Language Cert

A Comparability Study of Handwritten vs. Typed Responses to English Language Writing Tests

Irene Stoukou
Yiannis Papargyris
and
David Coniam

Abstract

This paper reports on a comparability study of Writing Test scores obtained from candidates who completed the writing test either by hand or typed, on a computer. The data comprise a comparatively large sample of candidates taking English language Writing Tests at four CEFR levels – B1 to C2. 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 script 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 over handwritten scripts. Differences observed on equivalence tests – an adaptation of the standard t-test – were not found to be statistically significant.

The paper concludes with the note 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.

Comparability of Scores Obtained from Handwritten or Computer-Processed Exam Scripts

There is a substantial literature on score equivalence obtained from handwritten (HW) and computer-processed (CP) scripts. Research dates back to the 1960s when the word-processing of scripts first began.

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. The following sections presents a review of the research.

Handwriting-based Studies Showing Advantage

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 & Cooper (1998) in a study involving Graduate Management Admissions Test scores reported higher scores with HW than with CP scripts. Klein & 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.

Computer-processed-based Studies Showing Advantage

An overall advantage for CP texts has been reported in certain studies (Sprouse & Webb, 1994; Peacock, 1988; Hughes & 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 & Haney (1997) and Russell and Plati (2001).

Neither Mode Conferring an Advantage

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 & 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.

Indeterminate Research Outcomes but Increasing Use of Computers

There is evidence then 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 & 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.

The IESOL Writing Test

The data in the study were drawn from three examinations – at CEFR levels B1–C1, which form part of LanguageCert's IESOL SELT suite of English language tests. In the LanguageCert Writing tests, candidates complete two writing tasks which elicit 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

All tasks are assessed on a four-point scale on four subscales double-marked with the final grade drawn from the mean of the two examiners' scores (see https://www.languagecert.org/en/language-exams/). 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 examiners must meet minimum requirements in terms of professional qualifications and experience in order to be eligible for consideration as an examiner (Papargyris and Yan, 2022). Prospective examiners go through a standardised training process before they are approved and allowed to mark. The training process includes marking sample scripts. Candidates for the examiner role must show they can mark accurately and consistently before they are certificated as examiners. During live marking, if an examiner is found to be marking inaccurately and/or inconsistently, they may be removed from the marking session and/or retrained or dismissed as an examiner. Examiners are then monitored on an ongoing basis and required to attend standardisation meetings on a regular basis.

The Current Study

Two sets of data for the Writing Test are presented. The first dataset contains descriptive statistics: means, standard deviations and effect size differences. The second test consists of equivalence independent samples t-tests ("equivalence tests"). Equivalence tests permit significance to be observed via specified upper and lower bounds, rather than regular t-tests, see Lakens (2017). The upper and lower bounds represent the extent of variation of t values regarding the two populations of the two samples being tested. If the t value of the equivalence test is within the estimated range, the two populations may be deemed to be equivalent.

Hypotheses

The overarching hypothesis in the current study is that mean scores obtained between the two modes of script production – computer-processed or handwritten – will not be significantly different. Specifically, the following three hypotheses are pursued:

1. That the difference between the mean scores for the two modes of script production will be less than 5% for any given CEFR level.
2. That only small effect size differences – if any – between the two modes will be observed.
3. That, on equivalence tests, significance will not emerge against specified upper and lower bounds for any given CEFR level.

Descriptive Statistics

Table 2 presents a summary of the effect size differences between the sets of means for the Writing Test total score (maximum 25) for each mode using Cohen's d. Cohen's d indicates standardised differences between two means, sharpening comparisons between two means. In general, a small effect is taken as 0.2, a medium effect as 0.5, and a large effect as 0.8 (Glen, 2021).

Table 2
Effect size and mode mean differences

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

Legend: CP=computer-processed; HW=Handwritten

As can be seen from Table 2, effect sizes are negligible for levels B1 and B2. While there is a small effect size at C1 level, the score difference between the two modes at C2 level is greater than 5%, with a medium effect size difference of 0.55. 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 3 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). The critical decision on equivalence, as stated earlier, is whether the estimated t values in Table 2 below are between the upper and lower bound. The p values for the t values (upper bound, t-test and lower bound) indicate significance where these go beyond the specified bounds.

