

## COMPOSITION MEDIUM COMPARABILITY IN A DIRECT WRITING ASSESSMENT OF NON-NATIVE ENGLISH SPEAKERS

Edward W. Wolfe and Jonathan R. Manalo

Michigan State University

### 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 performance on a direct writing assessment is comparable for examinees when given the choice to compose essays in handwriting versus word processing. We examined this relationship controlling for English language proficiency and several demographic characteristics of examinees using linear models. We found a weak two-way interaction between composition medium and English language proficiency with examinees with weaker English language scores performing better on handwritten essays while examinees with better English language scores performing comparably on the two testing media. We also observed predictable differences associated with geographic region, native language, gender, and age.

---

### INTRODUCTION

Increasingly, computers are being used to administer selection and certification tests. With the transition from a paper-based to a computer-based testing system comes a potential threat to the consequential basis of test use aspect of validity (Messick, 1989). That is, implementation of a computer-based testing program could result in unintended negative consequences for some examinees or for some societal components of the testing system. For example, differences in the performance of gender and ethnic groups exist on paper-based tests, and some fear that the shift toward a computer-based testing system may exacerbate existing social barriers to advancement opportunities for women, minorities, economically disadvantaged, and elderly individuals. Previous research comparing computer-based and paper-and-pencil tests has revealed only small differences between population means of multiple-choice tests administered in these two media (Mead & Drasgow, 1993). However, little is known about the influence of computerized testing on "at risk" groups of examinees or about the comparability of performance-based tests (e.g., direct writing assessments) administered in these two media, particularly for diverse populations of examinees. The purpose of this article is to compare computer-based and paper-based scores on the writing section of the Test of English as a Foreign Language (TOEFL) for a diverse population of international examinees.

### LITERATURE REVIEW

What evidence exists to support concerns about the potential negative impact of computer-based testing on some populations of examinees? First, it is clear that some groups of examinees are less likely to have access to, and hence experience and proficiency with, computers. In the US, minorities and women are less likely to have computers in their homes, and males are likely to dominate computer use at school -- the primary location within which some groups learn about and gain experience using computers (Campbell, 1989; Grignon, 1993; Keogh, Barnes, Joiner, & Littleton, 2000). Internationally, women, Africans, and Spanish speakers are less likely to have access to computers (Janssen Reinen & Plomp, 1993; Miller & Varman, 1994; Taylor, Kirsch, Eignor, & Jamieson, 1999). Similarly, one would expect

older individuals who learned how to use a computer later in life to have less experience using computers, although it is not clear whether these individuals would have restricted access.

Second, inequities in computer access and familiarity may lead to lower levels of confidence and higher levels of anxiety toward computer-based tasks. U.S. minorities and women (internationally) exhibit higher levels of computer anxiety and lower levels of confidence for performing computer-related tasks (Janssen Reinen & Plomp, 1993; Legg & Buhr, 1992; Loyd & Gressard, 1986; Massoud, 1992; Nolan, McKinnon, & Soler, 1992; Shashaani, 1997; Temple & Lips, 1989; Whitely, 1997). Interestingly, the magnitude of group differences in anxiety levels is greatly diminished when computer experience is held constant (Gressard & Loyd, 1987; Loyd & Gressard, 1986), indicating that, to some degree, non-cognitive influences on computer-based testing may lessen as computers become more commonplace in society.

Finally, scores from computer-based tests are already being used widely to make important decisions about individuals, and it is clear that affective responses, like computer anxiety, and proficiencies, like levels of computer experience, are correlated with computer-based test scores at non-trivial levels (Marcoulides, 1988). From previous research concerning computer-administered direct writing assessments with international populations, it is also clear that groups who have had fewer opportunities to use computers (e.g., females and individuals from developing countries) are less likely to choose a computer-based administration model when given the choice (Wolfe & Manalo, in press).

When scores on standardized multiple-choice computer-based and paper-based tests are compared, the differences in test performance at a population level tend to be small, but examinees perform slightly better on the paper-based versions of the tests (Mazzeo & Harvey, 1988; Mead & Drasgow, 1993). Obviously, population-level comparisons do not allow researchers to ascertain whether the influence of computer administration on test performance is stronger for small portions of the population (Wise & Plake, 1989). For example, analyses of several large-scale multiple-choice tests indicate that females may receive higher scores on paper-based tests, but that, contrary to what one might expect, African-Americans and Hispanics receive higher scores on computer-based tests (Gallagher, Bridgeman, & Cahalan, 2002).

