both prompt types used in the Analytical Writing measure, analyze an issue (issue) and analyze an argument (argument), were analyzed. An attempt was also made to compare the impact on DIF estimates of using different matching variables created by combining multiple-choice test scores and essay scores.

Procedure

The study used responses from 397,806 GRE General Test takers who took the test between October 2002 and October 2003. In total, 117 argument prompts and 109 issue prompts were administered to these test takers. Test takers who indicated that English was not their best language were not included in the analyses. Approximately 39% of the sample was male, 60% female, 61% White, 7% African American, 6% Asian, and 2% Hispanic.

Three alternative polytomous DIF detection techniques were used in the analyses:

- The Mantel test of linear association—a generalization of the Mantel-Haenszel procedure that accommodates polytomous questions (Agresti, 1990; Mantel, 1963; Zwick & Thayer, 1996)
- Logistic regression procedures, demonstrated by French and Miller (1996) and Zumbo (1999) to be appropriate for studying polytomous DIF
- Polytomous standardization (polySTAND) statistic, an extension of the standardization approach (Dorans & Kulick, 1986) for polytomous DIF analysis

Test takers were matched by ability level. Ability levels were determined by creating composite scores based on performance on the Verbal Reasoning measure plus performance on the other Analytical Writing prompt (i.e., when the argument prompt was studied, the *other prompt* was the issue prompt and vice versa). To create the composite score, the verbal score and the prompt score were first converted into z-scores. Then the z-scores were converted to a composite score using two different weights. First, the verbal and prompt z-scores were summed and the average of the two found. However, this meant that the prompt score was given equal weight to the verbal score, which was composed of many multiple-choice questions. Therefore, a second composite score was computed in which the weight of the prompt z-score was significantly decreased. Thus, each test taker had four ability estimates that were used as matching criteria: (a) simple average composite for argument, (b) simple average composite for issue, (c) weighted average composite for argument, and (d) weighted average composite for issue.

Because test takers' ability levels were based on their performance on the other prompt, which itself could contain DIF, basing the results on only one DIF detection method could further increase the risk of falsely flagging prompts. Therefore, the results of the three DIF methods were combined and then those prompts that had DIF indexes that exceeded values likely to have a practical impact on test performance were examined.

Six different DIF values were computed for each prompt by using the ability estimates (simple average composite and weighted average composite) with each of the three DIF detection techniques (i.e., Mantel, logistic regression, and polySTAND). DIF values for the prompts were examined in terms of magnitude and direction. These prompts were then ranked from highest to lowest in absolute DIF values separately for each of the six conditions (3 procedures × 2 ability estimates), with rank of 1 indicating the greatest DIF value. Then, an average rank was calculated for each prompt across all six conditions. Finally, an average rank of the procedures was calculated for those prompts that had large enough samples to be studied (or, in the case of the logistic regression procedure, where DIF values were nonzero).

Results

Results indicated that no DIF values were found that were large enough to warrant removal of the prompts from the question pool. The three DIF methods were in substantial agreement in terms of the extent to which they identified prompts having large DIF values. The correlations among DIF values from the three different methods ranged from .83 to .90 for the argument prompts and from .86 to .89 for the issue prompts.

In the gender comparison, most prompts showed low DIF values favoring females. Higher DIF values were observed in the White/African American comparison with some argument prompts being differentially more difficult for African American test takers. Moderate DIF values were also observed in the White/Hispanic comparison on a few argument prompts to the disadvantage of the Hispanic group. Moderate DIF was observed on some issue prompts in the White/African American comparison (favoring White test takers), the White/Asian American comparison (favoring White test takers), and the comparison of test takers who noted English as their best language/not their best language (favoring English-best test takers).

Even though prompts showed a range of DIF values, no prompts were found to have exhibited substantially higher DIF values than the others. However, the research indicates that some combinations of issue and argument prompts (i.e., argument prompts with higher DIF values paired with issue prompts with higher DIF values) should be avoided operationally, especially for African American test takers.

Preliminary analyses of the relationship between DIF values and specific prompt characteristics features (e.g., topic, type of required analysis) found low to moderate correlations. For example, DIF values correlated .29 with whether or not topics dealt with health and safety issues, with such prompts appearing to be differentially easier for women than other topics. Most prompt characteristics, however, showed no relationship at all with DIF values.

Conclusion

The content for all GRE prompts routinely undergo sensitivity review to ensure that prompts are appropriate for all test takers. Results of this study provide additional support for continuing such reviews as part of the GRE revised General Test development process for Analytical Writing measure prompts.

