

Score Equivalence between Paper-based and Computer-based College Students' English Writing

Yinxu Ji

College of Foreign Language Education
China West Normal University
Nanchong, P.R. China

Abstract—Paper-based and computer-based writings are two usual modes used by English teachers in teaching writing. Computer and assessment familiarity, however, may have influence on students' performance in their writings, so the reliability of the scores given in the two modes is challenged. This paper studied college students' scores in 298 writings on the two modes and found no significant difference in the score equivalence from overall view between paper-based and computer-based English writing though there was slight difference in scores given to Lexical Resources and Coherence and Cohesion. The research findings may ease some teachers anxiety about the reliability of scores for computer-based writing.

Keywords—score equivalence; reliability; paper-based; computer-based; writing assessment

I. INTRODUCTION

With China's English test's transition from paper to electronic mediums, the paper-based (PB) Writing, which has established a strong reputation over many years for its practicability and reliability by thousands of teachers and learners around the world, is now making way for computer-based (CB) writing and assessment. Newman, Couturier & Scurry points out in line with the increasingly important role of technology in all areas of higher education, computer-based assessment is becoming more and more common in most university disciplines[1]. With the popularity of pigai.org (an online writing and grading system), many English teachers in China have been incorporating this electronic tutoring or computer mode writing system to facilitate the teaching of writing by frequently asking students to submit electronic or computer-typed assignments and essays as part of standard ESL-program curricular requirements.. Compared with paper mode writing, the CB assessment does well to both learner and teachers in that it increases learners' motivation and enthusiasm and it promotes teachers assessing efficiency as well. However, with students' increased familiarity with computers in both learning and assessment, as well as the assessing criteria themselves, students begin to use advanced words and complex structures on purpose. Some researchers argue such scores can not objectively reflect their language proficiency, thus the scores are doubted in computer mode English writing. While this change in writing mode may not be new, and it may not be consistent in the context of the whole world, it is clear that computers, mobile technologies, and the Internet are

transforming the way people perceive and write essays, and this influence is quickly spreading. The creation of electronic text is starting to become a new normal for writing. This trend puts into question the authenticity of using paper-based writing assessments over computer-based ones, when the majority of students' writing is now produced by typing in electronic text[2]. This research study, therefore, aims to investigate whether meaningful differences would be observed in the score equivalence when writing was presented in two modes –paper and computer. It is anticipated that such research is a necessary pre-requisite to the prevailing introduction of a CB version of online writing practice in English teaching in China.

II. LITERATURE REVIEW

In terms of score equivalence, since the introduction of computer into language classroom, researchers had been concerned about the influence of computer on learning. Early research by Mazzeo and Harvey (1988) had suggested that CB tests at the time tended to be more difficult than PB versions, perhaps partly owing to test-takers' lack of familiarity with technology[3]. Many other researchers concluded that the scores across the CB and PB modes can be considered comparable depending on appropriate design (e.g. Puhan, Boughton & Kim, 2007; Taylor, Jamieson, Eignor & Kirsch, 1998; Weir et al, 2007; Wise & Plake, 1989). More recent studies have concluded, however, that there is substantial comparability between scores in the two modes, perhaps partially owing to test-takers' growing familiarity with computer use in education and daily life. This has also proved to be the case for research into language testing.

Taylor et al (1998), for example, studied the comparability of PB and CB versions for the 1996 administration of the TOEFL exam and found no significant differences in score for test-takers taking the two different versions[4]. Likewise, Wise and Plake (1989) contended that PB and CB versions of achievement tests yield very similar scores[5].

Wolfe and Manolo (2005) found that scores given to essays written in CB mode are in fact "slightly more reliable than scores assigned to handwritten essays and exhibit higher correlations with TOEFL multiple-choice sub-scores"[6]. Puhan et al (2007) examined over 1000 participants in a test of writing in CB and PB modes, and also found no significant difference in scores between the two modes[7]. Based on performance of 262 participants, Weir et al (2007) reported

Supported by the Talent Fund of China West Normal University (17YC069)

that the difference between the PB and CB versions was not significant[8].

