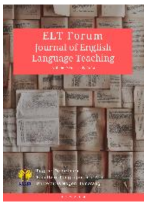
ELT FORUM 12(1) (2023)

Journal of English Language Teaching

http://journal.unnes.ac.id/sju/index.php/elt



A comparability study of handwritten versus typed responses in high-stakes English language writing tests

Irene Stoukou[1], Yiannis Papargyris[1], David Coniam✉[1]

[1]PeopleCert, Athens, Greece

Article Info

Article History:
Received in 15 February 2023
Approved on 30 March 2023
Published on 31 March 2023

Keywords: handwritten scripts; computer-processed scripts; effect size; equivalence t-tests

Abstract

This paper investigates fairness in writing test scores in terms of candidates who completed a writing test either by hand or typed, on a computer. The data for this large-scale comparability study comprise candidates taking English language writing tests at four CEFR levels – B1 to C2 in the period 2019–2022. The data were analysed via effect size differences and equivalence tests. Measured by effect size, a small amount of difference was apparent in scores obtained between the two production modes at B1, B2 and C1 levels. At C2 level, there was a medium effect size, indicative of a difference in favour of computer-produced scripts. Differences observed on equivalence tests – an adaptation of the standard t-test – were not found to be statistically significant. The contribution of the research to knowledge lies in the fact that (with the exception of C2 level) – whether writing tests are written by hand or on computer, while there is a slight skew towards higher scores with computer-processed texts, candidates generally receive similar scores in both modes. Practically, candidates may elect to write either on paper or on computer without fear of bias.

✉Correspondence Address:
3 Korai Street, Athens 10564, Greece
E-mail: david.coniam@peoplecert.org

p-ISSN 2252-6706 | e-ISSN 2721-4532

INTRODUCTION

There is a substantial literature on score equivalence obtained from handwritten (HW) and computer-processed (CP) scripts. Indeed, research into score equivalence between handwritten and computer-processed scripts stretches back to the 1960s when the word-processing of scripts first began. To put these issues into perspective, the current section first presents an overview of the research – which presents provide contrasting results. The section concludes with the research gap being explored in the current research.

While some studies have revealed better performance by candidates writing by hand; others have reported the opposite, with higher CP scores; and, in contrast, no significance has been found for either mode of delivery in other studies. A review of the research from the different angles is presented below.

Some of the earliest research was by Marshall and Powers (1969), in whose study neat handwritten essays scored higher than typed ones. Mazzeo and Harvey's (1988) study of handwritten and computer-processed scripts indicated better performance in HW mode, which they attributed, understandably at the time, to lack of familiarity with the technology.

Arnold et al. (1990) reported computer-processed scripts receiving lower scores than handwritten scripts. Sweedler-Brown (1991) reported likewise, although only with lower ability scripts. In Powers et al.'s (1994) and Russell and Tao's (2004) studies, students' HW scripts scored higher than the same students' comparable CP scripts. Bridgeman and Cooper (1998) in a study involving Graduate Management Admissions Test scores reported higher scores with HW than with CP scripts. Klein and Taub (2005) reported a teacher bias for legible HW scripts. In Breland et al.'s (2005) study of TOEFL candidates, HW scores, related to general English language ability, were reported.

While numerous studies have reported handwriting-based scripts to have received higher scores, there have also been many studies reporting computer-processed-based scripts to have received higher scores. Some studies showing such advantage are outlined below. An overall advantage for CP texts has been reported in certain studies (Sprouse and Webb, 1994; Peacock, 1988; Hughes and Akbar, 2010). On the issue of quality, Peacock (1988) reported an advantage for low-quality CP scripts. Peacock (1988) also reported an advantage regarding text type for CP essays where the essays were not related to external sources. In Canz et al.'s large-scale (2020) study, CP scripts received higher grades despite raters being highly trained raters. Russell and Plati (2000) reported lower secondary school students performing better under CP conditions. In Goldberg et al.'s (2003) meta-analysis of 26 writing studies of K-12 students writing in CP or HW modes, results indicated higher text quality for the CP scripts. Other confirmatory studies for students achieving higher grades in CP mode include Russell and Haney (1997) and Russell and Plati (2001).