References

Agresti, A. (1990). *Categorical data analysis.* New York, NY: Wiley.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3–23). Hillsdale, NJ: Erlbaum.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17*(1), 31–44.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23,* 355–368.

Dorans, N. J., & Schmitt, A. (1993). *Constructed response and differential item functioning: A pragmatic approach* (Research Report No. RR-91-47). Princeton, NJ: Educational Testing Service.

French, A. W., & Miller, T. R. (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement, 33,* 315–332.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Mantel, N. (1963). Chi-square test with one degree of freedom: Extensions of the Mantel-Haenszel procedure. *Journal of American Statistical Association, 58*, 690–700.

Zumbo, B. D. (1999). A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores. Ottawa, Ontario, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zwick R., & Thayer, D. T. (1996). Evaluating the magnitude of differential item functioning in polytomous items. *Journal of Educational Statistics, 21,* 187–201.

Notes

[1] Based on *Ensuring the Fairness of GRE® Writing Prompts: Assessing Differential Difficulty* (Research Report No. RR-05-11), by M. Broer, Y.-W. Lee, S. Rizavi, and D. Powers, 2005, Princeton, NJ: Educational Testing Service.

6.5 Effect of Extra Time on Verbal and Quantitative *GRE*® Scores[1]

Brent Bridgeman, Frederick Cline, and James Hessinger

Time limits serve at least two functions when used with a standardized test: (a) the rate at which an individual completes the test may be a critical component of what the test is measuring, and (b) even if time limits are not integral to the test, they may still be indispensable in helping to keep testing costs down. The *GRE*® General Test has specific time limits; thus, it is important to measure the impact, if any, of time limits on examinees' scores. Understanding the impact of time also helped inform the time limits for the GRE revised General Test.

Previous research showed that examinees who had greater time and fewer questions achieved an average score gain of 25 points on the 200–800 GRE scale (Wild, Durso, & Rubin, 1982). Research with the *SAT*® suggested that greater score gains are seen for high-ability level examinees (Bridgeman, Trapani, & Curley, 2004). However, these results were derived from studies conducted on other tests as well as an earlier, paper-and-pencil version of the GRE General Test and, therefore, are not directly applicable to a computer-delivered GRE General Test. Thus, the purpose of this study was to examine the extent to which the results of previous research are applicable to the computerized GRE General Test.

Procedure

Upon completion of the GRE General Test, examinees were invited to take an additional section of the GRE General Test as part of a research study. A total of 15,948 examinees provided usable data. Participation was voluntary, and all participants were offered an incentive to perform well.

Participants took either a Verbal Reasoning or Quantitative Reasoning measure as the research section. The research section was identical to the operational test with the exception of the timing. Participants were randomly assigned to a group using standard testing time or a group using 1.5 times the standard testing time on the research section. Thus, participants belonged to one of four groups (verbal standard, verbal 1.5 times, quantitative standard, and quantitative 1.5 times). Participants were also placed into one of three ability levels based on their operational GRE score: (a) low (200–500), (b) mid (510–700), and (c) high (710–800).

Results

The benefit from extra time for the total sample on the Verbal Reasoning and Quantitative Reasoning measures was only 7 points on average for each measure. In terms of participant race/ethnicity, the effects of extra time seemed to be relatively constant across all observed categories (African American, Asian American, Hispanic, and White) for both verbal and quantitative measures.

Participants in the low-ability group appeared to be impacted by additional time more so than those in the high-ability group. On the quantitative measure, the difference in average score was greater for participants in the low-ability group compared to those in the high-ability group (21 points vs. 6 points). Results were similar for the verbal measure, where participants in the low-ability group had an average difference of 13 points compared to an average difference of 4 points for the high-ability group. While the 21-point and 13-point differences may seem substantial, in the context of the 200–800 score scale, they are quite small; the 21-point difference on the quantitative measure reflects a standardized difference of only 0.17.

Conclusion

While the finding of greater impact of extended time on lower ability students is contradictory to findings in previous studies, an important difference between this study and previous ones is the test administration method. With a paper-and-pencil test, lower ability examinees will probably not see an increase in their scores because it is unlikely that they will be able to answer the more difficult questions that appear at the end of the test form, regardless of the time allotted. However, because the computer-delivered GRE General Test tailors questions to the examinees' ability level, there is an increased probability of them correctly answering questions at the end of the section. The minimal impact of the timing condition on the high-ability examinees for both the verbal and quantitative measures could be due to the fact that they came into the research study with very little room for improvement.