Studies concerning the impact of computers on the comparability of direct writing assessments are less common. The few studies that exist suggest that raters may be influenced by the appearance of essays in handwritten versus typed text. Specifically, raters may have higher expectations for word-processed text (Arnold, Legas, Obler, Pacheco, Russell, & Umbdenstock, 1990; Gentile, Riazantseva, & Cline, 2001), but they may also produce more reliable scores for word-processed text because handwriting effects are eliminated (Bridgeman & Cooper, 1998; Wolfe & Manalo, in press). Fortunately, readers can be trained to partially compensate for differential expectations they may have concerning the quality of handwritten and word-processed text (Powers, Fowles, Farnum, & Ramsey, 1994).

Regardless, the use of word processors seems to influence the quality of the writing produced by examinees. For example, handwritten essays contain shorter sentences (Collier & Werier, 1995), are better organized (Russell & Haney, 1997), are freer of mechanical errors (Gentile et al., 2001), and are neater, more formal in tone, and exhibit weaker voice (Wolfe, Bolton, Feltovich, & Niday, 1996) than word-processed essays. More important, however, there may be an interaction between computer experience or proficiency and composition medium with respect to essay quality. In studies conducted on school-aged children, examinees responding to direct writing assessment or performance assessment prompts who had less computer experience received higher scores when tested in handwriting, and examinees with higher levels of computer experience received higher scores when tested using computers (Russell, 1999; Russell & Haney, 1997; Wolfe, Bolton, Feltovich, & Bangert, 1996; Wolfe, Bolton, Feltovich, & Niday, 1996). We hypothesize that this relationship exists because the imposition of keyboard composition requires examinees with less computer experience to perform the equivalent of a translation in order to produce their text. These examinees may formulate their writing cognitively, but

then they are required to translate those thoughts into keyboard strokes -- a task that is not part of their natural written communication process. As a result, the use of word-processors by 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 for these examinees. It is likely that such an effect would be more pronounced for examinees for whom English is a second language because these examinees would perform a double translation -- native language to English and then English to keyboard strokes.

This article summarizes a study of the influence of composition medium on scores assigned to essays written for the TOEFL writing section. The study aims to determine the extent to which examinees with comparable levels of English language proficiency receive comparable scores on word-processed and handwritten TOEFL essays. Specifically, we addressed the following questions. Are there differences in the magnitudes of the scores assigned to essays composed in each mode of composition? Are there differences in the magnitudes 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 are other groups of examinees?

## **METHOD**

In this study, general linear modeling was employed to determine whether a main effect exists for computer medium and demographic characteristics with respect to essay scores when controlling for English language proficiency and whether an interaction exists between computer medium and English language proficiency with respect to essay scores for a large sample of TOEFL examinees.

### **Participants**

Participants were 133,906 TOEFL examinees who participated in regular administrations of the computer-based TOEFL between January 24, 1998, and February 9, 1999 -- a small portion of the total number of examinees tested during this period. Only those examinees who provided complete demographic data, multiple-choice scores, and writing assessment scores were selected for this study. Participants were from 200 countries and represented 111 different languages. There were slightly more males than females (54% vs. 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 academic programs (38% and 46%, respectively). In fact, 82% of the examinees indicated that they planned to pursue an academic degree. Only 15% of the examinees indicated that they were taking the TOEFL for reasons other than to satisfy academic requirements.

### **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 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.

The first three tests are fixed-length (i.e., 30 listening questions, 20 structure questions, and 44 reading questions) with a variable number of pretest questions. The listening and structure sections are

administered as computer-adaptive tests, and the reading section is administered as a linear on-the-fly test. 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 (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 to create a variable measuring English language proficiency (English). We used the English variable as a covariate in the model described in the next section because examinees were allowed to choose composition medium (our dependent variable, medium). Our reasoning was that English proficiency can serve as a proxy for unmeasured variables, such as the educational opportunities available to the examinee, so that differences in the ability levels of examinees could be removed from the comparison of essay scores from each composition medium. Otherwise, the host of factors that influence reasons for choosing handwriting over computer as the composition medium could not be disentangled from the influence of composition medium on examinee performance. The relationship between English proficiency and composition medium choice is non-trivial in strength with higher ability examinees being more likely to choose word-processing ( $r = .25, p < .0001$ ). A similar procedure was employed by Taylor et al. (1999) in their analyses of multiple-choice data from the TOEFL.

