THE EFFECT OF COMPUTERS ON THE TEST AND INTER-RATER RELIABILITY OF WRITING TESTS OF ESL LEARNERS

Dr. Selami AYDIN
Atatürk Üniversitesi
Dil Eğitimi-Öğretimi,
Uygulama ve Araştırma Merkezi
25240 Erzurum
saydin@atauni.edu.tr

ABSTRACT
This research aimed to investigate the effect of computers on the test and inter-rater reliability of writing test scores of ESL learners. Writing samples of 20 pen-paper and 20 computer group students were scored in analytic scoring method by two scorers, and then the scores were analyzed in Alpha (Cronbach) model. The results showed that the test and inter-rater reliability of the writing samples of the computer group students were significantly higher than the ones of the pen-paper group participants.

Key Words: English as a second language, computers, writing test, test reliability, inter-rater reliability

INTRODUCTION

Since the 1970s, computers have been in schools, in homes, and computer use has a considerable influence on education (Zandvliet and Farragher, 1997). Thus, for three decades, educational theorists and researchers have proposed many ways in which computers influence education. As a result of this influence, in recent years, there has been an explosion of interest in using computers in language teaching, learning and testing. Today, the role of computers in language instruction is a significant issue confronting large numbers of language teachers throughout the world (Warschauer and Healey, 1998).

The turning point on computer use in language testing is item response theory that has made individual test taking possible. The advances in item response theory and computer technology played a greater role in the development of language testing in 1990s, and extensive literature has been developed to examine the effectiveness of CALL (Brown, 1997). The literature on computers and language testing focused on four issues: item banking, computer-assisted language testing, computer-adaptive language testing, and the effectiveness of computers in language testing. However, computer use in language testing is still a specific area (Brown, 1997).

Computers have also become an accepted tool in writing classes, and research on various aspects of the writing process on computer has mushroomed in the last decade (Phinney, 1991). Researchers have argued that computer use helps students to prevent anxiety about writing and premature editing, to change revision strategies (Daiute, 1985), and improves attitudes towards writing (Dalton and Hannafin, 1987; Hawisher, 1987). However, little research has appeared on computer use with second language writers, although many studies on writing have been conducted for native speakers. Few studies on second language writing showed that second language writers are often assumed to have more apprehension than native language writers, to monitor their output (Krashen, 1982), to be more likely to edit prematurely, and to have more negative attitudes toward writing in their second language than first language writers. On the other hand, according to some studies, computer use seems to have positive effects on second language writers (Phinney, 1991), although research level in second language writing and computers does not come near the activity in first language writing. For instance, According to Phinney and Mathis (1991), the second language learners felt that the computer improved their attitudes toward writing in English. The learners also seemed to spend more time writing than the students who did not use a computer and produced longer papers (Phinney, 1988). Neu and Scarcella (1991) also noted similar results in their study. In sum, when these conflicting results are considered, it can be said that few researches on second language writing have not given an idea on composing on computer for second language learners, and there has not been a consensus on computer effects on writing test scores.

The research on the test and the inter-rater reliability of writing tests of ESL students shows that the results are also conflicting and not conclusive (McNamara, 1996). Some studies showed that scorers assigned lower scores to computer versions of the tests than the pen-paper ones (Bridgeman and Cooper, 1988; Sweedler-Brown, 1991). In another study, there was no difference between the typed and handwritten versions of the paper in the process of grading (Powers, Fowles, Farnum, and Ramsey, 1994).

Finally, this study was guided by the following reasons:

1. Although many studies have been conducted on computer use in native language writing, little research has appeared on second language writing.
2. The studies have not established a consensus on the computer effects in the testing of writing skills of ESL writers.
3. There is not certain empirical data on the effect of computers on the test and inter-rater reliability of writing tests of ESL learners.

In sum, these concerns show that it is a necessity to study the effects of computers on the test and inter-rater reliability of writing tests of ESL learners. In other words, the study has one research question: What is the effect of the computer on the test and inter-rater reliability of writing tests of ESL learners in analytic scoring?

