




TOEFL[®]

Research Reports

RR - 72

March 2005

A solid red vertical bar is positioned to the left of the main title text.

An Investigation of the
Impact of Composition
Medium on the Quality
of TOEFL Writing Scores

Edward W. Wolfe

Jonathan R. Manalo

**An Investigation of the Impact of Composition Medium
on the Quality of TOEFL Writing Scores**

Edward W. Wolfe
Virginia Tech, Blacksburg

Jonathan R. Manalo
ETS, Princeton, NJ

RR-04-29



ETS is an Equal Opportunity/Affirmative Action Employer.

As part of its educational and social mission and in fulfilling the organization's non-profit Charter and Bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

Copyright © 2005 by ETS. All rights reserved.

No part of this report may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission in writing from the publisher. Violators will be prosecuted in accordance with both U.S. and international copyright laws.

EDUCATIONAL TESTING SERVICE, ETS, the ETS logos, Graduate Record Examinations, GRE, Praxis Series: Professional Assessments for Beginning Teachers, TOEFL, and the TOEFL logo are registered trademarks of Educational Testing Service. The Test of English as a Foreign Language is a trademark of Educational Testing Service.

College Board is a registered trademark of the College Entrance Examination Board.

Graduate Management Admission Test and GMAT are registered trademarks of the Graduate Management Admission Council.

Abstract

This study examined scores from 133,906 operationally scored Test of English as a Foreign Language™ (TOEFL®) essays to determine whether the choice of composition medium has any impact on score quality for subgroups of test-takers. Results of analyses demonstrate that (a) scores assigned to word-processed essays are slightly more reliable than scores assigned to handwritten essays and exhibit higher correlations with TOEFL multiple-choice subscores; (b) female test-takers, examinees whose native language is not based on a Roman/Cyrillic alphabet, and examinees with lower English proficiency are more likely to choose the handwriting medium; (c) the probability of choosing handwriting as the composition medium increases with age for Asian examinees, but decreases with age for most European examinees; and (d) examinees with lower TOEFL multiple-choice scores tend to have higher handwritten than word-processed essay scores, while examinees with higher TOEFL multiple-choice scores tend to have similar scores in either medium.

Key words: Composition medium, computer-based testing, direct writing assessment, English proficiency, English as a Foreign Language, English as a Second Language, TOEFL writing test

The Test of English as a Foreign Language™ (TOEFL®) was developed in 1963 by the National Council on the Testing of English as a Foreign Language. The Council was formed through the cooperative effort of more than 30 public and private organizations concerned with testing the English proficiency of nonnative speakers of the language applying for admission to institutions in the United States. In 1965, Educational Testing Service® (ETS®) and the College Board® assumed joint responsibility for the program. In 1973, a cooperative arrangement for the operation of the program was entered into by ETS, the College Board, and the Graduate Record Examinations® (GRE®) Board. The membership of the College Board is composed of schools, colleges, school systems, and educational associations; GRE Board members are associated with graduate education.

ETS administers the TOEFL program under the general direction of a policy board that was established by, and is affiliated with, the sponsoring organizations. Members of the TOEFL Board (previously the Policy Council) represent the College Board, the GRE Board, and such institutions and agencies as graduate schools of business, junior and community colleges, nonprofit educational exchange agencies, and agencies of the United States government.



A continuing program of research related to the TOEFL test is carried out in consultation with the TOEFL Committee of Examiners. Its members include representatives of the TOEFL Board and distinguished English as a second language specialists from the academic community. The Committee advises the TOEFL program about research needs and, through the research subcommittee, reviews and approves proposals for funding and reports for publication. Members of the Committee of Examiners serve four-year terms at the invitation of the Board; the chair of the committee serves on the Board.

Because the studies are specific to the TOEFL test and the testing program, most of the actual research is conducted by ETS staff rather than by outside researchers. Many projects require the cooperation of other institutions, however, particularly those with programs in the teaching of English as a foreign or second language and applied linguistics. Representatives of such programs who are interested in participating in or conducting TOEFL-related research are invited to contact the TOEFL program office. All TOEFL research projects must undergo appropriate ETS review to ascertain that data confidentiality will be protected.

Current (2004-2005) members of the TOEFL Committee of Examiners are:

Catherine Elder (Chair)	Monash University
Deena Boraie	The American University in Cairo
Micheline Chalhoub-Deville	University of Iowa
Glenn Fulcher	University of Dundee
Marysia Johnson Gerson	Arizona State University
April Ginther	Purdue University
Bill Grabe	Northern Arizona University
Keiko Koda	Carnegie Mellon University
David Mendelsohn	York University
Tim McNamara	The University of Melbourne
Terry Santos	Humboldt State University

To obtain more information about the TOEFL programs and services, use one of the following:

E-mail: toefl@ets.org
Web site: www.ets.org/toefl

Acknowledgements

This project greatly benefited from the input of our colleagues. Specifically, Claudia Gentile provided input into the design and data collection for this study. Pat Carey, Robbie Kantor, Yong-Won Lee, Philip Oltman, and Ken Sheppard each provided guidance in obtaining and interpreting the data. In addition, James Algina, Paul Allison, Robert Brennan, Chris Chiu, Ken Frank, David Miller, and Mike Patetta gave us advice during various stages of data analysis. Finally, Dan Eignor, Paul Holland, Hunter Breland, Drew Gitomer, and Yong-Won Lee provided thoughtful suggestions for improving a previous version of this report. We also thank the Test of English as a Foreign Language (TOEFL) program for providing us with the funds to carry out this work. Finally, we want to point out that this report is based on data that have been presented in three different papers at professional conferences (Breland, Muraki, & Lee, 2001; Manalo & Wolfe, 2000a, 2000b).

Table of Contents

	Page
Introduction.....	1
Prior Research.....	1
Implications of Inequities in Technology Access.....	1
Impact of Word Processors on Writing Education.....	5
Impact of Computers on Testing.....	6
Impact of Word Processors on Testing.....	10
Implications for TOEFL.....	12
Purpose.....	14
Research Questions.....	14
Method.....	15
Participants.....	15
Instrument.....	15
Analysis.....	17
Measures of the Quality of the Ratings.....	17
Correlational Analyses.....	18
Group Characteristics.....	19
Group Comparisons.....	25
Results.....	29
Descriptive Statistics.....	29
Comparisons of the Quality of the Ratings.....	30
Correlational Analyses.....	31
Group Characteristics.....	31
Group Comparisons.....	43
Summary of Results.....	45
Quality of Ratings.....	45
Group Characteristics.....	45
Essay Performance.....	46
Discussion.....	47
Implications.....	48

Future Research	50
References.....	52

List of Tables

	Page
Table 1. Medium-Choice Model Selection Summary	24
Table 2. Essay Score Model Comparisons.....	28
Table 3. Essay Score Means for Each Covariate	29
Table 4. Descriptive Statistics.....	30
Table 5. Indices of Rating Quality	30
Table 6. Correlations Between TOEFL Sections by Composition Medium.....	31
Table 7. Characteristics of Examinees Who Chose Handwriting Versus Word-Processing	32
Table 8. Conditional Probabilities of Choosing Handwriting Versus Word-Processing.....	33
Table 9. Summary of the Parameter Estimates for the Final Model.....	41
Table 10. Model Comparisons of Handwritten and Word-Processed Essay Scores	44
Table 11. Essay Means Conditioned on Composition Medium by English Proficiency	44

List of Figures

	Page
Figure 1. Conventional Test Performance Model.	2
Figure 2. Computer-Based Test Performance Model.....	3
Figure 3. Medium-Choice Linearity of Logits for Age.....	21
Figure 4. Linearity of Essay Scores Across English Proficiency Levels.	27
Figure 5. Medium-Choice Continent-by-Gender Interaction.....	34
Figure 6. Medium-Choice Continent-by-Keyboard Interaction.....	35
Figure 7. Medium-Choice Age-by-English Interaction.	36
Figure 8. Medium-Choice Continent-by-English Interaction.	37
Figure 9. Medium-Choice Age-by-Continent Interaction.....	42

Introduction

Recently, the format of the Test of English as a Foreign Language™ (TOEFL®) examination changed in two ways: (a) the test is now administered via computer, and (b) the test includes a section requiring examinees to write an essay (i.e., a direct writing assessment). Taken together, these changes could introduce several potential sources of measurement error into TOEFL scores—sources of error that might reduce the reliability and validity of examination scores. This study aimed to identify the seriousness of these potential sources of measurement error.

Prior Research

Our discussion of the literature focuses on four topics: (a) the implications of inequities in access to technology, (b) the impact of word processors on writing education, (c) the impact of computers on testing, and (d) the impact of word processors on testing. Much of the literature upon which we draw is based native English-speaking populations from the United States. In cases where relevant research is available for English as a Second Language (ESL) or international populations, that fact is highlighted.

Implications of Inequities in Technology Access

Scores from standardized tests heavily influence selection decisions made by educational institutions and certification decisions made by professional organizations. Increasingly, selection and certification tests are administered via computer. There are several reasons for the pervasive shift from a conventional to a computer-based testing format: reduced testing time, standardized test administration, test content tailored to examinee ability, automated scoring, and faster score reporting (Wise & Plake, 1989). The implementation of computer-based selection and certification testing has improved the way tests are administered and test scores are reported. However, this shift toward a technology-based testing system may exacerbate existing social barriers to advancement opportunities for women, minorities, and economically disadvantaged individuals.

Consider the model of factors that influence performance on a conventional test shown in Figure 1. Performance is directly positively or negatively influenced by an examinee's degree of (a) achievement, (b) test preparation, and (c) test anxiety, along with a host of other unnamed variables that may contribute error to measures of ability as operationalized by test performance (e.g., health, testing environment). Previous research suggests relationships among the elements

of this model. Test performance is influenced by an examinee's academic achievement which, in turn, is influenced by opportunity to learn (Wiley & Yoon, 1995). Test preparation, which influences test performance, is also likely to influence test anxiety (Powers, 1993; Powers & Rock, 1999). Finally, it is clear that an examinee's social background influences opportunity to learn, test preparation, and achievement levels (Kim & Hocevar, 1998; Powers, 1993; Turner, 1993). Hence, it seems that social circumstances play an important role in determining performance on conventional tests.

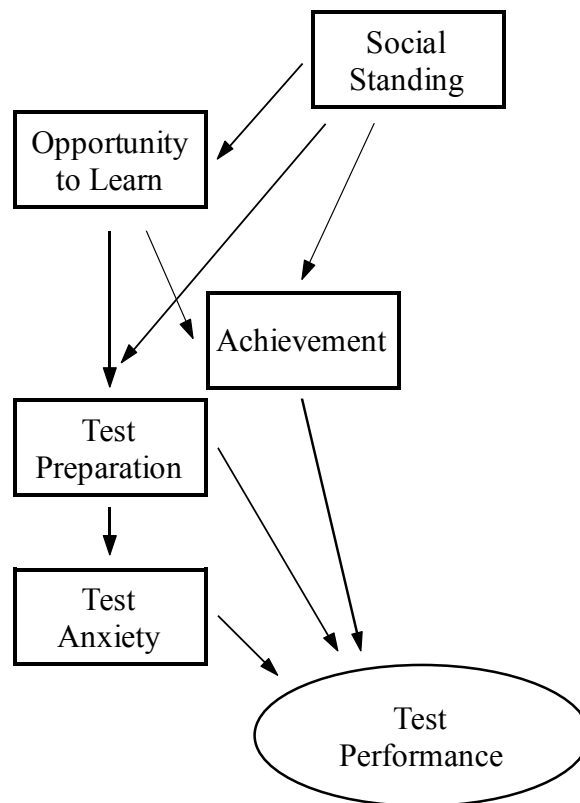


Figure 1. Conventional test performance model.

Now consider how the model changes for computer-based testing (see Figure 2). In this case, at least three variables in addition to achievement, test preparation, and test anxiety must be added as direct influences on test performance: (a) computer skill, (b) computer anxiety, and (c) computer attitudes. Prior research suggests that these three variables influence each other and that computer anxiety contributes to test anxiety (Shermis & Lombard, 1998). In addition, the computer variables are influenced by the examinee's exposure to computers, which is, in turn,

influenced by the examinee’s social background. Hence, computer-based testing may increase the influence of social background on an examinee’s test performance.

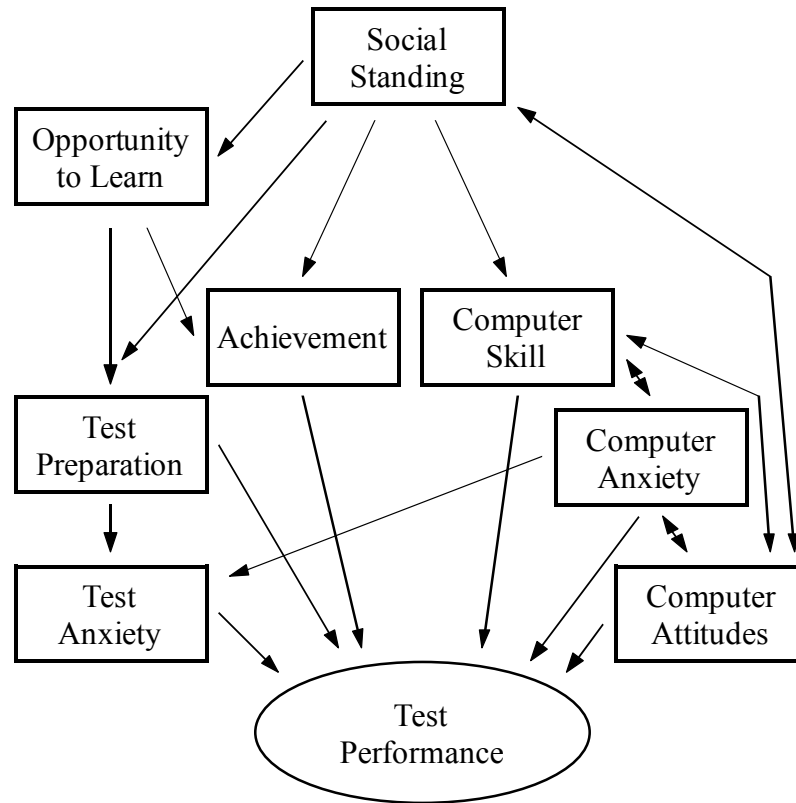


Figure 2. Computer-based test performance model.

What evidence exists to support this model? We already know that female and minority students have fewer opportunities to use computers and develop basic computing skills. Male students are more likely to have computers in their homes, and they dominate computer use in schools—the primary source of computer access for disadvantaged minority students and female students (Campbell, 1989; Grignon, 1993; Miller & Varma, 1994). Although the percentage of the TOEFL testing population having low levels of computer experience may not be large—perhaps as low as 16% (Taylor, Kirsch, Eignor, & Jamieson, 1999)—minority and female test-takers may make up the bulk of this group; hence, these test-takers may have considerably less experience and knowledge and may be less interested and confident in using computers than their White male counterparts (Grignon, 1993; Loyd & Gressard, 1986; Massoud, 1992; Shashaani, 1997; Temple & Lips, 1989). Unfortunately, experiential inequities may be magnified

by existing social norms that portray females as being less capable computer users (Temple & Lips, 1989; Whitely, 1997).

