

The Effect of Presentation Mode on Test Takers and Raters

Yuhua Liu^{1, 2}

¹ Center for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies, China

² Foreign Languages College, Jiangxi Normal University, China

Correspondence: Yuhua, Liu, Center for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies, Guangzhou, 510420, P. R. China. E-mail: 568955256@qq.com

The work was supported by the Postgraduate Research and Innovation Project of Guangdong University of Foreign Studies (21GWCXXM-067), China.

Received: February 23, 2023

Accepted: March 15, 2023

Online Published: March 24, 2023

doi: 10.5539/elt.v16n4p73

URL: <https://doi.org/10.5539/elt.v16n4p73>

Abstract

Technology development has led to computerization of language assessment. Many studies have centered on the comparability of test takers' cognitive processes and performance across the computer-based mode and the paper-based mode as well as the equivalence of raters' rating under the relevant modes. The effect of different modes on test-taking and essay-marking process and performance have been researched for decades. This paper critically reviews the effect of different modes on test takers' performance and raters' rating on the English writing part. The review indicates that there is no consensus on scores test takers receive under the two modes, or on the cognitive processes they experience. The computer familiarity contributes to their process and performance. On raters' rating, research has not reached a consensus either. The paper ends with discussing some issues worth consideration in future research of computerization in language test.

Keywords: comparability, delivery mode, computer-based, paper-based

1. Introduction

Significant technological development and increasing ease to computer access have led to computer-based language testing. Tests, especially the large-scale high-stakes test, have developed their computer-based testing system. For example, Cambridge English Language Assessment launched its first computer-based version of the International English Language Testing System (IELTS) in 2000. The Educational Testing Service (ETS) started its computer-based test (Internet-based Test of English as Foreign Language (TOEFL iBT) in 2005. Concurrently, the computer-based and paper-based versions coexist. Do the computer-based mode elicit the same language construct, especially for the writing assessment? Are test-takers' performance comparable between the two modes? Do raters score in the same way under different presentation modes? Research has been conducted on the effect of delivery mode on test-takers and raters when the two modes parallel or the computer-based version is to replace the paper-based version (Choi, Kim, & Boo, 2003).

This review has investigated the comparability studies published in recent three decades. The works reviewed are research reports and theoretical papers published in the international journals of applied linguistics. The aim was to present how researchers have explored the issue of computer use in writing assessment and compare their research findings.

The summary of the review is organized in two major sections as follows. The first section covers the effect of delivery mode on test-takers' process and performance. The second one focuses on relevant research on the effect of presentation mode on raters' marking.

2. The Effect of Delivery Mode on Test-takers

Previous research on the effect of delivery mode on test-takers can be majorly classified into three categories. There are the impact of delivery mode on test takers' performance, including scores and language features, the cognitive processes test takers will experience, the role of computer literacy in CB tests and test-takers' perception towards the modes.

2.1 The Effect of Delivery Mode on Test-takers' Performance

The research on the effect of delivery mode on writing scores has not reached a consensus. Studies have found test-takers can obtain non-significant overall scores under the two modes (Barkaoui & Knouzi, 2018; Brunfaut, Harding, & Batty, 2018; Chan, Bax, & Weir, 2018; Lee, 2002). Barkaoui & Knouzi (2018) asked 97 test-takers, with two levels of English language proficiency (high vs. low) and two levels of keyboarding skills (high vs. low) to write two equivalent independent writing tasks, on paper and on the computer. They found no significant differences on scores under the two modes. Unlike Barkaoui & Knouzi's (2018) and Lee's direct comparison of mean scores between two modes for the independent writing tasks (2018), Brunfaut et al. (2018) used multi-facet Rasch measurement (MFRM) method to analyze both independent and integrated writing tasks. They also found there was no significant difference on scores for both independent and integrated writing tasks. Chan et al. (2018) used MFRM and found that there was no significant difference on the holistic score. When analytic rating scale was adopted, significant difference was only found on one of the four analytic scores named Lexical Resources.

