Investigating the authenticity of computer- and paper-based ESL writing tests

Mingxia Zhi *, Becky Huang

Department of Bicultural Bilingual Studies, The University of Texas at San Antonio, San Antonio, USA

ARTICLE INFO

ABSTRACT

An assumption underlying replacing traditional paper-based (PB) writing tests with computer-based (CB) tests is that CB writing tests are more authentic than PB tests (Lessien, 2013). This study tested this assumption by examining the effect of test mode on the situational and inter-active authenticity of English as a second language (ESL) writing assessments using an embedded correlational model of mixed-methods design. Sixty international ESL students enrolled at a university in the United States completed two writing tasks in the CB and PB modes. The participants then filled out cognitive processing questionnaires and participated in a post-test interview. All essays were scored by two raters both holistically and analytically. The results indicated higher authenticity for the CB test but also uncovered some concerns with the effect of typing skills on CB test performance. The findings have practical implications for test development, administration policies, and stakeholders in choosing delivery mode for ESL writing assessments.

1. Introduction

The delivery mode of writing assessments has been identified as one of the key variables that potentially influence test-takers' performance (Shaw & Weir, 2007). Currently, there are two modes of standardized English as a second language (ESL) writing tests, the computer-based (CB) and paper-pencil-based (PB). Research remains inconclusive about the validity of the CB test and comparability between the CB and PB modes (Barkaoui, 2015; Chan, Bax, & Weir, 2017; Cheung, 2016; Weir, O'Sullivan, Jin, & Bax, 2007). The current study investigates test *authenticity,* referred to as the degree of correspondence of a given language test task's (LTT) characteristics to those of a target language use (TLU) task between the two modes (Bachman, Palmer, & Palmer, 2010). The authenticity of a test is important because it is related to *construct validity*, the meaningfulness and appropriateness of the inferences that we make from test scores (Bachman et al., 2010).

Authentic writing tests reflect the construct of writing in real life, including the language knowledge, skills, and strategies involved in real-life writing processes. Despite authenticity's critical role as a validity index (Messick, 1994), research on test authenticity is scarce, and the very few studies of test authenticity have focused on the test content rather than the test mode (Liu, 2005). The present study fills this research gap by investigating the authenticity of CB and PB ESL writing tests.

* Corresponding author.
  E-mail addresses: Zhimingxia1990@gmail.com (M. Zhi), becky.huang@utsa.edu (B. Huang).

2. Test authenticity

Authenticity is not an all-or-none quality; it exists in every test to varying degrees (Bachman et al., 2010). Building on previous theoretical discussions of authenticity (Bachman et al., 2010; Lewkowicz, 2000), Liu (2005) describes two types of evidence in authenticity – the *situational* and *interactive* aspects of authenticity. *Situational* authenticity relates to Bachman et al.'s (2010) original definition of authenticity by examining the correspondence of the characteristics of TLU tasks to the test tasks. Earlier research on test authenticity focused mainly on the test content, but not enough on how test-takers respond to the test content (Bachman, 1990; Green, 2014). While most research evaluated the authenticity of test materials in terms of resemblance to the real-life task, the performance conditions of the test should also be examined (Fulcher, 2010).

Interactive authenticity refers to the "interaction between test takers, test tasks/contexts, and discourse" (Liu, 2005, p. 39). Fulcher further argues that "the presence or absence of the construct" under performance conditions in a test task should represent what might happen in real life (p. 99). There are two aspects of interactive authenticity: test takers' and/or language users' internal interpretation of the authenticity and the involvement of test-takers' characteristics in accomplishing test tasks (Bachman et al., 2010, as cited in Liu, 2005, p. 39).

To operationalize the two types of authenticity evidence, Liu developed a conceptual model of five aspects of authenticity in language testing: TLU task, test task, learner interpretation, learner involvement, and learner response. TLU task and test task deal with the specific definitions of the features in the TLU and test tasks while learner interpretation, learner involvement, and learner response involve the learners/test-takers' interpretations, engagement, and response to the test tasks. In the present study, we examined the *situational authenticity* of the writing test in the CB and PB modes focusing on two features of TLU and test tasks – the writing mode and strategy use – during the writing processes to complete a task. Test mode is a part of the performance condition that may resemble or differ from the TLU task conditions (Fulcher, 2010). Different writers may respond to the same performance condition differently due to different training and daily writing habits. Hence, it is important to observe the writing process to examine how diverse test-takers respond to the test (Green, 2014). We also examined *interactive authenticity* through the learner factors by studying the relationships between test takers' computer literacy and their writing outcomes.

3. Comparing CB and PB ESL writing tests

Since the introduction of computerized tests, the comparability of the CB and PB writing tests has been the focus of research inquiry. Allowing the use of CB writing may increase the authenticity of the task for some test takers (Weigle, 2002). However, concerns remain due to the potential disadvantages such as a higher cost of the computer equipment at the testing site and the possible effect of test takers' computer literacy (Chan et al., 2017). If authenticity is one of the arguments for favoring CB writing tests, as Weigle suggested, empirical evidence should be provided to support this claim. However, previous research has mainly focused on test validity and reliability, rather than authenticity, by comparing the writing processes and products in CB and PB ESL writing tests.

3.1. Testing mode and writing process

Previous research has investigated the effect of test mode on writing processes by comparing the writing activities and strategies observed during the planning, transcribing (composing), and revising stages of writing (e.g., Hoomanfard & Meshkat, 2015; Weir et al., 2007; Zou & Chen, 2016). It seems that, although consensus has been made that test-takers took more time in pre-planning when writing in PB modes (Lee, 2002; Li, 2006), inconsistent results were shown in other aspects of the writing process. In terms of revisions, Li (2006) claimed that the ESL writers made significantly more revisions *after* composing in the CB mode while Hoomanfard and Meshkat (2015) found writers undertook more revisions in the CB mode *during* the process of writing. Furthermore, while some studies (Li, 2006) showed that writing in the PB mode led to more *higher level-revisions*, which are operationalized as revisions focusing on the meaning of the message, organization, and coherence, other studies (e.g., Lutz, 1987) indicated that students have revised more at a superficial editing level when using word processors.