In addition to inequities in computer access related to gender and ethnicity, U. S. populations exhibit differences in computer use. Specifically, males are more likely than females to use software and video games that contain sophisticated graphics (Grignon, 1993). In addition, males are more likely to enroll in computer courses and major in computer science than are females (Grignon, 1993; Temple & Lips, 1989). The tendency for males to dominate computer use in school may interfere with computer-based learning for females (Keogh, Barnes, Joiner, & Littleton, 2000). For example, males are more likely to learn computer-programming skills in school, while females are more likely to learn to program at home (Grignon, 1993).

Inequities in computer access and experience lead to higher levels of anxiety toward computer-based tasks (Loyd & Gressard, 1986). Although a large percentage of the population may suffer from computer anxiety—up to 30% by some accounts (Marcoulides, 1988; Miller & Varma, 1994; Tseng, Macleod, & Wright, 1997)—elevated levels of computer anxiety and negative attitudes toward computer use are found among minorities and females (Campbell, 1989; Gressard & Loyd, 1987; Legg & Buhr, 1992; Levin & Gordon, 1989; Loyd & Gressard, 1986; Whitely, 1997). Interestingly, when computer use and availability are held constant, gender differences in anxiety and attitude cease to exist (Campbell, 1989; Gressard & Loyd, 1987).

Although much of this research is based on U. S. populations, similar trends have been observed in international populations. Specifically, males tend to dominate computer use (Janssen Reinen & Plomp, 1993), although efforts have been made to equalize computer access opportunities based on gender, ability, and socioeconomic status (Janssen Reinen & Plomp, 1993; Nolan, McKinnon, & Soler, 1992). Still, gender differences exist with respect to computer access, with males having increased access to computers (Miller & Varma, 1994; Taylor et al., 1999), being more confident and interested in computers (Siann, Macleod, Glissov, & Durndell, 1990), and having better attitudes toward computers and computer-based instruction (Levin & Gordon, 1989). Gender-stereotyped attitudes about females being less capable computer users also exist at an international level, as does the relationship between computer access and computer anxiety and attitudes (Levin & Gordon, 1989; Miller & Varma, 1994; Siann et al., 1990). At least one study focusing on international learners suggests that increases in computer experience diminish gender differences in computer attitude (Miller & Varma, 1994; Siann et al.,

1990). Finally, geographic differences in computer access have also been noted (Taylor et al., 1999). Specifically, Spanish speakers exhibit relatively lower levels of computer familiarity and Japanese speakers exhibit higher levels. Interestingly, Latin Americans (many of whom speak Spanish) have the highest reported levels of computer familiarity while Africans have the lowest.

Impact of Word Processors on Writing Education

Before addressing the implications that differences in computer experience, attitudes, and anxiety have for test performance, consider the impact that word-processing has had on writing education practice with both English-speaking populations and ESL populations in the United States. U. S. writing educators have suggested that numerous instructional benefits may result from using word processors in writing education programs. In general, students of all ages enjoy using word processors for writing, and students' attitudes toward writing (both on and off of the computer) improve when students are given the opportunity to write with word processors (Cochran-Smith, 1991). These favorable attitudes may result from the ways that word processors facilitate editing and revising. While students may be discouraged from making extensive revisions to handwritten essays because of the difficulty of revising writing by hand (Daiute, 1986), the text editing tools available in most word-processing programs facilitate some aspects of the revision process. As a result, word-processed essays seem to be longer (Bradley, 1982; Broderick & Trushell, 1985; Collier, 1983; Hawisher, 1987; Kane, 1983), neater (Bridwell, Sirc, & Brooke, 1985), and contain fewer mechanical errors (Levin, Riel, Rowe, & Boruta, 1985) than handwritten essays. In addition, the public display of text on the computer monitor may facilitate social aspects of writing in the writing classroom, such as the sharing and discussion of ideas (Bruce, Michaels, & Watson-Gegeo, 1985; Dickenson, 1986; MacArthur, 1988).

Interestingly, prior research concerning the influence of word processors on writing instruction in the United States suggests that the use of word processors has mixed effects on the quality of student writing (Cochran-Smith, 1991). Even though word-processor features may make text revision easier, the availability of common editing tools found in word-processor packages may facilitate surface-level changes (e.g., spell-checking or page formatting) in student writing rather than deeper, meaning-based changes, particularly for students who are experienced with word processors (Dickenson, 1986; Hawisher, 1987; Kurth, 1987; Lutz, 1987; Oweston, Murphy, & Wideman, 1992). For students with less word-processing experience, the added cognitive demands of composing at a keyboard may divert students' attention away from the

quality of their writing (Cochran-Smith, Paris, & Kahn, 1991), and quality of writing may actually decrease when word processors are the medium of composition (MacArthur, 1988). That is, students with less word-processing experience may be distracted from the writing task because of a lack of keyboarding skills and may spend much of their preliminary writing time acquainting themselves with the layout of the keyboard and the functions of the word processor (Dalton & Hannafin, 1987; Porter, 1987).

Only a few studies have focused directly on ESL writing education. A summary of that research echoes the research on U. S. populations just described (Pennington, 1993). Specifically, Pennington concludes that using word processors to teach ESL writing (a) improves the quality of writing, (b) changes the activities in which writers engage, (c) increases revision behaviors, and (d) changes affective and social outcomes of writing education.

Three recent studies are consistent with these trends. Specifically, using networks in an ESL writing class improves the quality of writing and increases peer and teacher feedback (Braine, 1997), and word-processor use results in changes in revision behaviors (Kehagia & Cox, 1997). Finally, case studies on four ESL writers (high and low English proficiency crossed with one and two semesters of computer-based writing experience) reveal that experience with the computer is a stronger factor than writing proficiency in determining computer-based writing strategies (Phinney & Khouri, 1993).

For writing education, differences in writing quality that are due to mode of composition will likely decrease as technology becomes more commonplace in the home, schools, and workplace. But, currently, not all students have equal access to computers, and differences in writing quality that can be attributed to a student's lack of access to computers (rather than lack of writing ability or skill) constitute a source of construct-irrelevant variance in the context of a writing assessment (that is, a source of variability in test scores that has nothing to do with students' writing skills). Hence, an important issue is whether the use of word processors on an ESL writing test makes the writing task more difficult for some portions of this population.

Impact of Computers on Testing

More specifically, we should address whether (and, if so, how) differences in computer access influence scores on standardized direct writing assessments that are delivered by a computer. Little research has been performed to address this issue, particularly in the area of foreign language testing. However, a considerable amount of research has focused on the

influence of computers on scores from standardized tests composed of multiple-choice questions.

Generally, studies of overall differences in scores on computer-based and conventional tests designed to be interchangeable have been inconclusive. Some studies have found no differences between test scores from the two media (Ford, Romeo, & Stuckless, 1996). Only a few studies have revealed higher scores on computer-based versions of the tests (Mazzeo & Harvey, 1988; Russell, 1999). Other studies have shown that computer-based tests may be more difficult than conventional tests. For example, in a review of several studies of the interchangeability of computer-based and conventional tests, Mazzeo and Harvey found more studies in which conventional test scores were higher than their computer-based counterparts. Similarly, a meta-analysis of several studies in which computer and conventional test scores were compared revealed higher scores on conventional tests (Mead & Drasgow, 1993). However, in both of these cases, the magnitude of the medium effect was small. Interestingly, Russell found that students generally believe that they will receive higher scores on computer-based tests. In addition, a few studies have suggested that scores from computer-based and conventional tests may not be interchangeable without first applying some sort of statistical procedure, such as equating, to ensure score comparability. Correlations between parallel computer-based and conventional tests are found to be high in general, but in some cases may be below optimal levels (.72 to .79) for alternate forms (Lee, 1986; Mead & Drasgow, 1993).

It is important to point out, however, that these studies examined overall differences in the difficulty and association between computer-based and conventional tests. None of them examined the influence of the scores on individual examinees. It may be that although large overall effects have not been found, computer-based testing has a large impact on a small portion of the population (Wise & Plake, 1989). As suggested by previous sections of this review, the small portion of the population for whom this large impact might be expected would contain a disproportionate number of minority test-takers for U. S. populations and a disproportionate number of female, Spanish-speaking, and African test-takers for international populations. In fact, at least two studies focusing on U. S. populations have suggested that computer experience contributes to the interchangeability of computer-based and conventional tests (Lee, 1986; Spray, Ackerman, Reckase, & Carlson, 1989).

We also know that administering tests on a computer may alter examinees' cognitive and affective responses to the testing environment. For example, students commit different types of

mathematical errors in a computer-based medium than in a conventional medium (Ronau & Battista, 1988). Others have found similar differential affective influences between conventional and computerized tests (Lankford, Bell, & Elias, 1994; Signer, 1991). One of the few studies of the relationship between computer anxiety and experience on computer-based achievement test scores revealed that computer anxiety is an important predictor of computer-based achievement scores (Marcoulides, 1988). Experience was important too, but not as significant a predictor as anxiety.

However, a study of several large-scale examinations—the Graduate Record Examinations[®] (GRE[®]) General Test, the Graduate Management Admission Test[®], and the Praxis Series: Professional Assessments for Beginning Teachers[®]—conducted by Gallagher, Bridgeman, and Cahalan (2002) failed to indicate that racial/ethnic groups exhibit the performance deficit that the previously mentioned research implies. In fact, they concluded that, although the differences between groups were small, improved performance on computer-based tests of Hispanic and African American test-takers may be greater than it is for White test-takers. However, they also found that the scores of female examinees on computer-based tests were lower than scores from paper-and-pencil tests on some versions of these examinations.

In light of the trends outlined in this section, an important issue is whether training designed to familiarize examinees with the computer interface improves test performance. Fortunately, research suggests that it does. For example, elderly examinees who received one hour of training designed to familiarize them with the computer keyboard and screen received higher scores on a computer-based intelligence test than elderly examinees who did not receive the training (Johnson & White, 1980). Similarly, once examinees were trained to use a computer-based testing interface, additional assistance (e.g., proctors answering questions during test administration) did not influence test performance (Powers & O'Neill, 1993). However, the additional assistance did improve examinee attitudes toward computer-based testing.

Because these studies focus solely on U. S. examinee populations, their relevance to international language-testing populations may be questionable. What do studies of the impact of computer-based testing on ESL and international populations suggest about the comparability of computerized and conventional test formats? Two reviews relating to computerized language testing suggest that the same issues that are important for U. S. populations are of concern for ESL and international populations (Brown, 1997; Henning, 1991). Specifically, these reviews suggest

that (a) presentation of a test in a computerized medium may lead to different results than in a conventional medium, (b) differences in student computer familiarity, experience, and anxiety may lead to discrepancies in computerized and conventional test scores, and (c) computer-based testing tutorials diminish this effect. However, at least two smaller, more localized studies provide evidence contrary to these trends. ESL examinees at Brigham Young University overwhelmingly found a computer-based ESL test to be less stressful than a conventional test (Madsen, 1991). In addition, Madsen found that computer experience was not related to computer anxiety for these students. Similarly, computer-based ability estimates were actually *higher* than conventional ability estimates for ESL examinees (Stevenson & Gross, 1991).

However, the only large-scale study we were able to find (in addition to portions of Stevenson & Gross's 1991 study) jibes with the conclusions of the Brown (1997) and Henning (1991) reviews. Specifically, the Taylor et al. (1999) study of the TOEFL test reveals moderate-to-high correlations (ranging from .54 to .84) between scores on conventional and computer-based tests that were designed to be interchangeable. Their results also agree with the general observation that less computer familiarity leads to poorer performance on computer-based tests. However, in the Taylor et al. study, the difference in examination scores attributed to computer familiarity was confounded by the fact that examinees were allowed to choose their testing medium. It is reasonable to argue that examinees who chose to take the examination on a computer might have had higher levels of English language skill to begin with due to the general availability of educational opportunities that correlate with computer access. Interestingly, in the Taylor et al. study, the use of analysis of covariance (ANCOVA) procedures to partial out performance on a paper-and-pencil version of the TOEFL test revealed no meaningful relationship between the performance of computer-familiar and computer-unfamiliar examinees. However, as pointed out in that study, the ANCOVA procedures may have underestimated the actual difference between these groups.

Finally, the tutorial effects discussed previously might also generalize to ESL and international populations (Jamieson, Taylor, Kirsch, & Eignor, 1999). Specifically, Jamieson et al. revealed that most TOEFL examinees who participated in a study of a computer-based testing tutorial designed for nonnative English speakers were successful in completing the tutorial and thought that the tutorial was helpful. Interestingly, Jamieson et al. also found that computer familiarity and English ability both proved to be important in explaining differences in the time

required to complete the tutorial and the perceived effectiveness of the tutorial.

Impact of Word Processors on Testing

All of the studies cited in the previous section about the impact of computers on testing focused on content areas other than writing. Obviously, responding to a computer-based multiple-choice test is very different than composing an essay for a direct writing assessment in a computer-based environment. So, we must determine whether these trends generalize to computer-based direct writing assessments. Unfortunately, we could locate no studies of the influence of computers on the quality of large-scale direct writing assessment scores of international and ESL populations. Hence, the following discussion focuses only on studies performed on U. S. populations.

For years, those who develop and validate direct writing assessments have been concerned that the mere appearance of an essay in typed print may introduce a source of measurement error—differential reader perception. Prior research on U. S. populations suggests that readers do not treat essays that are written on a word processor in the same way they treat handwritten essays. Specifically, readers may perceive word-processed essays to be shorter and less developed, and raters may have higher expectations for word-processed essays than they do for handwritten essays (Arnold et al., 1990; Gentile, Riazantseva, & Cline, 2001). As a result, readers may tend to assign higher scores to handwritten essays. Fortunately, readers can be trained to compensate for their perceptions, reducing the influence of their individual preferences on the reliability and validity of examination scores (Powers, Fowles, Farnum, & Ramsey, 1994). And at least for examinees with high levels of experience using word processors for writing, the increased standardization offered by word-processed texts results in better agreement between readers (Bridgeman & Cooper, 1998).

However, even when reader effects are controlled for, there are differences between the qualities of essays written in these two composition media. Two common approaches to controlling for reader perception are to (a) transpose essays into a common text medium or (b) have examinees write essays in both media. Studies employing these strategies have resulted in discrepant conclusions about differences between scores of handwritten and word-processed essays. Some studies have favored word-processed essays (Russell, 1999; Russell & Haney, 1997); other studies have favored handwritten essays (Gentile et al., 2001; Wolfe, Bolton, Feltovich, & Bangert, 1996; Wolfe, Bolton, Feltovich, & Niday, 1996). Meanwhile, other studies

have revealed only small differences between the two media (Collier & Werier, 1995). While it did not examine a large-scale direct writing assessment, a companion study to the one reported here (Gentile et al., 2001) found that ESL students who write essays in both handwriting and on a word processor tend to create handwritten essays that are rated higher than their word-processed counterparts in the areas of development, organization, language use, and mechanics. Regardless of overall differences in essay quality, all studies that have focused on content have revealed qualitative differences in essays written by hand versus word processor. Specifically, word-processed essays may contain shorter sentences (Collier & Werier, 1995), be better organized in terms of paragraphing (Russell & Haney, 1997), contain fewer mechanical errors (Gentile et al., 2001), be neater, more formal in tone, and exhibit a weaker voice (Wolfe, Bolton, Feltovich, & Niday, 1996).