METHOD

The sample groups consisted of 40 second-year students in the English Language Teaching Department at the Faculty of Education at Atatürk University in Erzurum, Turkey. The reason why second year students were chosen was that they had writing and computer classes in the same term, spring 2002 – 2003. 20 students participated in the pen-paper tests in the classroom environment, and 20 wrote electronically their compositions in the computer lab. Two limitations, number of participants and the gender distribution (28 females 12 males) were closely related to the computer laboratory capacity at the faculty and the gender distribution of the student population at the ELT department.

Since writing ability between the participants in the pen-paper in computer groups seemed a significant variable that affects the reliability, the students were assigned according to their equal writing abilities. Thus, the final exam scores of the writing and computer classes of the previous term were used as criteria. Then, computer versions of the pre- and posttests were administered to the participants in the computer group. Similarly, pen-papers versions of the pre- and posttests were administered to the students in the pen-paper group.

All the participants were Turkish students who were ESL learners at upper-intermediate level. The three topics, chosen from the TOEFL practice tests (See Appendix 1), for the pretest and three for the posttest were given to the participants in the pen-paper and computer groups. The participants were asked to respond only one topic and to write in free writing style.

The computer lab consisted of 20 computers with the Windows operating system. The participants in the pen-paper group wrote their compositions in classroom environment, and the students in computer group produced texts on computers in computer laboratory. The participants in the computer group used Word 2002 to write their compositions.

Since the study focused on the test and inter-rater reliability of writing samples, the duration between the administration of the pre- and posttests was one week and the participants did not receive writing instruction during this time. In other words, students' progress was not a variable in the research. Then, pen-paper and computer versions of the tests after printing were delivered to the scorers.

The two scorers were teachers in the ELT department with PhD degrees in English language teaching. They have taught writing individually, administered and scored writing tests at ELT department for at least fifteen years. They scored the tests without seeing the ones given by the other. A scoring rubric for writing proficiency in a range of 0 – 100 points was developed (See Appendix 2). Analytical scoring procedure was applied by the scorers according to the writing proficiency grading table. Finally, after scoring, the raw scores were analyzed to find the test and inter-rater reliability coefficients in Alpha (Cronbach) model, a reliability analysis that allows to find the properties of measurement scales and that is used as a model of consistency. The Alpha (Cronbach) was computed to see the consistency between the scores of pre- and posttests and the reliability between the scores assigned by two scorers. The mean and standard deviations of the tests were also computed in order to see the consistency of the scores.

DATA ANALYSIS

Since the writing ability and computer familiarity of the participants could affect the reliability of writing tests administered in the study, the mean and standard deviations of the final examination scores of writing and computer classes in the previous instruction semester were analyzed and presented in Table 1. The mean differences between the previous semester scores of the participants were 1.3 in writing and 0.4 in computer

class in the scale of 0–100. The data showed that there were no significant mean differences between the groups on both writing ability and computer familiarity.

Table 1. The Mean and Standard Deviations of the Previous Writing and Computer Exam Scores

	Groups								
	Pen-paper			Computer			Total		
	N	Mean	Std. Deviation	N	Mean	Std. Deviation	N	Mean	Std. Deviation
Writing Test	20	69.3	4.01	20	70.6	5.59	40	69.95	4.85
Computer Test	20	65.95	9.19	20	66.35	9.22	40	66.15	9.09

The means of the pre- and posttest scores given by two scores were presented in Table 2. When the values in Table 1 were compared to the ones in Table 2, it was seen that the participants had lower scores in the pre- and posttests. As Phinney (1991) noted that computer use seemed to have positive effects on second language writers, the computer group participants had higher scores of which the mean differences between the groups, 0.53 for the pre- and 3.57 for the posttest.

Table 2. The Mean of the Pre- and Posttests

Groups		Pretest[a]	Pretest[b]	Pretest[c]	Posttest[d]	Posttest[e]	Posttest[f]
Pen-paper (N=20)	Mean	57.95	56.05	57.0	53.25	52.7	52.98
Computer (N=20)	Mean	58.3	56.75	57.53	56.8	56.3	56.55

a. First Scorer

b. Second Scorer

c. The average of the scores assigned by the first and second scorer

d. First Scorer

e. Second Scorer

f. The average of the scores assigned by the first and second scorer

The means of the text length were 226.5 for the pen-paper and 281.2 words for the computer group participants. Although the text lengths are related to the writing quality rather than the reliability of the tests, the significant point was that the computer group students produced longer texts, as Phinney (1988) noted.

