Title of Project:
Investigating the Authenticity of Paper- and Computer-Based ESL Writing Tests

Researcher:
Mingxia Zhi
zhimingxia1990@gmail.com

Current Affiliation
Northside Independent School District

Doctoral Institution
The University of Texas at San Antonio

Research Supervisor:
Dr. Becky Huang
The University of Texas at San Antonio

TIRF Research Topic Investigated:
Language Assessment

Final Report

Motivation for the Research
The use of technology in the administration of language assessments has received extensive research attention. Because the majority of students’ writing and learning activities are now mediated by computers (Endres, 2012; Kohler, 2015), many researchers have argued for using computer-based (CB) ESL writing tests, resting on the assumption that CB writing is more authentic than paper-based (PB) writing task (Cheung, 2016; Kohler, 2015; Lee, 2004; Lessien, 2013). Such assumptions, however, are supported by little empirical evidence. Little research has investigated the authenticity of English as a second language (ESL) writing assessments delivered in PB and CB formats. The authenticity of a test is defined as “the degree of correspondence of the characteristics of a given language test task to the features of a target language use (TLU) task” (Bachman & Palmer, 1996, p. 23). The correspondence of physical characteristics of the test task to TLU task is influential on test takers’ language use (Bachman & Palmer, 2010). Test authenticity is important because it relates to the domain of generalization of a test (Bachman & Palmer, 2010) and has the potential effect on test-takers’ perceptions of the test (perceived relevance). This study filled the research gap by examining the effect of test modality (PB and CB) on the authenticity of ESL academic writing tests.

Research Questions
Informed by Liu’s (2005) conceptual model of five aspects of authenticity (TLU task, test task, learner interpretation, learner involvement, and learner response), this dissertation study
compared the authenticity of CB and PB ESL writing tests. The research questions (RQs) are the following:

1. To what extent do second language (L2) writing processes reported in daily English Writing tasks correspond to writing processes reported in the CB and PB writing tests?
2. Does test-takers’ computer familiarity correlate with their CB and PB test essay scores?
3. What are the test-takers’ perceived connections between the CB and PB writing tests and daily English writing tasks?
4. Does test-takers’ perceived authenticity of the task medium correlate with their perceived validity and preferences of the testing mode?

Research Methodology
This study employed the embedded correlational model of mixed methods design (Creswell & Plano Clark, 2011). Sixty International ESL students (N = 60) were recruited using convenient sampling and snowball sampling methods. Data collection was carried out in three stages (Figure 1). Prior to the study, the participants were recruited using a pre-survey. Then, the participants were grouped into four groups, and group equivalence was attained by matching the participants on gender, age range, score of the standardized writing tests (TOEFL/IELTS), and self-ratings of computer familiarity. The counter-balanced measures are used to solve the issue of potential carryover effect, fatigue effect, and context effect (Price, Chiang, & Jhangiani, 2015). During the study, the participants first completed a one-minute online typing test and then completed the two assigned writing tasks. After the study, a post-survey and an interview were conducted with the participants to capture their perceptions of the two writing test tasks. To analyze the writing samples, the handwritten essays were transcribed verbatim into a typed form to ensure the same presentation of the essays to the rater. Each essay was rated by at least two raters using a holistic scale and an analytic rating scale. For RQ 1, bivariate correlation analysis (Spearman rho) was conducted to examine the relationships between the ratings on the process items in the pre-survey and the ratings on the cognitive writing questionnaire items in the CB and PB tasks, respectively. Correlation coefficients were compared with an updated version of Steiger’s Z test (Steiger, 1980) using a web-based calculator for testing the statistical significance of the difference between dependent correlations (Hoerger, 2013). For RQ 2, Spearman’s rank order correlation test was used to test the bivariate correlations between the survey Likert-scale items, such as computer use, and the essay scores. Two multiple regression analyses were conducted to predict the scores of CB- and PB-generated essays, respectively, using the typing speed, typing accuracy, after controlling for the standardized writing test scores. For RQ 3, descriptive statistical analysis was conducted with the ratings on post-survey item. For RQ 4, bivariate Spearman’s rank order correlation analysis was employed between the ratings on post-survey item. The two cycle coding method was used to analyze the qualitative data, with In-Vivo coding for the first cycle and focused coding for the second cycle.
Summary of Findings

RQ 1. Results suggested that both mediums of writing tests, to different degrees, yielded writing processes that were consistent with the TLU writing processes. Further test of the statistical significance of the difference between dependent correlations using Hoerger’s (2013) calculations indicated that the CB writing only yielded higher correlations than the PB writing test on three items: thinking of genre knowledge during planning ($Z_H = 2.68^{**}$, $p = .007$), making pauses during writing for reviewing ($Z_H = 2.189^*$, $p = .028$), and making pauses during writing for revising ($Z_H = 3.328^{***}$, $p < .001$). Therefore, the quantitative results suggested significant but limited evidence for a higher degree of authenticity of the CB writing test. Qualitative results suggested that more participants perceived the CB writing process to be more similar to TLU writing process, thanks to a frequent use of computer for studying and work. Further analysis revealed that many test-takers were “forced” by the PB medium to constantly check during composing due to the difficulties of conducting major revisions after writing. This might have contributed to the low correlation relationship between the PB writing and TLU tasks in the survey item -- d4 "I pause after a few sentences or a whole paragraph to read it again". Taken together, the results partially supported the claim that the CB writing test corresponded more to the TLU tasks than the PB writing task.

