Elsevier Editorial System(tm) for Assessing Writing

Manuscript Draft

Manuscript Number: ASW-D-17-00095R3

Title: Researching the comparability of paper-based and computer-based delivery in a high-stakes writing test

Article Type: SI: CB and PB

Keywords: test equivalence; score equivalence; cognitive validity; computer-based testing of writing; delivery mode; second language writing assessment

Corresponding Author: Dr. Sathena Chan, Ph.D

Corresponding Author's Institution: University of Bedfordshire

First Author: Sathena Chan, Ph.D

Order of Authors: Sathena Chan, Ph.D; Stephen Bax, Ph.D; Cyril Weir, Ph.D

Abstract: International language testing bodies are now moving rapidly towards using computers for many areas of English language assessment, despite the fact that research on comparability with paper-based assessment is still relatively limited in key areas. This study contributes to the debate by researching the comparability of a high-stakes EAP writing test (IELTS) in two delivery modes, paper-based (PB) and computer-based (CB). The study investigated 153 test takers' performances and their cognitive processes on IELTS Academic Writing Task 2 in the two modes, and the possible effect of computer familiarity on their test scores. Many-Facet Rasch Measurement (MFRM) was used to examine the difference in test takers' scores between the two modes, in relation to their overall and analytic scores. By means of questionnaires and interviews, we investigated the cognitive processes students employed under the two conditions of the test. A major contribution of our study is its use - for the first time in the computer-based writing assessment literature - of data from research into cognitive processes within real-world academic settings as a comparison with cognitive processing during academic writing under test conditions. In summary, this study offers important new insights into academic writing assessment in computer mode.

Researching the comparability of paper-based and computer-based delivery in a high-stakes writing test

Sathena Chan*, Stephen Bax, Cyril Weir

Centre for Research in English Language Learning and Assessment, University of Bedfordshire,
Hitchin Road, Luton, LU2 8LE, United Kingdom

*Corresponding author, Tel.: +44 01582 489795
E-mail address: sathena.chan@beds.ac.uk

***Brief author biography**

Dr. Sathena Chan is a Senior Lecturer in Language Assessment at the Centre for Research in English Language Learning and Assessment (CRELLA), University of Bedfordshire. Her research interests include integrated reading-into-writing assessment, cognitive processing of language use, criterial features of written performance, task design and rating scale development.

Prof. Stephen Bax was a Professor of Modern Languages and Linguistics at The Open University. He was an internationally recognised researcher in the areas of technology for language learning and assessment.

Prof. Cyril Weir is the Powdrill Research Professor in English Language Acquisition in the Centre for Research in English Language Learning and Assessment (CRELLA) at the University of Bedfordshire. His current interests include language construct definition, the validation of language tests and the history of English language testing in the UK.

***Highlights (for review)**

Highlights

- The study investigated score equivalence of IELTS AWT2 between PB and CB modes.
- The study investigated test takers' writing processes on the PB and CB IELTS AWT2.
- The examined the impact of test takers' characteristics on CB performance.

1. Introduction

In line with the increasingly key role of technology in all areas of higher education, computer-based (CB) assessment is becoming more and more common in most university disciplines (Newman, Couturier & Scurry 2010). In a similar fashion, many international language testing bodies now routinely use computers for various areas of Academic English writing assessment. In a study to compare ESL writers' performances on pen-and-paper and computer-delivered tests, Lee (2002) noted that test takers now believed a computer test to be more authentic and valid in relation to the target ESL contexts.

The International English Language Test System (IELTS) test, which is one of the most widely used tests of English language proficiency for educational, professional, and migration purposes, does not currently offer computer-based options. However, it seems more than likely, given the increased authenticity and other perceived benefits of CB testing, that in the near future IELTS will need to move towards offering computer-based options alongside traditional paper-and-pencil (PB) modes. In preparation for a possible move towards the CB assessment of IELTS, research was conducted some years ago to investigate differences between the CB and PB testing of IELTS writing (Weir, O'Sullivan, Yan and Bax, 2007). Although that research is still of relevance, in the intervening years students' increased familiarity with computers in both learning and assessment, as well as developments in test delivery technology, necessitate a fresh look at the questions of equivalence the study raised.

McDonald (2002) identified two fundamental types of equivalence which need to be examined when a pencil-and-paper writing test is offered alongside a computer delivered version and the two versions continue to co-exist side by side. The first, score equivalence, relates to the results of the test takers' performance and the concern is whether the scores obtained between the two modes are statistically equivalent and interchangeable. While score equivalence is often considered the most important issue in the delivery mode equivalence research, Mead and Drasgow (1993), who conducted a widely referenced meta-analysis of 159 correlations between paper-based and computer-based scores on writing tests, note that one should not assume that test takers use the same writing processes under different delivery conditions, especially when time-constraints are imposed. A second type of equivalence that needs to be examined, therefore, relates to the underlying construct that is being measured. Given that writing is a cognitively complex and socially situated activity, it is clearly impossible to achieve complete equivalence between the two conditions. However, in the context of direct writing assessment, it is essential to establish that the constructs operationalised by the tests are equally comparable between the two modes and in addition match as far as possible what students are expected to do in the target language use (TLU) domain (Bachman, 1996).