The inconsistent results may be explained by the individual differences of the L2 writers, such as different levels of computer literacy (Barkaoui, 2015; Lee, 2002; Zou & Chen, 2016). Computer literacy refers to the experience, familiarity, and skills with computers (Barkaoui, 2015). Test takers' computer literacy may have an effect on the composing stages of organizing ideas and translating/composing, indicating a bias against the test-takers with lower computer literacy (Zou & Chen, 2016). Barkaoui (2015) found that a high keyboarding skill group engaged in more planning, organizing, and revising activities than a lower keyboarding group in an independent writing task under the CB writing mode. However, the effect of computer literacy on writing performance may be negligible for the cognitive writing process (Weir et al., 2007) and the quality of revision behaviors (Barkaoui, 2015). In Barkaoui (2016) study, although there were observed differences in the overall quantity of the revision activities across the high- and low-keyboarding skill groups, there was no significant effect of keyboarding skills on the actual quality (types) of revisions. Most revisions were at the surface level (typography and language revisions) but not at the higher level (content revisions) across the two groups. Weir and colleagues (2007) found that only the accessibility of public computers and the frequency of word processing activities correlated with up to half a score band in their writing scores. A recent study by Chan and colleagues (2017) showed an increase in the participants' familiarity and comfort level with using computers. Some participants reported they spent less effort with pre-planning, were less careful with generating texts, and revised more at the higher level after composing. The increase in learners' daily use of computers may explain the differences between earlier and more recent studies. The writing process in a more authentic mode may require a less conscious effort with test use strategies and more engagement with L2 writing strategies.

3.2. Testing mode and written product

Studies comparing the written products of CB and PB modes have also shown mixed results (Cheung, 2012). Some studies found higher quality of and/or higher scores were assigned to the CB writing products in specific areas of writing, such as length (Cheung, 2016; Sarbakhshian & Saeidi, 2016), organization (Cheung, 2016), and grammatical accuracy (Sarbakhshian & Saeidi, 2016). In contrast, other studies reported higher scores in the PB writing scores in organization (Kohler, 2015). Still other studies suggested no significant differences in the quality of essays produced in the two modes (Kohler, 2015; Lee, 2002; Weir et al., 2007). Possible explanations for these mixed results include the different contexts of the studies (e.g. English as a second language versus a foreign language) and the variability of the participants' characteristics such as computer familiarity (Shirzad & Shirzad, 2017) and typing speed (Russell, 1999).

A body of research showed an effect of computer literacy on the CB writing performance. Shirzad and Shirzad (2017) found that test-takers with higher computer literacy performed better than a lower computer literacy group in a standardized writing test, although no difference was found when the two groups took the PB practice test. However, other studies suggested no significant relationship between typing speed (Kohler, 2015) and the CB writing test scores. These inconsistent results call for further investigation of the relationships between computer literacy and CB writing performance.

This study aimed to investigate the *situational* and *interactive* aspects of authenticity that link to construct validity. The two specific research questions are:

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. What are the effects of test-takers' computer literacy on their CB and PB test essay scores?

In this study, we conceptualized "computer literacy" as an umbrella construct that includes three subcomponents: computer availability (e.g., computer availability at home), computer usage (e.g., computer usage at home), and computer skills (e.g., typing speed and accuracy).

4. Materials and methods

The current study adopted the embedded correlational model of mixed methods design (Creswell & Plano Clark, 2011), which embedded a qualitative component within a correlational quantitative design. The qualitative data (post-test interviews) explained and complemented the quantitative measures in this study.

4.1. Participants

The study was conducted in a university in the Southwestern United States. Sixty international ESL students from both undergraduate and graduate programs were recruited via convenience sampling and snowball sampling. The screening criteria were: (1) have a non-English first language; (2) have taken a high-stakes English language proficiency test for admittance to the postsecondary institution within the past two years. The minimal number of participants was determined based on the power analysis results using G* Power 3.1 software. The result of the power analysis indicated that a large effect size ($r = .5$) for two-tailed bivariate correlational tests required a sample size of more than 46.

Fifty-two percent of the participants were males (N = 31) with an average age of 25 (SD = 4.2; range = 18–36). Eighty percent of the participants were pursuing a graduate degree. Fifty-two percent of these students were studying a computer-related major, such as computer science and information systems. On average, these students had been learning English for 13.58 (SD = 8.16) years. Based on their standardized English language proficiency test scores (e.g., IELTS and TOEFL), the participants were competent English language users with a mean IELTS total score of 6.62 (SD = .52) and IELTS writing score of 6.09 (SD = .70) (Understand and explain the IELTS scores, n.d.). The majority of these international students came from Asian countries, such as India (33 %), Bangladesh (12 %), China (10 %), Taiwan (3%), and Japan (3%).

4.2. Instruments

4.2.1. Pre-survey of writing strategy use in non-test settings

This survey consisted of 45 Likert-scale items (α = .821) and included three sections: background information (e.g., gender, L1, standardized test scores from IELTS or TOEFL iBT), writing strategy use (under non-testing conditions), and writing habits and computer literacy. The writing strategy component of the questionnaire was revised from a previously validated L2 writing strategy survey developed by Petrić and Czárl (2003) with permission of use. This questionnaire surveyed L2 writers about their use of strategies, defined as "actions or behaviors consciously carried out by writers in order to make their writing more efficient" (Petrić & Czárl, 2003, p. 189), following the structure of the writing process: before-, during-, and after writing.

The data elicited from this instrument were correlated with the activities reported during the composing processes (from the cognitive writing questionnaire) in the CB and PB tests. The purpose was to examine the correspondence between the writing processes and strategies used in the CB and PB test conditions and the writing processes in real-life academic language tasks in schools. The original questionnaire (38 Likert-scale items) was revised to be comparable with the cognitive writing questionnaire items. Some
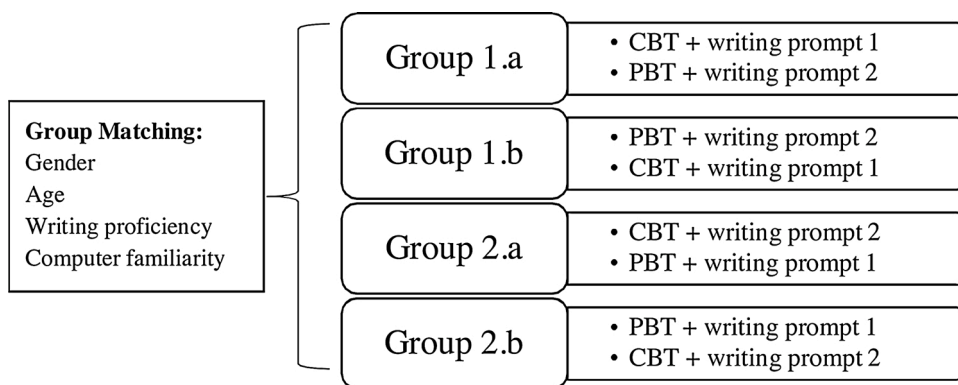
Fig. 1. Counter-balanced four group design.