We also created a composite writing score by averaging the independent scores (ranging from 0 to 6) that two raters assigned to the examinee's essay -- the dependent variable in this study (essay). The essay section of the TOEFL measures an examinee's ability to write in English. Because some examinees may not be accustomed to composing an essay on computer, they are given the choice of handwriting or word-processing the essay in the 30-minute time limit. Information concerning essay topics and scoring guidelines are made available to examinees through the [TOEFL Web site](#). Each examinee's essay is scored by trained essay readers who must meet several criteria, including passing a performance test in applying the scoring guidelines. Both typed essays and handwritten essay images are displayed to readers on a computer screen, and readers enter their ratings through this electronic interface. Each essay is rated independently by two readers. Neither reader knows the rating assigned by the other (TOEFL, 2003).

Examinees also provided self-report data about several demographic characteristics. From these data, we created four demographic covariate variables. Examinee age (recorded in years) was treated as a quantitative variable. Gender was treated as a dichotomous variable (0 = female, 1 = male). Countries were divided into the following regions, treated nominally, of course: North American, Africa, 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 language uses a keyboard containing an alphabet similar to the one used in English (e.g., Roman or Cyrillic, coded as 1) versus other systems (e.g., most Asian languages, labeled "other" and coded as 0).

### **Procedure**

Each examinee completed the examination in an operational administration of the TOEFL. This entailed completing the entire multiple-choice section of the examination in a computer-based testing environment. However, each examinee had the choice to respond to the single prompt for the direct writing assessment using a word processor (54%) or in handwriting (46%).

### **Analysis**

We utilized general linear modeling to address our research questions (Cohen, Cohen, West, & Aiken, 2003). That is, we evaluated the contribution of our independent variable (M = medium) and our covariate variables (E = English, G = gender, R = region, A = age, and K = keyboard) to the prediction of essay scores as a linear function, weighting the value of each independent variable by its parameter estimate (the predicted incremental increase in the value of essay scores for each one-point increase in the value of the independent variable).

We evaluated the assumptions required for this model. The dependent variable must be unbounded and continuous. Because there are 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 are required for each model we investigated: (a) normality of conditional distributions of essay scores for each level of 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 conditional distributions and variances indicated that the assumptions of normality and homogeneity of variances are satisfied. The linearity assumption was evaluated by computing conditional means for handwriting and word-processor composition media for each of 10 equal-interval bins of the 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 1, the relationship seems slightly non-linear, although not dramatically so.

