J·T·L·A

The Journal of Technology, Learning, and Assessment

Volume 2, Number 4 · December 2003

Examinee Characteristics Associated With Choice of Composition Medium on the TOEFL Writing Section

Edward W. Wolfe

www.jtla.org

A publication of the Technology and Assessment Study Collaborative
Caroline A. & Peter S. Lynch School of Education, Boston College

J·T·L·A

Volume 2, Number 4

Examinee Characteristics Associated With Choice of Compostion Medium on the TOEFL Writing Section

Edward W. Wolfe

Editor: Michael Russell
russelmh@bc.edu
Technology and Assessment Study Collaborative
Lynch School of Education, Boston College
Chestnut Hill, MA 02467

Copy Editor: Kathleen O'Connor
Design and Layout: Thomas Hoffmann

JTLA is a free on-line journal, published by the Technology and Assessment Study Collaborative, Caroline A. & Peter S. Lynch School of Education, Boston College.

Preferred citation:

Wolfe, E. W. (2003). Examinee characteristics associated with choice of composition medium on the TOEFL writing section. *Journal of Technology, Learning, and Assessment, 2*(4). Available from http://www.jtla.org

Abstract:

The Test of English as a Foreign Language (TOEFL) contains a direct writing assessment, and examinees are given the option of composing their responses at a computer terminal using a keyboard or composing their responses in handwriting. This study sought to determine whether examinees from different demographic groups choose handwriting versus word-processing composition media with equal likelihood. The relationship between several demographic characteristics of examinees and their composition medium choice on the TOEFL writing assessment is examined using logistic regression. Females, speakers of languages based on non-Roman/Cyrillic character systems, examinees from Africa and the Middle East, and examinees with less proficient English skills were more likely to choose handwriting. Although there were only small differences between age groups with respect to composition medium choice in most geographic regions, younger examinees from Europe and older examinees from Asia were more likely to choose handwriting than their regional counterparts.

J·T·L·A

Examinee Characteristics Associated with Choice of Composition Medium on the TOEFL Writing Section

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 being administered via computer, and this transition has improved the way tests are administered and test scores are reported. However, some fear that the shift toward a computer-based testing system may exacerbate existing social barriers to advancement opportunities for women, minorities, economically disadvantaged individuals, and the elderly. Previous studies of the comparability of computer-based and paper-and-pencil tests have identified only small differences between average scores on multiple-choice tests administered in these two media. However, two important additional issues have been given less attention – the influence of computer-based testing on the scores of "at risk" groups of examinees and the comparability of performance-based tests (e.g., direct writing assessments) administered in these two media.

It is possible that variables such as computer proficiency, comfort, and attitude differentially influence examinee performance on computer-based writing assessments. If so, then the adoption of computer-based direct writing assessments could introduce construct-irrelevant variance into the measurement of writing ability. A further complexity may be introduced into the study of cross-medium performance when examinees are allowed to choose the medium in which they compose their essays. An examinee is likely to consider a variety of factors when choosing a composition medium, and the weight given to each of these factors may vary between examinees. Further, the accuracy of one's beliefs concerning one's own computer ability and its influence on the quality of the writing one produces may vary between examinees. Hence, it seems that an important first step in understanding the interplay of composition medium choice and cross-medium performance on direct writing assessments is to identify characteristics that differentiate examinees who choose one composition medium versus the other. The purpose of this study is to identify whether, and if so, how, examinees who choose computer versus paper-and-pencil administration media for the written section of the Test of English as a Foreign Language (TOEFL) differ with respect to demographic characteristics.

Consider the model of factors that influence test performance shown in Figure 1. The left side of that figure identifies several variables that influence test performance, regardless of administration medium. Test performance is directly influenced by an examinee's (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 that these variables are related. Test performance is influenced by an examinee's academic achievement which, in turn, is influenced by opportunity to learn (Wiley & Yoon, 1995). Test preparation influences test performance and is likely to lower test anxiety (Powers, 1993; Powers & Rock, 1999). Finally, an examinee's social background influences opportunity to learn, test preparation, and achievement levels (Kim & Hocevar, 1998; Powers, 1993; Turner, 1993). Hence, it is clear that social circumstances play an important role in determining performance on conventional tests.

