DOCUMENT RESUME

ED 443 845 TM 031 490

AUTHOR Manalo, Jonathan R.; Wolfe, Edward W.
TITLE A Comparison of Word-Processed and Handwritten Essays Written for the Test of English as a Foreign Language.
PUB DATE 2000-04-00
NOTE 16p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 24-28, 2000). Funding provided by the TOEFL program.
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Adults; *Computer Assisted Testing; *Essay Tests; *Handwriting; Language Tests; Student Attitudes; *Test Format; *Validity; Word Processing
IDENTIFIERS Paper and Pencil Tests; *Test of English as a Foreign Language

ABSTRACT

Recently, the Test of English as a Foreign Language (TOEFL) changed by including a direct writing assessment where examinees choose between computer and handwritten composition formats. Unfortunately, examinees may have differential access to and comfort with computers; as a result, scores across these formats may not be comparable. Analysis of TOEFL results for 152,951 examinees reveals that when English language proficiency is controlled, handwriting composition scores are approximately one-third of a standard deviation higher than computer-based composition scores. It is suggested that this is a result of a double translation required to compose essays with word processors. (Contains 2 tables and 12 references.) (Author/SLD)

Running head: TOEFL ESSAY COMPARABILITY

A Comparison of Word-Processed and Handwritten Essays Written for the Test of English as a Foreign Language

Jonathan R. Manalo

Edward W. Wolfe

Michigan State University

Abstract

Recently, the Test of English as a Foreign Language changed by including a direct writing assessment where examinees choose between computer and handwritten composition formats. Unfortunately, examinees may have differential access and comfort to computers, and as a result, scores across mediums may be incomparable. Analysis reveals that when English language proficiency is controlled, handwriting composition scores are approximately one-third of a standard deviation higher than computer-based composition scores. We suggest that this is a result of a double translation required to compose essays with word-processors.

A Comparison of Word-Processed and Handwritten Essays Written for the Test of English as a Foreign Language

In 1998, the computer-based TOEFL examination began including a writing section (i.e., direct writing assessment). However, unlike the rest of the computer-based exam, the writing assessment is administered either in a computer-based or traditional paper and pencil-based medium. In addition, the examinee is given the option of choosing between these two formats. Allowing a choice between composition formats introduces potential sources of error that may degrade the quality of the scores and invalidate the inferences.

Two potential sources of this error are (a) differential rater perceptions and (b) differential examinee access and comfort. Differential rater perceptions may occur when different composition media influence a rater's judgement. That is, essays written by paper and pencil may be perceived to be inherently different from essays written by word-processors by a rater. For example, prior research has shown raters have higher expectations for computer-based essays than for handwritten essays (Arnold, Legas, Obler, Pacheco, Russell, & Umbdenstock, 1990). Differential examinee access and comfort may occur because subgroups of examinees have different levels of access to computer use and comfort using computers. That is, some groups of persons do not have the same opportunities to access computers, as do others. And, some persons (possibly resulting from the lack of access) do not have equal levels of comfort using computers, as do others. Specifically, prior research has suggested, at least in the United States, that females and lower socioeconomic status groups are less likely to have access to computers. In fact, frequently the most common place these groups use computers is at school (Campbell, 1989; Grignon, 1993). In addition, female and minority comfort levels (e.g.,

experience, interest, and confidence) with computers are lower than those of white males (Shashaani, 1997; Whitley, 1997).

Unfortunately lack of experience and reduced comfort with computers manifests itself in lower quality of essays written in that medium. Wolfe, Bolton, Feltovich, and Niday (1996) have shown that computer-based essays written by U.S. secondary students contain fewer words, substantively different content, and are generally judged to be lower quality than handwritten essays written by these same students. On the other hand, students who report having moderate or high levels of experience and comfort with computers compose essays on computer that are judged to be slightly better than essays composed in handwriting by the same student.