Some studies revealed higher scores (Li, 2006; Jin & Yan, 2017) under computer-based mode than under paper-based mode. Li (2006) found the advanced English second language (ESL) adult learners obtained higher scores under computer mode than paper mode on argumentative writing. Jin & Yan (2017) found that under computer mode, Chinese EFL college test-takers with moderate to high computer familiarity achieved higher scores under CB than PB; the test takers with low computer familiarity obtained a lower score under CB than PB. The computer familiarity took effect.

There are also studies suggesting higher scores under paper-based mode than the computer-based mode (Chen, White, McCloskey, & Chun, 2011; Yu, Livingston, Larkin, & Bonett, 2004). Chen et al.'s t-tests of adults' performance scores under the two modes suggested the advantage of paper mode over computer mode. Yu et al. (2004) found that the pre-service teachers performed better under paper-based mode than computer-based mode. Zou and Chen (2016) found that test takers with low computer familiarity had significantly lower scores under CB than PB, which suggests consideration of computer literacy when developing writing tests suitable for the computer mode.

Besides score comparability, studies also highlighted the comparison of language features under different modes. Kim et al. (2018) examined 26 examinees' essay quality in terms of linguistic (complexity, accuracy, fluency) and rhetorical features (integration of sources, progression of ideas, argument effectiveness) under the two modes in a counter-balanced order. They found no significant differences except that computer-based scripts were slightly longer. Barkaoui & Knouzi (2018) also found statistical significance across modes on fluency (the number of words per essay), lexical complexity, cohesion, and content. Jin & Yan (2017) investigated text complexity in terms of text length, sentence length, word length and difficulty level of the text, language errors in the essays written under the two modes. Results indicated that test-takers produced longer texts under the computer mode. Moreover, they made significantly fewer errors under the computer mode.

The previous research reveals no consistent conclusion on the effect of delivery mode on test takers' performance. Many confounding factors may contribute to the results: ESL or EFL, participants' demographic features, keyboard type, independent or integrated writing test. Further studies may take these factors into consideration. Meanwhile, some contextual factors should also be considered in the research.

2.2 The Effect of Delivery Mode on Test-takers' Cognitive Processes

Score Comparability, the reliability factor of writing test, is not sufficient to assure test validity for the computer-based mode (Chan et al., 2018). Recent literature has been highlighting the cognitive processes test-takers go through when writing under the two modes to decide whether the test activates the cognitive processes in the similar way to perform a writing assessment task (Shaw & Weir, 2007) under the two modes.

Lee (2002) adopted the retrospective interview to elicit six participants' recollection of their composing processes. The results show that individual participants will employ different processes depending on the writing prompt and delivery mode. The things in common were that all the participants tended to produce a rough writing script first and went back to expand it by adding sentences, even paragraphs on computer and they all went through invisible pressure to write any words continuously. Besides, they employed similar composing behaviors and strategies across the modes. On the time allotment, significantly more time was spent on planning on paper than on computer. One thing deserving attention is that the conclusions were made based on small samplings.

The research on cognitive processes between the two modes after Lee's (2002) study were most carried out quantitatively with the cognitive process questionnaire or a mixed design with cognitive process questionnaire and qualitative retrospective self-report protocols (Chamorro, 2022; Chan et al. 2018; Huang & Zhi, 2021; Weir, O'Sullivan, Jin, & Bax, 2007; Zou & Chen, 2016). Weir et al. found that test takers took similar cognitive processes when writing under the two modes. Zou & Chen (2016) found similarities as well as differences in the cognitive process among the test takers with different levels of computer familiarity. The similarities exist in the stages of goal-setting, generating ideas and reviewing. There were differences in organizing ideas and translating stages.

Chan et al.'s (2018) quantitative cognitive process questionnaire shows there were no significant differences in key writing processes used by participants under the two modes. However, the later qualitative interview data found some fine differences in the way test takers applied between the two modes. The differences in planning, generating texts, and monitoring and revising may influence test takers' performance.