Table 1
A Summary of Data Analysis to Research Questions.

Research Question	Data Sources	Analytic Methods
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?	a. Pre-survey b. Cognitive questionnaire in CB and PB mode c. Interview question 1.a, 1.b, 1.c	a. Bivariate correlational analysis (survey and questionnaires) b. Two cycle coding (interview transcript) c. Qualitative data to explain the quantitative data
2. What are the effects of test-takers' computer literacy on their CB and PB test essay scores?	a. Pre-survey b. Typing test c. Ratings of the writing sample d. Interview question 2	a. Bivariate correlational analysis (survey and questionnaires) b. Regression Analysis (typing and writing test scores) c. Two cycle coding (interview transcript) d. Qualitative data to explain the quantitative data

questions were added from the cognitive questionnaire from Chan and colleagues' (2017) article (see Appendix A).

4.2.2. Typing test
    To measure the typing speed of the participants as one of the computer literacy variables, an online one-minute speed-typing test was employed: https://www.speedtypingonline.com/typing-test. Both the typing speed (correct words per minute) and the typing accuracy (the percentage of correct words divided by all words typed) were automatically calculated.

4.2.3. Writing prompts
    The two writing prompts were adapted from the independent writing task topics for practice in the official guide to the TOEFL test (ETS, 2012*)*. Both prompts are academic topics and were composed in the same structure – "some people like… other people like… which do you prefer".

4.2.4. Cognitive processing questionnaire
    A cognitive processing questionnaire, adapted with permission from previous studies (Chan et al., 2017; Weir et al., 2007), was employed to capture writers' cognitive processes during writing. The questionnaire was administered immediately after the participants completed each writing task. The questions were comparable to those in the writing strategy questionnaire to examine the correspondence of writing processes in the LTT and the TLU tasks. All the items were composed in the past tense and the participants were instructed to rate the items based only on the writing task they completed right before the questionnaire.
    The original questionnaire (40 items) was revised to be comparable to the writing strategy questions in the pre-survey. After removing and adding new questions from Petrić and Czárl (2003) writing strategy questionnaire, the final instrument contained 45 Likert-scale items about the writing activities during the different stages of writing, including reading the test instructions, planning, writing, and revision stages. The Cronbach alpha reliability test showed that this instrument was reliable for both writing tasks (α = .869 for the survey used after CBT, α = .843 for the survey used after for PBT) (see Appendix B).

4.2.5. Post-test interview questions

A structured interview was undertaken with each participant at the end of the study. The interview included five main questions with several sub-questions; the authenticity questions were adapted from Lewkowicz (2000) (see Appendix C).

4.3. Procedures

The survey of writing strategy use in non-test settings was conducted with all potential participants prior to the writing tasks. Based on the survey responses, participants who shared similar characteristics in age, writing proficiency, and computer literacy were randomly assigned into four groups (see Fig. 1). We adopted a counterbalanced design to control for the potential effect of the task sequence and different writing prompts. Counterbalancing refers to the manipulation of the order of tasks/conditions in different groups. In this design, each participant completed two timed writing tasks (20 min each) responding to two different writing prompts in both CB and PB modes. In the CB mode, participants used word processor software where the autocorrect function was turned off, but they were able to copy, paste, delete, and undo at any time during the CB writing test. The cognitive processing questionnaire was administered immediately after the participants completed each writing task, and a post-test perception interview was conducted at the end of the study. The interviews were audio-recorded and later transcribed verbatim for qualitative analysis.

All PB essays were transcribed and entered into a computer. All essays were then scored by two raters both holistically and analytically in task achievement, organization, coherence and cohesion, lexical resources, as well as grammatical and accuracy. The holistic scale was adopted from the TOEFL writing rubric and the analytic scale was revised from the IELTS rubric. For the Prompt 1essays, two raters agreed within the same band at 86.67 % in holistic ratings and 83.33 % in analytic ratings. For Prompt 2 essays, the raters agreed at 91.67 % on both holistic and analytic ratings. Table 1 details the research questions and corresponding data sources and analytic methods. Because multiple comparison tests of correlations (n = 45) were conducted, Bonferroni adjustment was employed to adjust the alpha level (.05/45 = .001).

5. Results

5.1. Research Question 1. To what extent do L2 writing processes reported in daily English writing tasks correspond to writing processes reported in the CB and PB writing tests?

5.1.1. Quantitative results

The first research question examines the *situational* aspect of authenticity that focuses on the correspondence between LTT and TLU tasks. Table 1 presents the correlation coefficients between the writing process of TLU items measured in the pre-test survey and their corresponding items in the CB and PB test cognitive processing questionnaires. Most of the writing process items in both CB and PB test conditions correlated with the participants' self-reported writing processes in TLU tasks. The coefficients ranged from $rs = 215^*$ to $rs = .612^{**}$, suggesting weak to moderate relationships between the test writing process and reported real-life writing task processes. Twenty-seven of the 45 CB writing process items and 21 PB writing process items significantly correlated with the reported real-life academic writing process items, suggesting that both modes of writing tests, to different degrees, yielded writing processes that were consistent with the TLU writing processes. While the number of significantly correlated items were roughly the same for the *before- and after*-composing stages, there seemed to be more significant correlations in the CB condition than in the PB condition for the *during*-composition stage items, indicating that the CB test might resemble the TLU writing more than the PB test in the *during*-compositing process.

Further tests of the statistical significance of the difference between dependent correlations using Hoerger (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 (ZH = 2.68**, $p = .007$), making pauses during writing for reviewing (ZH = 2.189*, $p = .028$), and making pauses during writing for revising (ZH = 3.328***, $p < .001$). Because multiple comparison tests ($n = 45$) were conducted, Bonferroni adjustment was employed. After the Bonferroni adjustment, one item remained significant – making pauses during writing for revisions (ZH = 3.328***, $p < .001$). This result indicated that compared to the PB writing test, the CB test only yielded a higher correspondence between the LTT writing process and the TLU writing process in making pauses to revise during the composing stage. The majority of writing process items during PB and CB did not yield any significant difference when relating to the TLU writing processes. Therefore, the quantitative results suggested significant but limited evidence for a higher degree of authenticity of the CB writing test.