RQ 2. The results showed that the computer availability at school significantly and positively correlated with the CB essay holistic scores, specifically, with the grammar and accuracy scores. Nevertheless, computer availability at school variable did not correlate with the PB essay holistic scores. To further examine the relevancy of typing proficiency in the L2 CB and PB writing assessment, I conducted two multiple regression analyses to examine the effect of test-takers’ typing speed and accuracy on their CB-and PB-generated writing scores. Results suggested typing accuracy variable significantly and positively predicted the CB essay scores, but not PB essay scores, after controlling for their standardized writing scores. Qualitative results corroborated with the quantitative results in that the majority of the participants perceived an influence of their computer skills on their CB writing performance. Most test-takers reported this influence to be positive, for the reason that CB writing was a more familiar medium for more test-takers. Computers also provided writers with more time and convenience for revising. Nonetheless, some students felt they were penalized by this test medium due to a lack of typing.
skills in English. A small group of participants did not see a significant influence. Taken together, the results demonstrated an effect of computer familiarity and typing skills on test-takers' CB writing outcomes in this study. In particular, it was the computer availability and use at school, as well as the typing speed that made a statistical difference.

RQ 3. Both survey data and interview responses reported more participants perceived a higher correspondence between CB writing and TLU tasks compared to the PB test, indicating a higher perceived authenticity of the CB writing condition for assessing L2 writing. Only eight percent of the participants perceived the two tests as equally authentic. Further qualitative analysis revealed that although most students reported using computers in college studies, there was a discrepancy in assignment practices between U.S. colleges and schools in other countries. Some students shared that electronic assignments had not become a common practice in their home countries. Though having personal access to computers, computers were not always available or frequently used at schools. The results, though in favor of CBT for assessing writing, also brought concerns with the fairness of using computers for administering high-stakes writing test with test-takers from diverse backgrounds.

RQ 4. Results showed that test-takers who perceived the PB test more authentic rated it more valid, prefer to take the PB test, and rated CB test less valid. The ratings on the perceived authenticity of the CB writing, on the other hand, did not yield significant correlations to the perceived validity of CB or PB writing assessment. The perceived performance on the test was found to significantly and positively correlate with the perceived validity of the testing medium. Finally, typing speed significantly negatively correlated with the preference to the PB test. This means that the faster one typed, the less likely he/she preferred to take the PB test for writing. In the post interview, the participants explained that when they were more comfortable with the test medium, they could engage in a more routinized writing process, just like when they wrote in real-life academic English tasks in schools. Hence, they might perform better in the test. Taken together, the results supported a correlation relationship between the perceived authenticity of the testing medium and the consequential validity of the test.

Implications
This study uncovered both advantages and limitations of the CB writing test. The CB test was generally more authentic than the PB test for the majority of the test-takers. However, it also brought concerns in the construct relevance of computer literacy, fairness, and its consequence. The findings of this study have implications for test developers and university test administrators. First, test developers should consider the characteristics of the test-takers when developing high-stakes tests. Considering the various backgrounds of the ESL learners from different countries in the world, it is important to offer tests in more than one medium to adapt for all test-takers to avoid test biases. The digital divide within the test-taker population may be a reflection of the divide in socio-economic status (SES) of the international ESL learners. Second, when interpreting the test results, school admission personnel should keep in mind that the construct tested in CB and PB modes of the test may be different. Other than English language proficiency, the student candidates’ computer literacy skills may also be tested by the CB test of writing. In addition, test purpose is another factor to consider when choosing the medium for ESL writing tests. Different test mediums may be chosen for high-stakes English proficiency tests for college admission purposes and local in-house ESL assessments for diagnostic purposes. Finally, when comparing the equivalence of the two mediums of L2 writing tests, one key concern from the participants’ perspective was the time. For test-takers with higher computer
skills, typing seemed to earn more time for more deliberate planning and revisions. Similarly, for students who were more comfortable with handwriting, the lower their typing skills were, the lower their essay scores were on the CB test. It was perceived to be a key reason for lower self-rated performance on the test when participants had less time to plan, review, and revise due to a disadvantage in typing. Allowing more time in both testing conditions may mitigate this issue. More research is needed to investigate the ideal duration of the test with both objective measures and perspectives from the test-takers.
References


Hoa, N. T. N. (2018). Test takers’ attitudes to the test content of the two listening tests: IELTS and TOEFL iBT. *VNU Journal of Foreign Studies, 34*(4).