Figure 1

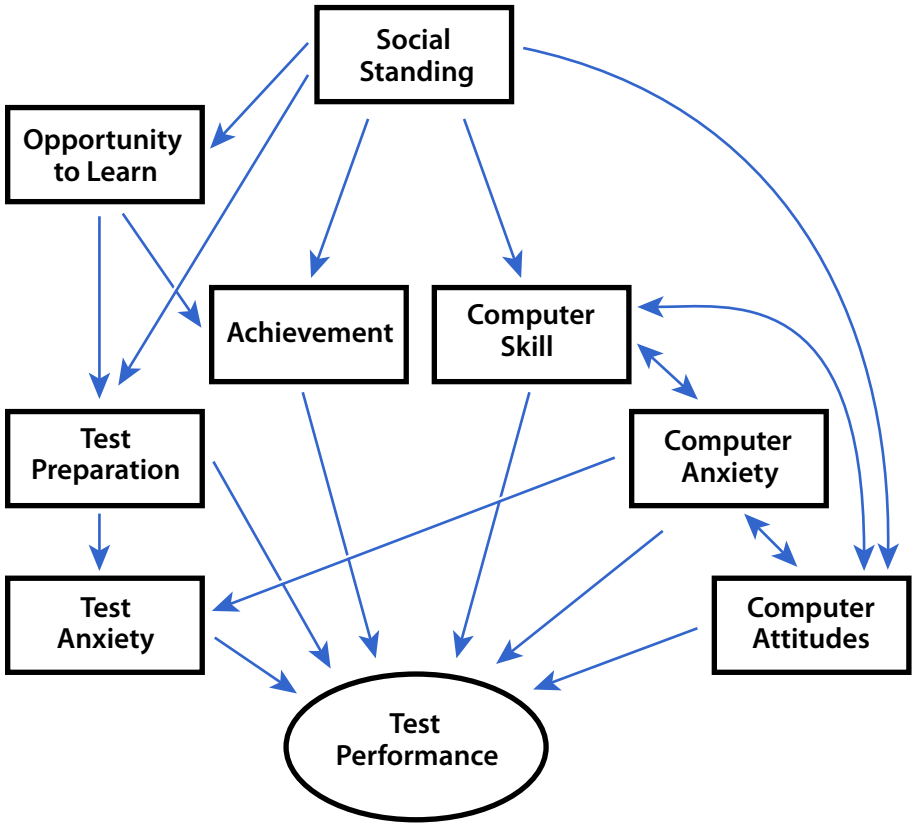


Figure 1. Computer-based test performance model.

The right side of Figure 1 contains additional variables that are relevant for computer-based tests: (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.

What evidence exists to support this model? First, previous research indicates that some groups have restricted access to, experience and proficiency with, and less favorable attitudes toward computers. In the U.S., school-aged minorities and women are less likely to have computers in their homes and males are more likely to dominate computer use at school (Grignon, 1993; Keogh, Barnes, Joiner, & Littleton, 2000). Internationally, women, Africans, and Spanish speakers have restricted access to computer use (Janssen Reinen & Plomp, 1993; Miller & Varma, 1994; Taylor, Kirsch, Eignor, & Jamieson, 1999). Not only do these groups differ in their access to computers, but they also differ in the way they use computers (Wenglinsky, 1998). For example, minorities are less likely to use computers for instructional purposes other than simple drill and practice. Although differences in computer access are likely to diminish over time, particularly in technology-rich countries, differences in computer use are likely to persist in regions that exhibit slower technological growth. Inequities in computer access, familiarity, and usage may lead to lower levels of confidence and higher levels of anxiety toward computer-based tasks. In all age groups, minorities (in the U.S.) and women (internationally) exhibit higher levels of computer anxiety and lower levels of confidence for computer-related tasks, and group differences in anxiety levels are greatly diminished when computer experience is held constant (Janssen Reinen & Plomp, 1993; Legg & Buhr, 1992; Massoud, 1992; Nolan, McKinnon, & Soler, 1992; Shashaani, 1997; Whitely, 1997). And, unfortunately, affective responses (e.g., computer anxiety) and proficiencies (e.g., levels of computer experience) are correlated with test scores on computer-based tests at non-trivial levels (Wolfe, Bolton, Feltovich, & Bangert, 1996; Wolfe, Bolton, Feltovich, & Niday, 1996).

Comparisons of standardized (typically multiple-choice) computer-based and paper-based tests suggest that there are no large differences in test performance at a population level. Generally, the small effect sizes indicate that examinees may receive slightly higher scores on paper-based tests (Mead & Drasgow, 1993). Interestingly, students may believe that they will receive higher scores on computer-based tests – a misperception that may drive some examinees to select a testing medium on which they will receive lower scores (Russell, 1999). It should be noted, however, that population-level (rather than group- or individual-level) comparisons fail to take into account the possibility that the influence of computer medium on test performance may be profound for small portions of the population (Wise & Plake, 1989). Empirical evidence suggests that females may receive higher scores on paper-based tests, but that, contrary to what one might expect, recent evidence suggests that some U.S. minority groups may actually receive higher scores on computer-based tests (Gallagher, Bridgeman, & Cahalan, 2002).

Unfortunately, most of the research concerning the comparability of paper-based and computer-based tests has focused on standardized multiple-choice tests. Studies concerning writing assessments suggest the following about the impact of computers on writing assessment scores. First, the appearance of essays in written versus typed text may influence raters. Specifically, raters may have higher expectations for word-processed text (Arnold et al., 1990; Gentile, Riazantseva, & Cline,

2001), but raters may also be better able to agree on scores for word-processed text because of the elimination of handwriting effects (Bridgeman & Cooper, 1998; Wolfe & Manalo, in press). The study of rater-by-medium effects is complicated by the difficulty of disentangling rater-by-medium interactions from examinee-by-medium interactions. Some researchers have attempted to evaluate the impact of textual appearance on essay ratings by comparing scores on originals to scores on essays that are transcribed to another medium (i.e., from word-processed to handwritten or from handwritten to word-processed) (Arnold et al., 1990; Harrington, Shermis, & Rollins, 2000; MacCann, Eastment, & Pickering, 2002; Powers, Fowles, Farnum, & Ramsey, 1994; Wolfe, Bolton, Feltovich, & Welch, 1994). However, several of these studies suggest that the interpreted medium effects may be due to a transcription effect because readers assign higher scores to the original essay, regardless of the presented medium (Harrington et al., 2000; MacCann et al., 2002; Wolfe et al., 1994). Regardless, readers can be trained to compensate for differential expectations they may have about the quality of handwritten and word-processed text, although these effects may not be completely removed (Powers et al., 1994).

Second, the use of word processors seems to influence essay content regardless of the quality of the writing. Compared to handwritten essays, word-processed essays tend to contain shorter sentences but more text, be better organized, freer of mechanical errors, and be neater, more formal in tone, and exhibit weaker voice (Collier & Werier, 1995; Gentile et al., 2001; Goldberg, Russell, & Cook, 2003; Russell & Haney, 1997; Wolfe, Bolton, Feltovich, & Niday, 1996).