5.1.2. Qualitative results

The post-test interviews were analyzed with the two-cycle coding method using initial coding in the first cycle and focused coding in the second cycle (Saldaña, 2016). In the initial coding process, we employed in vivo coding method by assigning a label to a section of data using a word or short phrase directly taken from that section of the data (Saldaña, 2016). The interviews were coded based on both explicit and implicit mentioning of particular themes. For the second cycle, we used focused coding to further analyze initial themes. The qualitative results were consistent with the quantitative results that the CB test indicated a slightly higher perceived authenticity compared to the PB writing test. Thirty-seven out of the 60 participants commented that their CB writing was more similar to their everyday English writing processes than the PB test. Eleven students preferred the PB test, and 12 students reported no preferences.

Table 2
Correlation Coefficients between Pre-test survey and Writing Process by Mode.

Pre-survey	CBT	PBT	Z_H
Before Composing			
b1.Before I start writing, I carefully read the requirement.	.038	−.007	.329
b2.I think about what I know about the topic.	.257*	.232	.196
b3.I think about the genre knowledge I learned in class.	.314*	.005	2.675**
b4.I think about the purpose of writing	.254*	.341**	−.776
b5.I think about the intended readers of my writing.	.355**	.359**	−.037
b6.I think about the ideas about the topic BEFORE writing.	.360**	.467**	−.828
b7.I write down words and ideas related to the topic.	.431**	.304*	1.133
b8.I start writing WITHOUT having a written or mental plan.	.202	.279*	−.540
b9.I think about the ideas I want to write and have a plan in my mind, but NOT on paper	−.120	−.007	−.0867
b10.I revise my ideas and outlines BEFORE writing.	.103	.303*	−1.463
b11.I write an outline in English.	.387**	.437**	−.454
b12.I write an outline in my native language	.237	.226	.086
During Composing			
d1.I come up most of my ideas WHILE writing.	.227	.454**	−1.896
d2.I always develop new ideas WHILE writing	.126	.203	−.520
d3.I pause after each sentence to read it again.	.489**	.230	2.189*
d4.I pause after a few sentences or a whole paragraph to read it again.	.409**	−.058	3.328***
d5.I reread what I have written to get ideas about how to continue.	.446**	.386**	.606
d6.I pause frequently to revise the spelling.	.220	.301*	−.578
d7.I pause frequently to revise the content of the essay.	.292*	.290*	.018
d8.I pause frequently to revise the organization of the essay.	.345**	.233	.939
d9.I pause frequently to revise the sentence structure.	.276*	.244	.251
d10.I pause frequently to revise the grammar.	.467**	.161	2.531*
d11.I pause frequently to revise the vocabulary.	.280*	.252	.230
d12.I go back to my outline and make changes in it.	.074	.099	−.183
d13.I always think about how to connect my ideas smoothly in the whole essay.	.107	.182	−.547
d14.I always think about how to make my ideas persuasive to the reader.	.255*	.146	.894
d15.I only use the grammar and words that I am sure are correct.	.466**	.368**	1.044
d16.I simplify what I want to write if I don't know how to express my thoughts in English.	.419**	.274*	1.427
d17.If I don't know a word in English, I write it in my native language and later try to find an English word.	.258*	.106	1.439
d18.If I don't know a word in English, I find a similar English word that I know.	.247	.289*	−.396
After Composing			
a1.When revising, I read my text aloud.	.091	.169	−.424
a2.I read what I have written only after I finish the whole paper.	.270*	.231	.269
a3.When I have written my paper, I hand it in without reading it	.173	.347**	−1.138
a4.I focus on one thing at a time when revising.	.480**	.388**	.871
a5.I always make changes in the content or ideas.	−.031	.138	−1.481
a6.I always make changes in the structure of the essay.	.296*	.381**	−.766
a7.I always make changes to make my writing coherent.	.342**	.265*	.598
a8.I always check if my own viewpoint on the topic is in my writing and revised accordingly.	.552**	.458**	.999
a9.I always check the possible effect of my essay on the intended reader and revised accordingly.	.612**	.548**	.871
a10.I always make changes in grammar	.257*	.087	1.341
a11.I always make changes in sentence structure.	.071	−.012	.661
a12.I always make changes in vocabulary.	.228	.105	.834
a13.I checked my spellings and revised accordingly	.134	.248	−.865
a14.I check if my essay matches the requirements.	.423**	.581**	−1.562
a15.I leave the text aside for a couple of days and then I can see it in a new perspective.	−.003	−.152	1.168

5.1.2.1. The pre-composing stage. More students reported planning ideas in the CB (N = 24) than in the PB mode (N = 14), though an equal number of students (N = 15) planned the structure of the essays in both tests. These results indicated that the test mode might have an influence on the writers' planning of ideas, but not structures. Participant 13 explained that he knew he could write (type) faster in the computer test, so he spent more time planning in the CB writing condition (Excerpt 1). PB writing, however, provided little time for him to consider extensive planning.

Excerpt 1

Participant 13: … When it comes to computer-based test, because I think I can **write faster in my computer**, so **I can plan more here in the computer-based test**, whereas I have very little time when it comes to pencil-based tests.

5.1.2.2. The during-composing stage. A similar number of participants in both CB and PB tests reported generating new ideas and revising during composing processes. Participant 28 reported having to stop more frequently to check, revise, or leave room for future revisions during PB writing (Excerpt 2). It was a different process from her CB writing process, in which she focused on typing ideas first before revising. She perceived this writing strategy in the CB test as more similar to real-life academic writing. This perception may

Table 3

Spearman rho Correlations Coefficients between Computer Familiarity and Scores by Mode.