Unlike Chan et al.'s study (2018) in the context of a high-stakes test, Chamorro (2022) set his study in the background of a classroom writing assessment. Chamorro also used a cognitive process questionnaire to collect data on what the processes the test-takers experience right after they finish the writing under the two modes. Thirty-eight participants, who were divided into two groups, took part in the research with a counter-balanced design. The findings confirmed the cognitive validity of both CB and PB tests, because both modes activated the key cognitive processes required to carry out a writing test. Meanwhile, each test mode had its own uniqueness. The CM mode triggered micro planning of text organization and after-writing revisions at both high and low levels. The PB mode activated more macro planning of content and text organization. No significant differences presented in task representation and translating processes.

The studies mentioned above conducted direct comparison on cognitive processes between CB and PB, Zhi & Huang (2018) employed the rationale test authenticity proposed by Bachman et al. (Bachman et al., 2010) to compare the corresponding items in the CB and PB test cognitive processing questionnaires with the writing process of target language use items measured in the pre-test survey. The quantitative results suggested significant but limited evidence for a higher degree of authenticity of the CB writing test.

To sum up, the literature indicates that the two modes may activate some different cognitive processes for the writing assessment while not altering the nature of writing test. The development on computer technology pose a question for test developers and administrators: which mode is more authentic to the target language use?

2.3 Test Takers' Computer Literacy & Perception towards the Computer- and Paper-based Modes

Research has found certain differences in test takers' performances in terms of scores and language features and their cognitive processes between CB and PB. Investigation was also done on the possible reasons for any inconsistency appearing in the studies. One major factor is the role of computer literacy. Individual differences such as computer familiarity, computer anxiety and computer attitude could be the potential factors to influence test takers' performance (McDonald, 2002).

Fulcher (1999) claimed that there was no significant effect of computer familiarity on language tests. In contrast, Taylor et al. discovered a significant, although small, effect of computer familiarity on computer-based TOEFL. The students with less computer familiarity got lower scores. Meanwhile, this effect could be eliminated through training. Some research found the role of keyboarding skills on test takers' performance. Russell (1999) revealed that students with low levels of keyboard skills would get lower scores when taking the computer-based test than the paper test and vice versa. Zhi & Huang (2021) found typing accuracy scores significantly predicted the CB essay scores but did not predict the PB essay scores. The qualitative results also indicated that most participants noticed the effect of their computer skills in the CB test. Familiarity in CB writing and proficient typing skills provided participants more time to plan their writing and make revisions. Jin & Yan (2017) found a high level of computer familiarity had a facilitative effect on test takers' performances. Barkaoui & Knouzi (2018) revealed that keyboarding skills had significant, but small, effects on measures of fluency, local cohesion.

On the perception towards the computer-based writing test, generally, the test takers who are familiar with computer use would tend to have a favorable attitude towards CB. Lee (2004) pointed out that habitual computer users prefer the computer-based modes. Barkaoui & Knouzi (2018) found a general preference for the computer-based mode across all test levels. Jin & Yan (2017) claimed that test takers with high level of computer familiarity had more positive perceptions towards their computer-based writing processes. Test takers with low to moderate levels of computer familiarity employed better planning strategies under PB.

3. The Effect of Presentation Mode on Raters

Apart from the deliver modes for test taking, raters also have different presentation modes for marking. According to Shaw (2008), we need to consider two presentation conditions, which are scoring mode and format. Test takers can submit their responses in handwritten or typewritten format. The handwritten scripts can be scored in paper version or scanned for computer-based marking. The word-processed responses can be scored under the computer mode, or printed out and scored in the paper version.

Currently, there are majorly three human rater relevant marking conditions – paper-based marking (PBM), on-screen marking (OSM), and online marking (OM). For PBM, raters score and comment on the writing scripts on paper. For a high-stake test, raters are asked to gather in an office to finish the marking task in order to assure the marking reliability and safety. On-screen marking (OSM) asks raters to mark the scanned scripts on the computer screen. Online marking means the marking of test takers' word-processed writing on a marking platform after test takers finish their writings through a word-processor program in a computer system.