Most important, however, there seems to be an interaction between computer experience or proficiency and the quality of essays written via keyboarding or handwriting. Specifically, examinees with lower levels of computer experience receive higher scores when composing in handwriting, and examinees with higher levels of computer experience receive higher scores when composing with keyboards (Harrington et al., 2000; Russell, 1999; Russell & Haney, 1997; Wolfe, Bolton, Feltovich, & Bangert, 1996; Wolfe, Bolton, Feltovich, & Niday, 1996). It may be that this interaction exists because the use of word-processors on examinees with weaker computer and keyboarding skills interferes with the production of writing, but no such interference is encountered by examinees with stronger computer skills because keyboarding has become an automated process – the *interference hypothesis*. Affording examinees the opportunity to choose between computer-based and paper-based administration media may help ameliorate this situation, but little is known about the appropriateness of decisions that are made by examinees when given such a choice. It is also important to note that, in most of these studies, the population of test takers is fairly homogeneous and computer literate having been drawn from the U.S. population – a population that is both fairly well educated and young. For example, Harrington and colleagues removed examinees who exhibited high levels of test anxiety and low levels of keyboarding skill from their sample causing that study to focus exclusively on examinees who have moderate-to-low levels of test anxiety and moderate-to-high levels of keyboarding skill

(Harrington et al., 2000). Hence, the results of such studies likely underestimate the magnitude of differential cross-medium test performance in more heterogeneous and less proficient international populations, like the population of TOEFL examinees. In fact, this seems to be the case. Analyses of similar data indicate that scores of essays produced using a word-processing composition medium tend to be lower for examinees once English language proficiency is taken into account (Wolfe & Manalo, in press).

The purpose of the research presented here is not to determine whether examinees make decisions that improve their chances of receiving high scores on direct writing assessments when given the opportunity to choose between composition media. Rather, the purpose is to identify the characteristics of examinees who choose handwriting versus keyboarding. Understanding these characteristics is an important first step in understanding the potential impact of composition medium choice in the context of direct writing assessments. By understanding the characteristics that differentiate examinees who choose handwriting from examinees who choose keyboarding, test developers will be better able to identify groups of examinees who are potentially "at risk" with respect to either making uninformed decisions concerning composition medium choice or exhibiting differential cross-medium performance. As a result, test developers may be better informed as they specify administration procedures and carry out validation studies, and they will be better able to evaluate whether examinees should be provided with choice of composition medium on direct writing assessments.

Method

In this study, logistic regression modeling was applied to several demographic variables as predictors of composition medium choice for a large sample of TOEFL examinees. Logistic regression is a generalized linear modeling procedure that allows analysts to evaluate whether linear combinations of continuous and/or categorical independent variables can be used as predictors of a dichotomous dependent variable (Allison, 1999; Hosmer & Lemeshow, 2000; Stokes, Davis, & Koch, 2000). Linearity of the relationship is achieved by scaling the dependent variable as the log of the odds of the event occurring versus not occurring. In the current study, the usefulness of several demographic characteristics of examinees as predictors of composition medium choice was evaluated. This section describes the examinees and the logistic regression procedures utilized in this study. Technical details concerning model selection and fit evaluation are presented in the Appendix.

Participants

Participants were 133,906 TOEFL examinees (a small portion of the total number) who participated in regular administrations of the computer-based TOEFL test between January 24, 1998 and February 9, 1999 and provided complete data (i.e., provided demographic data and valid multiple-choice and writing assessment scores). 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 was 24.26 years. The majority of examinees took the TOEFL for admittance into undergraduate or graduate studies (38% and 46%, respectively). Only 15% of the examinees indicated that they were taking the TOEFL for reasons other than to satisfy academic requirements. Each examinee completed the entire multiple-choice section of the examination in a computer-based testing environment but had the choice to respond to the single prompt for the direct writing assessment either using a word processor (54%) or in handwriting (46%).

Instrumentation

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 section is a direct writing assessment. The listening section measures the examinee's ability to understand English as it is spoken in North America. The structure section measures the examinee's ability to recognize language that is appropriate for standard written English using written stimuli. 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. 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.

As stated previously, examinees may choose to respond to the writing assessment in either handwriting or using a word processor. Scores from the listening and reading 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 (Educational Testing Service, 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. In addition, examinees provided self-report data about several demographic characteristics, including age, gender, native country, native language, reason for taking the TOEFL, and degree plans.

Analysis

The characteristics of individuals who chose each composition medium were compared via logistic regression. The explanatory variables were chosen for their substantive importance as mediators of an examinee's choice of composition medium. Older examinees – 36 years of age and older – (the *age* explanatory variable) were expected 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, females (*gender*) were expected to be more likely to choose handwriting because of their general lower levels of computer familiarity and the resulting higher levels of computer anxiety. Examinees from geographic *regions* in which there are several developing countries were expected to be more likely to choose handwriting because of the general lack of availability of computers in those regions. Examinees who exhibit lower levels of *English* proficiency (as indicated by their multiple-choice composite scores) were expected to be more likely to choose handwriting as the composition medium because of the double translation that would be required (thoughts-to-verbal expression and verbal expression-to-keyboard) to compose an essay using a word processor. Similarly, examinees who speak a native language that is not based on the Roman or Cyrillic alphabets were expected to be more likely to compose their essays in handwriting because of the difficulty of translating thoughts into words and words into keystrokes on a Roman alphabet keyboard (hence the variable name *keyboard*).