Table 3
Equivalence samples t-tests

Test Level	Statistic	t	df	p
B1	upper bound	9.36	22725	< .001
	t value	8.81	22725	< .001
	lower bound	8.26	22725	1.00
B2	upper bound	15.12	27588	1.00
	t value	14.23	27588	< .001
	lower bound	13.34	27588	< .001
C1	upper bound	9.99	10328	1.00
	t value	10.45	10328	< .001
	lower bound	10.91	10328	< .001
C2	upper bound	16.92	4361	1.00
	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 modes of writing scripts can be considered equivalent for the four CEFR levels examined in the study.

Discussion and Conclusion

This study has explored the comparability of scores obtained by candidates of LanguageCert's IESOL Writing Tests at CEFR levels B1 to C2 in the context of scripts produced by candidates in handwritten mode or in scripts written on computer.

The key hypothesis in the study was that mean scores and performance on the Writing test in either test production mode of delivery would not be significantly different from each other and that candidate scores would not be influenced by the mode in which they produced their test scripts. 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 size differences between the two modes would be observed. Negligible effect sizes were observed for levels B1 and B2, and a small effect size was observed at C1. For C2, however, a medium effect size was recorded, causing the hypothesis to be rejected.

The third hypothesis was that, on equivalence tests, significance would not emerge against specified upper and lower bounds for any given CEFR level. As significance was not observed for either bound in any of the test levels, it was determined that the two modes of test production may be considered broadly equivalent for the four CEFR levels examined, and the hypothesis was 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%.

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 lies in the candidates' background. In a survey (in mid-2022) of over 40 LanguageCert Writing test examiners, examiners 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 examiners as being older and more computer literate. Examiners perceived these two factors as helping to account for the skew towards higher scores achieved on computer-processed scripts.

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, as mentioned, is currently exploring Writing Test examiners' views regarding the effect of certain linguistic or textual features on candidates' scripts. Echoing examiners' 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-19 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.

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. 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.

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.

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. Retrieved from 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.

Nulty, D. D. (2008). The adequacy of response rates to online and paper surveys: what can be done? *Assessment & Evaluation in Higher Education*, 33(3), 301-314.

Ofqual. (2022). Ofqual corporate plan 2022 to 2025. Coventry, UK: Ofhttps://www.gov.uk/government/organisations/ofqual.

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. E., Fowles, M. E., 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, 220–233. https://doi.org/10.1111/j.1745-3984.1994.tb00444.x.

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

Russell, M., & Haney, W. (1997). Testing writing on computers: An experiment comparing student performance on tests conducted via computer and via paper-and-pencil. *Education Policy Analysis Archives*, 5(3), 1-19.

Russell, M., & Plati, T. (2001). Effects of computer versus paper administration of a state-mandated writing assessment. Teachers College Record. http://www.tcrecord.org/Content.asp?ContentID=10709

Russell, M., & Tao, W. (2004). The influence of computer-print on rater scores. *Practical Assessment, Research & Evaluation*, 9(10), 1–14.

Segovia, G. D., Jang, E. H., Manuel, C., & Staal, E. (2022). Uruguay: Rethinking teacher training and global education through Plan Ceibal. In Reimers, F.M., Budler, T.A., Irele, I.F., Kenyon, C.R., Ovitt, S.L., & Pitcher, C.E. *Reimagining our Futures Together. A New Social Contract For Education*, pp. 449-477. Paris: UNESCO.

Sprouse, J. L., & Webb, J. E. (1994). The Pygmalion effect and its influence on the grading and gender assignment on spelling and essay assessments. ERIC Document, ED 374096.

Steel, A., Moses, L. B., Laurens, J., & Brady, C. (2019). Use of e-exams in high stakes law school examinations: Student and staff reactions. *Legal Education Review*, 29, 1.

Sweedler-Brown, C. O. (1991). Computers and assessment: The effect of typing versus handwriting on the holistic scoring of essays. *Research and Teaching in Developmental Education,* 8, 5–14.

Taylor, C., Kirsch, I., Eignor, D., & Jamieson, J. (1999). Examining the relationship between computer familiarity and performance on computer-based language tasks. *Language Learning*, 49(2), 219–274.

Trobia, A. (2011). Cronbach's Alpha. In Lavraka, P. (ed.) *Encyclopedia of survey research methods*, Vols 1 & 2, pp. 168-169. Thousand Oaks, Ca.: Sage Publications.

Wise, S., & Plake, B. (1989). Research on the effects of administering tests via computers. *Educational Measurement: Issues and Practice*, 8(3), 5–10.

Wright, B., & Linacre, M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, 8(3), 370. http://www.rasch.org.

Language
Cert
languagecert.org