Authors' post-print version. Brunfaut, T., Harding, L. & Batty, A. (accepted/in-press, 2018). Going online: The effect of mode of delivery on performances and perceptions on an English L2 writing test suite. *Assessing Writing*.
.

Going online: The effect of mode of delivery on performances and perceptions on an English L2 writing test suite

Tineke Brunfaut[a], Luke Harding[b] and Aaron Batty[c]

[a]Lancaster University
Department of Linguistics and English Language
County South
Lancaster
LA1 4YL
United Kingdom
t.brunfaut@lancaster.ac.uk

[b]Lancaster University
Department of Linguistics and English Language
County South
Lancaster
LA1 4YL
United Kingdom
l.harding@lancaster.ac.uk

[c]Keio University
Shonan Fujisawa Campus
5322 Endo Fujisawa
Kanagawa 252-0882
Japan
abatty@sfc.keio.ac.jp

Corresponding author: Luke Harding, l.harding@lancaster.ac.uk, +44 (0)1524 593034

Authors' post-print version. Brunfaut, T., Harding, L. & Batty, A. (accepted/in-press, 2018). Going online: The effect of mode of delivery on performances and perceptions on an English L2 writing test suite. *Assessing Writing*.
.

Going online: The effect of mode of delivery on performances and perceptions on an English L2 writing test suite

Abstract

In response to changing stakeholder needs, large-scale language test providers have increasingly considered the feasibility of delivering paper-based examinations online. Evidence is required, however, to determine whether online delivery of writing tests results in changes to writing performance reflected in differential test scores across delivery modes, and whether test-takers hold favourable perceptions of online delivery. The current study aimed to determine the effect of delivery mode on the two writing tasks (reading-into-writing and extended writing) within the Trinity College London *Integrated Skills in English* (ISE) test suite across three proficiency levels (CEFR B1-C1). 283 test-takers (107 at ISE I/B1, 109 at ISE II/B2, and 67 at ISE III/C1) completed both writing tasks in paper-based and online mode. Test-takers also completed a questionnaire to gauge perceptions of the impact, usability and fairness of the delivery modes. Many-facet Rasch measurement (MFRM) analysis of scores revealed that delivery mode had no discernible effect, apart from the reading-to-writing task at ISE I, where the paper-based mode was slightly easier. Test-takers generally held more positive perceptions of the online delivery mode, although technical problems were reported. Findings are discussed with reference to the need for further research into interactions between delivery mode, task and level.

1. Introduction

Following significant technological developments and vast increases in computer accessibility, the past three decades have seen the introduction of several computer-based language testing systems. In some cases, tests have been conceptualised as computer-based from inception (e.g. Dialang, PTE Academic); in other cases, test developers aimed to replace a paper-based test with a computer-based system, or envisioned a parallel offer across the two modes of delivery (e.g. IELTS' concurrent paper-based and computer-based delivery of their reading, listening and writing components). Based on a comprehensive review of the computer-based testing literature, Davey (2011) concluded that the key motivations for adopting computer-based testing approaches are: (a) to target new constructs; (b) to achieve more accurate and efficient scoring; and (c) to make test administration more accessible, efficient and cost-effective. Factors such as market demand or policy requirements are also likely to play a role in the decision to move a paper-based test online. This article presents a comparative study within the context of motivation (c): the decision to add an online alternative to a primarily paper-based test (the Trinity College London ISE suite writing test) in order to make test delivery more efficient and accessible.[1]

At face value, writing, of all the language skills, may be the most suitable to test in a computer-based environment. Since the spread of word processing software, and more latterly mobile

[1] Note that in the remainder of this article we mostly use the term 'computer-based' instead of 'online' to clarify the nature of the device and because this is the more frequently used term in prior research. We appreciate the additional internet connectivity aspect of the online mode, but this did not constitute a specific focus of the present study (but see the trial project referred to in section 2).

technologies, a considerable proportion of day-to-day writing tasks are now completed on computers and other electronic devices. Scholars such as Jin and Yan (2017) have consequently argued for computer-based writing assessment. However, research has also shown that the writing medium may have an effect on the writing process (e.g., Van Waes & Schellens, 2003), and that writing processes may in turn influence text quality (e.g., Breetvelt, van den Bergh & Rijlaarsdam, 2009). A key question, therefore, is raised in those situations where a writing test offered in one delivery mode is replaced by another mode, or where two modes exist simultaneously and are intended to be parallel: does the mode of delivery affect test performance? This is especially important since task-related factors such as task layout, response mode or editing functions, and writer-related factors such as handwriting or computer skills may exert an influence. Depending on their effect and on their relevance, such factors may cause undesirable or construct-irrelevant variance. Therefore, researchers such as Choi, Kim and Boo (2003) have called for comparative studies on the impact of delivery mode as a prerequisite for the validation of computer-based tests (in contexts of paper-based replacements or parallel use of both modes).

1.1 Paper- versus computer-based testing of second language writing: Performance results

Previous research on the impact of delivery mode on writing test scores, which has primarily concentrated on independent writing tasks, has not led to uniform conclusions. For example, in an early study with a counterbalanced repeated measures design, Owston, Murphy and Wideman (1992) found that a group of computer-experienced eighth-graders obtained significantly higher scores on all writing criteria for their paper-based writing performances compared with their computer-based scores. Lei, Livingstone, Larkin and Bonett (2004) also found that pre-service teachers gained systematically higher scores on essays in a paper-based writing test than those they produced in a computer-based test, after controlling for essay topic, essay version, and test-taker characteristics. Similarly, in a relatively recent study, Chen, White, McCloskey, Soroui and Chun (2011) established that (young) adults had better overall results on the paper-based version of an adult written literacy assessment than on its computer-based counterpart (regardless of their gender or level of education). Analysing their data in more detail, Chen et al. observed different delivery mode effects on the writing tasks in their study (a complaint letter, opinion letter and request letter) and that these differences were related to test-takers' characteristics (employment status in the complaint letter task, and age and race/ethnicity in the opinion letter task). They concluded that the computer mode might negatively affect the performance level of specific groups of test-takers (e.g., unemployed or 65+). Other studies which also identified a delivery mode effect on writing test scores, however, observed the effect to go in the opposite direction. Li (2006), for instance, found that a group of advanced English second language (ESL), adult learners gained higher scores on their argumentative writing for performances composed on a computer as opposed to those written on paper. Jin and Yan (2017) found that Chinese ESL college students produced longer texts with fewer language errors when completing expository essay tasks from the College English Test in computer-based mode as opposed to paper-based mode.

By contrast, a number of studies have found no impact of delivery mode on test-takers' mean writing scores, but have revealed effects at a sub-group or individual level. For example, Endres (2012) found no impact of delivery mode on average writing scores. However, when looking at individual

test-takers' results, several obtained higher scores in the computer-based mode, and Endres warned that "writing in the two modes is a different experience for candidates and should not simply be considered comparable" (p.31). Breland and Muraki (2005) also found similar observed mean scores between handwritten versus computer-typed performances of TOEFL writing tasks overall and did not detect any real task-effect differences. However, at a sub-group level, a delivery mode effect was discovered, but in contrast to Endres (2012), with systematically lower scores on the computer-based mode for those test-takers with lower ESL ability. Wolfe and Manalo (2005), who also focussed on the TOEFL, found a similar association between ESL proficiency and score effects of writing exam delivery mode. More specifically, they observed that less proficient test-takers (as measured by their scores on TOEFL multiple-choice items) benefitted from the handwritten mode, while no score differences between the delivery modes were found for more proficient test-takers.

Finally, when exploring the potential role of rating approach on the scoring of performances produced in different modes, Lee (2004) did not observe systematic writing score differences between the writing section of a paper- versus computer-based ESL placement test when using the test's conventional holistic rating approach. However, when looking at analytic ratings of the written performances along the criteria of organization, content, linguistic expression, and use of sources (using a scale developed for research purposes), Lee found that raters' analytic scores on all criteria were significantly higher on the computer-based version than the paper-based version. Powers, Fowles, Farnum, and Ramsey (1994), on the other hand, who looked into the effect on scoring of the medium in which written performances were presented to raters, found that performances presented in handwritten mode received higher scores on average than those presented in word-processed format (regardless of whether the texts had originally been written in handwritten or word-processed mode and transformed into the other mode later). It should be kept in mind, however, that Powers et al.'s research was conducted at a time when word-processed writing was less widespread, and raters might have been more used to handwritten texts. Given rapid developments in accessibility, use of, and familiarity with computers and word-processing in the new millennium, it is thus uncertain how transferrable these findings are to the present time.