Medium, the outcome variable, was coded dichotomously, and composing an essay in handwriting rather than via word processor was the reference cell. As for 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 *regions*, treated nominally, of course: North American, Africa (reference cell), Asia and Pacific Islands, Central and South America, Europe, and the Middle East. *Keyboard* was treated as a dichotomous variable based on whether the examinee's language uses a keyboard containing an alphabet similar to the one used in English (e.g., Roman or Cyrillic) versus other systems (e.g., most Asian languages). The reference cell for keyboard was *other*. Note that, for each variable, the reference cell was the group believed to be most likely to choose handwriting as the composition medium.

The best fitting model took the following form:

$$\text{logit}(\textit{handwriting}) = 4.60 + \hat{\beta}A + \hat{\beta}R + \hat{\beta}E + \hat{\beta}G + \hat{\beta}K + \hat{\beta}A^2 + \hat{\beta}AR \quad \textbf{(1)}$$

where,

$\text{logit}(\textit{handwriting})$ = $\log(\pi_{\text{handwriting} \mid \text{covariates}} / \pi_{\text{word-processing} \mid \text{covariates}})$

$\pi_{\text{handwriting} \mid \text{covariates}}$ = predicted probability of the composition medium, given the examinee's demographic covariate pattern.

4.60 = the model's intercept – the value of logit(*handwriting*) when all covariates are set to equal zero.

$\hat{\beta}$ = the estimated contribution of each independent variable to the magnitude of logit(*handwriting*).

A = the examinee's age (in years)

R = the examinee's geographic region (generally, the continent on which the examinee took the TOEFL)

E = the examinee's English proficiency (i.e., the score on the multiple-choice portion of the TOEFL)

G = the examinee's gender (male versus female)

K = the keyboard of the examinee's native language (Roman/Cyrillic versus other)

Results

Table 1 displays the parameter estimates, standard errors, and the Wald statistics and their p-values for each variable in the best fitting model (see the Appendix for a discussion of these indices). Generally, as expected, the parameter estimates are negative, indicating that, in nearly all cases, the reference groups exhibited the greatest probability of choosing the handwriting medium. In all cases where the parameter estimates are positive, those estimates are close to zero. There are three main effects that are noteworthy. The *gender* main effect indicates that males were less likely than females to choose handwriting as the composition medium for their essays. As displayed in Table 2, the empirical proportion of females choosing handwriting was .49 while the proportion of males choosing handwriting was .43. The modeled probabilities (i.e., the expected proportions for females and males when the influence of other demographic variables on composition medium choice are held constant) were identical to the empirical probabilities. The *keyboard* main effect indicates that those who use a language based on the Roman/Cyrillic keyboard were less likely to choose handwriting as the composition medium for their essays. The empirical and modeled probabilities were .53 for choosing handwriting for non-Roman/Cyrillic language speakers and .38 for Roman/Cyrillic language speakers. Again, the modeled probabilities were identical to the empirical probabilities. 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 first decile of the multiple-choice section of the TOEFL to a low of .23 for the tenth decile. More specifically, the correlation between multiple-choice scores and composition medium is moderately large, $r_{point\text{-}biserial} = .25$. Generally, predicted probabilities were very similar to these values, with the largest discrepancies (about .07) in the first and tenth deciles.

Table 1 Summary of the Parameter Estimates for the Best Fitting Model

		Parameter		Statistical Significance	
Variable	**Level**	β	SE_β	χ^2_{Wald}	p
Intercept		4.60	0.15	919.02	<.0001
Age		-0.08	0.01	104.14	<.0001
Age²		0.00	0.00	142.12	<.0001
Region	Asia/PI	-2.25	0.13	319.38	<.0001
	Central/South America	-1.46	0.14	109.07	<.0001
	Europe	0.06	0.14	0.19	0.67
	Middle East	-0.07	0.14	0.24	0.62
	North America	-1.93	0.19	101.01	<.0001
Keyboard	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 × Region	Asia/PI	0.04	0.00	76.09	<.0001
	Central/South America	0.00	0.01	0.25	0.62
	Europe	-0.04	0.01	65.08	<.0001
	Middle East	-0.01	0.01	1.81	0.18
	North America	0.00	0.01	0.40	0.53

Note: β is the estimated parameter, and SE_β is the standard error of that estimate.

Table 2 Descriptive Statistics for Composition Media

Group	Handwriting	Word Processor	Overall
MC Total			
Mean	16.80	18.60	17.72
Gender			
Female	49%	51%	46%
Male	43%	57%	54%
Continent			
Africa	70%	30%	5%
Asia/PI	46%	54%	42%
Central/South America	30%	70%	14%
Europe	41%	59%	23%
Middle East	67%	33%	14%
North America	24%	77%	4%
Keyboard			
Roman/Cyrillic	38%	62%	50%
Other	53%	47%	50%
Medium			
Handwriting			46%
Word Processor			54%

Note: The handwriting and word processor columns show the conditional percent (or mean) of those categories for the group designated by the row. The overall column shows the marginal percent (or mean) of the total sample falling into the group designated by the row.