Research Questions

To summarize, large-scale examinations like the computer-based TOEFL that utilize both paper and pencil and computer media for direct writing assessments may be introducing sources of measurement error that hinder the consistency of the scores and the accuracy of the inferences. Different subgroups of examinees still do not have equal access to and do not feel comfortable with composing essays at a keyboard. Hence, differences in scores across writing assessment media are manifested. Unfortunately, all prior research has tended to focus on U.S. populations. We do not know whether and how these effects manifest themselves in for foreign language examinees. Nor does prior research have sufficiently large sample sizes to consider the findings of these studies to be robust. Therefore, the purpose of this paper is to provide evidence that computer and handwritten essays are not comparable using a large, foreign language examinee sample. Hence, our research question is: "Is performance on the TOEFL direct writing assessment equivalent for examinees who choose to compose their essays via handwriting and those who choose to compose their essays via word-processors?"

Method

Examinees

Participants in our study were 152,951 TOEFL examinees (males = 53.5%, females = 46.5%) from 223 countries that participated in regular TOEFL administrations of the computer-based and handwritten direct writing assessment between 1/24/98 and 2/9/99. Scores were available for the computer-based multiple-choice section of the TOEFL for each examinee. Most of the examinees indicated taking the test for admittance into undergraduate (39%) or graduate studies (45%). The others indicated non-academic reasons (16%). Each examinee was administered a computer-based multiple-choice section, while being offered the choice of word processor (51.5%) or handwriting (48.5%) for the single prompt of the writing section.

TOEFL

The computer-based TOEFL consists of four sections: (1) listening, (2) structure, (3) reading, and (4) writing. The listening section measures an examinee's proficiency in understanding North American-spoken English. Questions are administered after an examinee listens to recorded stimuli, and they require the examinee to do the following: (a) comprehend main and supporting ideas, (b) draw inferences, and (c) categorize topics or objects. Typically, examinees take 40-60 minutes to answer 30-50 multiple-choice items that are administered using a computer-adaptive design (i.e., items are selected to match the examinee's proficiency).

The structure section measures the examinee's proficiency in recognizing proper language in standard written English. Questions require the following from the examinees: (a) complete incomplete sentences; and (b) identify improper words or phrases. Typically, examinees take 15-20 minutes to answer 20-25 multiple-choice items that are administered using

a computer-adaptive design. The reading section measures the examinee's proficiency in reading and understanding short passages that are similar to academic texts used in North American colleges and universities. Questions require the following: (a) comprehend main ideas, facts, pronoun referents, and vocabulary, and (b) reason using inferences. Typically, examinees take 70-90 minutes to read 4 or 5 passages consisting of 250-350 words each and answer 10 or 14 multiple-choice items per passage that are administered using a linear-computer design (i.e., the paper and pencil multiple-choice version is administered via computer) (ETS, 1999).

The writing section measures the examinee's proficiency in writing English. A single prompt requires examinees to do the following: (a) compose a response that generates, organizes, and develops ideas; and (b) compose a response that supports those ideas using examples or evidence. Examinees are given 30 minutes to respond to the prompt that, as mentioned earlier, is administered via paper and pencil or computer depending on the examinee's choice.

Scoring

Operationally, scores for each the listening and reading sections are scaled to range from 0 to 30, while scores for the writing and structure sections are combined (and weighted equally) and are scaled to range from 0 to 30 (ETS, 1999). Here, scores for the structure section were scaled to range from 0 to 13, and were combined with the scores (scaled according to TOEFL's purposes) for the listening and reading section, where each section was weighted equally. The average of these three scores is the multiple-choice composite score. Operationally, the writing section is scored on a scale ranging from 1 to 6 by two trained and independent readers (i.e., trained in interpreting the response to TOEFL standards, to score across multiple topics, and to use the software which essays are distributed to the readers and scores are recorded; and independent from knowing the score assigned by the other reader). The average of the two

scores, i.e., the writing composite score, is assigned to the essay unless there is a discrepancy between the two readers. In that case, a third independent reader rates the essay, and the score is resolved (ETS, 1999). Here, the writing composite score is used.

Analysis

Because examinees chose the composition medium, it is not reasonable to assume that either treatment group is homogenous. For example, examinees who chose computers may have higher English language proficiency and comfort levels with computers than examinees who chose paper and pencil. This may potentially result in higher scores for persons with higher English language proficiency and comfort levels with computers. Hence, the multiple-choice composite score—a measure of English language proficiency—was used to control such effects. According to Taylor, Jamieson, Eignor, and Kirsch (1998), multiple-choice scores as the covariate seems appropriate. In addition, it seems appropriate to assume that computer experience has minimal influence on the computer-based multiple-choice scores.