So long as the test design is carefully moderated, and assuming appropriate familiarity, attitude and anxiety levels on the part of test-takers, score equivalence is possible across the two modes in large scale tests of writing. Nevertheless, the studies on score equivalence between PB and CB writing were mainly about TOEFL and IELTS which are two highly motivated tests and are for learners from foreign cultures. Chinese learners are now going through transition between PB and CB, and they mainly use PB writing as ways to practice. Given that score equivalence in writing on paper and writing on pigai.org in China's context remains largely under-researched, this study will be significant for teachers, researchers and test designers.

III. RESEARCH DESIGN

A. Participants

1) Test-takers

148 non-English majors from China West Normal University were recruited. 45.4% were male and 54.6% female. They came from a variety of major disciplines, including Business, Mathematics and Politics. They were then freshmen, preparing College English Test Band 4 that is a very popular English language proficiency test. There is one thing worth mentioning about them. They were familiar with both the two writing modes in that they were mainly trained by PB mode in high school and CB mode in college.

2) Raters

The scores reported were double rated by 2 highly trained and experienced colleague teachers. They marked all the scripts, and the average score of one script was used as the final score. The full mark is 15 according to the rating scale. The marks are given in four facets, 3.75 points for each.

B. Test Tasks

After pre-pilot test among another 13 students and advice from 3 experts and researchers, two tasks modeled from past CET tests were selected to be used in the research. Each participant completed two tasks, one under the traditional PB mode and the other in the experimental CB mode. In the CB mode, the participants composed the essay on pigai.org.

C. Data Collection

148 students participated in the study. A total of 298 scripts (148 paper scripts and 148 computer scripts) were collected. All scripts are valid.

D. Data Analysis

To understand the impact of grading criteria on scores and to investigate score equivalence across the two modes, this analysis compared test-takers' overall scores between the two modes and test-takers' performance on each analytic rating scale (i.e. Task Achievement, Coherence and Cohesion, Lexical Resources, and Grammatical Range and Accuracy) between the delivery modes, four further 4-facet analyses were conducted. In each of the 4-facet analyses, delivery mode was not designated as a facet. The individual analytic category between the modes, e.g. CB Task Achievement and PB Task

Achievement, were treated as separate items. In other words, the four pairs of analytic scales between the delivery modes were compared one by one.

In all analyses, the data were entered into the SPSS 19, which analyzes the correlation between the scores given to the two modes and the equivalence between the four rating scales.

IV. RESULTS AND DISCUSSION

Before reporting the results, two questions must be cleared, the reliability of raters and comparability of the tasks. With sample analysis, it was found the exact agreement between the first and second rater was 68.3%. With statistical analysis, we were confident that test-version effect should not invalidate the findings of this study.

As indicated by the fixed chi-square statistics in Table I, test scores obtained from the CB and PB mode were not statistically different in terms of difficulty. Test-takers' performance, in terms of mean and SD under the PB and CB conditions, were very close, and Sig 0.071>0.05. Thus, there was no significant difference in the overall PB and CB scores.

TABLE I. TEST MODE MEASUREMENT

Test mode	Mean score	Std Deviation	Std. Error Mean	Sig
CB	7.82	2.23	0.53	0.071
PB	7.44	2.41	0.49	

Now, score equivalence findings are reported in relation to the four analytic scores under both conditions based on the additional 4-facet analyses (see Table II).

TABLE II. ANALYTIC SCALES MEASUREMENT

Analytic scale	Mean score	Std Deviation	Std. Error Mean	Sig
CB Task Achievement	3.12	1.33	1.02	0.06
PB Task Achievement	2.98	1.31	1.04	
CB Coherence and Cohesion	2.31	0.94	0.83	0.03
PB Coherence and Cohesion	2.03	0.91	0.78	
CB Lexical Resources	1.86	0.92	0.79	0.02
PB Lexical Resources	1.34	0.88	0.73	
CB Grammatical Range and Accuracy	1.63	0.91	0.62	0.07
PB Grammatical Range and Accuracy	1.65	0.86	0.61	

According to Table II, as indicated by the fixed chi-square statistics, while differences in two of the four analytic scores (i.e. Task Achievement, Grammatical Range and Accuracy) were not significant, the difference reported in Lexical Resources and Coherence and Cohesion were significant. Here, students' familiarity with the CB writing system played a role since they knew careful selection of advanced and academic words and frequent use of cohesive words may contribute to writing scores. Computer-based writing could have a more positive effect on writers' attention to form and accuracy, while paper-based writing could have a more positive effect

on writers' attention to content and meaning. However, the difference was very small and, hence, as reported previously, did not contribute to a significant difference in the test-takers' overall scores between the two modes. It is, therefore, recommended that test providers should monitor closely test-takers' performance on Lexical Resources and Coherence and Cohesion means between the two modes[9].