On the whole, the timing for the GRE General Test had only a small effect on overall scores. These findings are comparable to those from previous research (Bridgeman et al., 2004; Wild et al., 1982). Additionally, time allotted seem to have no differential impact on test performance across race/ethnicity.

However, some of the results of this study were contradictory to findings in previous studies, and these differences were attributed to the design of the test. Since the GRE revised General Test uses a different testing approach (that is, multistage vs. computer-adaptive), close attention to understanding the impact of time limits for the GRE revised General Test is imperative. As a result, field trials were designed to ensure that appropriate time limits were established for the revised test (see Wendler, Chapter 1.2, this volume).

References

Bridgeman, B., Trapani, C., & Curley, E. (2004). Impact of fewer questions per section on SAT I scores. *Journal of Educational Measurement*, *41*, 291–310.

Wild, C. L., Durso, R., & Rubin, D. B. (1982). Effects of increased test-taking time on test scores by ethnic group, years out of school, and sex. *Journal of Educational Measurement*, *19*, 19–28.

Notes

[1] Based on "Effect of Extra Time on Verbal and Quantitative GRE® Scores," by B. Bridgeman, F. Cline, and J. Hessinger, 2004, *Applied Measurement in Education*, *17*, pp. 25–37.

6.6 Fairness and Group Performance on the *GRE*® revised General Test

Frédéric Robin

As a high-stakes standardized testing program, the *GRE®* General Test has long been concerned with fairness. Therefore, a comprehensive reexamination of fairness issues and approaches to address them was conducted throughout the development of the GRE revised General Test launched in August 2011. As a result, the GRE Verbal Reasoning, Quantitative Reasoning, and, to a lesser extent, the Analytical Writing measures have changed in a number of ways that enhance the fairness and validity of the test (Briel & Michel, Chapter 1.1, this volume). Most visibly, the test content, the score scale, and the testing experience, in which test takers can move forward and backward throughout an entire section, changed. The test design also changed from item-level[1] computerized adaptive tests (CAT) to section-level multistage adaptive (MST), which allows for a more controlled test assembly process and the approval of every test form before delivery (Robin & Steffen, Chapter 3.3, this volume).

This chapter first outlines the approaches and methods used to ensure fairness of the revised GRE Verbal Reasoning, Quantitative Reasoning, and Analytical Writing measures. It then provides a brief summary of the main pilot and field test studies conducted in preparation for the launch of the GRE revised General Test, focusing on fairness issues. Finally, this chapter documents important GRE and revised GRE testing outcomes related to fairness and shows the extent to which fairness goals have been achieved and are being maintained over time.

Approaches and Methods to Ensure Fairness

Following the professional standards for educational and psychological tests (American Educational Research Association [AERA], American Psychological Association [APA], & the National Council on Measurement in Education [NCME], 1999) and the ETS Standards for Quality and Fairness (Educational Testing Service [ETS], 2002), the GRE revised General Test was designed and developed to satisfy the following three critical fairness goals:

1. All tests are free of bias.
2. All test takers are given comparable opportunity to demonstrate their ability.
3. Observed group differences are not exacerbated.

Ensuring Nonbiased Tests

For tests to be free from bias, it is necessary that the questions included in the test be free from bias themselves. To achieve that goal, GRE questions go through a development process that includes following extensive fairness guidelines when authoring them and conducting comprehensive fairness reviews in order to avoid including content that may advantage or disadvantage any specific group (ETS, 2005, 2009a, 2009b). Then, before new questions are used to assemble new test forms, empirical pretest data are collected and statistical analyses are conducted to detect questions that do not perform the same across equally able groups. Because of limitations in the size of the pretest samples, these statistical analyses are only conducted on gender groups. However, as questions are used and reused in operational tests, additional data are collected and statistical analyses for the regional, gender, and racial/ethnic groups are conducted when enough data has accumulated (Robin & Steffen, Chapter 3.3, this volume).

The GRE program routinely performs two types of item analyses aimed at detecting potentially biased Verbal Reasoning and Quantitative Reasoning questions: differential item functioning (DIF) and item model data fit (Fit). DIF analyses compare question performance across two groups matched in ability (Holland & Thayer, 1988). As typically implemented at ETS, questions are classified into one of three categories according to the extent to which they are found to favor one group versus the other: negligible DIF (A-DIF), intermediate DIF (B-DIF), and large DIF (C-DIF; Dorans & Holland, 1993, p. 42). Fit analyses compare the performance of a group on a particular question with the performance expected for that group according to the model used to produce the reported score[2] (Hambleton, Swaminathan, & Rogers, 1991). Fit analyses are used to detect questions that exhibit significant misfit and, therefore, may introduce bias into the scoring process. Both methods are used for the GRE revised General Test in a complementary manner: Fit analyses are aimed at detecting questions that do not fit the scoring model for regional groups, and DIF analyses are aimed at detecting questions that show differential performance between domestic gender and between domestic racial/ethnic groups.