With respect to the influence of experience on writing quality, the synthesis of several studies suggests that there is an interaction between computer experience and the quality of essays written by examinees. Specifically, it seems that examinees with high levels of computer experience receive higher scores on word-processed essays, while examinees with lower levels of computer experience receive higher scores on handwritten essays. For example, professional writers who write using word processors and are asked to compose both handwritten and word-processed essays seemed to struggle to adapt their word-processor writing strategies to the handwriting medium (Collier & Werier, 1995). As a result, scores of the quality of initial handwritten essays were lower than scores on word-processed essays. Similarly, Russell and Haney (1997) found that, in contrast to scores from handwritten administration, computer administration of a writing test resulted in higher scores for students in a technology-oriented school. In a similar study, Russell (1999) found that students with slower keyboarding skills received lower scores on a computer-administered language arts test. Two studies (Wolfe, Bolton, Feltovich, & Bangert, 1996; Wolfe, Bolton, Feltovich, & Niday, 1996) revealed that students who have higher levels of computer experience score equivalently on handwritten and word-processed writing assessments, while students with lower levels of computer experience score considerably higher on handwritten essays. Interestingly, the imposition of a word-processing medium may influence both the quality of essays and the affective state of examinees with low levels of computer experience. For example, Wolfe, Bolton, Feltovich, and Niday (1996) discovered that these examinees might make self-

deprecating remarks or write about their emotional reactions to the test setting when exposed to a computer-based direct writing assessment.

Implications for TOEFL

This literature review reveals several important issues about the relationship between computer experience, computer attitudes, and computer skill that may have implications for TOEFL examinees. First, there are inequities in the degree to which some students have access to and familiarity with computers. Second, inequities in computer access and familiarity may lead to higher levels of anxiety toward computer-based tasks. Third, anxiety levels are greatly diminished when computer experience is held constant, and efforts to increase the computer experience of international learners decrease gender differences in computer attitudes.

The review also reveals two important issues related to the use of word processors in writing that may impact TOEFL test-takers. First, when U. S. students with less word-processing experience use word processors for writing, their attention may be diverted away from the quality of their writing. Second, for ESL students, keyboard familiarity may be more influential than writing proficiency on the quality of the writing produced for computer-based assessments. Both of these issues suggest that existing social inequities in advancement opportunities may be exacerbated if important selection or certification decisions are made about nonnative English speakers based on scores on writing assessments that are delivered using word processors as the composition medium.

With respect to the influence of computers on testing in general, there seem to be five important conclusions. First, computer-based tests are probably more difficult, on average, than conventional tests, though mean differences in test scores are not large between these two testing media. Interestingly, the commonly held belief of students that they will receive higher scores on computer-based tests may drive some students to select a testing medium on which they will receive lower scores. Second, even though average differences between test scores from computer-based versus conventional tests are not large, the impact may be great for portions of the examinee population, such as U. S. female and minority test-takers in the United States and female, African, and Spanish-speaking test-takers internationally. Third, we know that computer-based testing invokes different cognitive and affective responses on the part of examinees than does conventional testing, and unfortunately, affective responses—like computer anxiety, computer proficiencies, and levels of computer experience—are correlated with test scores on

computer-based tests at nontrivial levels. Fourth, fortunately, training examinees to use the computer interface reduces the negative impact of these variables on computer-based test scores. And finally, research about computer-based testing with international populations has revealed that lower levels of computer familiarity may lead to poorer performance on computer-based tasks. However, these differences are minimized when the influence of examinee ability is removed.

Our search for literature on this topic uncovered no studies of the influence of computers on the quality of large-scale direct writing assessment scores of international and ESL populations. However, given the parallel trends between U. S. populations and international and ESL populations in the areas of computer access, computer anxiety, use of word processors in writing instruction, and influence of computers on testing, we believe that there is good reason to believe that the observed trends in U. S. populations in the area of direct writing assessment will apply to ESL and international populations as well. That literature suggests the following. First, the appearance of essays as handwritten versus typed text may influence raters. Second, the use of word processors seems to influence the content of essays regardless of whether word-processed essays are of higher or lower quality. Finally, there seems to be an interaction between computer experience or proficiency and the quality of handwritten and word-processed essays.

We believe this literature suggests that the current practice of administering the TOEFL writing test using word processors should be carefully examined to determine whether some examinees may be unintentionally and unknowingly disadvantaged by the testing medium. Of course, one of the purposes of the writing test is to serve as a general indicator of the quality of an examinee's writing. If, for some examinees, scores on the two composition media differ by more than would be expected by random variation, we might conclude that the validity of scores from the writing test varies depending on the composition medium. The practical implications for the TOEFL program, if this type of differential prediction is uncovered in the current writing section, might result in either a reconsideration of administration procedures, test preparation practices, or directions to examinees for this section of the test.

Purpose

This report summarizes a study of the influence of composition medium on scores assigned to essays written for the TOEFL writing section. The study was designed to determine:

- whether word-processed and handwritten TOEFL essay scores exhibit similar levels of interrater reliability
- whether word-processed and handwritten TOEFL essay scores exhibit correlations with other components of the TOEFL that are of similar magnitude
- the characteristics of examinees who are “at risk” with respect to differential, cross-medium performance on the TOEFL writing assessment
- the extent to which examinees with comparable levels of English language proficiency receive comparable scores on word-processed and handwritten TOEFL essays

Research Questions

We addressed the following research questions:

- Are there differences in the interrater reliabilities of the scores assigned to essays composed in each composition mode?
- Are there differences in the degree to which scores assigned to essays that are composed in each medium correlate with scores from other sections of the TOEFL test?
- Are there differences between the characteristics of examinees who choose to compose essays using each mode of composition?
- Are there differences in the magnitude of the scores assigned to essays composed in each mode?
- Are there differences in the magnitude of the scores assigned to essays composed in each mode, once the influence of English language proficiency is taken into account?
- Are groups identified as being potentially “at risk” by prior research more likely to exhibit inconsistent performance in the two modes of composition than other groups of examinees?

Method

In this study, scores from a large number of operationally scored TOEFL essays were subjected to four types of analyses: (a) measures of the quality of the essay ratings, (b) correlational analyses of TOEFL subscales, (c) logistic regression modeling of demographic variables as predictors of medium choice, and (d) analysis of variance and analysis of covariance (and analogous logistic regression) procedures with medium choice as predictor of essay score.

Participants

In all, 133,906 TOEFL examinees—a small portion of the total number of test-takers who participated in regular administrations of the computer-based TOEFL test between January 24, 1998, and February 9, 1999—provided complete data for the study. Participants were from 200 countries and represented 111 different languages. There were slightly more males than females (54% versus 46%). Examinees ranged in age from 15 to 55 years, the average age being 24.26 years. The majority of examinees took the TOEFL test for admittance into undergraduate or graduate studies (38% and 46%, respectively); only 15% indicated that they were taking the TOEFL exam for reasons other than to satisfy academic requirements. Each examinee completed the multiple-choice section of the examination in a computer-based testing environment, but had the choice to respond to the single direct writing assessment prompt either using a word processor (54%) or in handwriting (46%).

Instrument

The computer-based TOEFL consists of four sections: (a) listening, (b) structure, (c) reading, and (d) writing. The first three sections are composed of multiple-choice items, and the fourth is a direct writing assessment. The first three tests are fixed-length with a variable number of pretest questions. The listening and structure sections are administered as computer-adaptive tests (i.e., the test is tailored so that items are selected to match the examinee's ability), and the reading section is administered as a linear on-the-fly test (i.e., a configuration of four reading comprehension sets from a pool of such sets is determined individually for examinees as they take the test, in such a manner that the test specifications are met for each examinee). As stated previously, examinees may choose to respond to the writing assessment in either handwriting or using a word processor.

The listening section measures the examinee's ability to understand English as it is spoken in North America. Questions require examinees to comprehend main ideas, the order of a process, supporting ideas, and important details; to draw inferences; and to categorize topics or objects after listening to information contained in recorded stimuli. Typically, an examinee takes about 40 to 60 minutes to respond to the 30 operational (and up to 20 pretest) questions from the listening section.

The structure section measures the examinee's ability to recognize language that is appropriate for standard written English using written stimuli. Questions require examinees to complete sentences and identify unacceptable words or phrases. Typically, an examinee takes about 15 to 20 minutes to complete the 20 operational (and up to 5 pretest) questions from the structure section.

The reading section measures the examinee's ability to read and understand short passages that are similar to those contained in academic texts used in North American colleges and universities. Examinees read each passage and answer questions that require comprehension of main ideas, factual information, pronoun referents, and vocabulary, as well as inferential reasoning from the written stimuli. Examinees take between 70 and 90 minutes to complete the 44 operational (and up to 16 pretest) questions in the reading section. Typically, there are four or five passages of 250 to 350 words each, with between 10 and 14 questions per passage (ETS, 1999).

The writing section measures the examinee's ability to write English, including the ability to generate, organize, and develop ideas; to support those ideas with examples or evidence; and to compose a response to a single writing prompt in written English. Examinees are given 30 minutes to complete the writing section. Trained TOEFL readers score the essay. Readers are trained to interpret TOEFL standards, score across multiple topics, and use the Online Scoring Network software—the vehicle through which essays are distributed to readers and scores are recorded. To be certified as a TOEFL reader, trained readers must pass a test to verify that they understand and can apply TOEFL scoring criteria. Prior to each operational scoring session, readers must score a set of calibration essays that contains both handwritten and word-processed essays to ensure that they are scoring accurately. During operational scoring, readers have access to the TOEFL scoring guide and training notes (ETS, 1999). For this study, 294 readers rated responses to 58 different prompts. No effort was made to equate these prompts

in terms of their difficulties; however, because the set of prompts was randomly allocated to examinees, this fact did not pose a serious problem for this study.

Readers were randomly selected to score packets of essays. Two readers independently rated each essay on a scale ranging from 1 to 6, each unaware of the score assigned by the other reader. Our analyses of reader agreement focused on the individual ratings assigned by these two readers.

Operationally, an examinee is assigned the average of the two readers' ratings unless there is a discrepancy between the ratings (a discrepancy is defined as scores that are more than two points apart). In the case of discrepant scores, a third "senior" reader independently rates the essay, and that third score replaces one of the two original ratings. Hereafter, the final score—whether derived from two or more ratings—is referred to as the *writing composite score*. Our correlation and group comparison analyses focused on this score.

Scores from the listening and reading TOEFL sections are scaled to range from 0 to 30. Scores for the structure and writing sections are combined, each contributing equally to the combined score, and are scaled to a range of 0 to 30 (ETS, 1999). For this study, the score for the structure section was scaled to range from 0 to 13 and was averaged with the TOEFL-scaled listening and reading scaled scores. This average is referred to hereafter as the *multiple-choice composite score*.

Analysis

As noted earlier, four general types of analyses were performed on these scores. The first focused on the quality of the ratings readers assigned for each composition medium. The second focused on the correlation of writing scores with scores from other sections of the TOEFL test. The third focused on characteristics of examinees who chose each composition medium. And the fourth focused on differences between scores assigned to examinees who chose to compose their essays in handwriting and scores of examinees who chose to use word processors.

Measures of the Quality of the Ratings

We analyzed the scores assigned to handwritten and word-processed essays separately using several reliability indices, particularly indices of reader agreement. Specifically, we computed (a) the Pearson product moment correlation between scores assigned to each essay by the two randomly-selected readers, (b) the proportion of perfect, adjacent, and outside-of-

adjacent agreement between these two scores, and (c) Cohen's coefficient κ .

For each pair of raters, two Pearson product moment correlation coefficients were computed—one for handwritten essays and one for word-processed essays. Because of the small number of essays some raters had in common, and because the Pearson product moment correlation is biased for small sample sizes (Howell, 2002), we corrected the correlations using a formula provided by Howell:

$$r_{adj} = \sqrt{\frac{(1-r^2)(N-1)}{N-2}}. \quad (1)$$

Adjusted correlations were then transformed using the Fisher z transformation, and weighted averages of the transformed correlations were computed and transformed back to the correlation metric.

The proportion of perfect, adjacent, and outside-of-adjacent agreement was computed by determining the proportion of absolute differences between the two scores for word-processed and handwritten essays: 0 when the ratings were the same, 1 when they differed by a single score point, and 2 when they differed by two or more score points. Cohen's coefficient κ indicates the degree to which readers agree beyond the level expected by chance:

$$\kappa = \frac{P_o - P_e}{1 - P_e}, \quad (2)$$

where P_o is the observed proportion of ratings in perfect agreement and P_e is the expected proportion of ratings in perfect agreement based on the marginal distributions of ratings assigned by the two readers.

Correlational Analyses

To determine the degree to which scores for handwritten and word-processed TOEFL essays are comparably correlated with scores from other sections of the computer-based TOEFL test, we computed Pearson product moment correlations between the writing composite score; scaled scores on the listening, reading, and structure multiple-choice sections; and the multiple-choice composite score.

Group Characteristics

The characteristics of individuals who chose each composition medium were compared using logistic regression modeling. The explanatory variables, provided in examinees' self-reports, were chosen for their substantive believability as mediators of an examinee's choice of composition medium. We expected older examinees (the *age* variable) to be more likely to choose handwriting because of their increased likelihood of having been raised in a home in which a computer was not present. As suggested in the literature review, we expected females (the *gender* variable) to be more likely to choose handwriting because of their general lower levels of computer familiarity and the resulting higher levels of computer anxiety. We expected examinees from continents in which there are several developing countries (the *continents* variable) to be more likely to choose handwriting because of the general lack of availability of computers in those regions. We reasoned that examinees who are less proficient with English (the *English* variable), as indicated by their scores on the TOEFL multiple-choice sections, would be more likely to choose handwriting as the composition medium because of the double translation that would be required (first from thoughts to verbal expression and then from verbal expression to keyboard) to compose an essay using a word processor. Similarly, we expected examinees who speak a native language that is not based on the Roman or Cyrillic alphabets to be more likely to compose their essays in handwriting because of the difficulty of translating thoughts into words and words into English-language keystrokes (the *keyboard* variable). Data were also available for each examinee's self-reported reason for taking the examination (e.g., graduate or undergraduate school admissions or business requirements).

Medium, the outcome variable, was coded as the dichotomous choice made by examinees to compose their essays on a computer or in handwriting, and we modeled the choice to compose an essay in handwriting rather than the choice to use a word processor. As for the treatment of the explanatory variables, age and English were treated as quantitative variables. Gender was treated as a dichotomous variable with females being the reference group. Countries were divided into the following continents, treated nominally, of course: North America, Africa (reference cell), Asia and Pacific Islands, Central and South America, Europe, and Middle East. Keyboard was treated as a dichotomous variable based on whether the examinee's native-language keyboard is based on an alphabet similar to the one used in English (e.g., Roman or Cyrillic) or on some "other" system (e.g., most Asian languages; the reference cell for keyboard

was “other”). For each variable, the reference cell was the group we believed would be most likely to choose handwriting as the composition medium.

Because of the very large sample size in this study, we utilized a modified version of the model selection strategy recommended by Hosmer and Lemeshow (Hosmer & Lemeshow, 2000). That strategy requires four steps. First, univariate analyses were performed on each potential explanatory variable to determine whether each one had enough predictive power by itself to warrant inclusion in a multivariate main effects model. Because this approach could potentially lead to rejection of an explanatory variable that would have good predictive power in a multivariate model, Hosmer and Lemeshow suggest adopting a large p-value for rejection; they suggest $p = .25$, which we adopted.

Once the contribution of each potential explanatory variable is evaluated in the context of univariate models, the second step of the model selection strategy fits a preliminary multivariate main effects model containing all of the variables selected for inclusion during the first step of the procedure. Because each of the potential explanatory variables demonstrated reasonable predictive power during the first step of the procedure, the preliminary multivariate main effects model contained all of the explanatory variables described previously.