In addition to studies citing an advantage for either mode, here have also been studies where neither mode has been reported as conferring an advantage, as outlined below. While positive findings have been reported for both modes, a number of studies have reported no significant difference in terms of grade received in either CP or HW mode. Among these are: Wise and Plake, 1989; Wright and Linacre, 1994; Taylor et al., 1999; Russell, 1999; MacCann et al., 2002; Horkay et al., 2006; Boulet et al, 2007; King et al., 2008; Mogey et al., 2010; Chan et al., 2018. As may be seen from the studies reported above, there is evidence for all positions: that under certain conditions CP scripts receive higher scores; under others that HW scripts score higher, with many studies also reporting no significant difference between modes.

Differences notwithstanding, it is nonetheless the case that with improvements in technology in terms of usability, speed and lower cost (see Lim and Wang, 2016), the use of a computer to produce essays in a variety of situations – classwork, homework and examinations – is increasing. Indeed, with the recent COVID-19 pandemic, greater acceptance has been observed of the use of computers and technology (Hodges et al., 2020).

In light of the above, it is worth considering the question of whether the ability or preference to use a computer in an examination is related to age. Older candidates do not necessarily opt for CB tests as such; it is simply the route they follow which leads them to an online-proctored environment (i.e., navigating the internet, selecting an exam provider online, registering, booking a slot and managing their time etc.). Against this backdrop, for more mature candidates, the CB component is simply part of the overall context.

Over the past three years, that is, during the period of COVID in 2019–2022, many examination bodies experienced exponential increases in online-administered examinations (see e.g., Ockey, 2021). LanguageCert English language tests are available in either traditional centre-based or online proctored (OLP) delivery modes (Coniam et al., 2021). During the COVID pandemic, LanguageCert saw a great increase in its OLP mode of delivery, and a concomitant increase in writing tests produced by computer as opposed to being handwritten. While the research outlined above has presented different perspectives on the two modes of delivery – computer-processed versus handwritten – and how the mode might confer an advantage on scores, little research has been conducted in the past few years – and certainly not in the context of the huge increase in computer-processed writing test scripts against the backdrop of COVID.

This is therefore the research gap that the current study fills in the context of computer-processed versus handwriting writing test scripts. Using comparatively large writing test datasets (a considerable number of which had been administered during the COVID pandemic period) at differing CEFR levels of ability, the study explores to what extent the mode of script production impacts on candidate score.

METHODS

This section outlines the study in terms of research design, the data and data analysis. Background to the LanguageCert Writing Test is first presented to situate the study.

The IESOL writing test

The data in the study come from four examinations – at CEFR levels B1–C2, which form part of the International ESOL (IESOL) suite of English language tests. The Writing Tests comprise two different writing tasks tapping a range of writing skills. Table 1 elaborates.

Table 1. IESOL writing test tasks

Level	Part 1 : Candidates produce	Word length	Part 2 : Candidates produce	Word length
B1	a neutral or formal text for a public audience	70-100	a letter using informal language	100-120
B2	a neutral or formal text for a public audience	100-150	a text using informal language	150-200
C1	a neutral or formal text for a public audience	150-200	a text using informal language	250-300
C2	a neutral or formal text for a public audience	200-250	a text using informal language	250-300

Each task is scored on four levels (0-3) against four subscales which for the most part are double-marked before final scores are amended or confirmed and signed off by a more senior member of the assessment team, usually a chief examiner. (Refer to LanguageCert).

Candidates may take the examination either at a physical centre or by online-proctored mode. If they take the examination at a centre, they generally handwrite. While it is possible to do a computer-based test at a physical centre, this option is not very popular; most candidates handwrite tests at centres. When tests are taken online, a locked-down computer is used. It should be noted that the term 'computer-processed' is used in the current paper to indicate that candidates write on a 'bare-bones' computer; they do not have access to a word processor or any of the more advanced facilities such as grammar/spellchecking that a word processor offers.