Table 3. The Word Length of the Texts

	Pen-Paper	Computer
N	20	20
Minimum	195.0	240.0
Maximum	279.0	341.0
Mean	226.5	281.2
Std. Deviation	23.2	30.1

The average of pre- and posttest scores given by two scorers for each paper were computed to find the test reliability in Alpha (Cronbach), a model of internal consistency, based on the average inter-item correlation. Depending on the means, standard deviations and pre- and posttest Alpha (Cronbach) values presented in Table 4, three results can be discussed: First, for both groups, the posttest means were lower than pretest means. However, since analytic scoring procedure was applied for both groups, scoring method was not the factor that affects the results. The different topics given for the pre- and posttests, writing medium and the scorers' experience on the scoring table could have been an influence on the scores. However, since the issue in the

research focused on the test reliability, the mean differences were significant to see the consistency between the tests. Second, the mean difference between pre- and posttest in the pen-paper group was higher than the one in the computer group. When the data in Table 1 and 4 was considered, it would be seen that the computer group participants had higher scores. Third, the reliability analysis showed that the computer group scores were more consistent when the Alpha (Cronbach) value and standard deviations were considered, and that the reliability coefficient of the computerized papers was significantly higher than the one of the hand-written ones. In sum, it seemed that the computer has a considerable effect on the test reliability in analytic scoring.

Table 4. Test Reliability Coefficients

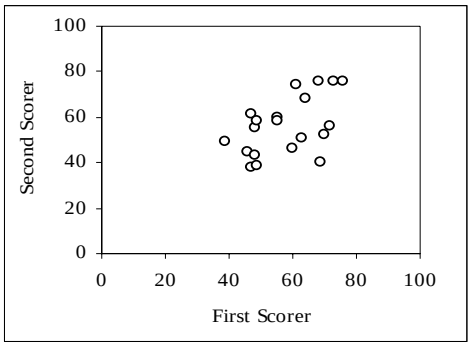
Groups	Tests	Mean	Standard Deviation	Alpha (Cronbach)
Pen-paper	Pretest	57.00	10.35	0.6111
	Posttest	52.98	12.20	
Computer	Pretest	57.53	17.63	0.9857
	Posttest	56.55	16.93	

The inter-rater reliability coefficients of the scores were computed between the scores given by the two scorers in analytic scoring. In Table 5 and Figure 1, the means, standard deviations and inter-rater reliability coefficients in Alpha model were compared among the pre- and posttests scores of the pen-paper and computer group participants. The scores given by the first and second scorers for each paper were used to compute the Alpha value. The findings presented in Table 5 and Figure 1 suggested that the inter-rater reliability coefficients of the computerized versions of the papers were considerably higher than the ones of hand-written papers. In sum, it seemed that the computer had a significant effect on the inter-rater reliability of the writing tests of ESL learners in analytic scoring, on the contrary of the studies that showed scorers assigned lower scores to computer versions of the tests than the pen-paper ones (Bridgeman and Cooper, 1988; Sweedler-Brown, 1991) and that found there was no difference between the typed and handwritten versions of the paper in the process of grading (Powers, Fowles, Farnum, and Ramsey, 1994).

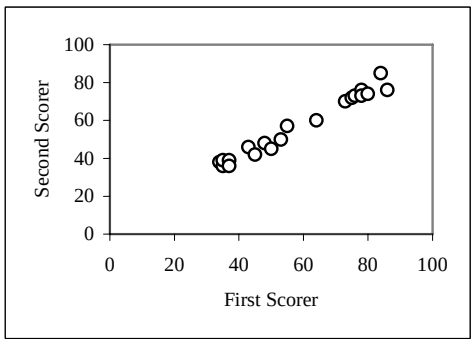
Table 5. Inter-rater Reliability Coefficients of the Tests

Groups	Tests	Scoring	Mean	Standard Deviation	Alpha (Cronbach)
Pen-paper	Pretest	First	57.95	11.03	0.6790
		Second	56.05	12.71	
	Posttest	First	53.25	13.15	0.8752
		Second	52.70	12.72	
Computer	Pretest	First	58.30	18.81	0.9892
		Second	56.75	16.57	
	Posttest	First	56.80	17.37	0.9929
		Second	56.30	16.52	