	CBT Holistic	CBT Task Achiev.	CBT Organization	CBT Coherence Cohesion	CBT Lexical Resources	CBT Grammar Accuracy	PBT Holistic	PBT Task Achiev.	PBT Organization	PBT Coherence Cohesion	PBT Lexical Resources	PBT Grammar Accuracy
Computer Availability												
Home	−.066	−.122	−.185	−.130	−.015	−.085	−.031	.174	.083	.112	.062	.046
School	.276*	.140	.089	.190	.251	.331**	.057	.152	.116	.130	.078	.097
Computer Usage												
Home	−.037	−.059	−.138	−.068	−.065	−.195	−.247	−.209	−.157	−.170	−.252	−.129
School	.262*	.271*	.199	.282*	.251	.191	−.035	.002	.087	.085	−.024	.191
Entertainment	−.092	−.047	−.072	−.040	.017	.029	−.151	−.094	−.228	−.096	−.078	−.101
Studying	.114	.125	−.009	.090	.138	.059	−.002	.148	.050	.205	.184	.217
Communication	−.066	−.141	−.022	−.120	−.112	−.076	−.129	−.179	−.084	−.176	−.218	−.160
Work	.129	.072	.040	.067	.105	.055	.008	.137	.112	.156	.138	.195
Learning Eng.	−.339**	−.256*	−.234	−.292*	−.130	−.209	−.310*	−.214	−.197	−.230	−.241	−.282*
Practicing Eng.	−.226	−.106	−.129	−.163	−.096	−.132	−.217	−.109	−.002	−.156	−.180	−.209
CBT frequency	−.182	.018	−.072	−.136	−.300*	−.307*	.031	−.027	.006	−.017	−.143	−.119
PBT frequency	.042	−.027	.087	.056	.153	−.031	−.277*	−.215	−.237	−.321*	−.248	−.231
Being Comfortable with …												
CB writing	.223	.248	.096	.127	.031	.043	.249	.258*	.205	.282*	.196	.185
PB writing	.011	.074	.045	.092	.139	−.020	−.042	.062	−.069	−.056	.009	.023
CB test	.026	−.027	−.052	.005	−.005	−.161	.108	.079	.110	.187	.115	.060
PB test	.103	.204	.121	.206	.098	.016	−.041	−.007	−.082	−.109	−.207	−.045
Habits and Attitudes												
Comp. is fun	−.077	.049	−.011	−.059	−.067	−.175	.186	.122	.199	.125	.152	.109
Reply on autocorrect	−.018	.019	.098	−.049	−.004	.038	.023	−.057	−.026	−.036	.074	−.117
Enjoy handwriting	.174	.147	.227	.215	.226	.143	.001	.060	.044	.028	−.005	−.049
Computer Skills												
Self-rated comp. skills	.098	.16	.002	−.018	.083	.114	.102	.062	.189	.122	.089	.066
Self-rated typing skills	.093	.131	.169	−.089	.110	.094	.141	.092	.126	.085	.057	−.051
Self-rated handwriting	.250	.219	.271*	.155	.100	.090	.133	.173	.170	.179	.176	.116
Typing Speed	.258*	.257*	.403**	.181	.426**	.387**	.159	.214	.146	.103	.118	.041
Typing Accuracy	.402**	.346**	.425**	.250	.314*	.346**	.293*	.284*	.232	.321*	.256*	.189

* $p < .05$, ** $p < .01$, *** $p < .001$.

[1]Typing Speed measured as number of correct words typed per minute.

[2]Typing Accuracy measured as the percentage of correct words among all typed words.

Table 4

Regression Results Predicting CB Essay Scores.

	B	S.E.	Beta	t	Sig
Constant	−6.507	3.145		−2.069	.043
Standardized Writing Scores (1-9)	.365**	.108	.366	3.362	.001
Typing Speed	.006	.007	.102	.883	.381
Typing Accuracy	11.357**	3.395	.390	3.345	.001

*$p < .05$, ** $p < .01$, *** $p < .001$.

explain the quantitative result that students' pause-making behaviors during CB writing yielded higher correspondence to students' real-life academic writing behaviors. Due to the limitations of major revisions in PB mode, the participant was forced to pause and revise during composing. This may be different from the writing behaviors in non-testing conditions where writers go through a process of drafting, revising, and publishing and have enough time to revise drafts and publish in a clean presentation.

Excerpt 2:

> Participant 28: (In CB writing), So, I just put my ideas on the computer and when I made up all contents, I started to revise on how I connected ideas. They if I understand or it is incorrect, I can revise it.

In handwriting, I cannot add so many pieces, so many words after I write it all, so when I write essay in handwriting, **I need to stop shortly after a paragraph** and I just think to, I tried to have margin to other words

5.1.2.3. The post-composing stage. More students engaged in reviewing and revising activities in the CB mode than in the PB mode. Ten test-takers revised the macrostructure (paragraphing) and/or microstructure (coherence, cohesions, transitions) in the CB test, as opposed to only four students who revised the microstructures in the PB test. Furthermore, 16 participants reported that they edited the essay in the CB mode, but only five students reported editing in the PB mode. In addition to the time-consuming aspects of revising handwritten essays (eight students), 15 students addressed their frustrations with revising the PB essays, such as making the essays "look messier". Difficulties with revisions in the PB test were a disadvantage for some participants as they could not follow their routine writing process of planning, writing, and revising. For example, Participant 59 complained that she was forced to only revise at the surface level because rearranging sentences was not possible in PB condition. It was a major disadvantage for her (Excerpt 3).

Excerpt 3

> Participant 59: Because **we could not go and erase a whole sentence** and **we could not rearrange sentences**. So, I didn't do all of that, because I didn't have much time, but yes, I look for spelling mistakes since it was a paper written test, You can't change, you can't change… i**t's a major disadvantage of writing in a paper than a computer.** Because we can go ahead and edit it, but that does not apply for paper tests.

5.2. Research Question2. What are the effects of test-takers' computer literacy on their CB and PB test essay scores?

5.2.1. Quantitative results

The second research question examined the *interactive* aspect of authenticity, and we focused on computer literacy variables. As shown in Table 2, computer availability at school and the frequency of using computers at school significantly and positively correlated with the CB essay holistic scores but did not correlate with the PB essay scores. Typing speed also significantly and positively correlated with the CB essay holistic scores and CB analytic scores in task achievement, organization, lexical resources, and grammar accuracy, but did not correlate with any measures of the PB scores. Typing accuracy, on the other hand, yielded significant correlations with the CB scores, but only weak correlations with some of the PB scores.

To further examine the relevance of typing proficiency in the L2 CB and PB writing assessment, we conducted two multiple regression analyses to examine the effects of test-takers' typing speed and accuracy on their CB and PB scores, controlling for their standardized writing test scores (participants' self-reported IELTS scores or converted IELTS scores from TOFEL iBT scores). The model predicting CB outcomes was statistically significant, $F(3, 55) = 10.309$, $p < .001$, and accounted for approximately 33.4 % of the variance of the CB scores ($R^2 = .360$, adjusted $R^2 = .325$) (See Table 3). Both the standardized writing scores and typing accuracy significantly predicted CB scores.