The studies surveyed above suggest that, while results of comparative studies remain mixed, there is value in considering the effect of delivery mode at different levels of proficiency. Also, since Lee (2004) is one of the only papers which has considered delivery mode effects across different analytic scoring criteria, further evidence is required to understand interactions between delivery mode and different aspects of written performance as reflected in scale criteria. Furthermore, there is value in considering the nature of the writing task; a gap which has not been fully addressed in the literature to date. Comparative studies have typically focused on investigating mode of delivery for "independent" writing tasks: those tasks which consist simply of a prompt or instructions for an extended piece of writing. However, an increasing number of language tests include integrated task types such as reading-to-write tasks whereby test-takers produce a piece of writing on the basis of reading input provided by the test developers. Although the research base on integrated writing tasks is expanding (see for example Cumming, 2014; Plakans and Gebril, 2017), hardly any studies so far have specifically explored the impact of delivery mode on such tasks. Furthermore, to our knowledge, there have been no comparative studies to date which compare delivery mode effects across independent and integrated tasks within the same study. Given the important role in computer-based

testing played by the user-interface (UI) (Fulcher, 2003), and the more complex UI that an integrated task requires (navigating prompt/instructions, text(s) and the writing space), investigating integrated tasks alongside independent tasks opens up the potential for building further theory of relevance to modern-day testing practices.

1.2 Paper- versus computer-based testing of second language writing: Perceptions

Apart from investigating the statistical equivalence of different modes of test delivery (see 1.1), McDonald (2002) recommended also exploring the 'experiential equivalence' of paper-and-pencil versus computer-based tests. Because the different delivery modes "provide test takers with qualitatively different experiences" (p.299), McDonald argued that individual differences between test-takers are likely to play a contributory/explanatory role in the test experience. One factor which McDonald proposed to systematically consider in combination with score analyses is computer familiarity. Although a number of studies have shown that computer familiarity may explain some variability in writing scores between delivery modes or between test-takers within computer-based modes (e.g., Jin and Yan, 2017; Taylor, Kirsch, Eignor and Jamison, 1999; Zou and Chen, 2016), fewer studies have considered the relationship between computer familiarity and perceptions of the test. Other individual difference factors listed by McDonald include computer anxiety and computer attitudes. At the same time, looking into the experiential equivalence for different delivery modes also responds to calls in the language testing literature to make the test-taker and their views a central component of the validation process (e.g. Weir, 2005). Indeed, Yu (2010) found a "psychological side" to delivery mode effects in the perceptual data of his research and urged to represent "the voices of students when investigating comparability of delivery modes" (p.119).

Although the empirical literature on individual differences in research on the delivery mode of second language writing tests is less prominent than studies on statistical equivalence, a few researchers have investigated aspects of experiential equivalence and described test-takers' views on the different delivery modes. Lee (2004), for example, concluded from a survey that test-takers who are more 'habitual computer writers' (p.4) favoured the computer-based writing test mode over the paper-based one. Maycock and Green (2005) found that test-takers generally liked the computer-based version of IELTS and that those who opted for this version felt relatively confident using the computer and able to compose text on computer without real issues. In fact, these test-takers appreciated having access to editing functions in the computer medium, which hints at potential usability issues of different delivery modes. The test-takers also suspected that a higher level of computer skills would lead to higher scores on the computer-based test, thus raising issues around impact and fairness of different delivery modes, although Maycock & Green point out that no actual impact on scores was observed in their study. In another study on the role of computer familiarity, which found that this variable played an explanatory role in writing test score differences between delivery modes, Jin and Wu (2010) additionally discovered that computer familiarity (and language proficiency) significantly affected test-takers' perceptions of the CET. Finally, Ling's (2017) study, which specifically explored the use of US keyboards on writing performance, indicated that most test-takers found the keyboard being used 'convenient' and 'efficient', but many still preferred 'a more familiar local keyboard' (p.36). Ling (2017) therefore recommended that there is "room for improvement in the keyboard-related test-taking experience" (p.36).

In sum, different delivery modes may lead to differential test-taker experiences (potentially mediated by a number of individual differences) and affect test-takers' perceptions of a test's impact, fairness, and usability across test conditions.

2. This study

Given the mixed nature of results from previous research, for high-stakes language tests there is a need for ongoing validation research to support the parallel use of paper-based and computer-based/online testing modes. Trinity College London – an international exam board for the performing arts and English language (see http://www.trinitycollege.co.uk/site/?id=263) – commissioned the current study in order to gain insight into the "translatability" of their current paper-based and face-to-face interactive and communicative-based English language tests into online formats. A small-scale practical and technical trial was first conducted by Trinity College London as an initial feasibility study of the online delivery of their Integrated Skills in English (ISE) exam suite – a four skills exam intended for youngsters and adults. As a second step, Trinity College London sought external academic research partners to investigate the potential impact of online delivery of the ISE exam suite in terms of exam construct, test performances, test-taker experience, and scoring.

The present study reports on one part of this programme, namely the exploration of delivery mode effects on the ISE writing tests in terms of performance results (scores) and test-taker perceptions. Although the study has a specific examination as its focus, the research presented has broader implications for the field as it explores a number of pressing issues in theorising the impact of delivery mode on writing test scores, namely by exploring effects at three different CEFR levels, by locating effects on individual rating scale criteria, by exploring the nature of effect on an integrated reading-into-writing task as well as an independent writing task, and by triangulating effects at the score level with test-takers' perceptions of impact, usability and fairness. Additionally, the study seeks to extend insights into the relationship between computer familiarity and perceptions of computer-based tests. The aims of the study were broken down into the following research questions:

RQ1. Is there a difference in test-takers' scores on the ISE writing test suite (including an integrated and independent task) depending on delivery mode?

RQ2. Is there a difference in test-taker's perceptions of the impact, usability and fairness of the ISE writing test suite depending on delivery mode?

RQ2a. To what extent is computer familiarity related to perceptions of the computer-based delivery mode?

Since the ISE exam suite consists of five different target language proficiency levels, each with their own specific test structure, the study was conducted at different exam levels separately. Specifically, since the market demands are highest for the CEFR B1, B2, and C1 levels, the study focused on the ISE I, ISE II and ISE III tests to help establish a comprehensive picture of the potential for online delivery.

3. Method

3.1 Research design

To investigate RQ1 – *whether there is a difference in test-takers' scores on the ISE writing test suite depending on delivery mode* – each participant was given two versions of their target level ISE exam: one version in paper-based mode (the current operational mode) and one version in online mode (the newly proposed mode). These were taken on the same day or on consecutive days, depending on test centre facilities and participant availability. To compensate for potential mode order effects, a counterbalanced design was used whereby at each ISE level some test-takers first completed the writing test in paper-based mode and then in online mode, whereas others first completed the online and then the paper-based mode.

Also, two different forms of the test at each ISE level were used to offset potential task effects on the findings and to avoid test-takers completing the same test form in each delivery mode (which would lead to potential learning effects). The different test forms – referred to as *Form A* and *Form B* in this paper – were selected by Trinity College London based on similarity in difficulty level (based on operational testing data) and topic areas. To avoid test form order effects, a counterbalanced design was used in which an approximately balanced number of test-takers first completed Form A and others first completed Form B, with further divisions by order of delivery mode. The research design is summarized in Table 1. Participants were randomly allocated a group, ensuring similar size groups across the study. Exam administration procedures replicated real-life ISE exam scenarios as much as possible.

[INSERT TABLE 1 HERE]

Furthermore, to reflect rater variation in operational testing, and to avoid rater-dependence, seven fully-trained raters were involved in evaluating the written performances. These experienced raters were allocated scripts in such a manner that each examined a number of performances in each delivery mode and on each test form within an ISE level. Also, 29% of the performances were double marked.

3.2 Instruments

3.2.1 Personal background questionnaire

To establish a profile of the test-takers participating in the study, a personal background questionnaire was designed with conventional biodata questions, including on participant's age, gender, first language, current activity (study, work,…), and years of learning English. In addition, given the study's focus on exploring the potential of the ISE writing test in computer-based online format, participants were also asked for how many years they had been using computers and to indicate on Likert scales (1) how often they had used computers in the last year, (2) how good they felt they were at using computers, and (3) how well they could type on a computer. They were also asked whether they had ever taken a language test on a computer before.