Figure 2 displays the *age-by-region* interaction – the only two-way interaction retained in the best fitting model. (See Appendix for information on other interactions.) This figure shows that there are differences between the *regions* with respect to the proportion of examinees choosing to write the essay in handwriting. Specifically, examinees from Africa and the Middle East are most likely to choose handwriting, while examinees from North America and Central/South America are least likely to choose handwriting. In addition, there are only small differences between *age* groups from these regions with respect to the proportion of examinees who chose handwriting, although the oldest (36+) and the youngest (under 21) examinees tend to be most likely to choose handwriting. The *age-by-region* interaction exists because the trends for the remaining two regions (Europe and Asia) diverged. Specifically, European examinees who are under 21 years of age were the most likely of examinees from that region to choose handwriting. On the other hand, Asian examinees who are under the age of 21 were least likely to choose handwriting as a composition medium. As shown in Figure 2, the predicted probability of choosing handwriting for Europeans decreased from a high of .50 to a

low of .31 across the age groups. Conversely, the predicted probability of choosing handwriting for Asians increased from a low of .44 to a high of .62 across the age groups. The empirical probability values were slightly less extreme – a decrease from a high of .53 to a low of .42 for Europeans and from a low of .43 to a high of .58 for Asians.

Figure 2

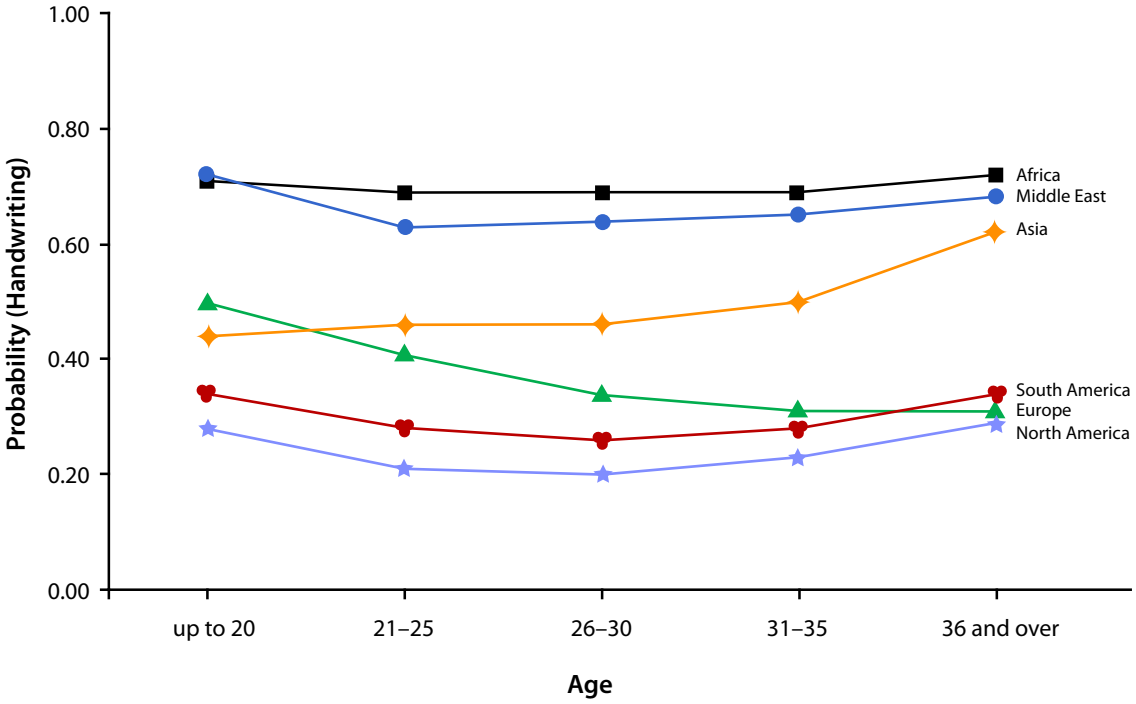


Figure 2. Age-by-Region Interaction.

Summary and Discussion

These results suggest the following.

- There are several main effects relating demographic characteristics to composition medium choice. Specifically, females are more likely than males to choose handwriting. Examinees who speak a language that is not based on a Roman/Cyrillic alphabet are more likely to choose handwriting than are examinees who speak a language that is based on a Roman or Cyrillic alphabet. Examinees with less proficient English language skills as measured by the multiple-choice sections of the TOEFL are more likely to choose handwriting than are examinees with more proficient English skills.
- An age-by-region interaction exists. There are only small differences between age groups in their tendencies to choose each composition medium for examinees from Africa, the Middle East, North America, and Central/South America. However, for Asian examinees, the probability of choosing handwriting as the composition medium increases with age. On the other hand, the probability of choosing handwriting decreases with age for examinees from Europe. If this interaction is ignored, the results indicate that there are large differences between regions with respect to composition medium choice, with examinees from Africa and the Middle East being most likely to choose handwriting and examinees from the Americas being the least likely to choose handwriting. In addition, the small main effect for age is slightly curvilinear with the youngest and the oldest examinees being most likely to choose handwriting.
- Several additional statistically significant two-way interactions exist (i.e., region-by-gender, region-by-keyboard, age-by-English, and region-by-English), but the small effect sizes make these interactions uninteresting from a substantive perspective.

Although there is no previous research concerning composition medium choice for second-language direct writing assessments, these 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 composition medium. Specifically, prior research suggests that females and individuals from regions where there are a greater number of developing countries would have fewer opportunities and higher levels of anxiety. However, the age-by-region interaction indicates that the anticipated influence of age on composition medium choices (i.e., that older examinees would be more likely to choose handwriting) may vary across geographic regions.

Discussion & Implications

Figure 1 posits a somewhat complex model that depicts the relationship between several characteristics of the examinee and performance on a direct writing assessment administered via computer (i.e., test anxiety, test preparation, achievement, computer skill, computer anxiety, and computer attitudes), and suggests that social standing may influence several of these variables. That model was posited based on conclusions drawn from previous research, and one goal of the research summarized in this report was to address questions that arose from consideration of that model. Specifically, this study sought to determine whether choice of composition medium is related to international examinees' memberships in "at risk" groups.