As a result of these analyses, Verbal Reasoning and Quantitative Reasoning questions are continuously screened to ensure that no potentially significant bias is present in any test. Pretested questions classified as C-DIF are discarded and, thus, will never be included in any operational test form. Operational questions classified as C-DIF or misfitting after enough operational test data have been collected to permit analyses are removed from the operational question bank. Thus, because the operational Verbal Reasoning and Quantitative Reasoning banks are very large and new questions are pretested in relatively small batches, the number of C-DIF and/or misfitting questions that may be included in a test will not be consequential.

The Analytical Writing measure is made up of two essay tasks scored by one or more human raters, depending on the degree of agreement with the *e-rater*® automated scoring engine (ETS, 2013). In this situation, the development of effective DIF analysis is a challenge, as it requires the availability of a reliable ability estimate (matching variable), which cannot be obtained from one essay alone (the other essay being the subject of the DIF analysis). Broer, Lee, Rizavi, and Powers (Chapter 6.4, this volume) developed a procedure to address this challenge[3] and were able to evaluate the pool of Analytical Writing essay prompts later revised for use with the GRE revised General Test (Robin & Kim, Chapter 2.3, this volume). Overall, their results indicated that for gender and racial/ethnic groups no DIF values were large enough to warrant removal of prompts from the pool of prompts. Such analyses will be repeated as the Analytical Writing pool is maintained.

Providing Comparable Testing Opportunity

Ensuring that all test takers are provided with a comparable opportunity to demonstrate their ability has several aspects (ETS, 2002). One is to make sure test information and practice materials are comprehensive and broadly available. For the GRE program, this is done through its website: https://www.ets.org/gre. Another is to make sure test centers are accessible to all potential test takers. This has been achieved through the development of a dense network of test centers with a worldwide reach[4] (Briel & Michel, Chapter 1.1, this volume). Another is to provide standardized and secure testing conditions at every test center, which is achieved by only using testing centers committed to following ETS's comprehensive test center procedures.

Ensuring that some test takers do not benefit from knowledge of previously administered questions is another aspect of comparability. For this purpose, the GRE program develops and maintains a large pool of questions and takes advantage of new automated test assembly processes and a new delivery infrastructure to produce large numbers of tests to be delivered and rotated out on a nearly continuous basis (Robin & Steffen, Chapter 3.3, this volume). In this way, it is very unlikely that knowledge of previously administered questions or tests provides an advantage to test takers.

Ensuring that all test forms provide comparable measurement[5] is also critical. To accomplish this, the GRE program has implemented extensive quality control procedures to check and approve every test form before it is delivered (Robin & Kim, Chapter 2.3, this volume; Robin & Steffen, Chapter 3.3, this volume). A close monitoring of test taker and test interactions[6] and testing outcomes has also been implemented to provide the necessary feedback to ensure that the time allocated remains sufficient and that the MST development process remains effective, as item resources are renewed and large numbers of MST forms are continually produced and delivered over time.

Investigating Group Differences

True differences in test performance among test takers grouped according to indicators of, for example, country of origin or educational background, are never known. Nevertheless, based on prior studies and testing experience, some levels of group difference are often expected. As stated in the joint standards (AERA, APA, & NCME, 1999, p. 75), "[W]hile group differences in testing outcomes should trigger heightened scrutiny for possible sources of test bias, outcome differences across groups do not in themselves indicate that a testing application is biased or unfair."

However, even though bias may not be present, it is still important to make sure that the diversity of the populations is taken into account when designing or revising a test (ETS, 2002, p. 20) or when monitoring testing outcomes to ensure that the observed differences continue to be the best estimates possible.

In the GRE General Test case, the testing population is very diverse in background and educational experience both in the United States and across the world. Therefore, test takers are categorized into different regional groups: U.S. citizens testing in domestic test centers (Dom); test takers testing in China, Hong Kong, Korea, or Taiwan (Asia); in India or Japan (Id/Jp); and in any other international test center (OInt). Several reasons justify this grouping, most importantly test takers' distinct cultural, language, and educational background. Practical reasons related to the frequency of testing and the reach of specific test center networks also justify these groupings.