3.2.2 ISE writing tests

For all three ISE levels explored in this study (ISE I-II-III, corresponding to CEFR B1-B2-C1, respectively), the writing section consists of two tasks: an integrated reading-into-writing (RIW) task and an independent writing (IW) task. In the reading-into-writing task, test-takers need to respond to a

writing prompt by using information from four input texts provided in the reading section of the exam. The task aims to assess test-takers' ability to:

- "identify (straightforward) information that is relevant to the writing task", "common themes and links across multiple texts" (all three levels; Trinity College London, 2016, p.37, 56, 75) and "finer point of detail, eg implied attitudes" (ISE III; p.75);
- "paraphrase and summarize short pieces of information" (ISE I; p.37), "factual ideas, opinions, argument and/or discussion" (ISE II; p.56), or "complex and demanding texts" (ISE III, p.75);
- or, "combine information to produce a short and simple response to suit the purpose for writing, eg to describe a problem and suggest solutions" (ISE I; p.37), "synthesise such information to produce coherent responses to suit the purpose for writing (eg to offer solutions to a problem and/or evaluation of the ideas)" (ISE II, p.56) or "synthesise such information to produce sophisticated responses with clarity and precision" (ISE III; p.75).

The reading-into-writing task is marked using four criteria: reading-for-writing, task fulfilment, organisation and structure, and language control. For each criterion, a score is awarded on a scale ranging from 0 to 4.

In the independent writing task, test-takers respond to a prompt relating to one of a range of possible subject areas, with differences in topic domains, range, and concreteness between the ISE levels. The task aims to evaluate test-takers' ability to produce: "a narrative, descriptive or instructional text following the instructions" (ISE I; Trinity College London, 2016, p.38), "a clear and detailed text in response to the prompt" (ISE II; p.57), or "a discursive, well-developed text following the instructions" (ISE III; p.76). The target language functions are expressing "simple facts and personal opinions in some detail coherently" (ISE I; Trinity College London, 2016, p.38), or "opinions, evaluating and making suggestions" (ISE II & III; p.57 & 76). The task is marked on a scale ranging from 0 to 4 for each of the following three criteria: task fulfilment, organisation and structure, and language control.

The targeted output genres for both writing tasks comprise: descriptive and discursive essays, articles, informal or formal emails and letters, and reviews (all three levels), and also argumentative essays and reports at ISE II & III. The word count targets for each of the writing tasks are 100-130 at ISE I, 150-180 at ISE II, 200-230 at ISE III. Sample tasks, test specifications, and further information on scoring can be found on the ISE website (http://www.trinitycollege.co.uk/site/?id=3192).

In both delivery modes explored in this study, test-takers were first presented with the RIW task and then the IW task. In the paper-based mode (the operational mode), the writing test forms part of a paper booklet containing the reading and writing sections of the ISE exam. Each writing task consists of a set of instructions, a space to make planning notes, and approximately 2.5 pages of lined space to compose one's written task response. At the end of each task, test-takers are invited to review their written composition. A notable feature of the integrated RIW task is that test-takers need to draw on information from four input texts presented in the second reading task of the ISE exam. Thus, test-takers need to turn back to the reading section which precedes the writing section while completing the RIW. Also, the paper-based mode allows for test-takers to underline or annotate parts of the input texts in the printed booklet. Sample booklets can be found via the exam level links on the ISE website (http://www.trinitycollege.co.uk/site/?id=3192).

In the computer-based mode, the ISE writing exam was delivered online through the assessment software Surpass. The task instructions and typing space for the written compositions (including automated word counters) were presented on one screen. Planning notes could be made on paper. In the RIW task, the four reading input texts were provided on the same screen as the instructions and composition space, with the additional option to open the texts as a pop-up and position them adjacent to the writing space for ease of reference. Text annotation functions were not available in this mode.

3.2.3 Perception questionnaire

A questionnaire, consisting of three sets of questions, was designed to establish test-takers' perceptions of the writing test depending on delivery mode. The first set elicited test-takers' views on the potential *impact* of delivery mode *on their emotional state* and was adapted from Boekaerts' (2002) on-line motivation questionnaire. It comprised six four-point Likert scale items (from 'strongly disagree' to 'strongly agree') that were repeated for each delivery mode. The emotional states explored were: nervousness, comfortability, frustration, confidence, boredom, and happiness. The second set of questions focused on *usability* of each delivery mode and consisted of four-point Likert scale statements (from 'strongly disagree' to 'strongly agree'). More specifically, the questions focused on understanding what to do, ease of writing using a pen/keyboard, ease of revising and editing on paper/screen, clarity of the test's layout, and ease of navigation through the test. This was followed by an open-ended question asking about any particular problems completing the writing test in a particular mode. The third set of questions asked for test-takers' views on the *fairness*[i] of the test depending on the delivery mode. It contained one four-point Likert scale item for each mode on how well the test assessed the test-taker's writing ability, and asked about estimated scoring level on each test version and potential preferences for a particular delivery mode for the writing test. A final open-ended question gave participants the opportunity to share any other thoughts on ISE writing test.

3.3 Participants

283 English second language learners who were based in either Ireland, Italy or Spain participated in the study: 107 took ISE I, 109 ISE II, and 67 ISE III. All participants were preparing for the ISE exam and/or were registered to take the exam, and were thus familiar with the paper-based exam. For the purposes of the study, they were also familiarised with the alternative online exam format through an introductory video. As an incentive to complete the tests to the best of their ability, the participants were provided with feedback and a diagnostic report on their performance (on the operational, paper-based exam format), and were provided with a retail voucher for their time and effort.

259 (92%) of the participants were willing to provide their biodata – presented per ISE level in Table 2. Table 2 also indicates that, overall, the participants' self-reported computer familiarity was high. The test-takers at all three test levels had been using computers for many years, and did so on a very frequent basis. Almost everyone also perceived their computer and typing ability as at least adequate, if not good or excellent. Fewer participants, however, had previously completed language tests on a computer.

[INSERT TABLE 2 HERE]

3.4 Analyses

3.4.1 ISE writing tests

To evaluate the comparability of paper-based versus computer-based delivery mode of the ISE writing exam suite, many-facet Rasch measurement (MFRM) was conducted using FACETS (Linacre, 2017). As the first and second sitting of the test was found to differ significantly in overall difficulty (irrespective of test form or delivery mode), it was included in the MFRM model. Therefore, a four-facet model was constructed for each ISE level including test-takers, order (first or second sitting), raters, and rating criteria on the tasks. Each rating criterion for each task was treated as a distinct rating scale (i.e., seven distinct scales). Order, raters, and rating criteria difficulty/severity estimates were centred at zero, which allowed the test-taker estimates to "float". Three further dummy facets were entered into the model. Dummy facets' estimates are anchored at zero and therefore do not contribute to the estimation of the remaining facets' elements' locations, but still allow for exploration of interaction effects through bias analysis. The dummy facets were: test form (A or B), delivery mode (paper-based or computer-based), and task. Summary statistics for each model at each ISE level are presented in Tables 3 to 5 below.

[INSERT TABLE 3 HERE]

At ISE I (Table 3), the mean measure for test-takers shows that the test was slightly easy (*M*=.31), although the standard deviation was relatively high suggesting a wide dispersion of measures across the logit scale. The average Infit and Outfit MS (mean square) statistics are near or at their expected value of 1, and standard deviations for the order, rater, and rating criteria facets indicate uniform fit to the model. The test-taker fit is less uniform, although the reliability of separation suggests that fit was not a serious problem. Separation statistics suggest moderate separation of test-takers between two statistically distinct levels. However these same statistics also suggest that raters were not a homogeneous group with a high reliability of separation. Nevertheless, raters were shown to have a higher than expected level of inter-rater agreement and the Rasch κ was approaching zero, which would suggest a good level of agreement. The order facet can be separated into two levels, which necessitated its inclusion in the model. The separation statistic for the rating criteria indicates that the seven criteria across the two writing tasks can be reliably separated into five distinct levels.

[INSERT TABLE 4 HERE]

The MFRM results for ISE II (Table 4) were similar to those of ISE I for test takers, although the test was slightly more difficult for candidates at the ISE II level (*M*=.51); however, the order facet did not exhibit the difference in difficulty separation observed in the ISE I, and raters were more uniformly severe. Rater agreement remained good, despite the lack of homogeneity, but the rating criteria could only be separated into three distinct levels.