The third step focused on justifying the chosen coding schemes for each quantitative variable; in our study, the relevant variables were age and English. To do so, we examined the relationship between the levels of these two explanatory variables and the logits relating to composition-medium choice. If an explanatory variable is modeled to be linear it is important to verify that the logits for that variable are indeed linear. Plots of English indicated that this explanatory variable was indeed linear in the logits. Hence, English was modeled to have a single parameter that indicated the incremental increase in composition-medium choice logits for each one-unit increase in English (measured in quartiles). Age, however, was not linear in the logits. As shown in Figure 3, examinees between the ages of 21 and 30 had lower logit values (lower probabilities of choosing handwriting) than examinees over 30 years of age and examinees under 20 years of age. Hence, we added a quadratic term for age to the model—a term that proved to be statistically significant and also to improve the fit of the model.

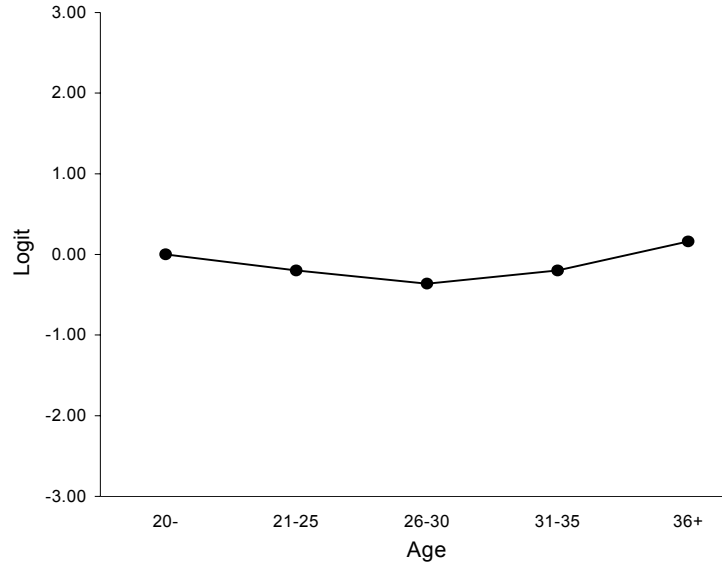


Figure 3. Medium-choice linearity of logits for age.

The fourth step of the model selection routine involved identifying statistically significant two-way interaction terms for a “preliminary final model.” Because we were unable to justify exclusion of any of the potential two-way interaction terms on substantive grounds, we evaluated the statistical contribution of all two-way interaction terms to the prediction of composition-medium choice. Hence, during each iteration of the forward selection routine, we identified the most powerfully predictive two-way interaction, then we fit two models—one without that interaction and one with it—and evaluated the relative fit of the two models. Interactions not included in the reduced model were identified as candidates for inclusion in the expanded model based on the p-values of their Wald χ^2 statistics in the expanded model.

The Wald statistic for a particular interaction equals the square of the value of its parameter estimate divided by its standard error:

$$\chi^2_{Wald} = \left(\frac{\beta}{SE_{\beta}} \right)^2. \quad (3)$$

This statistic is chi-squared distributed with 1 degree of freedom and tests the null hypothesis that the parameter estimate equals 0 (i.e., that the interaction in question makes no contribution to the prediction of the dependent variable). Type III Wald statistics can be created for interactions for which there are multiple parameters (i.e., nominal polytomous variables). These

Wald statistics simultaneously test the null hypothesis that all of the parameter estimates associated with the interaction equal 0.

In this study, interactions were selected for inclusion in the expanded model based on the p-value of the Type III Wald statistic—the smaller the p-value, the farther the parameter estimate(s) for that interaction is (are) from 0 and the better the interaction is as a candidate for inclusion in the expanded model. Our choice of how to interpret these p-values constitutes our first of two modifications to the procedure suggested by Hosmer and Lemeshow. We added interactions if the p-value of the Wald statistic exceeded .0002. This seemingly extreme p-value was chosen based on recommendation by Raftery (1995), who suggested using more restrictive p-values for model selection purposes when sample size is large (as it is in this study) to ensure that only terms exhibiting reasonable levels of association with the dependent variable are included in the final model. The reasoning here is similar to the rationale for promoting interpretation of effect sizes—with a sample size as large as the one in this study, we would almost certainly find statistically significant effects that have very little predictive power. Based on Raftery’s framework, the chosen p-value to include will only allow an interaction to enter the model if its predictive power is stronger than “moderate.”

Typically, in logistic regression, once an interaction is included in the expanded model, the parameters are re-estimated for the expanded model, and the decrease in the misfit of that model is compared to the reduced model using the likelihood ratio (*LR*) χ^2 statistic (which equals the difference between the deviance statistics, G^2 , of the reduced model and the full model, $\chi^2_{LR} = G^2_{reduced} - G^2_{full}$). χ^2_{LR} is approximately chi-squared distributed with degrees of freedom equal to the difference in degrees of freedom of the two models, $df = df_{reduced} - df_{full}$. If the likelihood ratio statistic is statistically significant, then the inclusion of the interaction in the expanded model reduces the misfit enough to justify its inclusion. Of course, with our large sample size, all progressively more complex models improved model fit to a statistically significant degree. These substeps—which identify an interaction for inclusion, re-estimate parameters for the expanded model, and evaluate relative fit of the expanded and reduced models—were performed until no additional interactions met the p-value criterion.

An alternative way of comparing the relative fit of each progressively more complex model is to examine the proportionality constant for each of those models. Normally, in logistic

regression, one would evaluate the overall fit of the data to each model by examining the deviance statistic (G^2), which compares the maximized log-likelihood of the model in question to the maximized log-likelihood value for the saturated model. For grouped data for which the model in question is appropriate, G^2 is distributed as a χ^2 statistic with degrees of freedom equal to the difference between the number of parameters in the saturated model and the model being tested. However, because our data are “individual-level” data (i.e., there are very few observations that are identical with respect to the independent variables), the deviance is not an appropriate measure of model-data fit (Allison, 1999).

An alternative statistical test is the Hosmer-Lemeshow test, but the sampling distribution of that test is not known, and it is very sensitive to large sample sizes. Faced with this problem, we chose to examine the value of the proportionality constant (PC), which equals the deviance divided by its degrees of freedom (G^2 / df). This was the second of our modifications to the routine suggested by Hosmer and Lemeshow (2000). For grouped data, values of the PC that are close to 1 are interpreted as indicating that the data contain about the same amount of misfit as would be expected due to random sampling, and values of the PC that are greater than 1 indicate that the data contain unmodeled variance. However, because our data are not grouped, the PC can only be interpreted as a relative index of fit between two models: a model with a smaller PC does a better job of accounting for the data than a model with a larger PC. Hence, we examined the improvement (decreases) in the PC for each iteratively more complex model we estimated.

Table 1 shows the progression of models investigated in the fourth step of the algorithm. The preliminary final model contained the following terms: age (linear and quadratic), continent, English, gender, keyboard, Age \times Continent, Continent \times English, Age \times English, Continent \times Keyboard, and Continent \times Gender. The 5 two-way interactions suggest that composition-medium choice: (a) varies across age groups between the continents, (b) varies across English proficiency groups between continents, (c) varies across age groups between English proficiency groups, (d) varies across language groups between continents, and (e) varies across continents between gender groups. Consistent with the recommendations of Hosmer and Lemeshow (2000), we evaluated the substantive contribution of these parameters, and we decided to eliminate all but one of the two-way interactions in the final model because the interaction was not deemed important enough for inclusion. We describe the eliminated two-way interactions and interpret the final model in the Results section of this report.

Table 1***Medium-Choice Model Selection Summary***

Model		Deviance			Likelihood ratio			Entry candidate	Wald statistic		
Iteration	Term added	G^2	df	G^2 / df	χ^2	df	p		χ^2	df	P
0	Main effects	27,520.60	20475	1.34	—	—	—	Age × Continent	783.70	5	< 0.0001
1	Age × Continent	26,505.81	20470	1.29	1,014.78	5	< 0.0001	Continent × English	218.90	5	< 0.0001
2	Continent × English	26,225.30	20465	1.28	280.51	5	< 0.0001	Age × English	62.42	1	< 0.0001
3	Age × English	26,145.55	20464	1.28	79.75	1	< 0.0001	Continent × Keyboard	63.27	5	< 0.0001
4	Continent × Keyboard	26,064.95	20459	1.27	80.60	5	< 0.0001	Continent × Gender	46.89	5	< 0.0001
5	Continent × Gender	26,005.34	20454	1.27	59.61	5	< 0.0001	English × Gender	9.66	1	0.0019

Group Comparisons

The final questions our analyses addressed were whether there were differences between scores on handwritten and word-processed essays, whether different groups of examinees performed differently on the essay test, and whether the magnitude of those differences depended on composition medium. To address these questions, we utilized general linear modeling. Specifically, we utilized six models, each of which is shown below:

- a three-way model with covariates in which the essay scores of examinees who chose handwriting and word-processing were compared while controlling for English proficiency (as defined by performance on the multiple-choice section of the TOEFL test), four covariates (gender, region, age, and keyboard), and all two- and three-way interactions

$$\begin{aligned}
 \hat{Y}_{essay} = & \hat{\alpha} + \hat{\beta}_{medium} X_{medium} + \hat{\beta}_{English} Z_{English} + \sum \hat{\beta}_{covariate} W_{covariate} + \\
 & \hat{\beta}_{medium \times English} X_{medium} Z_{English} + \\
 & \sum \hat{\beta}_{medium \times covariate} X_{medium} W_{covariate} + \\
 & \sum \hat{\beta}_{English \times covariate} Z_{English} W_{covariate} + \\
 & \sum \hat{\beta}_{medium \times English \times covariate} X_{medium} Z_{English} W_{covariate}
 \end{aligned} \tag{A}$$

- a two-way model with covariates in which the essay scores of examinees who chose handwriting and word-processing were compared while controlling for English proficiency, four covariates (gender, region, age, and keyboard), as well as all two-way interactions

$$\begin{aligned}
 \hat{Y}_{essay} = & \hat{\alpha} + \hat{\beta}_{medium} X_{medium} + \hat{\beta}_{English} Z_{English} + \sum \hat{\beta}_{covariate} W_{covariate} + \\
 & \hat{\beta}_{medium \times English} X_{medium} Z_{English} + \\
 & \sum \hat{\beta}_{medium \times covariate} X_{medium} W_{covariate} + \\
 & \sum \hat{\beta}_{English \times covariate} Z_{English} W_{covariate}
 \end{aligned} \tag{B}$$

- a restricted two-way model with covariates in which the essay scores of examinees who chose handwriting and word-processing were compared while controlling for English proficiency, the two-way interaction between English and medium, and the main effects for each covariate

$$\hat{Y}_{essay} = \hat{\alpha} + \hat{\beta}_{medium} X_{medium} + \hat{\beta}_{English} Z_{English} + \sum \hat{\beta}_{covariate} W_{covariate} + \hat{\beta}_{medium \times English} X_{medium} Z_{English} \quad \textcircled{C}$$

- a two-way model with no demographic covariates in which mean essay scores for each medium were compared while controlling for English language proficiency and the two-way interaction between these two variables

$$\hat{Y}_{essay} = \hat{\alpha} + \hat{\beta}_{medium} X_{medium} + \hat{\beta}_{English} Z_{English} + \hat{\beta}_{medium \times English} X_{medium} Z_{English} \quad \textcircled{D}$$

- a one-way model with no demographic covariates in which mean essay scores for each medium were compared while controlling for English language proficiency

$$\hat{Y}_{essay} = \hat{\alpha} + \hat{\beta}_{medium} X_{medium} + \hat{\beta}_{English} Z_{English} \quad \textcircled{E}$$

- a one-way model with no covariates in which mean essay scores for each medium were compared

$$\hat{Y}_{essay} = \hat{\alpha} + \hat{\beta}_{medium} X_{medium} \quad \textcircled{F}$$

Our primary interest lay with Models C, D, and E. We examined Models A and B to verify that we were not ignoring important interactions that might influence the accuracy of the predicted essay scores for a particular group, and we examined Model F for illustrative purposes—to determine the importance of English language proficiency as a covariate. To this end, we fit each of the specified models and interpreted four indices: the R^2 for the model, the F statistic associated with the effect in question (and its p-value), η^2 (the proportion of total variance accounted for by the effect in question), and the magnitude of the raw score group difference.

We anticipated that most of the F statistics would be statistically significant because of the large sample size. That is why we computed and interpreted the η^2 indices. We computed η^2 based on the Type III sums of squares for each effect ($\eta^2 = SS_{effect} / SS_{total}$). Although Cohen (1988) suggests that values of η^2 are small if they fall in the range of .01, we believe that this rule could lead to a dismissal of substantively important effect sizes as being trivial for the data analyzed

here—a belief supported by Rosenthal, Rosnow, and Rubin (2000). Hence, we applied a substantive criterion by declaring an effect size as being important if it results in a group mean difference of at least one-half of a point on the essay rating scale—the equivalent of one of the raters assigning a rating for an essay that is one point higher or lower than the other reader’s rating.

The linear modeling procedures we utilized require that the dependent variable be unbounded and continuous. Because there were 11 possible values for the dependent variable (1.0, 1.5, 2.0, 2.5, ..., 5.5, 6.0), and because the highest and lowest values of the rating scale were infrequently observed, we felt that this requirement was reasonably satisfied. In addition, three assumptions were required for each model we investigated: (a) normality of conditional distributions of essay scores for each composition medium, (b) homogeneity of the variances of those arrays, and (c) linearity of the relationship between multiple-choice scores and essay scores for each composition medium. Examination of group distributions and variances indicated that the assumptions of normality and homogeneity of variances were met. The linearity assumption was evaluated by computing conditional means for handwriting and word-processing for each of 10 equal-interval bins of multiple-choice scores and then examining the scatterplot of these mean essay scores for each composition medium. Again, these assumptions were met. As shown in Figure 4, the relationship seems slightly nonlinear, although not dramatically so.

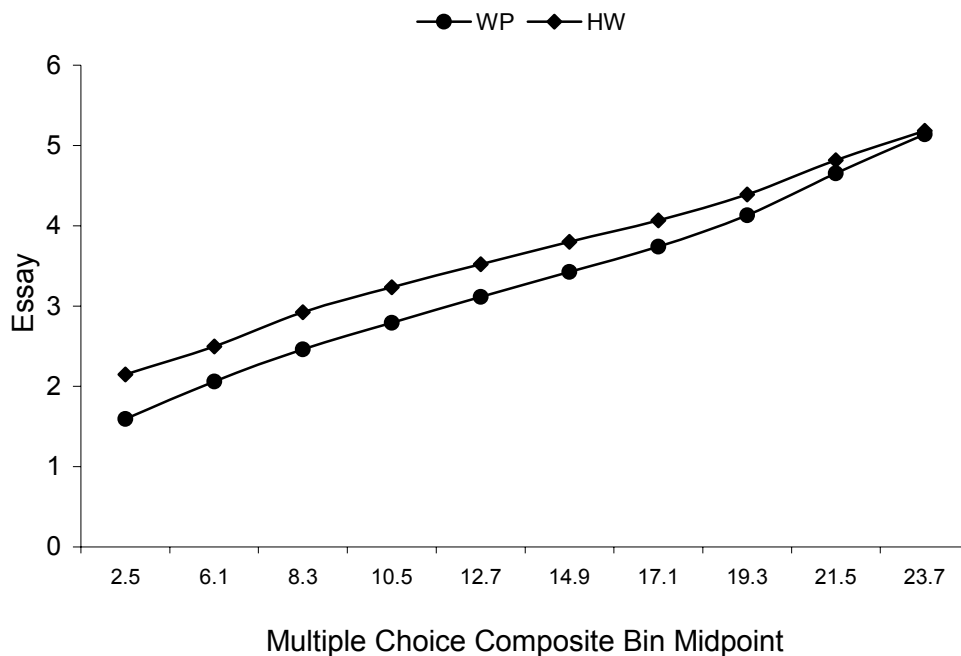


Figure 4. Linearity of essay scores across English proficiency levels.