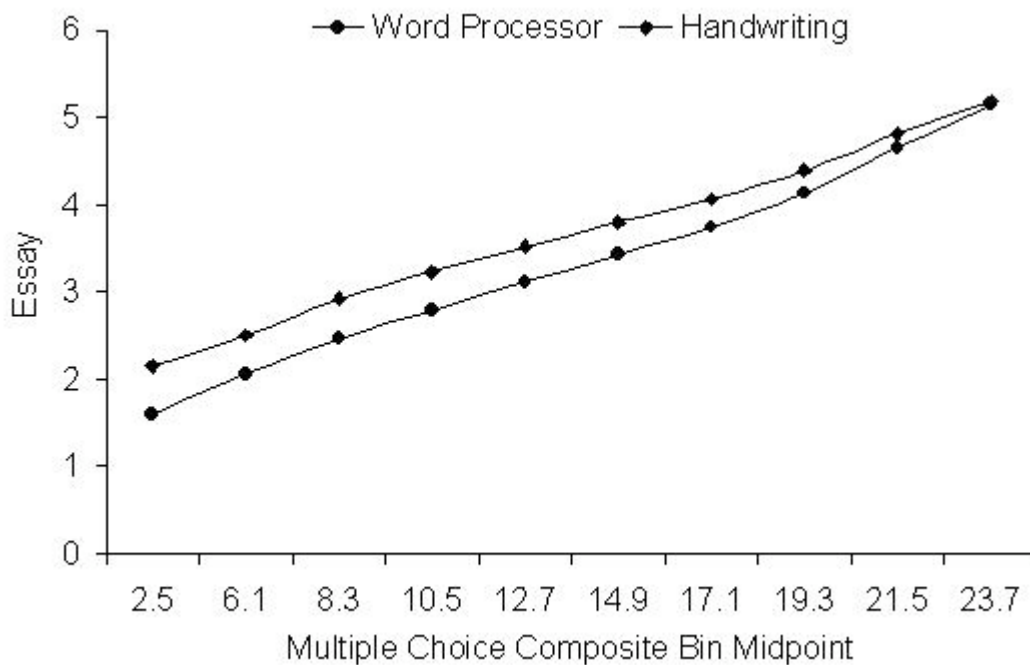


Figure 1. Linearity of essay scores across English proficiency levels

## RESULTS

Table 1 presents the essay and English variable descriptive statistics for each composition medium. From this table, it is clear that although there are clear differences in the multiple-choice scores (English) for examinees who chose different composition media, there are only small differences in the essay scores for those examinees. Both of these differences are statistically significant, but the effect size for the multiple-choice means is large while the effect size for the essay scores is small by the standards set forth by Cohen (1988):  $t_{\text{English}}(133,904) = 93.16, p < .0001, r^2 = .18$ ;  $t_{\text{essay}}(133,904) = 6.72, p < .0001, r^2 = .01$ . Also, recall that all examinees responded to the multiple-choice items using a computer, which, by the way, was found to be comparable to the paper-based version of the TOEFL (Taylor et al., 1999). Because examinees were allowed to choose composition medium and because composition medium choice and English language proficiency are correlated at a non-trivial level (recall that that correlation was  $r = .25$ ), it is clear that the composition medium groups are not comparable with respect to their English language abilities. Hence, we need to compare the scores obtained under each composition medium while

controlling for group differences in English language proficiency. That is, we need to include the English variable in our general linear model.

Table 1. Composition Medium Essay Scores

Variable	Handwriting	Word-Processor
Essay		
Mean	4.06	4.09
(SD)	(0.93)	(1.07)
English		
Mean	16.68	18.60
(SD)	(3.89)	(3.62)

Table 2 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. The  $r^2$  for this model indicates that 42% of the observed variance in essay scores is explained by the linear model containing these variables.<sup>1</sup> The values of the model's estimates (see the third column) indicate the expected incremental increase in essay scores, given a one-point increase in the continuous variables (e.g., English or age) or the existence of the characteristic in question for a qualitative variable (e.g., medium or gender) and assuming that all other variables are set to their zero values. For example, using a word-processor increases the expected essay score by 1.04 points, given all other variable values are set to zero. Being from South America, on the other hand, decreases the expected mean essay score by 0.35 points, given all other variable values are set to zero.

Table 2. Parameter Estimates for the General Linear Model

Parameter	Level	Estimate	SE	$\eta^2$ *
Intercept		1.12	0.02	
Medium				.01
	Handwriting	0.00		
	Word-Processor	1.04	0.02	
English		0.19	0.0008	.24
Age		-0.01	0.0003	.008
Gender				.004
	Female	0.00		
	Male	-0.12	0.004	
Region				.006
	Africa	0.00		
	Asia	-0.13	0.01	
	Europe	-0.27	0.01	
	Middle East	-0.23	0.01	
	North America	-0.28	0.01	
	South America	-0.35	0.01	
Keyboard				.006
	Other	0.00		
	Roman/Cyrillic	0.07	0.006	
English $\times$ Medium		-0.04	0.001	.007

\*NOTE: The  $\eta^2$  shown here is based on the Type III sum of squares. The model  $r^2 = .41$ . All effects were statistically significant at the .0001 level. The mean essay score is 4.08.

Given the large sample size, it is not surprising that the test comparing the observed estimate to a null value of zero is statistically significant for all variables. From the  $\eta^2$  effect size indices, it is clear that the two-way interaction between composition medium and English language proficiency is small, as are the

main effects for age, keyboard, region, and gender. Specifically, each of these variables accounts for less than 1% of the total variance in essay scores. However, there is a large effect for English proficiency (which accounts for 24% of the variance in essay scores), and there is a small effect for composition medium (which accounts for a little more than 1% of the variance in essay scores).