[INSERT TABLE 5 HERE]

Finally, the ISE III (Table 5) continued the trend of increasing difficulty for the test-takers (*M*=.21). The order facet once again can be separated into two levels, as can the raters. Agreement between the

raters, however, was good. The rating criteria, however, could only be separated into two levels of difficulty.

3.4.2 Perception questionnaires

To gain insights into test-takers' perceptions of delivery mode, descriptive and comparative statistics were run on the Likert scale and MC items of the perception questionnaire. More specifically, differences in test-takers' perceptions of the impact, usability and fairness of the paper-based versus computer-based writing test mode were explored through Wilcoxon-signed rank tests. The open-ended questions were analysed qualitatively using thematic analysis.

Additionally, the relationship between perceptions of computer-based mode (reported in the perception questionnaire) and computer familiarity (reported in the personal background questionnaire) was explored through a series of Spearman's correlations. Three measures of computer familiarity were investigated: frequency of use, self-assessed computer ability, and self-assessed typing ability. These analyses were conducted for each ISE level separately.

4. Results

4.1 ISE writing tests

The effect of delivery mode on ISE writing scores was examined in two ways: (1) through a bias/interaction analysis of mode and task (to provide an overall picture of the extent to which mode affected task performance), and (2) through a bias/interaction analysis of mode and rating scale category (to provide a more precise view of where any effect was located). Analyses were performed separately for the three ISE levels.

ISE I (B1)

Table 6 shows that there was a small, statistically significant difference in mean measure between delivery modes on the RIW task (d=.14, p<.001) in the direction of greater ease in the paper-based mode (PB) at the ISE I level. This effect did not hold for the IW task at ISE I.

Analysis of the pairwise bias report for mode and rating scale category (Table 6) revealed that the strongest effects were located on the Language Control (LC) scale (d=.23, p<.001), and the Reading-for-Writing (RfW) scale (d=.14, p=.021) for RIW. In each case, the bias was in the direction of greater ease in the paper-based condition. No statistically significant contrasts were observed on the three scales used to judge IW task performance.

[INSERT TABLE 6 HERE]

ISE II (B2)

Table 7 shows that there was no statistically significant difference in mean measures between delivery modes on either the RIW or IW task at ISE II. Any contrasts were negligible in light of the relevant standard error. This suggests that there was no discernible effect of delivery mode on either writing task.

Analysis of the pairwise bias report for mode and rating scale category confirmed that there were no statistically significant differences between mean measures across modes of delivery for scales used to judge either writing task (Table 7).

[INSERT TABLE 7 HERE]

ISE III (C1)

Similar to the ISE II results, Table 8 shows that there was no statistically significant difference in mean measures between delivery modes on either the RIW or IW task at ISE III. The findings reveal that, once again, there was no discernible effect of delivery mode on either writing task.

Analysis of the pairwise bias report for mode and rating scale category also revealed no statistically significant differences at the level of rating scale categories between mean measures across delivery modes (Table 8).

[INSERT TABLE 8 HERE]

4.2 Perceptions

ISE I (B1)

Impact. The descriptive analyses of test-takers' perceptions of their emotional state while completing the ISE I writing test in paper-based versus computer-based mode show that test-takers had quite similar views on both modes (Table 9). Test-takers slightly more frequently chose "agree" or "strongly agree" on "nervous", "frustrated" and "bored" in the paper-based mode, and "agree" or "strongly agree" on "comfortable", "confident" and "happy" in the computer-based mode. In most cases, however, the differences in mean responses between the two modes were not statistically significant. Nevertheless, ISE I test-takers reported being happier during the computer-based writing test (Z=-3.01, p=.001), with a small effect size (r=.22).

[INSERT TABLE 9 HERE]

Usability. The descriptive statistics in Table 10 indicate that the overall majority of participants were more positive about the usability of the ISE I writing test in computer-based mode compared with the paper-based mode. The difference between both delivery modes was statistically significant on two of the usability items, with test-takers finding it easier to revise and edit on a computer than on paper (Z=-4.28, p<.001; medium effect size, r=.30) and to navigate through the computer-based test (Z=-2.06, p=.039; small effect size, r=.15). Nevertheless, evaluations of the usability of the ISE I writing test were largely positive in both delivery modes.

[INSERT TABLE 10 HERE]

Fairness. Analysis of the items within the fairness section of the questionnaire showed a preference for the computer-based writing test format among the ISE I test-takers (Table 11). A Wilcoxon signed-rank test indicated that they thought that their writing ability was tested better through the computer-based test compared with the paper-based format (Z=-2.13, p=.033, r=.15). Many ISE I test-takers also

expected to have scored higher on the computer-based mode (46%) and would prefer this mode over the paper-based mode on future occasions (56%).

[INSERT TABLE 11 HERE]

Relationship between perceptions of computer-based mode and computer familiarity. A series of Spearman's correlations showed that, at ISE I, frequency of computer use had a weak, negative relationship with boredom during the computer-based test (r_s=-.23, p=.024). Self-assessed computer ability was weakly associated with typing ease (r_s=.25, p=.013). Self-assessed typing proficiency was positively associated with perceptions of typing ease (r_s=.37, p<.001) and revision and editing ease (r_s=.28, p=.006).

ISE II (B2)

Impact. The descriptive statistics in Table 12 show that somewhat more ISE II test-takers (strongly) disagreed that they had been nervous, frustrated or bored, and (strongly) agreed that they had been comfortable, confident and happy while completed the test in the computer-based mode as compared with the paper-based mode. Indeed, the comparative analyses revealed that test-takers' perceived emotional state during the computer-based writing test was statistically significantly more favourable on the positively formulated items: they were more comfortable (Z=-2.71, p=.007), confident (Z=-2.89, p=.004), and happier (Z=-2.56, p=.011) in the computer-based mode compared with the paper-based mode, albeit with small effect sizes (r=.19, r=.20, r=.18, respectively).

[INSERT TABLE 12 HERE]

Usability. In general, the ISE II test-takers held positive views on the usability of the writing test in both delivery modes, as can be seen from the proportion of participants (strongly) agreeing with the questionnaire items presented in Table 13. When contrasting the two modes, however, more favourable views were found for the computer-based version, with statistically significantly more test-takers finding it easier to: a) type on a keyboard (computer-based mode) than write on paper with a pen (paper-based mode) (Z=-2.15, p=.031; small effect size, r=.15), b) revise and edit in the computer-based mode compared with the paper-based mode (Z=-4.28, p<.001; medium effect size, r=.30), and c) navigate through the computer-based test (Z=-2.26, p=.024; small effect size, r=.16).

[INSERT TABLE 13 HERE]

Fairness. Although the overall majority of ISE II test-takers thought their writing ability was tested well by both delivery modes (see Table 14), more (44%) expected to have scored higher on the computer-based mode (versus 26% thought on the paper-based mode). The participants also expressed a clear preference for taking the writing test in computer-based mode: 61% stated that this would be their preferred mode in future tests.

[INSERT TABLE 14 HERE]

Relationship between perceptions of computer-based mode and computer familiarity. At ISE II level, self-assessed typing proficiency was negatively associated with perceptions of boredom during the

computer-based test (r_s=-.21, p=.036), and positively associated with typing ease (r_s=.20, p=.044) and revision and editing ease (r_s=.23, p=.021).

ISE III (C1)

Impact. At ISE II level, both the descriptive and inferential analyses of test-takers' perceptions of their emotional state during test completion indicate a more favourable evaluation of emotional impact in the computer-based mode (see Table 15). This is noticeable in test-takers' more frequent endorsement of "strongly disagree" on "nervous", "frustrated" and "bored", and of "strongly agree" on "comfortable", "confident" and "happy" in the computer-based mode compared with the paper-based mode. With the exception of the item "bored", the differences were confirmed as statistically significant through a series of Wilcoxon signed rank tests (the effect sizes were small). ISE III test-takers were less nervous (Z=-2.51, p=.012, r=.23) and less frustrated (Z=-2.56, p=.010, r=.23), and more comfortable (Z=-2.99, p=.003, r=.28), confident (Z=-2.45, p=.014, r=.23), and happy (Z=-2.32, p=.020, r=.20) in the computer-based mode compared with the paper-based mode.