All Writing Test markers hold professional accredited English language qualifications and experience as English language teachers. All prospective markers undergo a standardisation and training programme before being certified as qualified markers (for details, see Papargyris and Yan, 2022). The training programme involves marking sample scripts and prospective markers must demonstrate they can mark accurately and consistently before they are certificated as markers. Checking takes place by a group of chief examiners during live marking, and if markers are suspected of marking inaccurately and/or inconsistently, they may be removed from the marking session and/or retrained or even dismissed. Markers are monitored on an ongoing basis as well as attending standardisation sessions, again on a regular basis. LanguageCert markers mark across CEFR levels (Papargyris and Yan, 2022). At any one time, there may well be in the region of 200 markers marking different numbers of scripts at the different CEFR levels. While the scope of the current study does not involve an examination of marker performance, the reader is referred to Coniam et al. (2022) where an exploration using Many-Facet Rasch Analysis into marker performance can be found.

The current study

This section presents details on candidates' scores against the two modes of script production. Table 2 below provides detail on the number of candidates at each CEFR level for each mode. The data collection period extended over the three-year period from mid 2019 to mid 2022. Although not germane to the current study, it should be noted that the current study involved 143 different markers.

Table 2. Candidate sample sizes

Level	Mode	N	Level sample
B1	CP	3108	22727
	HW	19619	
B2	CP	14878	27590
	HW	12712	
C1	CP	7674	10330
	HW	2656	
C2	CP	2869	4363
	HW	1494	

Legend: CP=computer-processed; HW=Handwritten

At B1 level, the candidature comprises many school students. It is therefore not perhaps surprising that the majority of scripts at this level were handwritten. As one moves up the level, and demands of certification for study, work, immigration purposes come more to the fore, candidates tend to be slightly older and more computer literate. More online-proctored (OLP) examinations take place at this level, a situation exacerbated by COVID, and support for why computer-processed (CP) scripts outnumber handwritten (HW) scripts at B2-C2.

Hypotheses

The hypothesis pursued in the study is that scores awarded to either of the two modes of script production – computer-processed or handwritten – will not be significantly different. Three sub-hypotheses are pursued:

1. The difference between the mean scores for the two written script modes will be less than 5% for any given CEFR level.
2. Only small effect size differences will be noted between the two modes.
3. On equivalence tests, significance will not emerge against specified upper and lower bounds for any CEFR level.

FINDINGS AND DISCUSSION

Two sets of data for the Writing Test are presented. The first set of analyses contains descriptive statistics: means (maximum 25) for the two modes, standard deviations and effect size differences. The second set of analyses consists of equivalence independent samples t-tests ("equivalence tests"). Equivalence tests – as opposed to regular t-tests – permit for significance to be explored by specified upper and lower bounds (Lakens, 2017). The two bounds define the extent of variation of t values with respect to the populations being tested. If the t value falls within the estimated range, the two populations may be seen to be equivalent.

Descriptive statistics

Descriptive statistic results are in provided in Table 3 for the two types of writing for the four CEFR levels. The final two right-hand columns contain detail on score and effect size differences between the two modes. Effect size differences are reported in terms of Cohen's d, for which a small effect is generally 0.2, a medium effect 0.5, and a large effect 0.8 (Glen, 2021).