The model for PB outcomes was also significant, $F(3, 55) = 4.021$, $p = .012$, and accounted for approximately 17.7 % of the variance of the PB scores ($R^2 = .180$, adjusted $R^2 = .135$) (See Table 4). In contrast to the CB model, the standardized writing score was the only significant predictor in the model.

5.2.2. Qualitative results

The qualitative data corroborated the quantitative findings. Forty-five participants believed their CB writing performance was influenced by their computer skills. Thirty-three test-takers perceived this influence to be positive, while eight students thought it was negative. Some students shared that it was not easy for them to decide, as they were more familiar with typing in their L1, but not as fluent in English.

5.2.2.1. Typing speed and time. Thirteen out of the 60 test-takers attributed the positive influence of computer ability to the fact that

typing generated more time during their CB tests. For many test-takers, saving more time by typing is crucial when writing in a timed condition because it allows the writers to revise and improve the essay (see Excerpt 4). However, for test-takers who were not "good typers", the use of computers was perceived to have a reverse effect on their writing performance in the CB test. In Excerpt 5, Participant 30 was concerned that his typing speed was not as fast as "Americans", as the language being typed was not his L1.

Excerpt 4

Participant 3: Yes, absolutely because I had more time. I had more time to review and make a better essay.

Excerpt 5

Participant 30: Because **my typing speed is not that much faster than any other Americans do**.

5.2.2.2. More accustomed to computers. Four students commented that the positive effect of their computer skills on their writing performance was a result of being used to typing essays with computers. For example, Participant 59 commented that familiarity with the computer contributed to the writing performance because with a more routinized composing process, he could focus more on generating ideas.

Excerpt 6

Participant 59: Because since we are familiar with the format, **familiar with the functions or the abilities** that we could be doing when we are writing an essay in a Word document, how to format it, how to edit it and all, we have more ideas on what we are going to be doing.

5.2.2.3. Revise easily on a computer. Four students shared that their writing was better in CB because it was easier to revise on the computer. Participant 57 (excerpt 7) commented that the main advantage of CB writing was the easy and quick edits and revisions. However, this feature was not perceived positively by all test-takers. For example, in Excerpt 8, Participant 36 worried that the convenience of editing and revising had cost her deliberate planning and structuring before composing, which was a negative change on her writing process.

Excerpt 7

Participant 57: Oh, yes, it does, because you can, I mean, **easily edit** the content that you are writing, and **very quickly as** compared to the paper pencil does, so that is the main advantage, I think.

Excerpt 8

Participant 36: I think it's negative, because I don't do any structure before writing. So, I think **because we erase things quicker and faster, we don't think enough before writing.**

5.2.2.4. No influence. Nine students believed that there was no influence of computer skills on their writing performance. Four students reported to employing the same strategies for both CB and PB writing and engaging in the same writing process. As a result, they did not think there was any influence of the writing mode or computer skills. Interestingly, some students believed that there could be a "compensation effect" of the writing strategies in different writing conditions to obtain the same results. In Excerpt 9, Participant 31 shared that he had more time to revise with the CB writing, and it was easier to revise on a computer. However, this advantage "eventually took more from him" because he was not as concentrated during composing. As a result, there was more to revise on the CB essay. Eventually, the two essay products looked similar even though the writing processes were different.

Excerpt 9

Participant 31: When I write on the computer, I have the time, so I can make changes and that kind of thing actually enhances the writing to that level. **In handwriting, I put my full concentration and that's why when I revise the writing after handwriting, I see that there are not many changes …** I would say the revision process, no matter what I'm writing, the revision process is much easier in computer, but the revision time eventually takes more from me in computer because there is not a lot of revisions needed when I write using my hand, **a bigger number of revisions is needed when I write by my computer,** but still the revision process in computer is much easier, so that's why maybe altogether writing computer doesn't affect me. But eventually after revision, it is almost the same because I have revised the changes.

6. Discussion

This research followed Liu's model of authenticity to investigate the effect of testing mode on the authenticity of L2 writing assessments. Given the direct link between test authenticity and construct validity (Messick, 1994), examining test authenticity contributes to validity evidence in test interpretation and test use. Sixty participants took an ESL writing test in both CB and PB modes and reported their writing process using a cognitive questionnaire after each writing task. Finally, an interview was conducted to examine test-takers' perceptions of the tests.

6.1. Which mode is more authentic?

The results showed that both the CB and PB tasks replicated many features of real-life academic writing tasks in schools and hence

both were authentic to some degree. However, compared to the PB mode, more CB writing process items generated significant correlations with the self- reported TLU writing process items. The CB writing demonstrated a significantly higher resemblance to the writing processes in TLU tasks in *pause-making for revisions during writing*. Many participants reported having to modify their writing behaviors to compensate for the inconvenient revision processes in the PB test. Therefore, for the majority of the participants, the CB writing processes were more similar to their TLU writing processes. The results supported the claim that the authenticity of an ESL writing test may increase when switching to the CB mode from the traditional PB test mode (Weigle, 2002).

In addition, more participants planned ideas before composing in the CB mode, which contradicted previous research in which test-takers took more time in pre-planning in PB modes (Hoomanfard & Meshkat, 2015; Li, 2006). The discrepancies may be attributed to the concern about test time limits in the study. According to the interviews, some participants explained that they spent more time planning and revising because they knew their fast typing speed could allow them more time during the CB test. During the PB test, many students were concerned about running out of time and decided not to spend time on planning before writing. There were also more revision activities both at higher-level and in superficial editing in the CB writing. This particular finding aligned with recent literature by Cheung (2012), showing more superficial editing activities in CB mode. Overall, more writers in this study revised and edited their essays after composing in CB mode, corroborating the results in Li (2006).

Qualitative results further explained the differences in the writing processes in CB and PB tasks. First, more test-takers in this study were able to plan and revise more in the CB test because of a faster composing speed with the computer. Second, compared to handwriting, it was easier to revise with the computer when rearranging sentences, deleting, and reformatting. Therefore, some test-takers followed a process of brainstorming ideas before writing, transcribing ideas during writing, and revising the structure and coherence of the essay after writing. This writing process was perceived by many students to be more similar to their writing processes in daily academic writing. In PB mode, due to the difficulties with revising after completing the draft, many writers had to constantly stop to revise and check during composing, instead of revising after completing the draft. This finding differed from Hoomanfard and Meshkat's (2015) study in which writers involved more online revisions during writing in the CB mode. In this study, some test-takers considered the process of PB online revisions deviated from their daily writing routine, resulting in a lower perceived correspondence between the PBT process and their process in daily TLU tasks.