3.1 *The Effect of Delivery Mode on Raters' Scoring*

In order to maintain test validity, we need to assure the comparability among the different rating conditions. On writing assessments, review indicates that previous studies on the effect of presentation mode on raters' scoring has not led to uniform conclusions.

Some studies found there were significant differences between modes. One group of studies found test takers would obtain higher scores under PBM than either of OSM or OM. Arnold et al. (1990) first found that essays in original handwritten version received higher scores than in the transcribed word-processed form. Powers et al. (1994) transcribed handwritten writing samples to word-processed form and word-processed form to handwritten version. Results revealed that raters opted for awarding higher average scores for essays in the handwritten mode than for word-processed essays despite the mode in which they were initially written. Modified training could reduce this presentation mode effect. Russell & Tao (2004) replicated and extended the work of Powers et al. (1994) and made a similar finding. Various factors were proposed to explain the advantage of the PBM over OM. The visibility of errors and higher expectations for computer-printed responses may have caused the disadvantage of OM over PBM. Rater training could remove the effect (Russell & Tao, 2004). In contrast, raters might have a "reader empathy assessment discrepancy effect" when rating under PBM (Powers et al., 1994). Raters tended to show empathy with the authors of handwritten essays and thus give higher scores to them.

The other group of studies suggests the test takers would get higher scores under the computer-based modes compared with the paper-based mode. Peacock's study (1988) found that computer-assisted presentation methods significantly enhanced written responses' apparent quality compared to handwritten work. The relatively lower quality work tended to be improved more than somewhat higher quality work. Canz, Hoffmann & Kania (2020) explored the effects of rating media on scoring by mediating various factors, text quality, legibility, orthographic, grammatical punctuation erroneousness, and text genres. Raters were asked to rate two versions (originally handwritten and computer-typed transcripts) of 430 essays in German. The research found the main effect of presentation modes and concluded that raters gave higher scores to the computer-typed transcripts. The effect would be moderated by text quality, grammatical erroneousness, and genre. Canz et al. (2020) claimed contextualization could help explain why raters awarded lower scores in the paper-based mode. They contended that raters showed a deficit orientation for essays written by students under PBM as it was a high-stake test. Two points deserve attention: the raters marked the essays first under PBM and then OM, there was no counter-balanced research design. Moreover, they compared data under PBM from 19 raters with data from 5 raters under OM. The results may have been contaminated by learning effect and rater effect. Other studies found the effect of computer familiarity (Hughes & Akbar, 2010), the Pygmalion effect (Sprouse & Webb, 1994) and rater effect (He, 2019).

The third group of studies indicated the comparability across the modes. Zhang et al. (2003) found no significant statistical difference in internal reliability or inter-reader agreement between PBM and OSM. Raikes et al. (2004) found evidence to suggest that examiners' on-screen marking of short answer scripts was reliable and comparable to their marking of the paper originals. Meanwhile, they also indicated that more research was needed, particularly concerning extended responses, to ascertain under what circumstances rating with OSM was valid and reliable. Shaw (2008) employed a mixture of quantitative and qualitative methods to compare the effect of presentation mode on raters' judgment of participants' extended responses. The study found high inter-rater reliability on paper and screen, although lower on screen than on paper. We need to keep it in mind that the comparable inter-rater reliability between modes cannot signify the comparability of ratings. Coniam (2009, 2010) adopted classical and Rasch measurements to explore the possible difference in rating between

PBM and OSM. The two studies found that inter-rater reliability between the two marking media was comparable, and the marking medium would not affect scores awarded to test takers. Johnson et al. (2009) found no significant effect on the marking while variation in raters' screen marking behavior between OSM and PBM. The demerit of this study is that the marking accuracy was operationalized as scores awarded by the examination's Principal Examiner for each essay on paper. Johnson et al. (2009) and Shaw (2008) claimed that the raters considered the writing construct similarly under different modes, although there were differences in mental workload, spatial encoding, navigation, and annotation between different presentation modes.