We began our model selection procedures by first considering whether Models A or B indicated a need for a complex model containing three- or two-way interactions other than the medium-by-English interaction. Table 2 summarizes the model selection statistics for each of the six models we considered. From these figures, we decided not to consider the three-way interaction term or any of the two-way medium-by-covariate interaction terms because models containing those terms did no better than simpler models in terms of the proportion of variance explained (R^2) and because the effect size indices for the hierarchically highest-level terms in those models were all very small (the largest was less than .001). Examination of the statistics for Model F indicated that composition medium alone was not a satisfactory explanatory variable: The R^2 for this model was very small (.0003).

Table 2
Essay Score Model Comparisons

Model	Name	R^2	η^2 ^a
A	All three-ways	.41	.0001
B	All two-ways	.41	.001
C	Medium \times English, demographics	.41	.008
D	Medium \times English	.39	.005
E	Medium, English	.39	.39
F	Medium	.0003	.0003

^a The value shown here is that of the largest η^2 for the hierarchically highest level terms in the model. For Models C and D, this excludes the medium-by-English interaction.

This left us with a choice between three models. On substantive grounds, we preferred the model containing a two-way interaction between medium and English proficiency because the magnitude of the cross-medium means of examinees at the lower end of the English proficiency continuum seemed large enough to warrant reporting. As shown in Figure 4, the difference between the means of these examinees was about 0.40 score points, while there were no large differences for examinees with the highest multiple-choice scores. Using this logic, it seemed unreasonable to exclude the demographics main effects from the model, given that the

R^2 was slightly larger for this model and the largest effect size index was greater in value than the η^2 for the medium-by-English interaction.

Table 3 displays the group means for each level of the four covariates included in Model C. These figures reveal that the largest group differences within each covariate range from about one-eighth to about one-half of a raw score point on the TOEFL writing scale—values that are marginally within the range we defined as being substantively important (0.25 raw score points). As an aside, out of curiosity we fit a similar model that contained a quadratic term for age because the essay means for the age groups were slightly curvilinear. We decided not to include this quadratic term because the expanded model did exhibit better fit to the data and the effect size for the quadratic term was not large ($\eta^2 = .003$).

Table 3
Essay Score Means for Each Covariate

Variable	Group	Smallest mean	Largest mean	Difference	η^2
Age	> 35	3.94		0.17	.008
	21 – 25		4.11		
Continent	Middle East	3.92		0.43	.006
	Europe		4.35		
Gender	Male	4.02		0.12	.004
	Female		4.14		
Keyboard	Other	3.94		0.27	.0006
	Roman/Cyrillic		4.21		

Results

Descriptive Statistics

Table 4 lists descriptive statistics for each type of score. From these figures, it is clear that there were only small differences between the average scores of, and number of minutes required to complete (out of 30 minutes), handwritten and word-processed essays. On the other hand, there seem to be differences between the multiple-choice scores of examinees who chose to compose the essay in handwriting and those who chose to compose the essay on a word processor. Hence, it seems that there are differences in the English proficiency of these two groups, even though the essay scores are equivalent.

Table 4***Descriptive Statistics***

Score type	Composition medium				Overall	
	Handwriting		Word-processing		<i>M</i>	<i>SD</i>
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Writing composite	3.82	0.95	3.86	1.09	3.84	1.03
Minutes to complete essay	28.71	3.75	28.99	2.95	28.86	3.35
Structure subscale	8.40	3.12	9.57	2.92	9.03	3.07
Listening subscale	20.57	5.02	22.93	4.58	21.85	4.93
Reading subscale	21.06	4.84	23.31	4.52	22.27	4.81
Multiple-choice composite	16.68	3.89	18.60	3.62	17.72	3.87

Note. $N_{\text{handwriting}} = 61,650$; $N_{\text{word processor}} = 72,256$; $N_{\text{overall}} = 133,906$.

Comparisons of the Quality of the Ratings

Table 5 provides comparisons of the indices of rating quality. Specifically, for each composition medium, we show the following indices for the original pair of readers: (a) the average (weighted) adjusted (for sample size) Pearson product moment correlation r ; (b) Cohen's coefficient κ ; and (c) the percent of perfect, adjacent, and outside-of-adjacent agreement between the scores they assigned for each essay. The table reveals that, overall, it was easier for readers to agree on scores for word-processed essays than for handwritten ones, although the difference was only modest. This was true regardless of the index we considered.

Table 5***Indices of Rating Quality***

	Handwriting	Word-processing	Overall
<i>r</i>	.70	.78	.74
κ	.30	.34	.32
Perfect (%)	50	50	50
Adjacent (%)	45	44	44
Outside (%)	6	5	6

Correlational Analyses

Table 6 presents Pearson product moment correlations between writing composite scores, scaled scores for each multiple-choice section, and multiple-choice composite scores. The correlations in the upper-right of the table are based on examinees who chose to compose their essays in handwriting, and those on the lower-left are based on examinees who chose to use word processors. The corresponding off-diagonal correlations indicate that scores from various sections of the TOEFL test were slightly more consistent for examinees who used word processors than for examinees who handwrote their essays. However, these differences are very small—the largest being between listening scaled scores and writing-composite scores (.55 for word processor versus .51 for handwriting) and between structure and listening scaled scores (.70 for word processor versus .66 for handwriting). These increases are likely due to the increased reliability of word-processed scores.

Table 6
Correlations Between TOEFL Sections by Composition Medium

		— Handwriting —				
		Writing	Listening	Reading	Structure	Multiple-choice
Word- processing	Writing	—	.51	.54	.57	.60
	Listening	.55	—	.68	.66	.89
	Reading	.54	.69	—	.79	.92
	Structure	.59	.70	.79	—	.88
	Multiple-choice	.61	.90	.92	.89	—

Note. Upper-right entries are correlations between subtests for examinees who chose handwriting as the composition medium, and lower-left entries are correlations for examinees who chose word-processing as the composition medium.

Group Characteristics

Table 7 compares the characteristics of examinees who chose to compose their essays in handwriting with those who chose to use a word processor. Specifically, the table shows the average age of examinees who chose each mode, along with the joint and marginal probability distributions for each of several demographic variables that were collected as self-report data during the operational TOEFL testing. These data show that examinees from the Middle East or

Africa were more likely to compose essays in handwriting, while test-takers from Asia/Pacific Islands and Central/South America were more likely to compose essays using a word processor. Also, male examinees were more likely to choose word-processing over handwriting, while female examinees were evenly divided between the two media. Finally, examinees who speak character-based languages were more likely to choose handwriting, while those who speak languages based on the Cyrillic alphabet were more likely to choose the computer. There were only small differences between composition modes for age and reason for taking the examination.

Table 7
Characteristics of Examinees Who Chose Handwriting Versus Word-Processing

Variable	Group	Handwriting (% of total sample)	Word-processing (% of total sample)	Overall (% of total sample)
Age ^a	> 21 years	16	16	32
	21 – 25 years	15	18	33
	26 – 29 years	9	13	22
	30 – 35 years	4	5	8
	< 35 years	3	5	6
Continent	Africa	3	1	5
	Asia/Pacific Islands	20	23	42
	Central/South America	4	9	14
	Europe	9	13	23
	Middle East	9	5	14
	North America	1	3	4
Gender	Female	23	23	46
	Male	23	31	54
Keyboard	English	0.4	1	2
	Cyrillic	19	29	48
	Other	27	23	50
Reason	Undergraduate	17	20	37
	Graduate	23	26	49
	Business/other	6	8	14
Medium		46	54	

^a The mean age of test-takers who chose to handwrite their essays was 24.20 years; the mean age of those who chose to word-process their essays was 24.31 years; and the overall mean age of the sample was 24.26 years.

Table 8 displays the conditional proportions of examinees in each quartile of the multiple-choice composite scores who chose to handwrite and word-process their essays. From these figures, it is clear that examinees who received higher scores on the multiple-choice composite were considerably less likely to choose handwriting as the composition medium for their essays.

Table 8
Conditional Probabilities of Choosing Handwriting Versus Word-Processing

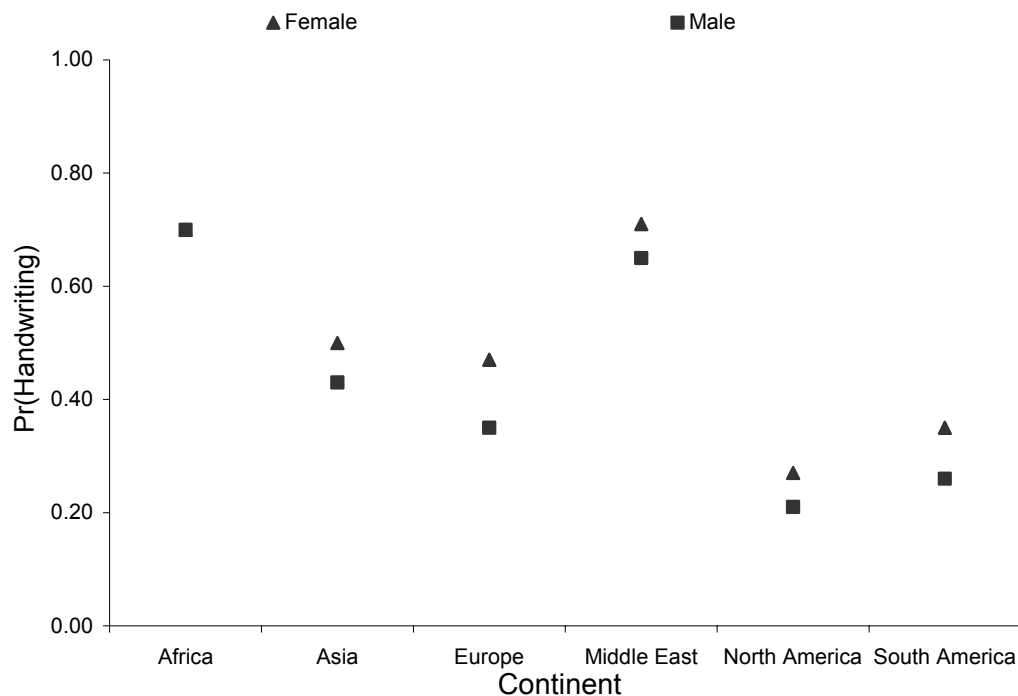
Multiple-choice composite	Composition medium	
	Handwriting	Word-processing
Quartile 1	.61	.39
Quartile 2	.53	.47
Quartile 3	.43	.57
Quartile 4	.29	.71

As mentioned previously, the preliminary final model contained the following two-way interactions: (a) age-by-continent (composition-medium choice varies across age groups between continents), (b) continent-by-English (composition-medium choice varies across English proficiency groups between continents), (c) age-by-English (composition-medium choice varies across age groups between English proficiency groups), (d) continent-by-keyboard (composition-medium choice varies across keyboard-language groups between continents), and (e) continent-by-gender (composition-medium choice varies across continents between gender groups).

The final step in model selection was to substantively evaluate the contribution of each of these terms. Specifically, because we chose a p-value to remove that, based on Raftery’s (1995) criteria, would allow “moderate” effects to remain in the model, we needed to determine whether the terms in the preliminary final model made a substantive contribution to the prediction of composition medium. We made these decisions by plotting the modeled probabilities for the various groups involved in each two-way interaction and then examining these plots to determine whether the detected interaction was large enough to be substantively important. Based on these judgments, we chose to eliminate all but one of the two-way interactions from the final model.

Figure 5 displays the continent-by-gender interaction—the first two-way interaction that

was removed from the preliminary final model. In this figure, we see that female examinees were slightly more likely than male examinees to choose handwriting across continents, with the exception of Africa, where the gender groups were equally likely to choose handwriting. Specifically, although empirical probabilities of choosing handwriting between African males and females were equal, females were about 8% more likely than males to choose handwriting across the remaining continents. We believe that this difference is small enough to be ignorable from a substantive perspective. Hence, we chose to remove this two-way interaction from our prediction model, even though it makes a statistically significant contribution to the prediction of examinee composition-medium choice. As shown in Table 1, the removal of this term from the model resulted in an increase of less than .01 in the proportionality constant (i.e., $PC = 1.27$ for models 4 and 5).

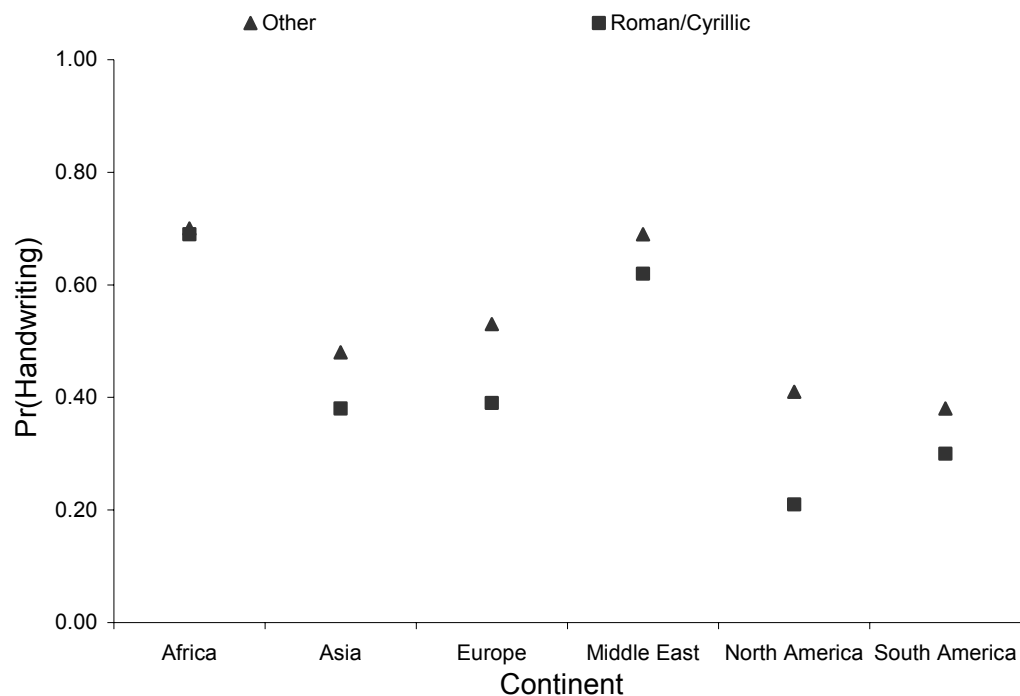


Note. “Asia” = “Asia/Pacific Islands;” “South America” = “Central/South America.”

Figure 5. Medium-choice continent-by-gender interaction.

Figure 6 displays the continent-by-keyboard interaction—the second term to be dropped from the preliminary final model. In this figure, we see that examinees who speak a native language based on a Roman or Cyrillic alphabet were slightly less likely to choose handwriting

than were examinees who speak other native languages. This was true in all continents except Africa. In addition, the difference was slightly larger in North America than in other continents. Specifically, although African examinees who use non-Roman/Cyrillic keyboards were only 1% more likely to choose handwriting than were African examinees who use Roman/Cyrillic keyboards, the analogous difference for examinees in North America was 20%. The average analogous difference across the remaining continents equaled 10%. Again, we saw this difference as being unimportant and chose to ignore it from a substantive perspective (i.e., we chose to remove this term from our prediction model), even though the difference was statistically significant.. This resulted in a .01 increase of the proportionality constant (i.e., from 1.27 to 1.28 between models 4 and 3, respectively).