In general, that model was supported. 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 that use alphabets different than a Roman or Cyrillic alphabet, examinees from developing regions, and the oldest of the examinees) are all more likely to choose to compose essays using handwriting than using a word-processor. It is somewhat surprising that the results added younger examinees to this list, and one may speculate that this is 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 varies across geographic regions. Generally, the curvilinear trend observed in most regions (higher probabilities of choosing handwriting for the youngest and the oldest 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). A possible explanation for these trends is that young Asians were exposed to computers during their education because schools in Asia were quick to adopt technology education. Asians who had already completed their education may not have been exposed to computers for several years as computers began to appear in the work place. If true, it seems possible that Asian youth would be more comfortable choosing computers for composing their compositions than would older Asians. Explaining the European trend, on the other hand, is more difficult. Perhaps the opposite emphasis occurred in Europe with computers being introduced in the workplace before being introduced into the school systems.

It is reasonable to attribute the remaining trends 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 that 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 than other groups on that demographic variable. Unfortunately, this study does not directly address this relationship – there was no direct

measure of computer familiarity or experience. But, evidence from other studies supports that notion (Breland, Muraki, & Lee, 2001; Russell & Haney, 1997; Wolfe, Bolton, Feltovich, & Bangert, 1996; Wolfe & Manalo, in press). I suggest that examinees with lower levels of language proficiency – examinees who are also likely to have less experience and less comfort using computers – may anticipate encountering additional cognitive demands when responding to a writing prompt using a keyboard. And, it is reasonable to claim that those additional cognitive demands, should they exist, may constitute construct-irrelevant variance in some contexts, potentially 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.

Hence, it seems that those who develop direct writing assessments and those who make decisions about examinees based on scores from direct writing assessments should carefully examine the demographic characteristics of their examinee population to determine whether at risk groups exist in that population. It is likely that the potential threat due to inequities in examinee computer experiences will diminish over time, but these results (which are admittedly somewhat dated – 5 years old at the time of publication) suggest that the impact on examinee choice is profound. Test developers and decision-makers should also evaluate the degree to which computer use is a valid aspect of the construct in question. Specifically, they should consider whether using a word processor to write is an important skill relating to the decision to be made. In addition, studies should be conducted to determine the comparability of computer-based and paper-based scores for direct writing assessments, and careful consideration should be given to the comparability of at risk populations. It seems that further research is warranted concerning the basis of decisions when composition medium choice is provided and the accuracy of examinee decisions with regard to choosing the medium that will allow the examinee to provide his or her best sample of writing.

References

Agresti, A. (1996). *An introduction to categorical data analysis.* New York: Wiley.

Allison, P.D. (1999). *Logisitic 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 paper). Whittier, CA: Rio Hondo College.

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.

Collier, R., & Werier, C. (1995). When computer writers compose by hand. *Computers and Composition, 12*, 47–59.

Educational Testing Service. (1999). Description of the computer-based TOEFL test [Internet Web Page, available: http://www.teofl.org/descbcbt.html. Princeton, NJ: Author.

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*. (TOEFL Research Council). Princeton, NJ: Educational Testing Service.

Goldberg, A., Russell, M., & Cook, A. (2003). The effects of computers on student writing: A meta-analysis of studies from 1992 to 2002. *Journal of Technology, Learning, and Assessment, 2*(1). Retrieved April 20, 2003, from www.jtla.org

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.

Harrington, S., Shermis, M. D., & Rollins, A. L. (2000). The influence of word processing on English placement test results. *Computers and Composition, 17*, 197–210.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York: John Wiley & Sons.

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.

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.

Legg, S. M., & Buhr, D. C. (1992). Computerized adaptive testing with different groups. *Educational Measurement: Issues and Practice, 11*, 23–27.

MacCann, R., Eastment, B., & Pickering, S. (2002). Responding to free response examination questions: Computer versus pen and paper. *British Journal of Educational Technology, 33*, 173–188.

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.

Massoud, S. L. (1992). Computer attitudes and computer knowledge of adult students. *Journal of Educational Computing Research, 7*, 269–291.

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 chidren'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.

Powers, D. E. (1993). Coaching for the SAT: A summary of the summaries and an update. *Educational Measurement: Issues and Practice, 12*, 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., & 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.

Russell, M. (1999). Testing on computers: A follow-up study comparing performance on computer and on paper. *Educational Policy Analysis Archives, 7*(20). Retrieved November 29, 2003, 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. *Educational Policy Analysis Archives, 5*(1). Retrieved May 1, 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.

Stokes, M. E., Davis, C. S., & Koch, G. G. (2000). *Categorical data analysis using the SAS system* (2nd ed.). Cary, NC: SAS Institute.

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.

Turner, B. G. (1993). Test anxiety in African American school children. *School Psychology Quarterly, 8*, 140–152.

Wenglinsky, H. (1998). *Does it compute? The relationship between educational technology and student achievement in mathematics*. Princeton, NJ: Policy Information Center, Educational Testing Service.

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*, 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.

Wolfe, E. W., Bolton, S., Feltovich, B., & Welch, C. (1994, April). *A comparison of word processor and handwriting essays from a standardized writing assessment*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Wolfe, E. W., & Manalo, J. R. (in press). *An investigation of the impact of composition medium on the validity of scores from the TOEFL writing section*. Princeton, NJ: Educational Testing Service.