Table 3. Writing mode descriptives

Level	Mode	N	Mean	SD	Raw score (%) difference	Effect size differences (Cohens's d)
B1	CP	3108	18.75	4.63	0.80 (3.20%)	0.17
	HW	19619	17.95	4.72		
B2	CP	14878	18.85	4.68	0.80 (3.21%)	0.17
	HW	12712	18.04	4.67		
C1	CP	7674	17.60	4.80	1.13 (4.53%)	0.23
	HW	2656	16.46	4.86		
C2	CP	2869	18.13	4.77	2.57 (10.28%)	0.55
	HW	1494	15.56	4.46		

Legend: CP=computer-processed; HW=Handwritten

Effect sizes reported are negligible for B1 and B2 levels, with only a small effect size at C1. At C2 level, however, the score difference is larger than 5%, and a medium effect size of 0.55 is reported. The implications of this are that C2 level candidates, who produce their Writing Test scripts on computer, score comparatively higher than C2 candidates who handwrite their tests.

Equivalence tests

Table 4 below presents equivalence test results comparing handwritten (HW) and computer-processed (CP) script production modes. Upper and lower bounds have been set at +/- 0.05 of the raw score (see Lakens, 2017). As mentioned, critical decisions regarding equivalence revolve around whether estimated t values are between the upper and lower bound. In Table 4 below, p values indicate significance with respect to upper and lower bound t values going beyond specified bounds.

Table 4. Equivalence samples t-tests

Test Level	Statistic	t	df	p
	upper bound	9.36	22725	< .001
B1	t value	8.81	22725	< .001
	lower bound	8.26	22725	1.00
	upper bound	15.12	27588	1.00
B2	t value	14.23	27588	< .001
	lower bound	13.34	27588	< .001
	upper bound	9.99	10328	1.00
C1	t value	10.45	10328	< .001
	lower bound	10.91	10328	< .001
	upper bound	16.92	4361	1.00
C2	t value	17.26	4361	< .001
	lower bound	17.59	4361	< .001

At none of the four levels was significance observed at both lower and upper bounds. This is an indication that the two writing script modes may be seen to be equivalent for the four CEFR levels examined in the study.

Discussion

The results above provide a consistent picture: at all levels, candidates who produced computer-processed scripts scored higher than did candidates who produced handwritten scripts. This finding echoes the study by Goldberg et al. (2003) who analysed studies of students writing in CP or HW modes, with results indicating CP scripts being rated more highly. As Lim and Wang (2016) report, the use of a computer to produce essays in many school situations is increasing. It may simply be the case that such increasing use of the computer results in a vicious, or virtual, cycle (depending on one’s point of view), whereby writing on computer becomes the norm and the mode to which people, examination candidates included, are simply becoming more accustomed.

The results for higher scores obtained on computers may be due to a number of factors. One consistent feature mentioned by LanguageCert markers in post-marking reports is that of the legibility (or lack of it) encountered in many handwritten scripts. Be that as it may, the main issue is that at CEFR levels B1-C1, the difference between the two modes is less than 5%, a figure generally taken as being indicative of significance.

What then might be the possible reasons for candidates using a computer to produce their script – in particular at the higher CEFR levels – to obtain comparatively higher scores? One possible explanation may be found in the candidates' background. In a survey (in mid-2022) of over 40 LanguageCert Writing Test markers, markers noted that, at the CEFR A and B levels, there were more younger candidates. These younger candidates were more used to writing on paper than using a computer. More proficient candidates – in particular those at C2 – were noted by some markers as being older and more computer literate. Markers perceived these two factors as helping to account for the skew towards higher scores achieved on computer-processed scripts.

CONCLUSION

This study has reported on comparability of scores awarded to candidates who completed Writing Tests either through handwriting or by using a computer at CEFR levels B1 to C2.

The key hypothesis in the study was that mean scores and performance on the Writing Test in either mode would not be significantly different from each other; i.e., that candidate scores would not be influenced by the writing mode. Specifically, three hypotheses were being investigated.

The first hypothesis was that differences between the mean scores for the two modes of test production would be less than 5% for any given CEFR level. This was the case for levels B1, B2 and C1. It was not the case for C2 where differences were greater than 5%. While the hypothesis was confirmed for B1, B2 and C1, it was rejected for C2.

The second hypothesis was that, at worst, only small effect sizes would be reported between the two writing modes. This was indeed the case with B1, B2 and C1. At the C2 level, however, a medium effect size was observed, causing the hypothesis to be rejected.