Note. “Asia” = “Asia/Pacific Islands;” “South America” = “Central/South America.”

Figure 6. Medium-choice continent-by-keyboard interaction.

Figure 7 displays the age-by-English interaction—the third two-way interaction dropped from the preliminary final model. In this figure, we see that, across all age groups, examinees with higher levels of English proficiency were less likely to choose handwriting than were examinees with lower levels of English proficiency. In addition, we see that there were only

small differences between age groups with respect to the choice of handwriting as the composition medium. The statistically significant interaction existed because the rate of decrease in the probability of choosing handwriting as English proficiency increased was shallower for examinees under the age of 21 than it was for other examinee age groups. Specifically, the probability of choosing handwriting for examinees under the age of 20 decreased by 33% between the lowest and highest deciles of the TOEFL multiple-choice scores, while the analogous decrease across the age categories was 46%. Again, we saw this difference as being unimportant and chose to ignore it from a substantive perspective, even though the difference was statistically significant. Removal of this two-way interaction from our regression model resulted in no change in the proportionality constant (i.e., $PC = 1.28$ for models 3 and 2).

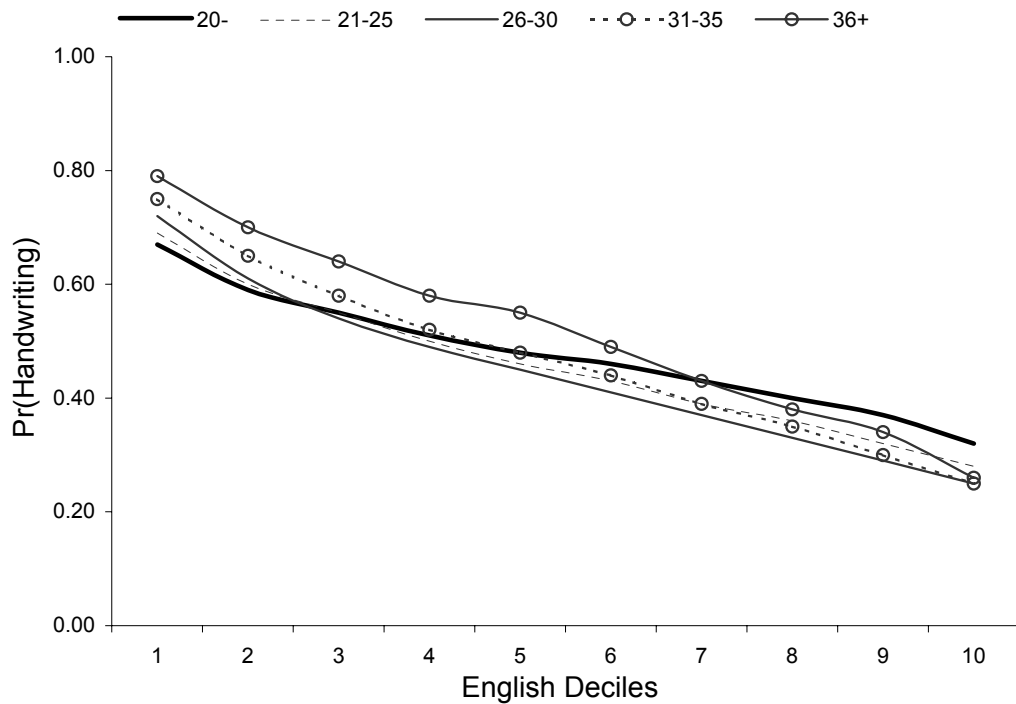
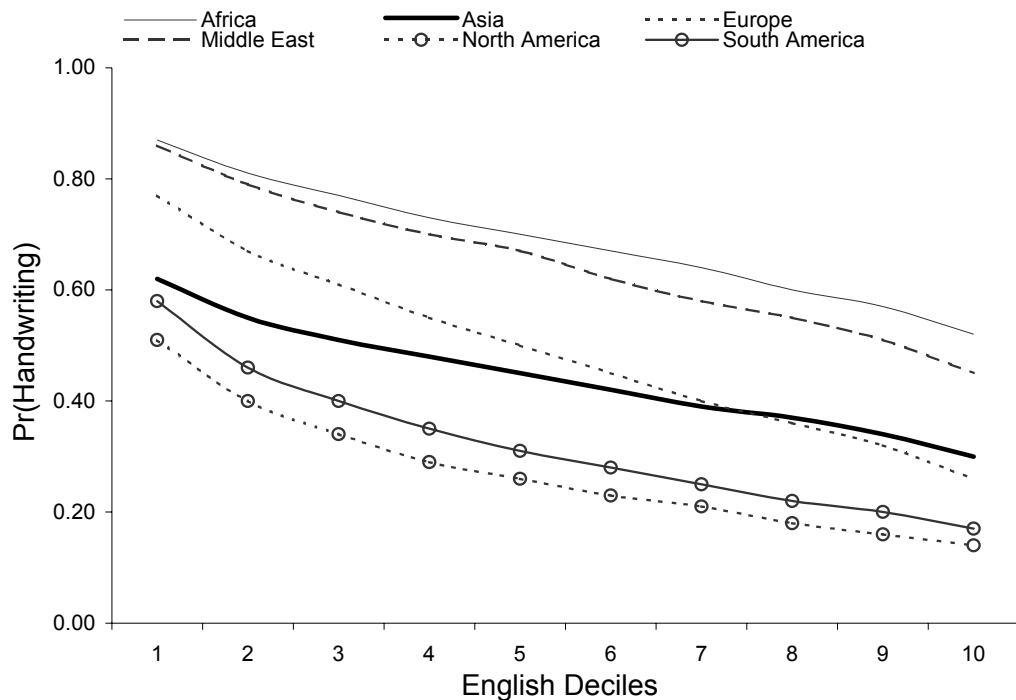


Figure 7. Medium-choice age-by-English interaction.

Figure 8 displays the continent-by-English interaction—the final term we dropped from the preliminary final model. In this figure, we see that, across continents, examinees with higher levels of English proficiency were less likely to choose handwriting than were examinees with lower levels of English proficiency. In addition, we see that there were fairly large differences

between continents with respect to the choice of handwriting as the composition medium. Specifically, examinees from Africa and the Middle East were most likely to choose handwriting, while examinees from North and Central/South America were least likely to choose handwriting. The statistically significant interaction was apparent because the rate of decrease in the probability of choosing handwriting as English proficiency increased was shallower for examinees from Asia/Pacific Islands than it was for examinees from the remaining continents. Specifically, the probability of Asian/Pacific Island test-takers choosing handwriting decreased by 32% between the lowest and highest deciles of the TOEFL multiple-choice scores, while the analogous decrease across the remaining continents was 41%. Again, we saw this difference as being unimportant and chose to ignore it from a substantive perspective (i.e., we chose to remove this term from our prediction model), even though the difference was statistically significant. This resulted in a .01 increase in the proportionality constant (i.e., from 1.27 to 1.28 for models 2 and 1, respectively).



Note. “Asia” = “Asia/Pacific Islands;” “South America” = “Central/South America.”

Figure 8. Medium-choice continent-by-English interaction.

Hence, the final model had the following form:

$$\begin{aligned} \text{logit}(\textit{handwriting}) = & \hat{\alpha} + \hat{\beta}_{age} X_{age} + \hat{\beta}_{continent} X_{continent} + \hat{\beta}_{English} X_{English} \\ & + \hat{\beta}_{gender} X_{gender} + \hat{\beta}_{keyboard} X_{keyboard} + \hat{\beta}_{age^2} X_{age^2} \\ & + \hat{\beta}_{age \times continent} X_{age \times continent} \end{aligned} \quad (4)$$

That is, the age-by-continent interaction was the only two-way effect in the final model. Linear main effects were included for English and age with a quadratic effect for age. Finally, gender, keyboard, and continent were included as nominal variables.

Once the final model was selected, we evaluated its overall predictive capacity. Unfortunately, because our data were individual-level data, there was no best index for evaluating model fit. Thus, we chose to examine several indices: the proportionality constant (PC), the proportion of concordant pairs, the max-rescaled R^2 , and the dissimilarity index.

The first index we used to evaluate model fit was the proportionality constant. As mentioned in the previous section, the deviance statistic is only chi-squared distributed for grouped data. As a result, the PC can only be interpreted as a relative index in this study. As shown in Table 1, the PC for the final model equals 1.29—a reduction of .05 from the main effects model, and only .02 points greater than the PC for a model that contains four additional two-way interactions. Because we cannot be certain about the degree to which the model sufficiently explains the data, and because the standard errors of the parameter estimates are inflated in logistic regression when there is unexplained heterogeneity in the data—a phenomenon known as *overdispersion* (Allison, 1999)—we chose to correct the standard errors by multiplying them by the square root of the PC (Agresti, 1996).

The second index we used to evaluate model fit was the proportion of concordant pairs ($P_{\text{concordant}}$), which indicates the degree to which the final model successfully predicts observed group membership on the dependent variable. Specifically, this index summarizes the proportion of pairs of observations with different outcomes (i.e., handwriting versus word-processing) for which the model-based expected value is consistent with the observed outcome. That is, a pair is concordant if the member of the pair who chose handwriting has a higher predicted value for handwriting than the member of the pair who chose word-processing. $P_{\text{concordant}}$ for the final model was .70, indicating that the model did a fairly good job of predicting group membership on the dependent variable.

The third index we considered was the maximized R^2 index, which is analogous to the R^2

adjusted index generated in ordinary linear regression. Two key differences, however, are that the R^2 index from nonlinear models cannot be interpreted in an absolute sense (e.g., as the proportion of variance in the dependent variable accounted for by the model, as demonstrated by its formula, $1 - \exp[-L/n]$) and that the R^2 index for binary outcomes is typically much smaller in magnitude than it is for ordinary linear models (the maximized R^2 adjusts, slightly, for the downward bias in the R^2 index). Because of these two shortcomings, the maximized R^2 can only be interpreted as a relative index for the models we investigated. Our final model produced a maximized R^2 of .15, while the main effects model and the model with two two-way interactions produced maximized R^2 indices of .14 and .15, respectively. Hence, it seemed that the chosen model performed comparably to the surrounding models investigated in the variable selection routine.

The final index we used to evaluate the overall model fit was the dissimilarity index. Agresti (1996) points out that a statistically significant misfit is not necessarily important—it may be an artifact of a large sample size. Hence, he suggests the dissimilarity index as a measure of effect size. Simply put, the dissimilarity index (D) represents the proportion of sample cases that would have to be moved to a different cell in the data matrix in order for the model to achieve a perfect fit. Agresti suggests an ideal value for the dissimilarity index of .03 or lower. The D index for the fitted model was .22, indicating marginally acceptable model fit. The D indices for the main effects and the two two-way interactions models were also both .22. Hence, the final model seemed to provide as good an explanation of the data as any of the surrounding more complex or simpler models.

Parameters for the logistic regression model are reported on the logit scale—the log of the odds of composition-medium choice. The odds equal the probability of a participant with a specific set of demographic characteristics choosing the handwriting medium divided by the probability of the participant choosing the word-processing medium:

$$\text{logit}(\textit{handwriting}) = \log\left(\frac{P_{\textit{handwriting}}}{1 - P_{\textit{handwriting}}}\right) = \log\left(\frac{P_{\textit{handwriting}}}{P_{\textit{word-processing}}}\right). \quad (5)$$

The odds range from 0 to ∞ , with (a) smaller values indicating that the probability of a handwriting choice is considerably less than the probability of a word-processing choice, (b) larger values indicating that the probability of a handwriting choice is considerably greater than

the probability of a word-processing choice, and (c) values of 1 indicating that a handwriting choice is as likely as a word-processing choice (i.e., $p_{\text{handwriting}} = p_{\text{word-processing}}$). The log transformation of the odds creates the logit scale, which ranges from $-\infty$ to $+\infty$, with negative values indicating that handwriting is less likely to be chosen than word-processing, and positive values indicating that handwriting is more likely to be chosen than word-processing. Of course, values of 0 [$\log(1)$] occur when handwriting and word-processing choices are equally likely. This points out the direct connection between logits and probabilities. Specifically, given the logit of a handwriting choice, one can determine the probability of a handwriting choice as:

$$p_{\text{handwriting}} = \frac{\exp\left(\sum_{t=0}^T \beta_t X_t\right)}{1 + \exp\left(\sum_{t=0}^T \beta_t X_t\right)}, \text{ where } \sum_{t=0}^T \beta_t X_t \quad (6)$$

is the portion of the logistic model relevant to the group in question.

Table 9 displays the parameter estimates, standard errors, and Type III Wald statistics and their p-values for each variable in the final model. Generally, as expected, the parameter estimates are negative, indicating that, in nearly all cases, the reference-cell groups exhibited the greatest probability of choosing the handwriting medium. In all cases where the parameter estimates are positive, the parameter estimate is close to zero. Three main effects are noteworthy.

First, the gender main effect indicated that male examinees were less likely than female examinees to choose handwriting as the composition medium for their essays. The empirical proportion of female participants choosing handwriting was .49, while the proportion of male participants choosing handwriting was .43. The modeled proportions (i.e., the expected proportions of female and male test-takers when the influence of other demographic variables on composition-medium choice were held constant) were identical.

Second, the keyboard main effect indicated that those examinees whose native language is based on the Roman/Cyrillic alphabet were less likely to choose handwriting as the composition medium for their essays. The empirical and modeled probabilities for choosing handwriting were .53 for non-Roman/Cyrillic-language speakers and .38 for Roman/Cyrillic-language speakers. Again, the modeled probabilities were identical.

Third, the main effect for English indicated that the probability of an examinee choosing handwriting decreased as English proficiency increased. The empirical probability of choosing

handwriting decreased from a high of .66 for the 1st decile of TOEFL multiple-choice scores to a low of .23 for the 10th decile. Generally, predicted probabilities were very similar to these values, with the largest discrepancies (about .07) in the 1st and 10th deciles.