The studies surveyed above demonstrated several points worth attention. Attention has majorly been paid to the reliability issue in terms of mean score comparison and inter-rater agreement. Proper attention should be assigned to the more important validity issue. For the reliability issue, MFRM is a good way to explore raters' severity, or use of various scales etc compared with the traditional method. Furthermore, the three modes can be compared together to get a clear understanding of mode effect. Besides reliability issue, qualitative methods, such as think-aloud, retrospective recall are among the list to figure out the possible reasons from the perspective of validity to explain patterns found in the quantitative research.

3.2 Rater Behavior under Different Rating Modes

Investigation into raters' behavior is in great need to find an answer to the inconsistent conclusions on scoring comparability. The possible inhibition of on-screen assessment on reading comprehension leads to studies on the influence of presentation mode on raters' judgement on constructed written performances (Johnson et al., 2009).

Previous research adopted interview to investigate raters' behavior under different modes. Whithaus et al.'s study (2008) found OM make raters noticed more spelling and other surface errors. However, the discovery of the errors did not lead to any overall score loss. The comparison on the navigation and annotation behaviors between PBM and OSM indicated that raters tended to navigate more iteratively under PBM. The iterative navigation ease the rough estimation of the writing quality. Johnson et al. (2012) suggested that Raters also tended to annotate more often under PBM. The different practices in navigation and annotation may influence raters' comprehension while marking, and sequentially influence their marking.

4. Conclusion

The above review shows that no complete consensus on the effect of delivery mode on test takers' process and performance and raters' marking. On the effect of delivery mode on test takers, literature indicates significant or no significant score difference. Test takers may go through somewhat different processes. Computer familiarity may play a role in the process or performance. From the perspective of rater marking, raters may exhibit different marking behavior and award different scores to the same essay under different modes.

Resolving the above-mentioned consensus from the perspectives of test taker and raters calls for attention from researchers, test designers and administrators.

On the delivery mode, since many studies observed the effect of computer on test takers' process and performance, more research is need to figure out to what extent it may influence test taker. For example, how can the keyboarding skills disrupt the test taker's cognitive process? What kind of computer literacy is needed to cater for the computer-based assessment? Do the writing type (independent vs. integrated), English language proficiency take effect. Test administrators should be informed with the factors which would affect test takers' process and performance in order to ensure the computer-based writing assessments. They need to consider the issues of screen size, keyboard types etc. in order not to invite any construct-irrelevant variance. Meanwhile, they should also consider the test takers who are not used to computer-based assessments. Choices should be given on computer-based or paper-based modes.

On the presentation mode, more attention should be paid to raters' scoring process to figure out any patterns appearing for raters under different modes. On the score analysis, MFRM is a good choice to consider test taker ability, rater severity, writing prompt difficulty, mode difficulty etc. comprehensively. For rater training, raters need to have a clear understanding of the writing construct. They need to be told to separate writing quality from test takers' handwriting, traces of revision. They also need to know the perceived shorter length of word-processed essays than handwritten essays.