[INSERT TABLE 15 HERE]

Usability. In both delivery modes, ISE III test-takers understood what they had to do to complete the writing test. They also generally held positive views on the usability of the writing test in both delivery modes, as can be seen from the proportion of participants (strongly) agreeing with the questionnaire items presented in Table 16. Comparisons between the two delivery modes, however, indicated more favourable evaluations of the usability of the computer-based test. More specifically, more participants found it easier to type on a keyboard (computer-based mode) than write on paper with a pen (paper-based mode) (Z=-2.35, p=.019; small effect size, r=.21), and to revise and edit in the computer-based mode compared with the paper-based mode (Z=-4.36, p<.001; medium effect size, r=.40). Also, some found the test layout clearer in the computer-based mode (Z=-2.40, p=.016; small effect size, r=.22).

[INSERT TABLE 16 HERE]

Fairness. Analysis of the fairness items demonstrated a strong preference for the computer-based format over the paper-based format for writing at ISE III level (Table 17). A Wilcoxon signed-rank test confirmed that the paper-based mode was considered to test writing ability less well compared with the computer-based mode (Z=-2.64, p=.008; small effect size, r=.24). Additionally, half of the test candidates (50%) felt that they had performed better in the computer-based writing mode (versus 22% in the paper-based mode), and two-thirds (67%) would prefer to take the ISE in computer-based mode next time.

[INSERT TABLE 17 HERE]

Relationship between perceptions of computer-based mode and computer familiarity. At ISE III level, self-assessed computer ability was found to be positively associated with feeling comfortable during the computer-based test (r_s=.27, p=.046).

Qualitative feedback

Test-takers were also given the opportunity to describe any usability problems in the two delivery modes or to share other thoughts regarding their test-taking experience. In total, 87 meaningful comments were provided: 21 at ISE I, 33 at ISE II, and 33 at ISE III. Similar comments were made across ISE levels and so the thematic analysis results are reported for the three levels together. The largest number of responses (52) support the more favourable usability evaluations of the computer-based mode found for the closed questions, with participants emphasizing the advantages for editing, speed and neatness of typing, and keeping track of the word count (or the opposite for the paper-based version). Illustrative comments are:

> CB:
>
> *The writing test on the computer was better because I could see the letter counter in every moment* (ISE I)
>
> *I prefer the CB: it's fast and clean. You can change words in any time without putting crosses over the mistakes* (ISE I)
>
> *It's much better to write with a keyboard and to change something, to correct mistakes, and the word count is a great help that allows you to save plenty of time* (ISE III)
>
> PB:
>
> *It is so difficult to edit if you commit mistake or if you have a new and better idea* (ISE II)
>
> *It is difficult to edit what you have written, if you make mistakes, you don't have the possibility to correct them in a 'nice way'* (ISE III)

Other participants more generally questioned the dated nature of a handwritten test (2 comments).

> *In my opinion, PB writing test is an old fashion way to do an exam* (ISE II)

However, two key issues were also reported with the computer-based test: interference from typing noise from neighbouring test-takers (5 comments), and technical problems with the internet connection and hardware (6 comments).

> *When have many people in the same time in the room, typing on keyboard the noise is very loudly and isn't good for concentration* (ISE I)
>
> *At first the internet does not go* (ISE I)
>
> *My keyboard didn't work properly, in particular letter A and S so I had to press them more than once* (ISE I)

Finally, a number of comments specifically concerned the reading-into-writing task (17). A common issue in both modes was the need to go back-and-forth between the reading input texts and the writing space, although this comment was particularly salient in the computer-based format.

> PB:

We need to have a look in the texts and for that, it's hard to find the right pages (ISE I)

CB:

It was uncomfortable to scroll down and up the page every time I had to source some information in the texts while I was writing the article about the previous texts (ISE II)

During the CB writing test we could not underline parts of the text as we could do during the reading part. I would have personally used that function again (ISE III)

5. Discussion

This study first addressed the question of whether there is a difference in test-takers' scores on the ISE writing test suite depending on delivery mode (RQ1). The findings showed that at ISE levels II and III there was no discernible effect of delivery mode on scores at the task level and at the individual rating category level. This would suggest that, on average, an argument could be supported to use the two delivery modes interchangeably at these levels. However there was a clear, though relatively small, effect of delivery mode at the ISE I level on the reading-into-writing task with the paper-based mode found to be easier for candidates. This effect was observed on two rating scale categories:

Reading for Writing:

- Understanding of source materials
- Selection of relevant content from source texts
- Ability to identify common themes and links within and across the multiple texts
- Adaptation of content to suit the purpose for writing
- Use of paraphrasing/summarising

Language Control

- Range and accuracy of grammar and lexis
- Effect of linguistic errors on understanding
- Control of punctuation and spelling

(Trinity College London, 2016, p.39-40)

Given the nature of these two criteria, one possible explanation for the observation of an effect on these specific categories is that the cognitive load required to complete an integrated task at the ISE I level (B1) in the unfamiliar online mode led to costs in terms of linguistic accuracy, and in terms of test-takers' ability to draw effectively on source materials in their response. This explanation is in alignment with previous comparative research (e.g., Noyes, Garland and Robbins, 2004) which has found that computer-based tasks can require more "cognitive workload" for test-takers, particularly those with lower levels of comprehension. It is also congruent with Skehan's "trade-off" hypothesis, which would predict that limited attentional resources in a context of communicative stress (e.g., an unfamiliar delivery mode) could lead to decreased performance in terms of form and complexity in favour of fluency (note that task fulfilment, for example, was not affected by mode). Whether or not this communicative stress was a symptom of a problematic user interface (see Fulcher, 2003) is possible given some of the problems noted in the perceptions data.

Setting aside task type, overall, our findings of delivery mode impact on test scores align with those of Breland and Muraki (2005) and Wolfe and Manalo (2005) who found score differences depending on test-takers' English language ability, i.e. at lower levels of proficiency the paper-based mode was beneficial, whereas at higher levels of proficiency no delivery mode effects were observed.

The second research question sought to triangulate the score comparison findings by taking account of test-takers' perceptions of the two test modes. Findings suggested a general preference for the computer-based mode across all test levels, though with more muted findings for ISE I compared with ISE II and III. With respect to impact, test-takers reported feeling happier in the computer-based mode across all levels, and more comfortable and confident in that mode at ISE II and III. Usability was also rated more highly for the computer-based mode, with test-takers particularly endorsing the ability to revise and edit offered in the online mode. However, ease of navigation – which may have been expected to be more favourably endorsed in the paper-based mode at the ISE I level (according to the analysis of scores) – was in fact more clearly preferred in the computer-based mode at that level. Open-ended responses confirmed the usability advantages perceived in the online version of the writing test, although it is noteworthy that the usability glitches identified mostly concerned navigation in the integrated task. Finally, global judgements of preference and face validity showed that the majority of test-takers at all levels preferred the computer-based mode. While it was not within the scope of the current study to investigate the relationship between perceptions and scores across modes at the level of individual learners, drawing together the findings from both parts of the study we would speculate that, at an aggregate level, positive perceptions of computer-based delivery may work independently of any actual advantage at the score level, and may indeed function as positive *in spite of* small negative effects at the score level for lower-proficiency learners.

Overall, perceptions in the current study reflect those observed by Maycock and Green (2005): test-takers were clearly more comfortable in the computer-based mode and appreciated the additional features which would support their writing. Also, test-takers generally self-rated their computer and typing skills as solid, and their perceptions of the delivery modes were mostly independent of any differences in computer familiarity (RQ2a). This is congruent with the generally high levels of computer familiarity observed among the sample as a whole.

6. Conclusion

This article has investigated the comparability of paper-based and online modes of delivery of the Trinity College London ISE writing suite (Levels I, II and III). Our study is unique in its exploration of the impact of delivery mode on both independent and integrated writing tasks. An additional strength is that we have explored this topic in terms of statistical as well as experiential equivalence by gauging delivery mode effects on test scores and test-taker perceptions.

At a practical level, the study provides useful evidence for Trinity College London to support the context validity of ISE writing exams delivered interchangeably in paper-based and online format, in particular at ISE II and III level, provided technical guarantees. At ISE I level, however, a small effect

of delivery mode on writing scores was found, with test-takers performing slightly better in the paper-based mode. In practice, this impact of delivery mode was situated in the reading-into-writing task. Therefore, before the use of this integrated task type in online mode can be fully supported at ISE I level, further research is required.