Through observation of participants' completing the tasks, there are also very interesting findings in the study. Some participants seemed very confident that they were able to create more content and better organization with a paper-based essay than with a computer-based essay. Though quantity of typed words and paragraphs in PB mode had a strong positive correlation toward computer-based scores of Content, Organization (for word count, not for paragraph count), and language use, quantity of handwritten words and paragraphs did not correspond with better scores in any category. This may have been due to the increased probability that writers with pen and paper would feel less anxious to create more paragraphs, thus eliminating any discrepancy between number of paragraphs and scores. Typists, on the other hand, were less likely to create more paragraphs due to its timed condition and pressures from typing speed, so the difference in scores was more pronounced when they did produce more.

V. CONCLUSION

The most important conclusion from the study is that there were no significant differences in the scores awarded by two independent raters for candidates' performances on the tests taken under two conditions, one paper-and-pencil and the other computer. In terms of the mean of the Lexical Resources scores and Coherence and Cohesion scores, test-takers scored 0.52 and 0.28 higher in the PB than CB conditions but the difference was very small. It's true that computer familiarity has some positive effect on students' performances on the two modes.

Though the scope of this study was small, and the small sample size may not be enough to explain cases in a larger context, it's expected that the information in this investigation can provide administrators, instructors, learners, and raters alike with important points to consider when it comes to making choices about choosing and using a computer or a pencil and paper for writing. Hopefully the differences between these modes of writing will continue to be explored and discussed in the future, as a better understanding of the

strengths, weaknesses, and differences in computer-based and paper-based writing can ultimately lead toward better utilization and performance in both. The question then changes from asking when we should choose PB or CB mode to how we can best use both modes to help learners with English essay writing.

In summary, the difference of scores between the two modes is at an acceptable level. So, both modes can be used as reliable means to teach writing and a combination of the two modes may be more productive in class. They tended to be more careful when generating texts in the PB than the CB mode. In terms of aspects of the revisions, some participants tended to focus more at the word level in the PB mode and more at the levels of clauses and sentences to improve coherence or argument in the CB mode.

REFERENCES

- [1] Newman, F., Couturier, L. and Scurry, J. *The Future of Higher Education: Rhetoric, Reality, and the Risks of the Market*. San Francisco: John Wiley & Sons. (2010).
- [2] Endres, H. A comparability study of computer-based and paper-based writing tests. *Cambridge ESOL: Research Notes*, 49, 26-32 (2012)
- [3] Mazzeo, J. and Harvey, A. L. *The equivalence of scores from automated and conventional educational and psychological tests: A review of the literature*. College Entrance Examination Board, New York. (1988).
- [4] Taylor, C., Jamieson, J., Eignor, D. and Kirsch, I. *The relationship between computer familiarity and performance on computer-based TOEFL test tasks*. Research Reports 61. Princeton, NJ: Educational Testing Service. (1998).
- [5] Wise, S. and Plake, B. Research on the effects of administering tests via computers. *Educational Measurement: Issues and Practice*, 8(3), 5-10. (1989).
- [6] Wolfe, E.W. & Manalo, J.R. Composition medium comparability in a direct writing assessment of non-native English speakers. *Language, Learning, and Technology*, 8(1), 53-65. (2004).
- [7] Puhon, P., Boughton, K. and Kim, S. Examining Differences in Examinee Performance in Paper and Pencil and Computerized Testing. *Journal of Technology, Learning, and Assessment*, 6(3), 1-21. (2007).
- [8] Weir, C. J, O'Sullivan, B., Yan, J. and Bax, S. Does the computer make a difference? Reaction of participants to a computer-based versus a traditional handwritten form of the IELTS writing component: effects and impact. *IELTS Research Report*, Vol 7, (pp. 1-37). IELTS Australia, Canberra and British Council, London. (2007).
- [9] Sathena Chan. Researching participants taking AWT2 in paper mode and in computer mode in terms of score equivalence, cognitive validity and other factor. *IELTS Research Report*, Vol 4, (pp. 1-47) (2017)