Generally, the model suggests that examinees who chose to compose essays in handwriting are predicted to receive higher scores than examinees who did not when controlling for differences due to demographic characteristics and English language proficiency. For example, the model predicts that an African female who speaks a language based on a Roman or Cyrillic alphabet, who has an average multiple-choice test score (17.72 points), and who produces an essay in handwriting will receive an essay score equal to 4.83 while her word-processing counterpart will receive an essay score equal to 4.56. However, the two-way interaction indicates that this difference is greater for examinees who receive low scores on the multiple-choice test than it is for examinees who receive higher scores on the multiple-choice test. For example, an examinee with the same characteristics described previously who receives a fairly low multiple-choice score (5.00 points) has predicted essay scores in handwriting and word-processing equal to 2.94 and 2.12, respectively, while her counterpart who receives a fairly high multiple-choice score (20.00 points) has predicted essay scores in handwriting and word-processing equal to 5.16 and 5.00, respectively.

Table 3 reports the largest observed essay score mean differences between various demographic groups. Here, we see that the examinee's geographic region and the alphabet used for the examinee's native language (keyboard) produce noticeable differences in essay scores (0.43 and 0.27 points on the six-point scale, respectively). On the other hand, age and gender produce only small differences in mean essay scores.

Table 3. Largest Essay Score Mean Differences for Each Demographic Group

Variable	Smallest Mean	Largest Mean	Difference
Age*	> 35 3.94	21 - 25 4.11	0.17
Region	Middle East 3.92	Europe 4.35	0.43
Gender	Male 4.02	Female 4.14	0.12
Keyboard	Other 3.94	Roman/Cyrillic 4.21	0.27

\*NOTE: For these computations, age was divided into five groups of five-year widths, beginning at age 16 and ending with those above age 35.

The more interesting question, however, is whether composition medium influences essay scores when controlling for differences in the demographic characteristics and English language proficiency of examinees who chose each composition medium. Figure 2 graphically depicts the predicted essay score for a typical examinee (i.e., an examinee who has the average value of all continuous and demographic variables in the model) with various English proficiency levels and who chooses either of the composition media. From this figure, it is clear that examinees who have low levels of English proficiency are expected to receive higher essay scores on handwritten essays while there is no difference between composition media for examinees with high levels of English proficiency.