In terms of the test-taker experience, our findings show that there is a clear preference for computer-based writing, which is especially pronounced at the higher levels of proficiency. This signals a shift towards typing/writing on screen as being the norm, and handwriting increasingly being the unusual format – in particular for the production of continuous pieces of text or lengthier prose. It also suggests that the applicability of other earlier research (e.g., from the 1990s) may be less relevant to testing contexts today; thus, the continuation of research on test delivery mode effects is pertinent.

It should be noted, however, that the present study was conducted within a European test administration context with participants originating from countries (mostly Italy, Spain and Brazil) in which computer literacy will have increased at a rapid rate in recent decades. Familiarity and practices regarding paper versus online forms of writing (as well as practicalities) may be vastly different in some other contexts, with potential implications for test delivery mode effects. Therefore, additional research is needed to confirm the generalizability of our conclusions and the scope for online ISE delivery in other regions of the world.

Overall, since we found a delivery mode effect on an integrated task at a lower level of test-taker proficiency, our results suggest that a fruitful line for further research would look further at the interaction between task complexity, proficiency level, and delivery mode. In a follow-up phase, we intend to investigate whether differences observed at the score level also manifest within the discourse of the written performances.

References

Breetvelt, I., van den Bergh, H., & Rijlaarsdam, G. (2009). Relations between writing processes and text quality: When and how? *Cognition and instruction, 12*(2), 103-123.

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.

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.

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.

Cumming, A. (2014). Assessing integrated skills. In A.J. Kunnan (ed.), *The companion to language assessment* (Vol.1, pp.216-229). Hoboken: John Wiley & Sons, Inc.

Davey, T. (2011). *Practical considerations in computer-based testing*. ETS White Paper. Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/Media/Research/pdf/CBT-2011.pdf

Endres, H. (2012). A comparability study of computer-based and paper-based writing tests. *Research Notes, 49*, 26-33.

Fulcher, G. (2003). Interface design in computer-based language testing. *Language Testing, 20*(4), 384-408.

Jin, Y., & Wu, J. (2010). A preliminary study of the validity of the internet-based CET-4: Factors affecting test takers' perception of and performance on the test. *Computer-Assisted Foreign Language Education, 2*, 3–10.

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.

Lee, H.K. (2004). A comparative study of ESL writers' performance in a paper-based and a computer-delivered writing test. *Assessing Writing, 9*(1), 4-26.

Lee, H.K. (2004). A comparative study of ESL writers' performance in a paper-based and a computer-delivered writing test. *Assessing Writing, 9*(1), 4-26.

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

Linacre, J.M. (2017). *Facets computer program for many-facet Rasch measurement, version 3.80.0*. Beaverton, Oregon: Winsteps.com. Retrieved from http://www.winsteps.com/facets.htm

Ling, G. (2017). Is writing performance related to keyboard type? An investigation from examinees' perspectives on the TOEFL iBT. *Language Assessment Quarterly, 14*(1), 36-53.

Noyes, J., Garland, K., & Robbins, L. (2004). Paper-based versus computer-based assessment: is workload another test mode effect? *British Journal of Educational Technology, 35*(1), 111-113.

Maycock, L., & Green, A. (2005). The effects on performance of computer familiarity and attitudes towards CB IELTS. *Research Notes, 20*, 3–8.

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.

Owston, R.D., Murphy, S., & Wideman, H.H. (1992). The effects of word processing on students' writing quality and revision strategies. *Research in the Teaching of English, 26*(3), 249-276.

Plakans, L., & Gebril, A. (2017). Exploring the relationship and organization and connection with scores in integrated writing assessment. *Assessing Writing, 31*, 98-112.

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*(3), 220-233.

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics, 30*(4), 510-532.

Stoynoff, S. (2013). Fairness in language assessment. In C.A. Chapelle (ed.), *The encyclopedia of applied linguistics*. Blackwell Publishing.

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.

Trinity College London (2016). *Integrated Skills in English (ISE). Specifications – Reading & Writing. ISE Foundation to ISE III* (3rd ed.). London: Trinity College London.

Van Waes, L., & Schellens. P.J. (2003). Writing profiles: the effect of the writing mode on pausing and revision patterns of experienced writers. *Journal of Pragmatics, 35*(6), 829-853.

Weir, C.J. (2005). Language testing and validation: An evidenced-based approach. Basingstoke: Palgrave Macmilan.

Wolfe, E.W., & Manalo, J.R. (2005). *An investigation of the impact of composition medium on the quality of scores from the TOEFL writing section: A report from the broad-based study*. TOEFL Research Report RR-72. Princeton, NJ: Educational Testing Service.

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. Princeton, NJ: Educational Testing Service.

Yu, G. (2010). Effects of presentation mode and computer familiarity on summarization of extended texts. *Language Assessment Quarterly, 7*(2), 119–136.

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.

.

Table 1: Counterbalanced research design

		ISE test		Mode	
		Form A	**Form B**	**Paper-based**	**Computer-based**
Time 1	Group A	X		X	
	Group B		X		X
	Group C		X	X	
	Group D	X			X
Time 2	Group A		X		X
	Group B	X		X	
	Group C	X			X
	Group D		X	X	

.

Table 2: Participant background information and computer familiarity

		ISE I (n=99)	ISE II (n=101)	ISE III (n=59)
BiODATA	Gender	46% male 54% female	49% male 52% female	37% male 63% female
	Age	13-43 years (M=22.95; SD=7.02)	14-58 years (M=26.05; SD=8.49)	17-52 years (M=26.59; SD=8.63)
	L1	42% Italian 41% Spanish 15% Portuguese 1% Russian	32% Italian 55% Spanish 11% Portuguese 1% French, Polish, Moldavian	42% Italian 41% Spanish 15% Portuguese 2% Polish
	Current activity	40% secondary school 24% university 23% employed 13% other	16% secondary school 34% university 33% employed 16% other	10% secondary school 41% university 36% employed 14% other
	Years learning English	0-29 years (M=7.78; SD=5.72)	0-23 years (M=8.06; SD=5.05)	2-40 years (M=11.90; SD=7.06)
		ISE I (n=96)	**ISE II (n=97)**	**ISE III (n=59)**
COMPUTER FAMILIARITY	Years of computer use	2-30 years (M=11.17; SD=5.88)	3-33 years (M=13.29; SD=5.17)	7-30 years (M=13.86; SD=5.11)
	Computer use frequency	3% couple of times 4% monthly 26% weekly 67% daily	2% couple of times 5% monthly 16% weekly 77% daily	0% couple of times 5% monthly 14% weekly 81% daily
	Perceived computer ability	3% not good 20% adequate 50% good 27% excellent	2% not good 14% adequate 49% good 36% excellent	3% not good 22% adequate 44% good 31% excellent
	Perceived typing ability	0% not good 21% adequate 54% good 25% excellent	1% not good 18% adequate 56% good 25% excellent	0% not good 19% adequate 49% good 32% excellent
	Prior computer-based language test experience	16%	20%	29%

.

Table 3: Summary statistics for MFRM analysis – ISE I

	Test-takers	**Order**	**Raters**	**Rating criteria**
N	107	2	7	7
Measures				
M	0.31	0.00	0.00	0.00
SD (pop.)	1.40	0.16	0.42	0.60
SE	0.46	0.06	0.11	0.11
Infit MS				
M	1.01	1.01	0.96	1.00
SD (pop.)	0.54	0.10	0.14	0.14
Outfit MS				
M	1.00	1.01	0.96	1.00
SD (pop.)	0.56	0.10	0.14	0.14
Homogeneity index (χ2)	993.60*	16.00*	75.30*	226.50*
df	106	1	6	6
Separation (pop.)	2.84	2.65	3.47	5.50
Reliability of separation (pop.)	0.89	0.88	0.92	0.97
Inter-rater reliability				
Observed exact agreement %			56.6	
Expected %			47.4	
Rasch κ			0.17	

*p<.001

.