The third hypothesis was that, for any given CEFR level, significance between upper and lower bounds would not be observed on equivalence t-tests. Significance was not observed for either bound at any test level. Consequently, the two script writing modes can be taken as broadly equivalent, and the hypothesis can be accepted.

While differences at B1 and B2 were minimal, it could be seen that as one moved up the CEFR levels, the relative score gain conferred by using a computer increased. At B1 and B2 the difference was 3%. At C1, it was 5%, and at C2, 10%.

As mentioned above, use of a computer in an examination may be seen to be related to age in that older candidates simply follow an online path which leads to an online-proctored environment (i.e., navigating the internet, selecting an exam provider online, registering, booking a slot and managing their time etc.). For older candidates, the CP component in terms of how a test is taken may well be seen as simply a part of an online path they have followed.

The current study has been purely quantitative. A further study is currently exploring Writing Test markers' views on the effect of certain linguistic or textual features on candidates' scripts. Echoing markers' comments alluded to above, a more fine-grained examination lies in determining to what extent demographic factors such as age might have an effect on results obtained from writing tests by hand versus on computer.

Another aspect of the interaction between digital environment and textual production, worth exploring in the future, is that of task requirements vis-a-vis the support each environment allows. In a digital environment for instance, candidates have the option of employing a variety of content control features (provided these are made available by the test provider). Such features may significantly contribute to the authoring, editing, and proofreading of longer, complex and structurally challenging texts and thus account for the increasing discrepancy between scores, which culminates at C2.

The research literature revealed support for all modes: for handwritten scripts, for scripts written on computer, and for there being no difference. The current study, however, lends support to the view that, while differences remain, it is computer-processed scripts that certain candidates tend to score higher on.

A generally greater uptake of the use of computers is seen in the production of text – for all purposes, not just examinations. In the light of such uptake, one potential solution to the discrepancy score situation, as one looks to the future, is that all scripts be computer processed. Indeed, many professional examinations – law examinations, for example (Steel et al., 2019) – are now required to be done solely on computer as are the Association of Chartered Certified Accountants' (ACCA) financial and accounting examinations.

The COVID pandemic has accelerated the computer processing of scripts, with many more candidates taking exams online rather than on paper (Fuller et al., 2020; Abduh, 2021). For such a move to be accepted more widely, however, school students in particular need to have easy access to a computer and to be computer literate. This is contingent upon schools moving increasingly towards total computer-based work, with each child having their own laptop for continual school and home use, as with Uruguay's Plan Ceibal (see Segovia et al., 2022), for example. In the UK, the government Office of Qualifications and Examinations Regulation (Ofqual) has recently announced a three-year plan to explore the possibility of across-the-board online testing for students (Ofqual, 2022). Indeed, in the long run, what Mogey and Fluck (2015) describe as "post-paper assessment" is possibly what education and assessment authorities should be considering. Whether these changes will happen quickly will be observed and reported on in due course.

FUNDING STATEMENT

The authors received no financial support for the research and/or authorship of this article.

REFERENCES

Abduh, M. Y. M. (2021). Full-time online assessment during COVID-19 lockdown: EFL teachers' perceptions. *Asian EFL Journal*, 28(1.1), 26-46.

Association of Chartered Certified Accountants Association. *Computer-based exams.* (n.d.). https://www.accaglobal.com/vn/en/student/exam-support-resources/fundamentals-exams-study-resources/f1/technical-articles/computer-based-exams.html.

Arnold, V. (1990). Do students get higher scores on their word-processed papers? A study of bias in scoring hand-written vs. word-processed papers. The Educational Resources Center. Whitter, CA Rio Hondo College.

Boulet, J. R., McKinley, D. W., Rebbecchi, T., & Whelan, G. P. (2007). Does composition medium affect the psychometric properties of scores on an exercise designed to assess written medical communication skills?. *Advances in Health Sciences Education*, 12(2), 157-167.