Domestic test takers, who constitute about two thirds of the GRE population, are further divided into self-reported gender (female and male) and racial/ethnic (Asian, Black, Hispanic, and White) groups. For these groups, particular attention is required, as indicated by the AERA, APA, and NCME joint standards: "There [are] legal requirements to investigate differences in outcomes among [such] subgroups. Those requirements further may provide that, other things being equal, a testing alternative that minimize outcome differences…should be used" (AERA, APA, & NCME, 1999, p. 76).

Test fairness for much smaller groups, such as those taking the test on paper or receiving testing accommodations (together less than 2% of the total population), needs to be paid attention to as well. As with the previous version of the GRE General Test, a large majority of questions on the GRE revised General Test do not require modifications when delivered on paper. Also, the ability for test takers to skip and revisit questions on the GRE revised General Test means that the computer and paper versions are now more similar than previously. Therefore, previous research on the older paper and computer versions of the test, which concluded that the two versions were sufficiently similar (Gallagher, Bridgeman, & Cahalan, 2000; Schaeffer, Reese, Steffen, McKinley, & Mills, 1993; Schaeffer, Steffen, Golub-Smith, Mills, & Durso, 1995), provided some support for the comparability of the revised computer and paper versions. The fairness of test accommodations was also an important consideration in the development of the GRE revised General Test. For more information, see Cahalan Laitusis, Frankel, Loew, Midouhas, and Minsky (Chapter 6.2, this volume) and Frankel and Kirsh (Chapter 6.3, this volume).

Pilot and Field Test Studies

From 2003 to 2010, a number of pilot and field test studies were conducted to examine proposed new question types and new test configurations for each of the three GRE General

Test measures (Wendler, Chapter 1.2, this volume). From a fairness point of view, the evaluations were conducted to ensure that (a) the new item types were appropriate for all groups of examinees, (b) the workload (as a result of test length and time allocated) was appropriate for all test takers, and (c) the new test did not exacerbate differences in performance between gender and between racial/ethnic groups.

The primary goal of the pilot studies was to evaluate the functioning of proposed new question types. Pilot data were collected at the end of regular operational test administrations by asking volunteer test takers to take a 30 to 45 minute research section approximating a half-length linear test form. As is customary with such studies, data from participants whose test results showed clear lack of motivation were screened out before analyses were conducted. After screening for motivation, the remaining pilot data sets were large enough[7] to obtain stable analyses results.

DIF analyses, as described earlier in this chapter, were conducted for gender and racial/ethnic groups when sample sizes were sufficient. Any question that was classified as B-DIF or C-DIF was reviewed by content specialists, and those thought to be problematic were eliminated from the pool of questions that were to be used with the GRE revised General Test.

The primary goal of the subsequent field test studies was to evaluate the potential of alternative full-length linear test configurations that had been refined based in part on the pilot study results. The field test administrations were conducted independently from operational administrations, mostly in the United States and at a small number of overseas locations. Field test participants who had taken or were planning to take the GRE General Test were recruited for these special administrations and were administered one or more complete forms. After screening for motivation, the remaining data sets were large enough to include in the analysis for most groups except the Black and Hispanic groups.[8]

Pilot and field test analyses included the computation of standardized group differences[9] for the male–female (M-F), White–Black (W-B), and White–Hispanic (W-H) groups. The operational standardized group differences based on the study participants' operational scores were also computed to serve as a benchmark for evaluating alternative test configurations.

Table 6.6.1 shows the standardized group differences obtained across the 2003 to 2005 data collection events. Because the changes in the Verbal Reasoning measure were the most extensive, more data collection events occurred for this measure (four pilots and two field tests) than for other measures (for a detailed description of these changes, see Briel & Michel, Chapter 1.1, this volume). For the Quantitative Reasoning measure, four pilots and one field test were conducted, and for Analytical Writing, one field test was conducted. At the bottom of the table, the results obtained from the full year 2006 and 2007 operational administrations are provided for comparison.