Pen-paper Group
Pretest (Alpha=0,6790)



Computer Group
The Pretest (Alpha=0,9892)



Posttest (Alpha=0,8752)

Posttest (Alpha=0,9929)

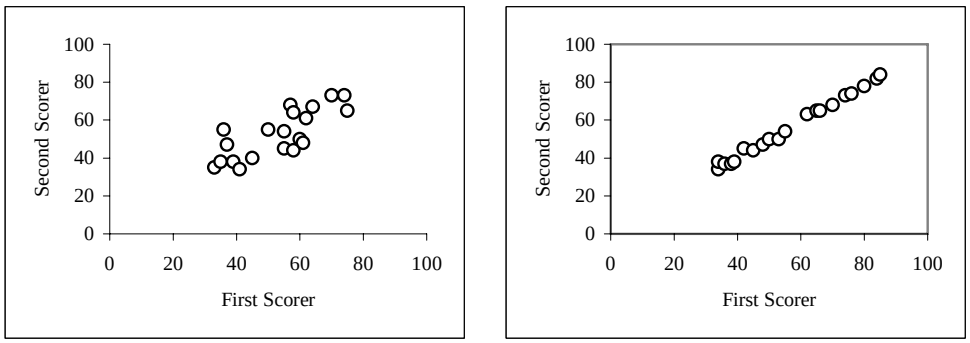


Figure 1. The Consistency between the Scorers

CONCLUSION AND DISCUSSION

One of the results was that the scores of the computer versions were higher than the pen-paper ones. However, in some studies (Bridgeman and Cooper, 1988; Sweedler-Brown, 1991), it was found that scorers assigned lower scores to computer version of the tests than the pen-paper ones. In other studies (Powers, Fowles, Farnum, and Ramsey, 1994), there was no difference between the computer and hand-written versions of the tests in the process of scoring. In another study, Russell and Haney (1997) compared students' responses on writing assessment items and found that writing on computer had a positive impact on students' writing scores. Finally, although the research in this area is not conclusive, and has not established a consensus on test medium on scores (Bunderson, Inouye, and Olsen, 1989), in this study, it was found that the scores of the computer versions were higher than the pen-paper ones.

The findings in the study showed that the test and inter-rater reliability of the writing test scores of ESL learners in analytic scoring were significantly higher than the ones of the pen-paper group participants. Indeed, the reliability of a test depends on some factors; scoring method, scale length, text length, writing approach or method, topic, writing abilities and progress level of writers, and raters (Penny, Johnson and Gordon, 2000). Two of the variables that affect reliability were scoring method and raters' react to writing on computer. Breland (1983) found the higher levels of inter-rater reliability were associated with analytic scoring. Since the same scoring procedure was used for both versions of the tests, the scoring method was not a factor that affects the scores. On the other hand, in Hee-Kyung's (2004) study on comparison among the hand-written, transcribed and computer generated essays in analytic and holistic scoring of writing tests of ESL writers, it was found that hand-written essays were more reliable than transcribed and computer-generated essays. In sum, in this study, the scores of both the handwritten and computer versions of the tests in analytic scoring were reliable. However, the reliability coefficient of the tests administered on computer was significantly higher than the ones of pen-paper tests. On the other hand, since the duration between pre- and posttests was one week and the students did not receive any writing instruction, the progress level of the participants was not a variable that affects the reliability. Finally, writing ability was not also a factor that affects the scores since the sample group consisted of the students that have equal writing abilities. In conclusion, the results in this study showed that computer use in the writing tests of ESL writers had an effect that increases the test and inter-rater reliability when the writing tests of ESL learners are scored analytically. However, the research on the test and inter-rater reliability of writing tests of ESL students seems conflicting and not conclusive as McNamara (1996) points out that the reliability is an unresolved issue in writing assessment.

Some limitations of the research can be noted. First of all, the study is limited to the ESL learners at ELT Department of the Education Faculty of Ataturk University, Erzurum, Turkey. Second, the compositions were written in free writing approach, and the tests were scored analytically. Third, the different topics presented as pre- and posttest might be a factor that affects the scores. In sum, the results in the study are limited to the ESL writers of upper-intermediate level, free writing approach, the scale presented below, and analytic scoring.