Breland, H., Lee, Y. W., & Muraki, E. (2005). Comparability of TOEFL CBT essay prompts: response-mode analyses. *Educational and Psychological Measurement*, 65(4), 577-595.

Bridgeman, B., & Cooper, P. (1998). Comparability of scores on word-processed and handwritten essays on the Graduate Management Admissions Test. http://eric.ed.gov/?id=ED421528.

Canz, T., Hoffmann, L., & Kania, R. (2020). Presentation-mode effects in large-scale writing assessments. *Assessing Writing*, 45, 100470.

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.

Coniam, D., Lampropoulou, L., & Cheilari, A. (2021). Online proctoring of high-stakes examinations: A survey of past candidates' attitudes and perceptions. *English Language Teaching*, 14(8), 58-72. https://doi.org/10.5539/elt.v14n8p58.

Coniam, D., Stoukou, I., Lee,. T., & Milanovic, M. (2023). SELT IESOL Writing Test quality. London, UK: LanguageCert.

Fuller, R., Joynes, V., Cooper, J., Boursicot, K., & Roberts, T. (2020). Could COVID-19 be our 'There is no alternative'(TINA) opportunity to enhance assessment?. *Medical Teacher*, 42(7), 781-786.

Goldberg, A., Russell, M., & Cook, A. (2003). The effect of computers on student writing: A meta-analysis of studies from 1992 to 2002. *Journal of Technology, Learning, and Assessment,* 2(1).

Hodges, C., Moore, S., Lockee, B., Trust, T., & Bond, A. (2020).The difference between emergency remote teaching and online learning. *EDUCAUSE Review*. https://er.educause.edu/articles/2020/3/the-difference-between-emergency-remote-teaching-and-online-learning.

Horkay, N., Bennett, R. E., Allen, N., Kaplan, B., & Yan, F. (2006). Does it matter if I take my writing test on computer? An empirical study of mode effects in NAEP. *Journal of Technology, Learning, and Assessment*, 5(2). https://ejournals.bc.edu/index.php/jtla/article/view/1641.

Hughes, J., & Akbar, S. (2010). The influence of presentation upon examination marks. 11th Annual Conference of the Subject Centre for Information and Computer Sciences, 178–182.

King, F.J., F. Rohani, C. Sanfilippo, N. White. (2008). Effects of handwritten versus computer-written modes of communication on the quality of student essays. Center for Advancement of Learning and Assessment (CALA Report). http://www.cala.fsu.edu/files/writing_modes.pdf, 2008.

Klein, J., & Taub, D. (2005). The effect of variations in handwriting and print on evaluation of student essays. *Assessing Writing*, 10, 134–148. https://doi.org/10.1016/ j.asw.2005.05.002.

Lim, C. P., & Wang, L. (Eds.). (2016). Blended learning for quality higher education: Selected case studies on implementation from Asia-Pacific. Bangkok: UNESCO Bangkok Office.

MacCann, R., Eastment, B., & Pickering, S. (2002). Responding to free response examination questions: Computer versus pen and paper. *British Journal of Educational Technology*, 33(2), 173-188.

Marshall, J. C., & Powers, J. C. (1969). Writing neatness, composition errors, and essay grades. *Journal of Educational Measurement*, 6, 97–101. https://doi.org/10. 1111/j.1745-3984.1969.tb00665.x.

Mazzeo, J., & Harvey, A. L. (1988). The equivalence of scores from automated and conventional educational and psychological tests: A review of the literature. New York: College Entrance Examination Board.

Mogey, N., & Fluck, A. (2015). Factors influencing student preference when comparing handwriting and typing for essay style examinations. *British Journal of Educational Technology*, 46(4), 793-802.

Mogey, N., Paterson, J., Burk, J., & Purcell, M. (2010). Typing compared with handwriting for essay examinations at university: Letting the students choose. *ALT-J Research in Learning Technology*, 18, 29–47. https://doi.org/10.1080/09687761003657580.