To clarify and summarize, an ANCOVA design was used, where the covariate was multiple-choice composite score, the independent variable was composition medium, and the dependent variable was the writing composite score. To examine the difference between the scores of the composition medium, we first tested the assumption of equal slopes between the covariate and the independent variable. If the slopes were equal, the main effect between the treatment groups was examined.

Results

Table 1 presents the ANCOVA results. The test for equal slopes between composition medium and multiple-choice composite score was statistically significant. That is, the slope of the computer-based scores and the slope of the handwriting scores are not equal across levels of

multiple-choice score. More specifically, there is a larger difference for examinees with low levels of English language proficiency and a smaller difference for examinees with higher levels of English language proficiency between computer-based and handwriting scores. Although the interaction effect is statistically significant, it is well known that statistical significance may be a statistical artifact of large sample sizes. Here, we had a considerably large sample size—153,531 participants. It is also well known that the scores at the tails of the distributions have lower reliability; as a result, it is possible the increased error distorts the actual slopes. In addition, the interaction effect is not necessarily meaningful. That is, the proportion of variance explained by this interaction is nearly zero, $\hat{\eta}^2 = .01$. Cohen (1988) refers to effect sizes of this magnitude as small. Taking these into consideration, we decided to continue the analyses and interpret a model with main effects only. The main effect of computer medium revealed a statistically significant difference. That is, when English language proficiency is controlled, there is a statistically significant difference between computer-based and handwriting scores. This difference favors handwriting and has a somewhat moderate effect size, $\hat{d} = .30$ (Cohen, 1988). Table 2 presents the unadjusted (before controlling for English language proficiency) and adjusted means (controlling for English language proficiency) for each composition media. In terms of the TOEFL scoring scale, this difference favors handwritten essays by approximately 1/3 standard deviation or approximately 0.30 points on the 6-point TOEFL essay rating scale.

Discussion

By examining computer-based and handwritten TOEFL compositions while controlling for English language proficiency, we found that scores assigned across the two media are inequivalent. This is consistent to Arnold et al. (1990) and Powers, Fowles, Farnum, & Ramsey (1994) who use U.S. samples. That is, foreign language groups score unequally across

handwritten and computer-based compositions. Specifically, results suggested that average scores for handwritten compositions were 1/3 standard deviation higher than scores for computer-based compositions. For the TOEFL, which uses a 6-point scale, foreign language examinees would score a third of a point higher if they were to use paper and pencil rather than a computer to compose an essay.

Unlike Arnold et al. (1990) and Powers et al. (1994) who suggest that the difference between the scores is the result of increased rater expectations for the computer-based compositions, we suggest [as do Dalton and Hannafin (1987) and Gentile (1999)] that the difference is the result of a double translation required to work with computers. For example, an examinee must first translate the composition from the native language to the English language. Next, the examinee must translate the composition from the English language to the computer language. Hence, the examinee is required to make two cognitive translations—one between the native language and the English language and another between the English language and the computer language. It is apparent that examinees with lower levels of English language proficiency have increased difficulty when composing an essay with computers. Although, we chose not to interpret the statistically significant interaction effect because of possible statistical artifacts, the interaction effect does suggest that this is the case. That is, the difference between computer-based and handwritten compositions scores increases as examinee English language proficiency decreases. Others, like Wolfe et al. (1996), demonstrate this effect using U.S. samples. Specifically, they demonstrate that examinees who are competent with computers actually produce computer-based compositions that are better than their handwritten compositions.

One limitation of the current study is that we found a significant interaction effect between the independent variable, composition medium, and the covariate, English language proficiency. Although we chose to ignore this because of the large sample size, unreliability of the scores at the tails of the distribution, and the small amount of variance that was explained by the effect, it is unclear exactly how meaningful this interaction truly is. Future research is needed to determine if the interaction effect can be ignored or must be interpreted. Another limitation is that we only investigated TOEFL examinees. Although this was part of our purpose, we acknowledge that this limits the scope of generalization for our results. Future research is needed to examine how robust these findings are with examinees of other foreign language examinations.