Table 6.6.1

Standardized Group Differences

	Verbal			Quantitative			Writing		
Data collection event	W-B	W-H	M-F	W-B	W-H	M-F	W-B	W-H	M-F
2003–2004									
Pilot	**0.68**	**0.41**	**0.26**	**0.95**	**0.46**	**0.57**			
Operational	1.12	0.63	0.27	1.15	0.51	0.56			
Pilot				**0.73**	**0.45**	**0.56**			
Operational				1.07	0.55	0.61			
2004–2005 a									
Pilot (33, 35) [a]	**0.9**	**0.55**	**0.25**	**0.76**	**0.36**	**0.64**			
Operational	1	0.66	0.3	1.03	0.47	0.63			
Pilot (42, 40) [a]	**0.82**	**0.57**	**0.23**	**0.74**	**0.5**	**0.61**			
Operational	0.95	0.71	0.27	0.99	0.68	0.61			
2004–2005 b									
Pilot (33, 35) [a]	**0.8**	**0.57**	**0.21**	**0.88**	**0.43**	**0.55**			
Operational	1	0.7	0.23	1.08	0.56	0.58			
Pilot (42, 40) [a]	**0.88**	**0.51**	**0.23**	**0.85**	**0.52**	**0.53**			
Operational	1.03	0.66	0.21	1.14	0.62	0.55			
2004–2005 c									
Pilot	**0.82**	**0.38**	**0.33**						
Operational	0.97	0.49	0.3						
2005									
Field Test-A	**0.94** [b]	**0.5** [b]	**0.12** [b]						
Operational	0.91	0.51	0.16						
Field Test-B	**0.95**	**0.46**	**0.19**						
Operational	0.88	0.5	0.16						
2005									
Field Test-L	**1.07** [b]	**0.38** [b]	**0.19** [b]	**1.04** [b]	**0.29** [b]	**0.55** [b]	**0.76** [b]	**0.45** [b]	**-0.09** [b]
Operational	1.04	0.28	0.1	1.03	0.27	0.43	0.87	0.48	-0.06
Field Test-M	**1.16** [b]	**0.55** [b]	**0.27** [b]	**0.85** [b]	**0.37** [b]	**0.47** [b]			
Operational	1.03	0.43	0.17	0.9	0.47	0.45			
2006 Operational	0.94	0.55	0.26	1.05	0.54	0.56	0.86	0.41	0.09
2007 Operational	0.93	0.56	0.25	1.05	0.54	0.55	0.85	0.43	0.08

Note. Adapted from "A Review of Subgroup Differences for the Revised GRE," by M. Golub-Smith, 2008, unpublished manuscript, Statistical Analysis Center, Educational Testing Service, Princeton, NJ. W-B = White–Black; W-H = White–Hispanic; M-F = male–female.

[a] The first and second numbers in parenthesis indicate the number of minutes allotted for the Verbal Reasoning and Quantitative Reasoning measures, respectively. [b] Each set of pilot and field test results, which is in boldface to facilitate comparisons with the set of results below, obtained based on the operational scores obtained by the same test takers.

While, given the studies' limitations, no single result presented in Table 6.6.1 can be seen as conclusive, the pattern of Verbal Reasoning and Quantitative Reasoning results suggested relatively little sensitivity to the pilot timing and content configurations for all the groups. Then, focusing more on the field test results that were based on data collected from more refined full-length test administrations, the pattern of results suggested that only small increases in group differences may occur for the Verbal Reasoning M-F, W-B, and W-H groups. The Quantitative Reasoning and Analytical Writing results suggested that no changes would occur. It was concluded, therefore, that the GRE revised General Test could be configured to produce the desired level of measurement without exacerbating the differences in performance across gender and major racial/ethnic groups.

When it was decided in 2008 to move to an MST design for the Verbal Reasoning and Quantitative Reasoning measures, the test configuration was further adapted and additional studies were conducted to finalize the GRE revised General Test configuration (Wendler, Chapter 1.2, this volume). This resulted in the configuration in use since the test was launched in 2011. For a description of the final test configuration, see Briel and Michel (Chapter 1.1, this volume).

Testing Outcomes

Testing outcomes were extensively scrutinized as soon as data were available following the launch of the GRE revised General Test. When sufficient data were collected, the new Verbal Reasoning and Quantitative Reasoning scales were established, and the new interpretive information such as score percentile ranks, reliability, and standard error of measurement statistics were made available for all three measures (ETS, 2011). Since the launch, critical test outcomes related to test fairness continue to be monitored on a monthly, quarterly, and yearly basis. This section provides a summary of such outcomes before and after the launch of the GRE revised General Test based on data collected during the 2009 and 2012 years (avoiding the potentially atypical 2010–2011 period from the announcement of the revised tests to several months after its launch).

Comparable Opportunity

One aspect of providing a comparable testing experience to test takers involves ensuring that, within the allowed section time, they are able to manage their time, revisit questions, and answer most, if not all, of them. Table 6.6.2 provides a summary of time management results obtained across the regional groups for the Verbal Reasoning and Quantitative Reasoning measures in 2012. Issues related to Analytical Writing time configurations are discussed elsewhere (see Broer et al., Chapter 6.4, this volume; Robin & Zhao, Chapter 1.8, this volume).