Taken together, the findings supported the claim that the CB task was more authentic than the PB writing task for this group of test-takers. Hence, the CB writing test results might yield a slightly higher level of validity when making inferences on ESL writing proficiency based on the test results. Nevertheless, both tests demonstrated correspondence to the TLU writing tasks to varying degrees.

6.2. Construct relevancy of computer literacy

Although both CB and PB writing tests purported to assess L2 English academic writing proficiency, CB writing tests may involve another learner characteristic – computer literacy. In the present study, after controlling for the standardized writing test scores, the typing accuracy scores significantly predicted the CB essay scores but did not predict the PB essay scores. This finding corroborated Russell's (1999) study in which the computer mode had a positive effect on students with higher keyboarding speed but a negative effect on students with low keyboarding speed. The results also aligned with Shirzad and Shirzad's (2017) study that the higher computer literacy group performed better than the lower computer literacy group in the TOEFL writing test but not in the PB practice test.

Qualitative results also revealed that most participants noticed an influence on their computer skills in the CB test. Familiarity in CB writing and proficient typing skills allowed participants more time for planning and revisions. This result supports the findings by Chan and colleagues (2017), who observed a high level of test-takers' computer literacy skills in the IELTS writing test. Participants in Chan et al. (2017) also reported taking less effort with pre-planning, being less careful with generating texts, and having more revisions at the higher level. Overall, test-takers tended to be in favor of the testing procedure that resembled their actual real-life academic writing situations in schools and had less anxiety over the time limit.

The findings indicated that the test-takers' computer literacy played a role in the ESL CB writing test, but not in the PB test. The CB and PB tests of writing actually measured two slightly different constructs. The key question is whether computer proficiency is a part of the English academic writing construct in higher education. On one hand, for domestic students who are already admitted to college, many U.S. higher education institutions offer computer literacy courses or require a computer competence placement exam. This practice can help ensure the computer literacy skills needed for students to succeed in U.S. colleges (Richards-Mealy, 2018). On the other hand, for international students, if CB academic English proficiency tests are biased against students with limited access to computers, these students may be denied a place in their dream universities due to their low socio-economic backgrounds.

6.3. A digital divide

With the development of technologies, many "digital natives" test-takers – students who were equipped with proficient computer literacy appeared to benefit from taking CB tests. Nevertheless, some participants reported a negative influence of their computer skills on CB tests due to their relatively slow typing speed. This result was consistent with Noubandegani (2012), who found that computer anxiety and keyboard skills were the key disadvantages of CB writing tests. As a result of a slow composing speed, these students did not have time to plan or revise on their CB writing tests. CB writing tests may thus disadvantage test-takers who are not computer literate.

These findings have uncovered a potential "digital divide" among test-takers. *Digital divide* refers to "the gap between individuals (and societies) that have the resources to participate in the information era and those that do not" (Chen & Wellman, 2004, p. 40), and

Table 5

Regression Results Predicting PB Essay Scores.

	B	S.E.	Beta	t	Sig
Constant	−1.224	3.721		−.329	.743
Standardized Writing Scores (1−9)	.340*	.128	.326	2.647	.011
Typing Speed	.006	.008	.100	.764	.448
Typing Accuracy	5.703	4.017	.187	1.420	.161

*$p < .05$, ** $p < .01$, *** $p < .001$.

implies inequalities in access to the digital space and its associated cultural, social, and economic capitals (Lee & Wang, 2019). While some cities, such as Hong Kong, have grown into a technologically advanced space (Lee & Wang, 2019), areas from other countries may still be in need of resources. In the case of a high-stakes ESL test for college admission purposes, a student's education opportunity in an English-speaking country may be denied due to the lack of access to digital resources. As a result, the inclusion of computer literacy in the test construct may negatively influence the consequential validity and fairness of the test. Nevertheless, for lower-stakes ESL tests such as mid-term exams and quarterly quizzes in an ESL writing course, including computer literacy in the test construct may help diagnose the learners' need for improvement in computer skills and motivate a positive change. Therefore, test developers and administrators should take into consideration the purpose of the assessment and its consequence when choosing a test mode for ESL writing tests.

7. Conclusion and future directions

To conclude, CB tests may be more authentic for ESL writing assessment because of a higher resemblance in the writing process between tests and real-life academic tasks. Overall, this study has shown that testing mode, as part of the performance conditions, indeed has an influence on the authenticity of ESL writing tests. However, the findings also uncovered the dilemma of whether switching from PB to CB writing improves the validity of ESL writing tests, and whether this change is fair for test-takers.

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. Given the normalization of technologies in language education (Bax, 2011), PB ESL writing tests may not resemble real-life academic writing tasks in schools for the majority of the ESL learners. However, 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 take account of all test-takers. 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. The validity of the test is threatened if the high-stakes test systematically biases against the students from lower SES backgrounds.

Second, when two validated forms of English proficiency tests are available, school admission personnel should keep in mind that the construct tested in CB and PB modes of the test may be different when interpreting the test results. Other than English language proficiency, the student candidates' computer literacy skills may also being tested by the CB test of writing.

Third, test purpose is another factor to consider when choosing the medium for ESL writing tests. For high-stakes English proficiency tests for college admission purposes, due to the potential bias of the CB writing test, a PB version of the test should always be an option available for test-takers who are not comfortable with the CB tests. On the other hand, for ESL program assessments and local in-house assessments for diagnostic purposes, the use of CB writing tests may inform learners about their computer literacy skills, motivating them to improve the digital competence necessary for future academic success. Since the CB test better resembled the academic tasks at the tertiary level in the U.S., administering diagnostic tests in the CB medium may help diagnose the students' computer competence to succeed in postsecondary education in the U.S (Table 5).