Appendix

A modified version of a common model selection strategy employed in logistic regression was utilized in this study (Hosmer & Lemeshow, 2000). First, univariate analyses were performed on each potential explanatory variable to determine whether each variable had sufficient predictive power to warrant inclusion in a multivariate main effects model using a large p-value for variable rejection ($p = .25$). Second, a preliminary multivariate main effects model containing all of the variables selected for inclusion during the first step of the procedure was fit to the data. All of the potential explanatory variables demonstrated reasonable predictive power during the first step of the procedure, so the preliminary multivariate main effects model contained all of the explanatory variables described previously. Third, the linear coding of quantitative variables (*age* and *English*) was evaluated. The plot of *English* with the empirical logits relating to composition medium choice indicated linearity, but that plot for *Age* was cuvilinear. Examinees between the ages of 21 and 35 had lower logit values (lower probabilities of choosing handwriting) than examinees under 21 years of age and examinees over 35 years of age. Hence, a quadratic term for *age* was added to the model – a term that proved to be statistically significant and also to improve the fit of the model.

Fourth, statistically significant two-way interaction terms were identified for inclusion in a "preliminary final model." Interactions were identified for inclusion in the expanded model based on the p-values of their Type III Wald statistics. Interactions were added only if the p-value of the Wald statistic was less than 0.0002, based on a recommendation by Raftery (1995), who suggested using more restrictive p-values for model selection purposes when sample size is large to ensure that only terms exhibiting reasonable levels of association with the dependent variable are included in the final model. The chosen p-value would only allow an interaction to enter the model if its predictive power is at least "moderate," based on Raftery's framework.

With the large sample size in this study, all progressively more complex models improved model fit to a statistically significant degree via the Likelihood Ratio chi-squared statistics, so the value of the *proportionality constant* (PC), which equals the deviance divided by its degrees of freedom (G^2 / df), was examined. 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. Because these 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 better accounts for the data than a model with a larger PC.

Table A1 shows the progression of models considered using this algorithm. The preliminary final model contained the following terms: Age (linear and quadratic), Region, English, Gender, Keyboard, Age × Region, Region × English, Age × English, Region × Keyboard, and Region × Gender. That is, the five two-way interactions suggest that composition medium choice: (a) varies across age groups

between the regions (AR), (b) varies across English proficiency groups between regions (RE), (c) varies across age groups between English proficiency groups (AE), (d) varies across language groups between regions (RK), and (e) varies across regions between gender groups (RG).