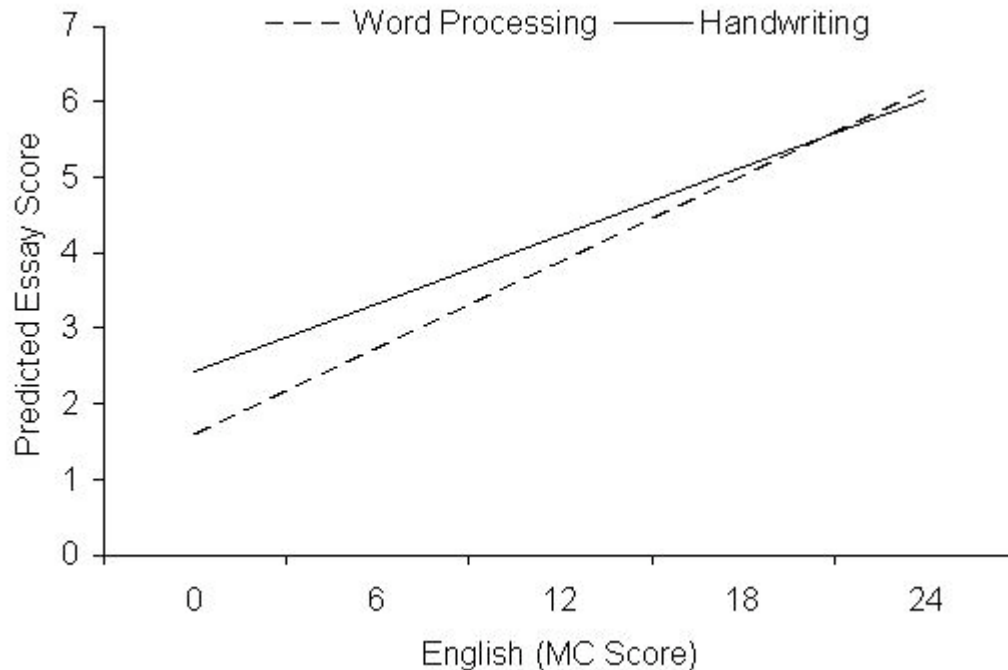


Figure 2. Predicted essay scores for a "typical" examinee

### SUMMARY AND DISCUSSION

In this section, we summarize the results of the study and discuss the theoretical and practical implications and the limitations of those results. We draw the following conclusions. Overall, there is only a small difference between essay scores of examinees who choose to compose their essays in handwriting versus word-processing, but when differences in overall English proficiency between composition medium groups are controlled, an interaction emerges. Specifically, examinees who have lower scores on the multiple-choice section of the TOEFL tend to have higher essay scores when essays are composed in handwriting, and examinees who have higher scores on the multiple-choice section of the TOEFL tend to have similar scores on essays composed in handwriting versus word-processing. For examinees who only answer a few multiple-choice questions correctly (those with very low English language proficiency), the predicted difference is about one essay score point (on a six-point scale). At the highest end of the multiple-choice scale, the scores are about equal. There are no substantively important medium-by-covariate interactions. 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 as evidenced by the very small effect sizes for these variables (e.g., the largest  $\eta^2$  equals .007). However, small main effects for covariates exist on essay scores, even when English language proficiency is taken into account. Specifically, region and native language mean differences tend to be moderately large while differences are small for age and gender.

Obviously, one problem with these results is the fact that examinees were given a choice of composition medium, and examinee characteristics associated with that choice are likely related to performance on the assessment. However, these results are consistent with previous research concerning test medium differences in direct writing assessment that did control for composition medium choice. Specifically, Wolfe, Bolton, Feltovich, and Niday (1996) found that secondary-level English-speaking junior-high students in the US who have considerable experience using computers and report 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. Similarly, Russell and Haney (1997) have 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.

Another limitation of this study is the fact that we did not directly control for the possibility of rater-by-medium interactions. That is, it is possible that differences in computer-based and paper-based essay scores may have been influenced by differences in the standards raters held for each composition medium. Unfortunately, it is nearly impossible to disentangle rater-by-medium and examinee-by-medium effects. One technique employed by researchers addressing this issue is to transcribe the essay from its original format (e.g., computer-based) to the alternate format (e.g., paper-based) preserving the original content to the degree possible. However, researchers have concluded that it may not be the case that observed differences are due to rater preferences for one medium over another. Rather, raters may show a preference for (i.e., assign higher scores to) the original over the transcribed version (MacCann, Eastment, & Pickering, 2002). Regardless, because prior research has shown that raters may be trained to partially compensate for medium preferences (Powers et al., 1994), and because TOEFL raters are highly trained and monitored, we believe that this threat to the internal validity of this study is minimal.

Hence, it seems that social conditions influence scores on computer-delivered direct writing assessments in ways that are predictable and to a degree that may warrant the attention of those developing or employing computer-based direct writing assessments. Previous research using these data has suggested that 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 when given a choice of composition medium (Wolfe & Manalo, in press). The research reported here indicates that not only is composition medium choice on direct writing assessments related to social conditions, but also to test performance.

In the introduction of this article, we speculated that this may be due to a "double translation" effect. Specifically, we suggested that examinees who have less experience and comfort with computers may encounter an additional cognitive demand when composing essays using a word-processor -- not only may they be performing a translation from their native language into English, but they may also be performing a translation from English language into keystrokes. The data reported here suggest that demographic characteristics that one would suspect to be related to computer exposure are indeed related to performance on a computer-based direct writing assessment, albeit only weakly. We believe that additional research is warranted to determine the cognitive mechanisms through which examinees compose essays during a direct writing assessment in each of these composition media.

In addition, our results present some interesting practical considerations for testing programs that utilize word-processors with direct writing assessments. Our interpretation of these results is that examinees with lower levels of language proficiency -- examinees who are also likely to have less experience and less comfort using computers -- may encounter additional cognitive demands when responding to a writing prompt using a keyboard. And, it is reasonable to claim 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. Hence, those involved in developing, administering, and reporting results of direct writing assessments designed for examinees with diverse language backgrounds should think seriously about providing examinees with a choice of composition medium (which is current practice with the TOEFL), particularly when high-stakes decisions will be made based upon the test results. We find it troubling that little is known 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 his results were based on U.S. populations, it is disconcerting that Russell (1999) found that examinees generally believe that they will receive higher scores on computer-based examinations. Hence, another important implication of these results is that further research should be performed to determine whether it is advisable to inform examinees of potential differences in performance on computer-based and pencil-and-paper writing assessments and the interaction between computer facility and test performance.