Since this is a small-scale study, the results cannot be over-generalized and studies with a larger sample size are needed. Additionally, we used a retrospective cognitive questionnaire to elicit the writing process data and a writing strategy questionnaire to collect information about participants' writing habits. Although retrospective questionnaires are frequently used and have demonstrated their validity in measuring writing processes (Weir et al., 2007), they limits participants' response to the items in the questionnaires. Future research should consider incorporating more objective observation measures, such as using online keystroke tools or concurrent think-alouds, to examine cognitive processes. Furthermore, we only examined the argumentative essay task in language proficiency tests for academic admissions. Other genres of academic compositions, such as expository writing, are also important in academic composition and should be studied in future research.

Funding

Declaration of Competing Interest

The study does not have potential conflict of interest.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.asw.2021.100548.

References

Bachman, L. F., Palmer, A. S., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world.* Oxford: Oxford University Press.

Barkaoui, K. (2015). Test takers' writing activities during the TOEFL iBT® writing tasks: A stimulated recall study. *ETS Research Report Series, 2015*(1), 1–42.

Barkaoui, K. (2016). What and when second- language learners revise when responding to timed writing tasks on the computer: The roles of task type, second language proficiency, and keyboarding skills. *The Modern Language Journal, 100*(1), 320–340.

Chan, S., Bax, S., & Weir, C. (2017). Researching participants taking IELTS Academic Writing Task 2 (AWT2) in paper mode and in computer mode in terms of score equivalence, cognitive validity and other factors. *IELTS research reports online series, No. 4.* Australia: British Council, Cambridge English Language Assessment and IDP: IELTS.

Chen, W., & Wellman, B. (2004). The global digital divide-within and between countries. *IT & Society, 1*(7), 39–45.

Cheung, Y. L. (2012). Critical review of recent studies investigating effects of word processing- assisted writing and pen-and-paper writing on the quality of writing and higher level revisions. *Procedia-Social and Behavioral Sciences, 46*, 1047–1050.

Cheung, Y. L. (2016). A comparative study of paper-and-pen versus computer-delivered assessment modes on students' writing quality: A Singapore study. *The Asia-Pacific Education Researcher, 25*(1), 23–33.

Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research* (pp. 53–106). Los Angeles, CA: Sage Publications.

Fulcher, G. (2010). *Practical language testing.* London, UK: Hodder Education.

Green, A. (2014). *Exploring language assessment and testing.* New York: Routledge.

*G*POWER. UCLA: Statistical Consulting Group (n.d.). https://stats.idre.ucla.edu/other/gpower/.

Hoerger, M. (2013). *ZH: An updated version of Steiger's Z and web-based calculator for testing the statistical significance of the difference between dependent correlations.* Retrieved from http://www.psychmike.com/dependent_correlations.php.

Hoomanfard, M. H., & Meshkat, M. (2015). Writing on a computer and using paper and pencil: Is there any difference in the internal cognitive processes? *GEMA Online® Journal of Language Studies, 15*(2).

Kohler, B. (2015). Paper-based or computer-based essay writing: Differences in performance and perception. *Linguistic Portfolios, 4*(1), 13.

Lee, Y. J. (2002). A comparison of composing processes and written products in timed-essay tests across paper-and-pencil and computer modes. *Assessing Writing, 8*(2), 135–157.

Lee, A. Y., & Wang, K. J. (2019). A perspective of digital use divide. In E. Morrell, & J. Rowsell (Eds.), *Stories from inequity to justice in literacy education: Confronting digital divides.* New York, NY: Routledge.

Lewkowicz, J. A. (2000). Authenticity in language testing: Some outstanding questions. *Language Testing, 17*(1), 43–64.

Li, J. (2006). The mediation of technology in ESL writing and its implications for writing assessment. *Assessing Writing, 11*(1), 5–21.

Liu, H. M. (2005). An investigation of methods for assessing authenticity in computer-assisted language learning and assessment. *Retrospective theses and dissertations* (p. 269). Iowa State University.

Lutz, J. (1987). A study of professional and experienced writers revising and editing at the computer with pen and paper. *Research in the Teaching of English, 21*, 398–421.

Messick, S. (1994). Standards-based score interpretation: Establishing valid grounds for valid inferences. *ETS Research Report Series, 1994*(2), 291–305.

Noubandegani, P. A. (2012). Students' perceptions of computerized TOEFL test. *Language Testing in Asia, 2*(2), 73.

Petrić, B., & Czárl, B. (2003). Validating a writing strategy questionnaire. *System, 31*(2), 187–215.

Richards-Mealy, L. (2018). *A quantitative study to analyze new student digital literacy knowledge and skills at a four-year institution.* Unpublished Doctoral dissertation. Northcentral University.

Russell, M. (1999). Testing on computers: A follow-up study comparing performance on computer and on paper. *Education Policy Analysis Archives, 7*(20), 1–47.

Saldaña, J. (2016). *The coding manual for qualitative researchers.* Washington, DC: Sage.

Sarbakhshian, B., & Saeidi, M. (2016). The comparison of typed and handwritten essays of Iranian EFL students in terms of length, spelling, and grammar. *The Journal of Applied Linguistics, 9*(19), 104–118.

Shaw, S. D., & Weir, C. J. (2007). *Examining writing: Research and practice in assessing second language writing* (Vol. 26). Cambridge University Press.

Shirzad, M., & Shirzad, H. (2017). The effect of computer literacy on the participants' writing ability in TOEFL iBT. *Theory and Practice in Language Studies, 7*(2), 134–139.

Understand and Explain the IELTS Scores. (n.d.). Retrieved from https://takeielts.britishcouncil.org/teach-ielts/test-information/scores-explained.

Weigle, S. (2002). *Assessing writing.* Cambridge University Press.

Weir, C., O'Sullivan, B., Jin, Y., & Bax, S. (2007). Does the computer make a difference? The reaction of candidates to a computer-based versus a traditional hand-written form of the IELTS writing component: Effects and impact. *IELTS Research Reports, 7*, 1–37.

Zou, X. L., & Chen, Y. M. (2016). Effects of test media on different EFL test-takers in writing scores and in the cognitive writing process. Technology. *Pedagogy and Education, 25*(1), 79–99.

Mingxia Zhi is a Program Evaluator at NorthSide Independent School District. She received her Ph.D. in Culture, Literacy, and Language at The University of Texas at San Antonio. Her research area includes language assessment, educational measurement, and program evaluation.

Becky H. Huang is Associate Professor in the Department of Bicultural-Bilingual Studies at the University of Texas in San Antonio. She received her PhD in Educational Psychology from the University of California, Los Angeles. Her research areas are language and literacy development, assessment and evaluation, and test bias and fairness.