Table 6.6.2

Group Time Management by Section

	Verbal				Quantitative			
	S1	S21	S22	S23	S1	S21	S22	S23
Domestic test center								
% answering 80% items	99	99	100	100	95	98	97	97
Mean # rapid response	0	0	0	0	0	0	0	0
Mean # questions revisited	7	6	8	10	6	5	6	6
Asia								
% answering 80% items	99	100	100	100	100	98	99	100
Mean # rapid response	1	1	0	0	1	1	0	0
Mean # questions revisited	8	8	9	11	12	6	8	10
India or China								
% answering 80% items	99	100	100	100	99	99	99	99
Mean # rapid response	1	0	0	0	0	0	0	0
Mean # questions revisited	9	9	10	11	9	9	9	9
Other international test center								
% answering 80% items	99	99	100	100	96	97	97	98
Mean # rapid response	1	1	0	0	1	1	0	0
Mean # questions revisited	7	6	8	11	7	5	6	7

Note. Each Verbal Reasoning or Quantitative Reasoning multistage administration includes a first middle-difficulty section (S1), followed by either an easy (S21), medium (S22), or difficult (S23) section, depending on the level of performance on S1. Asia = China, Hong Kong, Korea, or Taiwan.

For the Verbal Reasoning measure, more than 97% of test takers in each group were able to complete at least 80% of the questions in the first and second sections of the test without apparent recourse to rapid responses (Mean # rapid responses—spending less than 10 seconds on a question—was 1 or less). For the Quantitative Reasoning measure, more than 95% of test takers in each group were able to complete at least 80% of the questions without apparent recourse to rapid responses. This is a slightly lower completion rate than the one observed for Verbal Reasoning. But this slight decrease appears to be related to the use of numerical entry questions, rather than insufficient testing time. Quantitative Reasoning numerical entry questions require creating an answer, while the other question types only require selecting an answer from a number of available alternatives, which makes it easy to guess. In fact, we do see in the data that numerical entry questions are omitted at a noticeable rate, while it is not the case with any of the other question types.

With both measures, test takers appeared to have enough time to revisit about 5 to 11 questions per section, on average (Mean # questions revisited), depending on their level of ability.

Altogether these results provide evidence that the test provided enough time for all test takers to demonstrate their ability.

Group Differences

As indicated earlier, standardized differences allow us to examine performance differences between two groups. For the international groups (Asia, India/Japan, and other international), differences were computed with reference to the domestic group, with positive numbers indicating better performance for the reference group (Dom). Similarly, gender differences were computed with reference to domestic male test takers and racial/ethnic differences with reference to domestic White test takers.

Table 6.6.3 shows a clear increase in standardized differences on all measures for the Asia group between 2009 and 2012, with lower average Verbal Reasoning and Analytical Writing performance relative to that of the domestic reference group in 2012 (0.85 vs. 0.65, and 1.26 vs. 1.05) and with higher Quantitative Reasoning relative performance (-1.83 vs. -1.46). These results coincided with a 50% increase in Asia testing volumes between 2009 and 2012. It is conceivable, therefore, that changes in the Asia population, as well as the revisions in the Verbal Reasoning content, may have contributed to that result. Also, measurement at the top of the Quantitative Reasoning scale has been enhanced. It is likely, therefore, that having more points to distinguish between test-taker performance toward the top of the scale contributed to larger standardized differences between the domestic and Asia groups. A similar pattern was found with the India/Japan group and, to a much lesser degree, with the other international group. The India/Japan volumes also increased (by about 25%) between 2009 and 2012.

Table 6.6.3 also shows that the changes in the domestic gender and ethnicity standardized differences between 2009 and 2012 were small or nonexistent. This means that the revisions to the GRE General Test did not affect the magnitude of group differences between males and females and among racial/ethnic groups.

Conclusion

Achieving and maintaining a high degree of fairness for all individuals and groups of individuals requires the implementation of an appropriate test design and an effective infrastructure for delivery, the ongoing development of large numbers of questions, the ongoing assembly and quality control of large numbers of test forms, and the consistent monitoring of testing outcomes. This chapter outlined how the GRE program has been striving and continues to strive to meet these requirements. The summary results reported here document the extent to which the GRE revised General Test has fulfilled the GRE program's goals in the relatively short period since its launch. Now, because of the ongoing assembly and delivery of large numbers of MST forms, a lot of attention is devoted to monitoring and maintaining the level of fairness achieved and to further investigate the potential for further fairness improvements. For example, with longitudinal data becoming available over a longer period of time, more attention is now being devoted to investigating such issues as identifying and addressing the drift in the measurement properties of some questions that could occur over time, as well as the drift in the meaning of the reporting score scale.