Table 9
Summary of the Parameter Estimates for the Final Model

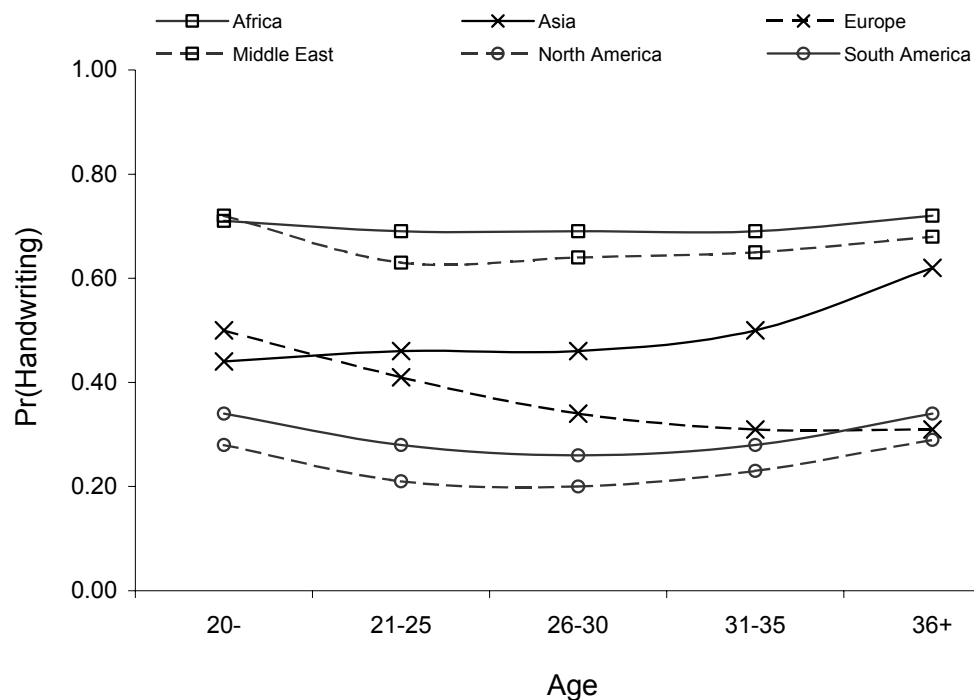
Variable	Level	Parameter		Statistical significance	
		β	SE_{β}	χ^2_{Wald}	p
Intercept		4.60	0.15	919.02	< .0001
Age		-0.08	0.01	104.14	< .0001
Age ²		0.00	0.00	142.12	< .0001
Continent					
	Asia/Pacific Islands	-2.25	0.13	319.38	< .0001
	Europe	0.06	0.14	0.19	.6700
	Middle East	-0.07	0.14	0.24	.6200
	North America	-1.93	0.19	101.01	< .0001
	Central/South America	-1.46	0.14	109.07	< .0001
Keyboard	Other (Non-Roman/Cyrillic)	-0.29	0.02	239.69	< .0001
English		-0.13	0.00	4651.87	< .0001
Gender	Male	-0.35	0.01	684.12	< .0001
Age × Continent					
	Asia/Pacific Islands	0.04	0.00	76.09	< .0001
	Europe	-0.04	0.01	65.08	< .0001
	Middle East	-0.01	0.01	1.81	.1800
	North America	0.00	0.01	0.40	.5300
	Central/South America	0.00	0.01	0.25	.6200

Note. β is the estimated parameter and SE_{β} is the standard error of that estimate. The Wald chi-squared value is the Type III (omnibus) version.

Figure 9 displays the age-by-continent interaction—the only two-way interaction entered into the logistic regression model. This figure shows that there were differences between the continents with respect to the proportion of examinees choosing the handwriting medium. Specifically, examinees from Africa and the Middle East were most likely to choose

handwriting, while examinees from North and Central/South America were least likely to choose handwriting. In addition, there were no substantial differences between age groups on these continents with respect to the proportion of examinees who chose handwriting, although the oldest and the youngest examinees tended to be most likely to choose handwriting.

The age-by-continent interaction exists because the trends for the remaining two continents (Europe and Asia/Pacific Islands) differ from these trends. Specifically, European examinees under 21 years of age were the most likely examinees from that continent to choose handwriting as a composition medium. On the other hand, Asian/Pacific Island examinees under the age of 21 were the least likely examinees from that continent to choose handwriting as a composition medium. Specifically, the empirical probability of choosing handwriting for Europeans decreased from a high of .53 to a low of .42 across the age groups. Conversely, the empirical probability of choosing handwriting for Asians increased from a low of .43 to a high of .58 across the age groups. We believe that this difference is large enough to be of substantive importance, so we chose to include it in the final model.



Note. “Asia” = “Asia/Pacific Islands;” “South America” = “Central/South America.”

Figure 9. Medium-choice age-by-continent interaction.

Group Comparisons

Table 10 presents the parameter estimates from the general linear model predicting essay scores based on composition medium while controlling for English language proficiency and examinee demographic characteristics. From the η^2 indices, it is clear that the variables with the greatest impact on essay scores are the two-way interaction between composition medium and English proficiency and the main effects for English proficiency, age, and gender.

Overall, when controlling for differences due to demographic characteristics, examinees who chose to compose essays in handwriting were predicted to receive lower scores than examinees who did not. For example, the model predicted that an African female test-taker who speaks a language based on a Roman or Cyrillic alphabet, has an average multiple-choice test score (17.72), and produces an essay in handwriting would receive an essay score equal to 4.83, while her word-processing counterpart would receive an essay score equal to 4.56. However, the two-way interaction indicated that this difference was greater for examinees who received lower scores on the multiple-choice test than it was for examinees who received higher scores on the multiple-choice test. For example, an African female examinee who speaks a language based on a Roman or Cyrillic alphabet but receives a fairly low multiple-choice score (e.g., 5.00 points) had predicted handwriting and word-processing essay scores equal to 2.94 and 2.12, respectively, while her higher-scoring counterpart (e.g., 20.00 points) had predicted handwriting and word-processing essay scores equal to 5.16 and 5.00, respectively.

Table 11 shows the empirical essay means of each composition medium, conditioned on multiple-choice decile (English proficiency). From these values, it is apparent that scores assigned to handwritten essays tended to be higher than scores assigned to word-processed essays, but that the difference was greater for examinees who received lower scores on the multiple-choice section of the TOEFL. The least square means for examinees who received the lowest multiple-choice scores was almost one-half of a point on the six-point rating scale. For examinees who received the highest scores on the multiple-choice section, there was almost no difference between the predicted handwriting and word-processing essay scores.

Table 10***Model Comparisons of Handwritten and Word-Processed Essay Scores***

Parameter	Level	Estimate	SE	η^2
Intercept		1.68	0.020	
Medium		1.04	0.020	.001
English		0.19	0.001	.242
Age		-0.01	0.003	.008
Gender		-0.12	0.004	.004
Continent				.001
	Asia/Pacific Islands	-0.13	0.010	
	Europe	-0.27	0.010	
	Middle East	-0.23	0.010	
	North America	-0.28	0.010	
	Central/South America	-0.35	0.010	
Keyboard		0.07	0.006	.001
English \times Medium		-0.04	0.001	.007

Note. The η^2 shown here is based on the Type III sum of squares.

Table 11***Essay Means Conditioned on Composition Medium by English Proficiency***

English proficiency decile	Word-processing mean score	Handwriting mean score	Difference
1	2.70	3.11	0.40
2	3.17	3.59	0.41
3	3.42	3.80	0.38
4	3.63	3.98	0.35
5	3.82	4.13	0.31
6	3.98	4.28	0.30
7	4.18	4.44	0.27
8	4.43	4.66	0.23
9	4.70	4.88	0.18
10	5.09	5.14	0.05

In addition, meaningfully large main effects were found for both age and gender, even after taking English proficiency into account. Hence, there seem to be age- and gender-related differences in essay scores for non-English speakers beyond those accounted for by composition medium and familiarity with the English language.

Summary of Results

Quality of Ratings

Relating to the quality of the ratings, we draw the following conclusions:

- Scores assigned to word-processed essays are slightly more reliable than scores assigned to handwritten essays.
- Scores assigned to word-processed essays exhibit slightly higher correlations with the multiple-choice components of the TOEFL than do scores from handwritten essays.

These results are consistent with previous studies that compared the psychometric qualities of word-processed and handwritten essays (Bridgeman & Cooper, 1998; Powers, Fowles, Farnum, & Ramsey, 1994). Bridgeman and Cooper hypothesized that the lack of handwriting variation in word-processed essays results in increased standardization and agreement between raters.

Group Characteristics

We found several main effects relating demographic characteristics to composition-medium choice. Specifically, we draw the following conclusions:

- Females are more likely than males to choose handwriting as the composition medium.
- Examinees who speak a language that is not based on a Roman/Cyrillic alphabet are more likely to choose handwriting than are examinees who do speak a language that is based on a Roman or Cyrillic alphabet.
- Examinees with poorer English language skills, as measured by the multiple-choice sections of the TOEFL test, are more likely to choose handwriting than are examinees with better English skills.

In addition, an age-by-continent interaction exists. Specifically, we draw the following conclusions:

- For examinees from Africa, the Middle East, and Central/South America, only small differences exist between age groups in their tendencies to choose each composition medium.
- For Asian examinees, the probability of choosing handwriting as the composition medium increases with age.
- The probability of choosing handwriting decreases with age for examinees from Europe—except for the oldest age group, which also has the highest probably of choosing handwriting.
- The small effect for age is slightly curvilinear, with the youngest and the oldest examinees being most likely to choose handwriting.

We found several additional statistically significant two-way interactions—namely, continent-by-gender, continent-by-keyboard, age-by-English, and continent-by-English—but small effect sizes made these interactions uninteresting from a substantive perspective. Although no previous research concerning composition-medium choice for foreign-language writing tests exists, our results are consistent with the expectations that groups who have historically exhibited lower levels of computer experience and higher levels of computer anxiety are less likely to choose the word processor as their composition medium. However, the age-by-continent interaction we observed indicates that the anticipated influence of age on composition-medium choice (i.e., that older examinees would be more likely to choose handwriting) may vary across geographic regions.

Essay Performance

Relating to the group comparisons of essay performance, we draw the following conclusions:

- Overall, there is no difference between essay scores of examinees who choose to compose their essays in handwriting and word-processing.
- When differences in overall English proficiency between composition-medium groups are controlled, a small interaction emerges. Specifically, examinees who have lower scores on the TOEFL multiple-choice sections tend to have higher handwritten essay scores, and examinees who have higher scores on the TOEFL multiple-choice sections tend to have similar scores on handwritten and word-processed essays.

- No substantively important medium-by-covariate interactions exist. That is, an examinee's geographic region, gender, age, and native language do not influence the comparability of scores on handwritten and word-processed essays, once overall English proficiency is taken into account.
- Substantively important main effects for covariates exist on essay scores, even when English language proficiency is taken into account. Specifically, the age and gender main effects produce a meaningfully large effect size even after taking into account both composition medium and English language proficiency.

These results are consistent with previous research concerning testing-medium differences in direct writing assessment. Specifically, Wolfe, Bolton, Feltovich, and Niday (1996) found that secondary-level English-speaking junior-high-school students in the United States who had considerable experience and above-average levels of comfort using computers exhibited no differences between scores on handwritten and word-processed essays, while students with lower levels of computer experience and comfort scored considerably higher on handwritten essays. Russell and Haney (1997) demonstrated a predictably similar effect for examinees with very high levels of computer experience and comfort. Specifically, that study demonstrated that students from technology-oriented schools received higher scores on a computer-based writing assessment than on a paper-and-pencil version of the assessment. In addition, the magnitude of our effect size at the lower end of the multiple-choice score distribution was consistent with that demonstrated in an analysis of data similar to those reported here (Breland, Muraki, & Lee, 2001).

Discussion

What is most interesting about the results of this study, in comparison to the results of studies of similar populations concerning cross-medium performance on multiple-choice tests, is that our study revealed a predictable differential cross-medium performance, while most studies of cross-medium performance on multiple-choice tests have revealed no large group differences. For example, in a study of a variety of large-scale assessments, Gallagher, Bridgeman, and Cahalan (2002) found only small differences (beyond those observed for paper-and-pencil tests) between racial/ethnic or gender groups on computer-based tests. In fact, African American and Hispanic examinees—examinees who would be expected to have less exposure to computers—

actually seemed to fair better on computer-based versions of multiple-choice examinations. Similarly, Taylor et al. (1999) concluded that, once ability differences between TOEFL examinees who chose computer-based and paper-and-pencil examinations were removed, a negligible relationship existed between test medium and test performance.

Implications

In general, we found support for the somewhat complex model we posited earlier, which depicts the relationship between several examinee characteristics (i.e., test anxiety, test preparation, achievement, computer skill, computer anxiety, and computer attitudes) and examinee performance on a computer-based direct writing assessment (see Figure 2). Groups that have traditionally been associated with lower levels of computer experience and higher levels of computer anxiety (most notably, females), or who could be predicted to exhibit these characteristics (e.g., examinees with lower levels of English proficiency, examinees who speak languages based on alphabets other than a Roman or Cyrillic alphabet, examinees from developing regions, and the oldest and the youngest examinees), are all more likely to choose to compose essays using handwriting than a word processor. We found it somewhat surprising that our results added younger examinees to this list, which we speculate may be due to that group being more heterogeneous in the TOEFL examinee population.

In addition, the relationship between composition-medium choice and an examinee's age also varies across geographic regions. Generally, the curvilinear trend we observed in most regions (indicating higher probabilities of choosing handwriting for the oldest and for the youngest examinees) is not followed by Asian examinees (for whom the oldest examinees are most likely to choose handwriting) and European examinees (for whom the youngest are most likely to choose handwriting). We also have difficulty explaining this trend on substantive grounds. However, if this interaction is ignored, the results indicate that there would be large differences between continents with respect to composition-medium choice, with examinees from Africa and the Middle East being most likely to choose handwriting and examinees from North and Central/South America being the least likely to choose handwriting.

We believe that the remaining trends can be attributed to the notion that examinee choice of composition medium is driven by that examinee's comfort and familiarity with using computers for writing tasks. Each of the groups who exhibited higher probabilities for choosing handwriting was identified as potentially being "at risk" with respect to computer familiarity and

comfort, and each of these groups exhibited a lower tendency to choose word-processing as a composition medium for a direct writing assessment. Unfortunately, this study does not directly address this relationship: We did not directly measure computer familiarity or experience. Nonetheless, we believe our evidence supports that notion.

In addition, our results suggest that there is a relationship between composition medium and essay ratings, even when group differences in English language proficiency are removed. Specifically, after employing procedures similar to the analysis of covariance performed by Taylor et al. (1999) and the logistic regression performed by Breland, Muraki, and Lee (2001), we determined that average scores assigned to handwritten essays were almost one-half of a point higher for examinees with low scores on the TOEFL multiple-choice sections. However, there were no differences between scores from the two composition media for examinees who received high scores on the TOEFL multiple-choice sections.

We believe that this result suggests that examinees with lower levels of English proficiency—examinees who are also likely to have less experience and less comfort using computers—may encounter additional cognitive demands when responding to the writing prompt using a keyboard. It is reasonable to claim that that additional cognitive demand constitutes construct-irrelevant variance, rendering the writing assessment to be a less valid indicator of the examinee's written communication skill when the essay is generated in a computer-based environment.

We liken this to the comparability problem that arises in the translation of tests from one language to another in international testing. In the case of the TOEFL exam, the first translation takes place when the examinee translates a thought from the native language to English. A second translation takes place when the examinee translates the English version of the thought into keyboard presses so that the words appear on the computer monitor. Of course, the first translation is more difficult for examinees who have poorer English skills. Similarly, the second translation is more difficult for examinees who have poorer computer skills—a skill that is not relevant for measuring English language proficiency on examinations like the TOEFL test.

Obviously, the practical implication for these results is that examinees should be afforded choice of composition medium when high-stakes decisions will be made based upon scores from direct writing assessments that are offered in a computer-based format. Fortunately, this is current practice for the written section of the TOEFL examination. What is troubling, however, is the fact that we know little about the accuracy of examinees' beliefs about their own levels of

computer skill and the factors that examinees consider when choosing between composition media on a direct writing assessment. Although Russell's results (1999) were based on U.S. populations, it is disconcerting that he found that examinees generally believe that they will receive higher scores on computer-based examinations. Hence, another important implication of these results for testing practice is that examinees should be made aware of potential differences in performance on computer-based and pencil-and-paper writing assessments and the interaction between computer facility and test performance.