References

- Arnold, V., Legas, J., Obler, S., Pacheco, A., Russell, C., & Umbdenstock, L. (1990). *Do students get higher scores on their word-processed paper? A study of bias in scoring handwritten vs. word-processed papers*. The Educational Resources Center. Whitter, CA Rio Hondo College.
- 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., & Knouzi, I. (2018). The effects of writing mode and computer ability on L2 test-takers' essay characteristics and scores. *Assessing Writing*, 36, 19-31. <https://doi.org/10.1016/j.asw.2018.02.005>
- Bennett, R. E. (2003). *Online Assessment and the Comparability of Score Meaning (ETS RM-03-05)*. Princeton, NJ: Educational Testing Service.
- Brown, C. (1991). Computers and Assessment: The Effect of Typing Versus Handwriting on the Holistic Scoring of Essays. *Research and Teaching in Developmental Education*, 8(1), 5-14.
- Brunfaut, T., Harding, L., & Batty, A. O. (2018). Going online: The effect of mode of delivery on performances and perceptions on an English L2 writing test suite. *Assessing Writing*, 36, 3-18. <https://doi.org/10.1016/j.asw.2018.02.003>
- Canz, T., Hoffmann, L., & Kania, R. (2020). Presentation-mode effects in large-scale writing assessments. *Language Assessment Quarterly*, 18(2), 83-106. <https://doi.org/10.1080/15434303.2020.1799222>
- Chamorro, M. E. G. (2022). Cognitive validity evidence of computer-and paper-based writing tests and differences in the impact on EFL test-takers in classroom assessment. *Assessing Writing*, 51, 100594. <https://doi.org/10.1016/j.asw.2021.100594>
- Chan, S., Bax, S., & Weir, C. (2018). Researching the comparability of paper-based and computer-based delivery in a high-stakes writing test. *Assessing Writing*, 36, 32-48. <https://doi.org/10.1016/j.asw.2018.03.008>
- Chen, J., White, S., McCloskey, M., Soroui, J., & Chun, Y. (2011). Effects of computer versus paper administration of an adult functional writing assessment. *Assessing Writing*, 16(1), 49-71. <https://doi.org/10.1016/j.asw.2010.11.001>
- Coniam, D. (2009) Marking essays on screen: An investigation into the reliability of marking extended subjective texts. *British Journal of Educational Technology*, 41(5), 814-826. <https://doi.org/10.1111/j.1467-8535.2009.00979.x>
- Coniam, D. (2010). A comparison of on-screen and paper-based marking in the Hong Kong public examination system. *Educational Research and Evaluation: An International Journal on Theory and Practice*, 15(3), 242-263. <https://doi.org/10.1111/j.1467-8535.2009.00979.x>
- Choi, I.-C., Kim, K. S., & Boo, J. (2003). Comparability of a paper-based language test and a computer-based language test. *Language Testing*, 20(3), 295-320. <https://doi.org/10.1191/0265532203lt258oa>
- Fulcher, G. (1999). Computerizing an English language placement test. *ELT Journal*, 53, 289-299. <https://doi.org/10.1093/elt/53.4.289>
- Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, 23(1), 17-27. <https://doi.org/10.1111/j.1745-3992.2004.tb00149.x>
- He, T. H. (2019). The impact of computers on marking behaviors and assessment: A many-facet Rasch measurement analysis of essays by EFL college students. *SAGE Open*, 9(2), 2158244019846692. <https://doi.org/10.1177/2158244019846692>
- Hughes, J., & Akbar, S. (2010). *The influence of presentation upon examination marks*. 11th Annual Conference of the Subject Centre for Information and Computer Sciences.
- Jin, Y., & Yan, M. (2017). Computer literacy and the construct validity of a high-stakes computer-based writing assessment. *Language Assessment Quarterly*, 14(2), 101-119. <https://doi.org/10.1080/15434303.2016.1261293>
- Johnson, M., Hopkin, R., & Shiell, H. (2012). Marking extended essays on screen: Exploring the link between marking processes and comprehension. *E-Learning and Digital Media*, 9(1), 50-68. <https://doi.org/10.2304/elea.2012.9.1.50>