Future research could also focus on the qualitative differences between essays composed by paper and pencil and computer as well as the cognitive processes that are unique in using each composition medium. In other words, research can examine the textual components of the essays and the examinee's mental processes that are used when composing with paper and pencil and computer-based mediums. Gentile (1999) has completed a pilot for such a study. In addition, future research can examine the differences between examinee characteristics for those that compose essays in handwriting and those that compose essays with computers. Although our descriptive statistics show approximately half of the examinees chose each medium; we also show that the handwritten medium has slightly higher scores. Hence, if we can identify the examinee characteristics that are different across mediums, then we can begin to determine the implications of the characteristic differences and the fairness of the decisions concerning those examinees.

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. word-processed papers (unpublished paper). Whittier, CA: Rio Hondo College.

Campbell, N.J. (1989). Computer anxiety of rural middle and secondary school students. Journal of Educational Computing Research, 5, 213-220.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. Hillsdale, NJ: Lawrence Erlbaum.

Dalton, D., & Hannafin, M. (1987). The effects of word processing on written composition. Journal of Educational Research, 50, 223-228.

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

Gentile, C. (1999). An investigation of the impact of composition medium on the quality of scores from the TOEFL writing section: A report from the pilot focused study. Research report submitted to the TOEFL research council. ETS: Princeton, NJ.

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.

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.

Shashaani, L. (1997). Gender differences in computer attitudes and use among college students. Journal of Educational Computing Research, 16, 37-51.

Taylor, C., Jamieson, J., Eignor, D., & Kirsch, I. (1998). The relationship between computer familiarity and performance on computer-based TOEFL test tasks (ETS Research Report RR 98-8). Princeton, NJ: ETS.

Whitely, B.E. Jr. (1997). Gender differences in computer-related attitudes and behavior: A meta-analysis. Computers in Human Behavior, 13, 1-22.

Wolfe, E.W., Bolton, S., Feltovich, B., & Niday, A.W. (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.

Table 1

ANCOVA Comparisons for Composition Medium for Examinees of TOEFL

Purpose	Source	SS (df)	F (p)	d (η^2)
Test for Common Slopes	Mode × MC	919.90 (1)	1486.61 (< .0001)	-- (.01)
	Error	94639.47 (152942)		
Test for Intercept of Covariate	MC	61469.05 (1)	98381.36 (< .0001)	-- (.64)
	Error	95559.37 (152943)		
Test for Mode Main Effect	MC	61469.05 (1)	98381.36 (< .0001)	.30 (.03)
	Error	95559.37 (152943)		

Note. N = 152,951. MC = multiple-choice composite score.

Table 2

Unadjusted and Adjusted Means for Each Composition Media

	Unadjusted Mean	Adjusted Mean
Handwriting	4.0	4.18
Word Processor	4.1	3.89

Note. N = 74,630 for Handwriting. N = 78,961 for Word Processor. Scores range from 1 – 6.

Acknowledgements

This project greatly benefited from the input of our colleagues. Specifically, Claudia Gentile provided input into the design and data collection for this study. Pat Carey, Robbie Kantor, Yong-Won Lee, Philip Oltman, and Ken Sheppard each provided guidance in obtaining and interpreting the data. In addition, James Algina, Robert Brennan, Chris Chiu, and David Miller gave us advice during various stages of data analysis. We also thank the TOEFL program for providing us with the funds to carry out this work.

AERA

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

TM031490

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title:

A COMPARISON OF WORD PROCESSED AND HANDWRITTEN ESSAYS WRITTEN FOR THE TEST OF ENGLISH AS A FOREIGN LANGUAGE

Author(s): JONATHAN R. MANALO & EDWARD W. WOLFE

Corporate Source:

AERA 2000 IN NEW ORLEANS

Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☑

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) and paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

☐

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

☐

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature:	Printed Name/Position/Title: JONATHAN R. MANALO / RESEARCH ASSISTANT
Organization/Address: 401 ERICKSON HALL, MICHIGAN STATE UNIVERSITY, EAST LANSING, MI 48824	Telephone: 517-351-8459 / FAX: —
	E-Mail Address: MANALOJ@MSU.EDU / Date: 5/13/00

(over)