Table 4: Summary statistics for MFRM analysis – ISE II

	Test-takers	**Order**	**Raters**	**Rating criteria**
N	109	2	7	7
Measures				
M	0.51	0.00	0.00	0.00
SD (pop.)	1.39	0.03	0.25	0.39
SE	0.42	0.05	0.10	0.10
Infit MS				
M	0.98	1.00	1.01	1.00
SD (pop.)	0.52	0.05	0.15	0.16
Outfit MS				
M	0.99	1.01	1.02	1.00
SD (pop.)	0.51	0.05	0.16	0.16
Homogeneity index (χ2)	1099.00*	0.90	42.00*	104.10*
df	108	1	6	6
Separation (pop.)	3.09	0.00	2.11	3.73
Reliability of separation (pop.)	0.91	0.00	0.82	0.93
Inter-rater reliability				
Observed exact agreement %			52.9	
Expected %			45.0	
Rasch κ			0.14	

*p<.001

.

Table 5: Summary statistics for MFRM analysis – ISE III

	Test-takers	**Order**	**Raters**	**Rating criteria**
N	67	2	7	7
Measures				
M	0.21	0.00	0.00	0.00
SD (pop.)	1.18	0.20	0.40	0.34
SE	0.42	0.06	0.13	0.12
Infit MS				
M	0.97	1.00	0.96	1.00
SD (pop.)	0.61	0.06	0.17	0.08
Outfit MS				
M	0.97	1.00	0.95	1.00
SD (pop.)	0.62	0.08	0.17	0.09
Homogeneity index (χ2)	648.40*	18.90*	69.10*	54.00*
df	66	1	6	6
Separation (pop.)	2.55	2.91	2.92	2.60
Reliability of separation (pop.)	0.87	0.89	0.90	0.87
Inter-rater reliability				
Observed exact agreement %			50.8	
Expected %			46.0	
Rasch κ			0.09	

*p<.001

.

Table 6: Pairwise bias report for (1) mode & task, and (2) mode & rating scale category – ISE I

			PB		CB				Rasch-Welch			
	Task	**Scale**	**Meas**	***SE***	**Meas**	***SE***	**Contrast**	***SE***	***t***	***df***	***p***	***d***
MODE & TASK	RIW	/	-0.25	0.08	0.25	0.07	0.50	0.11	4.74	1077	0.000	0.14
	IW	/	-0.09	0.08	0.09	0.08	0.17	0.12	1.45	807	0.148	0.04
MODE & RATING SCALE CATEGORY	RIW	RfW	0.04	0.14	0.52	0.15	0.48	0.20	2.33	267	0.021	0.14
		TF	-0.74	0.15	-0.40	0.15	0.34	0.21	1.61	267	0.108	0.10
		O&S	-1.02	0.16	-0.63	0.16	0.40	0.23	1.73	267	0.085	0.11
		LC	-0.96	0.15	-0.19	0.15	0.77	0.21	3.73	267	0.000	0.23
	IW	TF	0.56	0.14	0.62	0.14	0.06	0.20	0.30	267	0.768	0.02
		O&S	0.09	0.16	0.42	0.16	0.33	0.22	1.49	267	0.136	0.09
		LC	0.76	0.14	0.92	0.14	0.16	0.20	0.80	267	0.424	0.05

.

Table 7: Pairwise bias report for (1) mode & task, and (2) mode & rating scale category – ISE II

			PB		CB				Rasch-Welch			
	Task	**Scale**	**Meas**	***SE***	**Meas**	***SE***	**Contrast**	***SE***	***t***	***df***	***p***	***d***
MODE & TASK	RIW	/	-0.06	0.07	0.06	0.07	0.12	0.1	1.14	1085	0.256	0.03
	IW	/	-0.09	0.07	0.09	0.07	0.18	0.1	1.79	813	0.074	0.05
MODE & RATING SCALE CATEGORY	RIW	RfW	0.18	0.15	0.21	0.15	0.03	0.21	0.15	269	0.880	0.01
		TF	-0.47	0.14	-0.40	0.14	0.07	0.20	0.35	269	0.728	0.02
		O&S	-0.65	0.15	-0.61	0.15	0.04	0.21	0.20	269	0.840	0.01
		LC	-0.27	0.14	0.03	0.14	0.31	0.20	1.54	269	0.126	0.09
	IW	TF	0.15	0.12	0.30	0.12	0.15	0.16	0.90	269	0.369	0.05
		O&S	0.11	0.14	0.28	0.14	0.17	0.19	0.85	269	0.394	0.05
		LC	0.44	0.13	0.69	0.13	0.25	0.18	1.36	269	0.176	0.08

.

Table 8: Pairwise bias report for (1) mode & task, and (2) mode & rating scale category – ISE III

			PB		CB				Rasch-Welch			
	Task	**Scale**	**Meas**	***SE***	**Meas**	***SE***	**Contrast**	***SE***	***t***	***df***	***p***	***d***
MODE & TASK	RIW	/	0.02	0.09	-0.02	0.09	-0.04	0.12	-0.3	757	0.764	0.01
	IW	/	0.02	0.09	-0.02	0.09	-0.04	0.13	-0.29	567	0.775	0.01
MODE & RATING SCALE CATEGORY	RIW	RfW	0.21	0.17	0.00	0.17	-0.21	0.24	-0.86	187	0.389	0.06
		TF	-0.42	0.18	-0.37	0.18	0.05	0.25	0.21	187	0.833	0.02
		O&S	-0.48	0.18	-0.52	0.18	-0.05	0.26	-0.19	187	0.853	0.01
		LC	-0.19	0.17	-0.13	0.16	0.05	0.23	0.23	187	0.817	0.02
	IW	TF	0.24	0.16	0.12	0.16	-0.12	0.22	-0.53	187	0.598	0.04
		O&S	0.29	0.17	0.19	0.18	-0.10	0.25	-0.42	187	0.675	0.03
		LC	0.48	0.16	0.58	0.16	0.10	0.23	0.45	187	0.653	0.03

.

Table 9: Impact across ISE I writing test modes

	Mode	***N***	***f*(%)**				***M* (SD)**	***Z***	***p***	***r***
			1	**2**	**3**	**4**				
Nervous	PB	101	13(13%)	40(40%)	43(43%)	5(5%)	2.40 (.78)	-1.71	.088	.12
	CB	101	18(18%)	43(43%)	33(33%)	7(7%)	2.29 (.84)			
Comfortable	PB	100	4(4%)	28(28%)	52(52%)	16(16%)	2.80 (.75)	-1.59	.133	.11
	CB	99	2(2%)	20(20%)	58(58%)	19(19%)	2.95 (.70)			
Frustrated	PB	96	25(26%)	45(47%)	23(24%)	3(3%)	2.04 (.79)	-1.53	.127	.11
	CB	98	29(30%)	50(51%)	18(18%)	1(1%)	1.91 (.72)			
Confident	PB	99	0(0%)	31(31%)	61(61%)	7(7%)	2.76 (.57)	-1.40	.162	.10
	CB	98	1(1%)	24(24%)	61(62%)	12(12%)	2.86 (.63)			
Bored	PB	98	23(23%)	52(53%)	17(17%)	6(6%)	2.06 (.81)	-1.59	.113	.11
	CB	98	32(33%)	47(48%)	15(15%)	4(4%)	1.91 (80)			
Happy	PB	99	4(4%)	35(35%)	54(55%)	6(6%)	2.63 (.66)	-3.09	.002	.22
	CB	98	2(2%)	26(27%)	57(58%)	13(13%)	2.83 (.67)			

.

Table 10: Usability across ISE I writing test modes

	Mode	***N***	***f(%)***				***M* (SD)**	***Z***	***p***	***r***
			1	**2**	**3**	**4**				
Understood what I had to do	PB	101	1(1%)	2(2%)	67(66%)	31(31%)	3.27 (.55)	-.83	.405	.06
	CB	100	1(1%)	3(3%)	61(61%)	35(35%)	3.30 (.58)			
Easy to write with a pen/keyboard	PB	101	6(6%)	9(9%)	61(60%)	25(25%)	3.04 (.76)	-1.38	.168	.10
	CB	101	2(2%)	12(12%)	50(50%)	37(37%)	3.21 (.73)			
Easy to revise and edit on paper/computer	PB	101	10(10%)	19(19%)	57(56%)	15(15%)	2.76 (.83)	-4.28	.000	.30
	CB	101	0(0%)	12(12%)	47(47%)	42(42%)	3.30 (.67)			
Test layout was clear	PB	101	1(1%)	9(9%)	68(67%)	23(23%)	3.12 (.59)	-1.55	.122	.11
	CB	101	0(0%)	9(9%)	60(59%)	32(32%)	3.23 (.60)			
Easy to navigate	PB	100	5(5%)	13(13%)	62(61%)	20(20%)	2.97 (.73)	-2.06	.039	.15
	CB	101	1(1%)	16(16%)	49(49%)	35(35%)	3.17 (.72)			

.