#### **NOTE**

1. We also examined hierarchical models containing all two-way and all three-way interactions and found that these models do little to improve the prediction of essay scores. Because of the large sample size, many of the possible two-way and three-way interaction parameter estimates were different from a null value of zero to a statistically significant degree. However, inclusion of these terms had only a very small impact on the values of  $r^2$  and the parameter estimates in comparison to the values in the model chosen for this study.

---

#### **ACKNOWLEDGMENTS**

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 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 papers at professional conferences. Specifically, see the following works: Breland, Muraki, & Lee, 2001; Manalo & Wolfe, 2000a, 2000b.

#### **ABOUT THE AUTHORS**

Edward W. Wolfe is an Assistant Professor of Measurement and Quantitative Methods at Michigan State University. His research focuses on validation of assessments that utilize innovative formats; such as performance assessments, portfolio assessment, and computer-based assessment; and applications of latent trait measurement models to those assessment formats.

E-mail: [wolfee@msu.edu](mailto:wolfee@msu.edu)

Jonathan R. Manalo is a doctoral candidate in Measurement and Quantitative Methods at Michigan State University. His research interests include the measurement of group differences using classical and latent trait based methodology, the detection of person and item misfit, and the selection of models using information-theoretic procedures.

E-mail: [manalojo@msu.edu](mailto:manalojo@msu.edu)

**REFERENCES**

- 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.
- 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.
- Campbell, N. J. (1989). Computer anxiety of rural middle and secondary school students. *Journal of Educational Computing Research*, 5(2), 213-220.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.
- Collier, R., & Werier, C. (1995). When computer writers compose by hand. *Computers and Composition*, 12, 47-59.
- ETS. (1999). *Description of the computer-based TOEFL test*. Retrieved October 22, 2003, from <http://www.toefl.org/educator/eddescbcbt.html>
- Gallagher, A., Bridgeman, B., & Cahalan, C. (2002). The effect of computer-based tests on racial/ethnic and gender groups. *Journal of Educational Measurement*, 39(2), 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: ETS.
- 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(3), 1-15.
- Janssen Reinen, I., & Plomp, T. (1993). Some gender issues in educational computer use: Results of an international comparative survey. *Computers in Education*, 20(4), 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(1), 33-43.
- Legg, S. M., & Buhr, D. C. (1992). Computerized adaptive testing with different groups. *Educational Measurement: Issues and Practice*, 11(2), 23-27.
- Loyd, 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. *Association for Educational Data Systems Journal*, 18(4), 302-311.
- MacCann, R., Eastment, B., & Pickering, S. (2002). Responding to free response examination questions: Computer versus pen and paper. *British Journal of Educational Technology*, 33(2), 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.
- Marcoulides, G. A. (1988). The relationship between computer anxiety and computer achievement. *Journal of Educational Computing Research*, 4(2), 151-158.
- Massoud, S. L. (1992). Computer attitudes and computer knowledge of adult students. *Journal of Educational Computing Research*, 7(3), 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 (CBR 87-8). Princeton, NJ: Educational Testing Service.
- Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin*, 114(3), 449-458.
- Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed.; pp. 13-103). New York: Macmillan Publishing.
- Miller, F., & Varman, N. (1994). The effects of psychosocial factors on Indian children's attitudes toward computers. *Journal of Educational Computing Research*, 10(3), 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(3), 299-314.
- 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.
- Russell, M. (1999). Testing on computers: A follow-up study comparing performance on computer and on paper. *Education Policy Analysis Archives*, 7(20). Retrieved August 15, 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* 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(1), 37-51.
- 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(2), 219-274.
- Temple, L., & Lips, H. M. (1989). Gender differences and similarities in attitudes toward computers. *Computers in Human Behavior*, 5, 215-226.
- TOEFL. (2003). *Description of the Computer-Based TOEFL Test*. Retrieved September 3, 2003, from <http://www.toefl.org/educator/eddescbcbt.html>
- Whitely, B. E. (1997). Gender differences in computer-related attitudes and behavior: A meta-analysis. *Computers in Human Behavior*, 13(1), 1-22.
- 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(3), 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(2), 123-147.

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: ETS.