- Johnson, M., Nadas, R., & Bell, J. (2010). Marking essays on screen: An investigation into the reliability of marking extended subjective texts. *British Journal of Educational Technology*, 41(5), 814-826. <https://doi.org/10.1111/j.1467-8535.2009.00979.x>
- Johnson, M., Nádas, R., & Shiell, H. (2009). An investigation into marker reliability and other qualitative aspects of on-screen essay marking Martin. *Paper presented at the British Educational Research Association annual conference*. Manchester University.
- Kim, H. R., Bowles, M., Yan, X., & Chung, S. J. (2018). Examining the comparability between paper-and computer-based versions of an integrated writing placement test. *Assessing Writing*, 36, 49-62. <https://doi.org/10.1016/j.asw.2018.03.006>
- 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. [https://doi.org/10.1016/S1075-2935\(03\)00003-5](https://doi.org/10.1016/S1075-2935(03)00003-5)
- Li, J. (2006). The mediation of technology in ESL writing and its implications for writing assessment. *Assessing Writing*, 11(1), 5-21. <https://doi.org/10.1016/j.asw.2005.09.001>
- McDonald, A. S. (2002). The impact of individual differences on the equivalence of computer-based and paper-and-pencil educational assessments. *Computers & Education*, 39, 299-312. [https://doi.org/10.1016/S0360-1315\(02\)00032-5](https://doi.org/10.1016/S0360-1315(02)00032-5)
- Peacock, M. (1988). Handwriting versus word processed print: An investigation into teachers' grading of English language and literature essay work at 16+. *Journal of Computer Assisted Learning*, 4, 162-172. <https://doi.org/10.1111/j.1365-2729.1988.tb00173.x>
- Powers, D., Fowles, M., Farnum, M., & Ramsey, P. (1994). Will they think less of my handwritten essay if others word process theirs? Effects on essay scores of intermingling handwritten and word-processed essays. *Journal of Educational Measurement*, 31(3), 220-233. <https://doi.org/10.1111/j.1745-3984.1994.tb00444.x>
- Raikes, N., Greateorex, J., & Shaw, S. (2004). *From Paper to Screen: some issues on the way*. Paper presented at the IAEA Conference, Vilamoura, Portugal.
- 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. <https://doi.org/10.14507/epaa.v7n20.1999>
- Russell, M., & Tao, W. (2004). Effects of Handwriting and Computer-Print on Composition Scores: A Follow-up to Powers, Fowles, Farnum, & Ramsey. *Practical Assessment, Research, and Evaluation*, 9, Article 1. <https://doi.org/10.7275/9g7k-yr32>
- Shaw, S. (2008). Essay Marking On-Screen: implications for assessment validity. *E-learning*, 5, 256-274. <https://doi.org/10.2304/elea.2008.5.3.256>
- Shaw, S., & Weir, C. (2007). Examining writing: Research and practice in assessing second language writing. *Studies in language testing*, Vol. 26. Cambridge: Cambridge University Press.
- Shiell, H., Johnson, M., Hopkin, R., Nadas, R., & Bell, J. (2011). Extended essay marking on screen: Does marking mode influence marking outcomes and processes? *Research Matters: A Cambridge Assessment publication*, 11, 2-7. <https://doi.org/10.1080/13803611.2012.659932>
- Sprouse, J., & Webb J. (1994). *The Pygmalion Effect and Its Influence on the Grading and Gender Assignment on Spelling and Essay Assessments*. Retrieved from <http://www.eric.ed.gov>.
- 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.
- Whithaus, C., Harrison, S., & Midyette, J. (2008). Keyboarding compared with handwriting on a high-stakes writing assessment: Student choice of composing medium, raters' perceptions, and text quality. *Assessing writing*, 13(1), 4-25. <https://doi.org/10.1016/j.asw.2008.03.001>
- Yu, L., Livingston, S. A., Larkin, K. C., & Bonett, J. (2004). Investigating differences in examinee performance between computer-based and handwritten essays. *ETS Research Report RR-04-18*. Princeton, NJ: Educational Testing Service. <https://doi.org/10.1002/j.2333-8504.2004.tb01945.x>

- Zhang, Y., Powers, D., Wright, W., & Morgan, R. (2003). *Applying the online scoring network (OSN) to advanced program placement program (AP) Tests (ETS Research Report RR-03-12)*. Princeton, NJ: Educational Testing Service. <https://doi.org/10.1002/j.2333-8504.2003.tb01904.x>
- Zhi, M., & Huang, B. (2021). Investigating the authenticity of computer-and paper-based ESL writing tests. *Assessing Writing*, 50, 100548. <https://doi.org/10.1016/j.asw.2021.100548>
- 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. <https://doi.org/10.1080/1475939X.2014.954140>

Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/4.0/>).