Table 11: Fairness across ISE I writing test modes

	Mode	***N***	***f(%)***				***M* (SD)**	***Z***	***p***	***r***
			1	**2**	**3**	**4**				
Tested my writing ability well	PB	101	1(1%)	8(8%)	78(77%)	14(14%)	3.04 (.51)	-2.13	.033	.15
	CB	101	0(0%)	13(13%)	57(56%)	31(31%)	3.18 (.64)			

	N	**PB**	**CB**	**the same / no preference**
Scored higher	100	28%	46%	26%
Preference	100	27%	56%	17%

.

Table 12: Impact across ISE II writing test modes

	Mode	***N***	***f(%)***				***M* (SD)**	***Z***	***p***	***r***
			1	**2**	**3**	**4**				
Nervous	PB	102	14(14%)	36(35%)	49(48%)	3(3%)	2.40 (.76)	-1.82	.068	.13
	CB	100	21(21%)	39(39%)	35(35%)	5(5%)	2.24 (.84)			
Comfortable	PB	101	9(9%)	20(20%)	62(61%)	10(10%)	2.72 (.76)	-2.71	.007	.19
	CB	100	1(1%)	20(20%)	53(53%)	26(26%)	3.04 (.71)			
Frustrated	PB	99	23(23%)	56(56%)	19(19%)	1(1%)	1.98 (.69)	-1.15	.252	.08
	CB	99	32(32%)	47(47%)	19(19%)	1(1%)	1.89 (.74)			
Confident	PB	101	4(4%)	28(28%)	62(61%)	7(7%)	2.71 (.65)	-2.89	.004	.20
	CB	99	1(1%)	23(23%)	52(53%)	23(23%)	2.98 (.71)			
Bored	PB	101	25(25%)	57(56%)	15(15%)	4(4%)	1.98 (.75)	-1.82	.069	.13
	CB	100	36(36%)	46(46%)	17(17%)	1(1%)	1.83 (.74)			
Happy	PB	102	8(8%)	35(34%)	56(55%)	3(3%)	2.53 (.69)	-2.56	.011	.18
	CB	100	9(9%)	24(24%)	51(51%)	16(16%)	2.74 (.84)			

.

Table 13: Usability across ISE II writing test modes

	Mode	***N***	***f(%)***				***M* (SD)**	***Z***	***p***	***r***
			1	**2**	**3**	**4**				
Understood what I had to do	PB	103	1(1%)	2(2%)	60(48%)	40(39%)	3.35 (.57)	.00	1.000	0
	CB	103	1(1%)	6(6%)	52(50%)	44(43%)	3.35 (.64)			
Easy to write with a pen/keyboard	PB	102	9(9%)	13(13%)	53(52%)	27(26%)	2.96 (.87)	-2.15	.031	.15
	CB	103	3(3%)	14(14%)	39(38%)	47(46%)	3.26 (.80)			
Easy to revise and edit on paper/computer	PB	102	18(18%)	28(27%)	32(31%)	24(24%)	2.61(1.04)	-4.28	.000	.30
	CB	103	4(4%)	13(13%)	33(32%)	53(51%)	3.31 (.84)			
Test layout was clear	PB	102	2(2%)	6(6%)	63(62%)	31(30%)	3.21 (.64)	-1.21	.225	.08
	CB	103	2(2%)	9(9%)	46(45%)	46(45%)	3.32 (.72)			
Easy to navigate	PB	103	6(6%)	14(14%)	54(52%)	29(28%)	3.03 (.81)	-2.26	.024	.16
	CB	103	1(1%)	15(15%)	40(39%)	47(46%)	3.29 (.75)			

.

Table 14: Fairness across ISE II writing test modes

	Mode	***N***	***f(%)***				***M* (SD)**	***Z***	***p***	***r***
			1	**2**	**3**	**4**				
Tested my writing ability well	PB	102	2(2%)	10(10%)	70(69%)	20(20%)	3.06 (.61)	-1.32	.188	.09
	CB	102	3(3%)	13(13%)	51(50%)	35(34%)	3.16 (.75)			

	N	**PB**	**CB**	**the same / no preference**
Scored higher	103	26%	44%	30%
Preference	103	29%	61%	10%

.

Table 15: Impact across ISE III writing test modes

	Mode	***N***	***f(%)***				***M* (SD)**	***Z***	***p***	***r***
			1	**2**	**3**	**4**				
Nervous	PB	60	9(15%)	25(42%)	19(32%)	7(12%)	2.40 (.89)	-2.51	.012	.23
	CB	59	17(29%)	24(41%)	15(25%)	3(5%)	2.07 (.87)			
Comfortable	PB	57	5(9%)	17(30%)	31(54%)	4(7%)	2.60 (.75)	-2.99	.003	.28
	CB	57	4(7%)	4(7%)	33(58%)	16(28%)	3.07 (.80)			
Frustrated	PB	58	11(19%)	33(57%)	13(22%)	1(2%)	2.07 (.70)	-2.56	.010	.23
	CB	58	18(31%)	33(57%)	6(10%)	1(2%)	1.83 (.68)			
Confident	PB	58	0(0%)	17(29%)	39(67%)	2(3%)	2.74 (.52)	-2.45	.014	.23
	CB	58	2(3%)	7(12%)	41(71%)	8(14%)	2.95 (.63)			
Bored	PB	58	17(29%)	30(52%)	9(16%)	2(3%)	1.93 (.77)	-1.89	.059	.18
	CB	58	19(33%)	34(59%)	3(5%)	2(3%)	1.79 (.70)			
Happy	PB	58	2(3%)	28(48%)	27(47%)	1(2%)	2.47 (.60)	-2.32	.020	.22
	CB	58	4(7%)	17(29%)	30(52%)	7(12%)	2.69 (.78)			

.

Table 16: Usability across ISE III writing test modes

	Mode	***N***	***f(%)***				***M* (SD)**	***Z***	***p***	***r***
			1	**2**	**3**	**4**				
Understood what I had to do	PB	60	1(2%)	2(3%)	29(48%)	28(47%)	3.40 (.64)	-.38	.705	.03
	CB	59	1(2%)	1(2%)	32(54%)	25(42%)	3.37 (.61)			
Easy to write with a pen/keyboard	PB	60	5(8%)	8(13%)	25(42%)	22(37%)	3.07 (.92)	-2.35	.019	.21
	CB	60	2(3%)	1(2%)	27(45%)	30(50%)	3.42 (.70)			
Easy to revise and edit on paper/computer	PB	60	9(15%)	24(40%)	16(27%)	11(18%)	2.48 (.97)	-4.36	.000	.40
	CB	60	1(2%)	6(10%)	19(32%)	34(57%)	3.43 (.75)			
Test layout was clear	PB	60	1(2%)	6(10%)	35(58%)	18(30%)	3.17 (.67)	-2.40	.016	.22
	CB	60	1(2%)	0(0%)	36(60%)	22(37%)	3.34 (.58)			
Easy to navigate	PB	60	3(5%)	10(17%)	33(55%)	14(23%)	2.97 (.78)	-1.12	.262	.10
	CB	60	3(5%)	6(10%)	32(53%)	19(32%)	3.12 (.78)			

.

Table 17: Fairness across ISE III writing test modes

	Mode	***N***	***f(%)***				***M* (SD)**	***Z***	***p***	***r***
			1	**2**	**3**	**4**				
Tested my writing ability well	PB	60	2(3%)	13(22%)	29(48%)	16(27%)	2.98 (.79)	-2.64	.008	.24
	CB	60	0(0%)	4(7%)	34(57%)	22(37%)	3.30 (.59)			

	N	**PB**	**CB**	**the same / no preference**
Scored higher	60	22%	50%	28%
Preference	60	17%	67%	17%

[i] The concept of fairness in this study was restricted to a psychometric dimension: absence of bias in measurement, or "fairness as it pertains to the test itself" and not "as it relates to the social consequences associated with test use" (Stoynoff, 2013, p.1). More specifically, the potential sources of unfairness explored through the questionnaire concerned perceived accuracy of English writing ability assessment as associated with mode of delivery, and preference of delivery mode in relation to future delivery mode offer.