Future Research

We can identify three important areas for further research concerning the comparability of handwritten and word-processed essays that are written for the TOEFL writing section. First, it is important to determine to what degree qualitative differences exist between handwritten and word-processed essays and the processes that examinees use to produce these essays. That is, we not only need to study differences in the textual components of essays written in the two composition media, but we also need to study how the writing process differs for these two modes of composition. Gentile (2001) has begun focusing on this issue.

Second, it is important to understand the reasoning that goes into the choice of composition medium on the part of the examinee. Our analyses have demonstrated that, in general, examinees who have lower levels of English proficiency tend to choose handwriting as their medium, and they write better essays in handwriting than they do in word-processing. However, almost 40% of the examinees who scored in the lowest quartile on the TOEFL multiple-choice sections chose to word-process their essays. For these examinees, the predicted difference between handwritten and word-processed essay scores is almost one-half of a score point. However, we believe that further research, as described in the next paragraph, is necessary prior to making recommendations to examinees regarding which medium to choose.

Third, we believe that it is important to understand whether—and if so, why—some examinees may exhibit differential performance when handwriting and word-processing TOEFL essays. The biggest limitation of this study is the fact that we simply compared the quality of essays written by examinees who chose handwriting and word-processing, and we made projections concerning average score differences for individuals based on those intact groups. Our results may have varied somewhat if examinees had been randomly assigned to composition medium or if repeated measures had been made for each examinee—once under each

composition medium.

A serious shortcoming of most research concerning score differences attributable to test delivery medium is the fact that most of these studies examine group differences rather than individual differences. These studies have suggested that, on average, there are only small differences between scores on computer-based and pencil-and-paper tests. Unfortunately, to our knowledge no studies have attempted to ascertain the magnitude of the impact of testing medium on individual examinees, particularly those who are members of groups who may be expected to be “at risk” due to lower levels of computer familiarity and comfort or higher levels of computer anxiety.

References

- Agresti, A. (1996). *An introduction to categorical data analysis*. New York: Wiley.
- Allison, P. D. (1999). *Logistic regression using the SAS system: Theory and application*. Cary, NC: SAS Institute.
- Arnold, V., Legas, J., Obler, S., Pacheco, M. A., Russell, C., & Umbdenstock, L. (1990). *Direct writing assessment: A study of bias in scoring hand-written vs. wordprocessed papers*. Unpublished manuscript, Rio Hondo College, Whittier, CA.
- Bradley, V. (1982). Improving students' writing with microcomputers. *Language Arts*, 59, 732-743.
- Braine, G. (1997). Beyond word processing: Networked computers in ESL writing classes. *Computers and Composition*, 14, 45-58.
- Breland, H., Muraki, E., & Lee, Y. W. (2001, April). *Comparability of TOEFL CBT writing prompts for different response modes*. Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.
- Bridgeman, B., & Cooper, P. (1998, April). *Comparability of scores on word-processed and handwritten essays on the graduate management admissions test*. Paper presented at the American Educational Research Association, San Diego, CA.
- Bridwell, L., Sirc, G., & Brooke, R. (1985). Case studies of student writers. In S. W. Freedman (Ed.), *The acquisition of written language* (pp. 172-194). Norwood, NJ: Ablex.
- Broderick, C., & Trushell, J. (1985). Word processing in the primary classroom. In J. Ewing (Ed.), *Reading and the new technologies* (pp. 119-128). London: Heinemann Educational Books.
- Brown, J. D. (1997). Computers in language testing: Present research and some future directions. *Language Learning and Technology*, 1, 44-59.
- Bruce, B., Michaels, S., & Watson-Gegeo, K. (1985). How computers can change the writing process. *Language Arts*, 62, 143-149.
- Campbell, N. J. (1989). Computer anxiety of rural middle and secondary school students. *Journal of Educational Computing Research*, 5, 213-220.
- Cochran-Smith, M. (1991). Word processing and writing in elementary classrooms: A critical review of related literature. *Review of Educational Research*, 61, 107-155.

- Cochran-Smith, M., Paris, C. L., & Kahn, J. (1991). *Learning to write differently*. Norwood, NJ: Ablex.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.
- Collier, R. (1983). Writing processes and revising strategies: Study of effects of computer-based text editors on revising strategies for independent writers. *College Composition and Communication, 34*, 149-155.
- Collier, R., & Werier, C. (1995). When computer writers compose by hand. *Computers and Composition, 12*, 47-59.
- Daiute, C. (1986). Physical and cognitive factors in revising: Insights from studies with computers. *Research in the Teaching of English, 34*, 149-155.
- Dalton, D., & Hannafin, M. (1987). The effects of word processing on written composition. *Journal of Educational Research, 50*, 223-228.
- Dickenson, D.K. (1986). Cooperation, collaboration, and a computer: Integrating a computer into a first-second grade writing program. *Research in the Teaching of English, 20*, 357-378.
- ETS. (1999). Description of the computer-based TOEFL test. Princeton, NJ: Author. Retrieved October 10, 1999, from: <http://www.ets.org/toefl/description.html>
- Ford, B. D., Romeo, V., & Stuckless, N. (1996). The effects of computer versus paper-and-pencil administration on measures of anger and revenge with an inmate population. *Computers in Human Behavior, 12*, 159-166.
- 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.
- Gentile, C., Riazantseva, A., & Cline, F. (2001). *A comparison of handwritten and word processed TOEFL essays: Final report*. Unpublished manuscript.
- Gressard, C. P., & Loyd, B. H. (1987). An investigation of the effects of math anxiety and sex on computer attitudes. *School Science and Mathematics, 87*(2), 125-135.
- Grignon, J. R. (1993). Computer experience of Menominee Indian students: Gender differences in coursework and use of software. *Journal of American Indian Education, 32*, 1-15.
- Hawisher, G. (1987). The effects of word processing on the revision strategies of college freshmen. *Research in the Teaching of English, 21*, 145-159.

- Henning, G. (1991). Validating an item bank in a computer-assisted or computer-adaptive test using item response theory for the process of validating CATS. In P. Dunkel (Ed.), *Computer-assisted language learning and testing: Research issues and practice* (pp. 209-222). New York: Newbury House.
- Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York: John Wiley & Sons.
- Howell, D. C. (2002). *Statistical methods for psychology* (5th ed.). Pacific Grove, CA: Duxbury.
- Jamieson, J., Taylor, C., Kirsch, I., & Eignor, D. (1999). *Design and evaluation of a computer-based TOEFL tutorial* (ETS Research Report 99-1). Princeton, NJ: ETS.
- Janssen Reinen, I., & Plomp, T. (1993). Some gender issues in educational computer use: Results of an international comparative survey. *Computers in Education, 20*, 353-365.
- Johnson, D. F., & White, C. B. (1980). Effects of training on computerized test performance in elderly. *Journal of Applied Psychology, 65*, 357-358.
- Kane, J. (1983). *Computers for composing* (Technical Report No. 21). New York: Bank Street College of Education.
- Kehagia, O., & Cox, M. (1997). Revision changes when using wordprocessors in an English as a foreign language context. *Computer Assisted Language Learning, 10*, 239-253.
- Keogh, T., Barnes, P., Joiner, R., & Littleton, K. (2000). Gender, pair composition, and computer versus paper presentation of an English language task. *Educational Psychology, 20*, 33-43.
- Kim, S., & Hocevar, D. (1998). Racial differences in eighth-grade mathematics: Achievement and opportunity to learn. *Clearing House, 71*, 175-178.
- Kurth, R.J. (1987). Using word processing to enhance revision strategies during student writing activities. *Educational Technology, 127*(1), 13-19.
- Lankford, J.S., Bell, R.W., & Elias, J.W. (1994). Computerized versus standard personality measures: Equivalency, computer anxiety, and gender differences. *Computers in Human Behavior, 10*, 497-510.
- Lee, J. A. (1986). The effects of past computer experience on computerized aptitude test performance. *Educational and Psychological Measurement, 46*, 727-733.
- Legg, S. M., & Buhr, D.C. (1992). Computerized adaptive testing with different groups. *Educational Measurement: Issues and Practice, 11*(2), 23-27.

- Levin, J., Riel, M., Rowe, R., & Boruta, M. (1985). Muktuk meets jacuzzi: Computer networks and elementary school workers. In S. W. Freedman (Ed.), *The acquisition of written language* (pp. 160-171). Norwood, NJ: Albex.
- Levin, T., & Gordon, C. (1989). Effect of gender and computer experience on attitudes toward computers. *Journal of Educational Computing Research*, 5, 69-88.
- Lloyd, B. H., & Gressard, C. P. (1986). Gender and amount of computer experience of teachers in staff development programs: Effects on computer attitudes and perceptions of usefulness of computers. *AEDS Journal*, 19, 302-311.
- Lutz, J. A. (1987). A study of professional and experienced writers revising and editing at the computer and with pen and paper. *Research in the Teaching of English*, 21, 398-421.
- MacArthur, C. A. (1988). *Write to learn*. New York: Holt, Rinehart, and Winston.
- Madsen, H. (1991). Computer-adaptive testing of listening and reading comprehension: The Brigham Young University approach. In P. Dunkel (Ed.), *Computer-assisted language learning and testing: Research issues and practices* (pp. 237-257). New York: Newbury House.
- Manalo, J. R., & Wolfe, E. W. (2000a). *A comparison of word-processed and handwritten essays written for the Test of English as a Foreign Language*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.
- Manalo, J. R., & Wolfe, E. W. (2000b). *The impact of composition medium on essay raters in foreign language testing*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.
- Marcoulides, G. A. (1988). The relationship between computer anxiety and computer achievement. *Journal of Educational Computing Research*, 4, 151-158.
- Massoud, S. L. (1992). Computer attitudes and computer knowledge of adult students. *Journal of Educational Computing Research*, 7, 269-291.
- Mazzeo, J., & Harvey, A. L. (1988). *The equivalence of scores from automated and conventional versions of educational and psychological tests: A review of the literature* (College Board Report 87-8). Princeton, NJ: ETS.
- Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin*, 114, 449-458.

- Miller, F., & Varma, N. (1994). The effects of psychosocial factors on Indian children's attitudes toward computers. *Journal of Educational Computing Research, 10*, 223-238.
- Nolan, P. C. J., McKinnon, D. H., & Soler, J. (1992). Computers in education: Achieving equitable access and use. *Journal of Research on Computing in Education, 24*, 299-314.
- Oweston, R. D., Murphy, S., & Wideman, H. H. (1992). The effects of word processing on students' writing quality and revision strategies. *Research in the Teaching of English, 26*, 249-276.
- Pennington, M. C. (1993). A critical examination of word processing effects in relation to L3 writers. *Journal of Second Language Writing, 2*, 227-255.
- Phinney, M., & Khouri, S. (1993). Computers, revision, and ESL writer: The role of experience. *Journal of Second Language Writing, 2*, 257-277.
- Porter, R. (1987). Writing and word processing in year one. *Australian Educational Computing, 1*, 18-23.
- Powers, D. E. (1993). Coaching for the SAT: A summary of the summaries and an update. *Educational Measurement: Issues and Practice, 12*(2), 24-30.
- Powers, D. E., Fowles, M. E., Farnum, M., & Ramsey, P. (1994). Will they think less of my handwritten essay if others word process theirs? Effects on essay scores of intermingling handwritten and word-processed essays. *Journal of Educational Measurement, 31*, 220-233.
- Powers, D. E., & O'Neill, K. (1993). Inexperienced and anxious computer users: Coping with a computer-administered test of academic skills. *Educational Assessment, 1*, 153-173.
- Powers, D. E., & Rock, D. A. (1999). Effects of coaching on SAT I: Reasoning test scores. *Journal of Educational Measurement, 36*, 93-118.
- Raftery, A. E. (1995). Bayesian Model Selection in Social Research. *Sociological Methodology, 25*, 111-163.
- Ronau, R. N., & Battista, M. T. (1988). Microcomputer versus paper-and-pencil testing of student errors in ratio and proportions. *Journal of Computers in Mathematics and Science Teaching, 8*, 33-38.
- Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research: A correlational approach*. New York: Cambridge University Press.

- Russell, M. (1999). *Testing on computers: A follow-up study comparing performance on computer and on paper*. Education Policy Analysis Archives. Retrieved June 8, 1999, from: <http://epaa.asu.edu/epaa/v7n20>
- Russell, M., & Haney, W. (1997). *Testing writing on computers: An experiment comparing student performance on tests conducted via computer and via paper-and-pencil*. Education Policy Analysis Archives. Retrieved January 15, 1997, from: <http://epaa.asu.edu/epaa/v5n3.html>
- Shashaani, L. (1997). Gender differences in computer attitudes and use among college students. *Journal of Educational Computing Research, 16*, 37-51.
- Shermis, M. D., & Lombard, D. (1998). Effects of computer-based test administrations on test anxiety and performance. *Computers in Human Behavior, 14*, 111-123.
- Siann, G., Macleod, H., Glissoy, P., & Durndell, A. (1990). The effect of computer use on gender differences in attitudes to computers. *Computers in Education, 14*, 183-191.
- Signer, B.R. (1991). CAI and at-risk minority urban high school students. *Journal of Research on Computing in Education, 24*, 189-203.
- Spray, J. A., Ackerman, T. A., Reckase, M. D., & Carlson, J. E. (1989). Effect of the medium of item presentation on examinee performance and item characteristics. *Journal of Educational Measurement, 26*, 261-271.
- Stevenson, J., & Gross, S. (1991). Use of a computerized adaptive testing model for ESOL/bilingual entry/exit decision making. In P. Dunkel (Ed.), *Computer-assisted language learning and testing: Research issues and practice* (pp. 223-235). New York: Newbury House.
- Taylor, C., Kirsch, I., Eignor, D., & Jamieson, J. (1999). Examining the relationship between computer familiarity and performance on computer-based language tasks. *Language Learning, 49*, 219-274.
- Temple, L., & Lips, H. M. (1989). Gender differences and similarities in attitudes toward computers. *Computers in Human Behavior, 5*, 215-226.
- Tseng, H. M., Macleod, H. A., & Wright, P. (1997). Computer anxiety and measurement of mood change. *Computers in Human Behavior, 13*, 305-316.
- Turner, B. G. (1993). Test anxiety in African American school children. *School Psychology Quarterly, 8*, 140-152.

- Whitely, B. E. (1997). Gender differences in computer-related attitudes and behavior: A meta-analysis. *Computers in Human Behavior, 13*, 1-22.
- Wiley, D. E., & Yoon, B. (1995). Teacher reports on opportunity to learn: Analyses of the 1993 California Learning Assessment System (CLAS). *Educational Evaluation and Policy Analysis, 17*, 355-370.
- Wise, S. L., & Plake, B. S. (1989). Research on the effects of administering tests via computers. *Educational Measurement: Issues and Practice, 8*(3), 5-10.
- Wolfe, E. W., Bolton, S., Feltovich, B., & Bangert, A. W. (1996). A study of word processing experience and its effects on student essay writing. *Journal of Educational Computing Research, 14*, 269-284.
- Wolfe, E. W., Bolton, S., Feltovich, B., & Niday, D. M. (1996). The influence of student experience with word processors on the quality of essays written for a direct writing assessment. *Assessing Writing, 3*, 123-147.



**Test of English as a Foreign Language
PO Box 6155
Princeton, NJ 08541-6155
USA**

To obtain more information about TOEFL programs and services, use one of the following:

**Phone: 1-877-863-3546
(US, US Territories*, and Canada)**

**1-609-771-7100
(all other locations)**

Email: toefl@ets.org

Web site: www.ets.org/toefl

* America Samoa, Guam, Puerto Rico, and US Virgin Islands