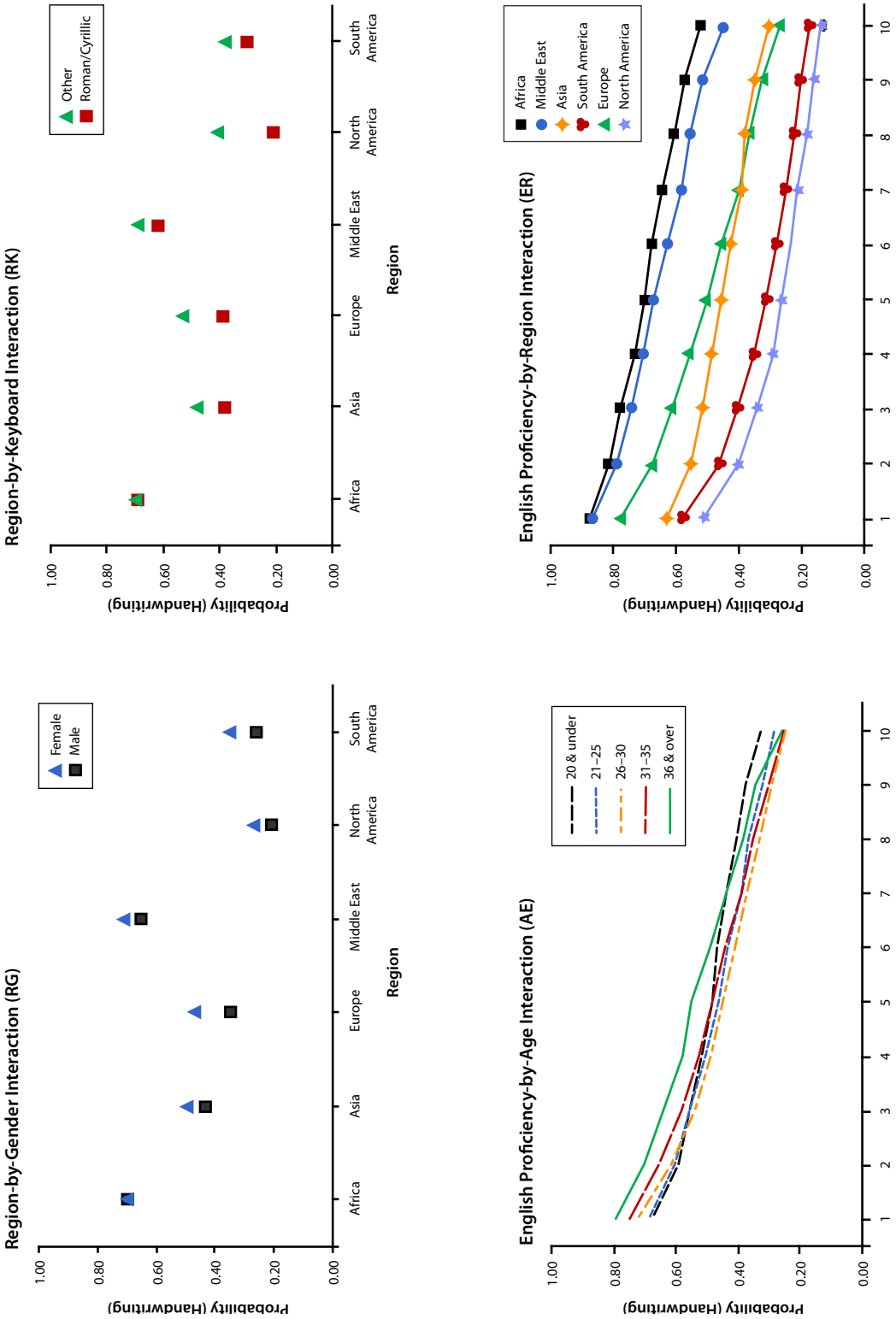
Table A1 Model Selection Summary

Model		Deviance		
Iteration	**Terms Included**	**G^2**	**df**	**PC**
0	A,R,E,G,K,A^2	27520.60	20475	1.34
1	AR,E,G,K,A^2	26505.81	20470	1.29
2	AR,RE,G,K,A^2	26225.30	20465	1.28
3	AR,RE,AE,G,K,A^2	26145.55	20464	1.28
4	AR,RE,AE,RK,G,A^2	26064.95	20459	1.27
5	AR,RE,AE,RK,RG,A^2	26005.34	20454	1.27
6	AR,RE,AE,RK,RG,EG,A^2	25993.06	20453	1.27

Note: A = Age, R = Region, E = English proficiency, G = Gender, and K = Keyboard.
The final model is shown as Iteration 1.

Consistent with the recommendations of Hosmer and Lemeshow (2002), the substantive contribution of these parameters was evaluated, and this led to the elimination of all but one of the two-way interactions in the final model. As shown in Figure A1, the RE, AE, RK, and RG interaction terms produced only marginally important effects. Specifically, the RG effect indicates that females are slightly more likely to choose handwriting than are males across regions (about 8% more likely) except in Africa, where the gender groups are equally likely to choose handwriting. The RK effect indicates that examinees who speak a native language that is based on a Roman or Cyrillic alphabet are about 10% more likely to choose word processor than are examinees who speak other native languages. This is true in all regions except Africa, where the language groups are equally likely to choose handwriting versus word-processing, and North America, where the difference is slightly larger (about 20% more likely). The AE interaction reveals that the rate of decrease in the probability of choosing handwriting as English proficiency increases is shallower for examinees who are under the age of 21 (about a 33% decrease) than it is for other examinee age groups (the bolded line – about a 46% decrease). Finally, the RE interaction indicates that the rate of decrease in the probability of choosing handwriting as English proficiency increases is shallower for examinees from Asia (about a 32% decrease) than it is for examinees from the remaining regions (about a 41% decrease). Note that the removal of these terms from the model resulted in only a 0.01 increase in the proportionality constant.

Figure A1 Interaction Terms Removed from the Preliminary Final Model



Overall, the predictive capacity of the AR, E, G, K, A^2 model is adequate. As shown in Table 1, the PC for that model equals 1.29 – a reduction of 0.05 from the main effects model, and only 0.02 points greater than the PC for a model that contains five additional two-way interactions. Because the standard errors of the parameter estimates may be slightly inflated due to the unexplained heterogeneity in the data – termed *overdispersion* (Allison, 1999) – the standard errors were corrected by multiplying them by the square root of the PC (Agresti, 1996). The proportion of concordant pairs ($P_{concordant}$), which indicates the proportion of pairs of observations with different outcomes (i.e., handwriting versus word processing) for whom the model-based expected value is consistent with the observed outcome (i.e., 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), equals .70 indicating that the model does a fairly good job of predicting group membership on the dependent variable. Similarly, the maximized R^2 index, which is analogous to the R^2 adjusted index generated in ordinary linear regression, equals .15, while the main effects model and the model with two two-way interactions produced maximized R^2 indices of .14 and .15, respectively. As is the case with the PC index, the maximized R^2 index can only be interpreted as a measure of the relative fit of two models, not as the proportion of variance explained. Hence, it seems that the chosen (more parsimonious) model performs comparably to the surrounding models investigated in the variable selection routine. Finally, the dissimilarity index (D; Agresti, 1996), 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, equals .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 seems to provide as good an explanation of the data as any of the more complex models and a better depiction of the data than the simpler model.

Author Note

Edward W. Wolfe
Measurement & Quantitative Methods
Michigan State University

This project greatly benefited from the input of my colleagues. Specifically, Claudia Gentile provided input into the design and data collection for this study. Pat Carey, Xiaoting Huang, Kevin Joldersma, Robbie Kantor, Yong-Won Lee, Philip Oltman, and Ken Sheppard each provided guidance in obtaining and interpreting the data. In addition, Paul Allison, Ken Frank, and Mike Patetta gave me advice during various stages of data analysis, and Jonathan Manalo assisted with the data analysis. Finally, the Dan Eignor and Paul Holland provided thoughtful suggestions for improving a previous version of the report from which this article was generated. I also thank the TOEFL program for providing me with the funds to carry out this work. Finally, I want to point out that this article is based on data that have been presented in three different papers at professional conferences and one research report. Specifically, see the following work (Breland et al., 2001; Manalo & Wolfe, 2000a, 2000b; Wolfe & Manalo, in press).

Correspondence concerning this article should be addressed to the author at 450 Erickson Hall, Michigan State University, East Lansing, MI 48824 or via electronic mail at wolfee@msu.edu.

J·T·L·A

The Journal of Technology, Learning, and Assessment

www.jtla.org

Technology and Assessment Study Collaborative
Caroline A. & Peter S. Lynch School of Education, Boston College