Considering that the study is limited to the test and inter-rater reliability of writing tests of ESL writers, further research should be focused on the factors that affect the attitudes of scorers and writers. The scoring scale, the comparison of holistic and analytic scoring, different writing approaches and methods, and the topics of writing exams are other areas to be investigated. Finally, the writing abilities and progress level of participants are also other factors that should be researched.

REFERENCES

Breland, H. (1983). The Direct Assessment of Writing Skills: A Measurement Review, Technical Report No: 83-6, Princeton, NJ: College Entrance Examination Board.

Bridgeman, B., & Cooper, P. (1988). Comparability of Scores on Word-processed and Handwritten Essays on the Graduate Management Admissions Test, Paper Presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.

Brown, J. S. (1977). Uses of Artificial Intelligence and Advanced Computer Technology in Education, in Robert J. Seidel & Martin Rubin, (Eds.) Computers and Communication: Implications for Education, New York, NY: Academic Press Inc.

Brown, J. D. (1997). Computers in Language Testing: Present Research and Some Future Directions. *Language Learning and Technology*, Vol. 1, No. 1, pp. 44-59.

Bunderson, C. V., Inouye, D. K. & Olsen, J. B. (1989).The Four Generations of Computerized Educational Measurement. In R. L. Linn (Ed.), Educational Measurement, London, Collier Macmillan, 367-407.

Daiute, C. (1985).*Writing and Computers*. Menlo Park: Addison-Wesley.

Dalton, D. W., & Hannafin, M. J. (1987). The Effects of Word Processing on Writing Composition. *Journal of Educational Research*, No: 80, 338-382.

Dunkel, P. (1991).The Effectiveness Research on Computer-assisted Instruction and Computer-assisted Language Learning. In P. Dunkel (Ed.), Computer-assisted Language Learning and Testing: Research Issues and Practice, New York, Newbury House, 5-36.

Hee-Kyung, L. (2004). A Comparative Study of ESL Writers' Performance in a Paper-based and Computer-delivered Writing Test. *Assessing Writing*, Published by Elsevier Inc.

Krashen, S. (1982). *Principles and Practice in Second Language Acquisition.* New York: Pergamon.

McNamara, T. (1996). *Measuring Second Language Performance.* Longman, London.

Neu, J. & Scarcella, R. (1991). Word Processing in the ESL Writing Classroom: A Survey of Student Attitudes. In P. Dunkel (Ed.) Computer-assisted Language Learning and Testing: Research Issues and Practice, New York: Newbury House, 169-187.

Penny, J., Johnson, R. L., & Gordon, B. (2000). The Effect of Rating Augmentation on Inter-rater Reliability: An Empirical Study of Holistic Rubric. *Assessing Writing*, 7, 143-164.

Phinney, M. (1988). Computers, Composition and Second Language Learning. In M. C. Pennington, (Ed.), Teaching Language with Computers: The State of Art, 81-96, San Francisco, Athelstan.

APPENDIX 1: Writing Topics

1. Pretest items for the pen-paper and computer group participants:
   a. When choosing a place to live, what do you consider most important: location, size, style, number of rooms, types of rooms, or other features? Use reasons and specific examples to support your answer.
   b. Films can tell us a lot about the country in which they were made. What have you learned about a country from watching its movies? Use specific examples and details to support your response.
   c. Because of developments in communication and transportation, countries are becoming more and more alike. How is your country becoming more similar to other places in the world? Use specific examples and details to support your answer.
2. Posttest items for the pen-paper and computer group participants:
   a. People attend colleges or universities for many different reasons (for example, new experiences, career preparation, and increased knowledge). Why do you think people attend colleges? Use specific reasons and examples to support your answer.
   b. If you could change one important thing about your hometown, what would you change? Use specific reasons and examples to support your answer.
   c. If you had the time and money to invent something new, what product would you develop? Use specific details to explain why this product is needed.

APPENDIX 2: Writing Proficiency Scoring Table

Writing Proficiency Scoring Table			
Student's Name		Scorer's Name	
Student's Number			

Criteria / Point	10	9	8	7	6	5	4	3	2	1	0
Vocabulary											
Accuracy (Grammar and structure)											
Organization											
Originality and Creativity											
Unity and Coherence											
Relevance											
Mechanics											
Fluency											
Content											
Language Control											
Score											
Total Score											