Table 6.6.3

Standardized Differences for Testing Groups

	% of total sample [a]		Verbal		Quantitative		Writing	
	2009	2012	2009	2012	2009	2012	2009	2012
Dom (reference)	69	63	-	-	-	-	-	-
Asia	8	12	0.65	0.84	-1.46	-1.83	1.05	1.26
Id/Jp	8	10	0.60	1.07	-0.94	-0.64	1.20	1.20
OInt	16	16	0.56	0.63	-0.51	-0.45	0.75	0.75
Male (reference)	36	34	-	-	-	-	-	-
Female	64	58	0.29	0.34	0.57	0.59	0.08	0.06
White (reference)	73	70	-	-	-	-	-	-
Asian	6	6	0.05	0.15	-0.41	-0.46	0.07	0.07
Black	9	8	0.92	1.00	1.06	1.01	0.84	0.83
Hispanic	7	7	0.52	0.50	0.51	0.46	0.45	0.41

Note. Dom = domestic test centers; Asia = China, Hong Kong, Korea, or Taiwan; Id/Jp = India or Japan; OInt = other international test center.

[a] For the gender and racial/ethnic groups, the total sample is the domestic sample. For these groups, percentages do not add up to 100% because of missing responses and the smaller racial/ethnic groups that are not included.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Dorans, N. J., & Holland, P. W (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer, H. (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale, NJ: Erlbaum.

Educational Testing Service. (2002). *ETS standards for quality and fairness*. Princeton, NJ: Author.

Educational Testing Service. (2005). *Constructed-response guidelines*. Princeton, NJ: Author.

Educational Testing Service. (2009a). *ETS fairness review*. Princeton, NJ: Author.

Educational Testing Service. (2009b). *ETS international principles for fairness review of assessments*. Princeton, NJ: Author.

Educational Testing Service. (2011). *GRE guide to the use of scores*. Princeton, NJ: Author.

Educational Testing Service. (2013). *GRE guide to the use of scores*. Princeton, NJ: Author.

Gallagher, A., Bridgeman, B., & Cahalan, C. (2000). *The effect of computer-based tests on racial/ethnic, gender, and language groups* (GRE Board Report No. 96-21P). Princeton, NJ: Educational Testing Service.

Golub-Smith, M. (2008). *A review of subgroup differences for the revised GRE*. Unpublished manuscript, Statistical Analysis Center, Educational Testing Service, Princeton, NJ.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Erlbaum.

Schaeffer, G. A., Reese, C. M., Steffen, M., McKinley, R. M., & Mills, C. N. (1993). *Field test of a computer-based GRE General test* (GRE Board Report No. 88-08P). Princeton, NJ: Educational Testing Service.

Schaeffer, G. A., Steffen, M., Golub-Smith, M. L., Mills, C. N., & Durso, R. (1995). *The introduction and comparability of the computerized adaptive GRE test* (GRE Board Report No. 88-08aP). Princeton, NJ: Educational Testing Service.

Notes

[1] Because of the technical nature of this document, the psychometric term *item* is used instead of *test question* when discussing analyses, processes, or technical results.

[2] For GRE, the model used to produce the reported score is the 2-parameter item response theory model (Robin & Steffen, Chapter 3.3, this volume).

[3] They found that an effective matching variable could be obtained by using a weighted combination of the test-taker scores on the Verbal Reasoning measure and the Analytical Writing task not investigated for DIF.

[4] A list of test centers and testing dates are available at http://www.ets.org/gre. Prometric test centers are used in most of the world (https://www.prometric.com) and ETS accredited test centers in some countries.

[5] See the *GRE Guide to the Use of Scores* (ETS, 2013, Tables 5, 6A, and 6B) for more details on reliability and standard error of measurement.

[6] Test taker and test interactions included the number of questions answered, number of times each question was visited, time spent on a specific question, and so forth.

[7] Sample sizes were larger than 1,000 for White, male, and female groups and larger than 250 for Hispanic and Black groups.

[8] Sample sizes were larger than 500 for White and female groups, larger than 250 for male groups, but only around 45 for Hispanic and Black groups.

[9] The standardized group difference is equal to the difference between the average scores of the two groups divided by the pooled standard deviation. Therefore, a standardized difference of 1.0 corresponds to approximately 1 standard deviation on the operational scale: about 120 points on the prior Verbal Reasoning and Quantitative Reasoning scales and about 8 points on the revised scales.

The Research Foundation for the *GRE®* revised General Test:

A Compendium of Studies

Listening. Learning. Leading.®

www.ets.org