Some research has been conducted to examine the cognitive processes of writers completing IELTS writing tasks (Yu, Rea-Dickins & Kiely, 2011 on AWT1), but evidence of the cognitive validity of IELTS writing between the two modes is lacking, as is any comparison of either with the constructs underlying real life writing activities. Our aim is to examine the extent to which the results of computer-based IELTS, as a direct writing assessment, are statistically equivalent and construct valid as compared to the results from the paper-and-pencil IELTS. We also compare writing in both modes to real life writing in a university setting. The findings will contribute to establishing an evidence base of comparability that is a necessary pre-requisite to the introduction of a CB version of the IELTS writing test.

2. Literature Review

2.1 Computer-based assessment of (academic) writing

In line with the shift towards computer-based academic writing in real life, many international high-stakes language testing organisations are moving towards the CB testing of writing, in some cases, abandoning the PB mode altogether. Cambridge English Language Assessment (http://www.cambridgeenglish.org/) offers CB versions of KET, PET, FCE, CAE and CPE in more than 350 test centres in 64 countries. The TOEFL iBT (https://www.ets.org/toefl/ibt/about) has already been taken in 1355 test centres in 149 countries, with the PB format now being phased out completely. Pearson (https://pearsonpte.com/) offers the PTE Academic test in CB mode only, and states that "more than 27 million test questions making use of this technology have been delivered, responded to, and automatically scored for individuals from over 100 countries around the world" (Pearson 2012, p.7). The British Council has launched a CB test – Aptis (https://www.britishcouncil.org/exam/aptis). As almost all major academic writing assessments offer some forms of CB essay tasks, the momentum towards the need of CB writing test is compelling and those who do not follow in this direction risk being left behind and losing market share. Drawing on McDonald's (2002) work on the impact of individual variables on test equivalence, and on Mead and Drasgow's (1993) meta-analysis, we will now consider two types of equivalence, scoring and cognitive in turn.

2.2 Score equivalence

The literature of score equivalence in writing test between paper-based and computer-based modes presents a varied picture as regards outcomes. Early research by Mazzeo and Harvey (1988) suggested that CB tests at the time tended to be more difficult than PB versions, perhaps partly owing to test takers' lack of familiarity with the technology involved. Some found inconsistent results of the effect of delivery mode on performance. For example, Burke and Cizek (2006), in their study examining eighty 6th grade students, found that score equivalence was dependent on the prompt variable.

However, more recent research show that a CB test may elicit better performance from writers. Russell and Plati (2000) found that grades 8 and 10 students performed significantly better when they composed extended composition items under CB conditions. Wolfe and Manolo (2005) found that scores given to essays written in CB mode are in fact "slightly more reliable than scores assigned to handwritten essays and exhibit higher correlations with TOEFL multiple-choice sub-scores". Goldberg, Russell, and Cook (2003) performed a meta-analysis of 26 writing studies that were conducted from 1992 to 2002 concerning grades K-12 students. The results showed that students produced significantly better texts in terms of quality (effect size=.41, n=15) and quantity of writing (effect size=.50, n=14) under CB conditions. However, it is worth noting that only six of the studies examined the effects of writing mode on revisions, (which will be discussed more fully later), and they yielded inconclusive results.

On the other hand, a large body of more recent research concerning large-scale language tests showed that, depending on appropriate design, the scores across the CB and PB modes can be considered comparable (Puhan, Boughton, & Kim, 2007; Taylor, Jamieson, Eignor & Kirsch, 1998; Wise & Plake, 1989). Taylor et al (1998), studied the comparability of PB and CB versions for the 1996 administration of the TOEFL exam and found no significant differences in score for test takers taking the two different versions. Likewise, Wise and Plake (1989) contended that PB and CB versions of writing tests yield very similar scores. Puhan et al. (2007), who examined over 1000 participants in a test of basic proficiency in reading, writing, and mathematics in CB and PB modes, found no significant difference in scores between the two modes. Based on performance of 262 participants, Weir et al. (2007) reported that the difference between the PB and experimental CB versions of IELTS was not significant.

It appears that, provided that the test design is carefully constructed, score equivalence is achievable across the two modes in large-scale tests of writing. However, the mixed findings especially in relation to writers' varied performance in sub-criteria of writing quality under the PB and CB conditions could suggest that the different modes of the writing test were eliciting different processes.

2.2 Cognitive equivalence

Score equivalence is insufficient in itself to ensure the equivalence of CB and PB test modes. Weir, et al. (2007) argue that to establish equivalence between the same level examinations across testing modes, comprehensive specification of the cognitive processes elicited is as essential as demonstrating statistical alternativeness. Both tests must be developed according to a rigorous specification of cognitive and contextual parameters. The implication of this is that language test providers need to establish for both modes, CB and PB, that the cognitive processes which a candidate draws on when completing the test writing task(s) constitute an equally accurate and comprehensive representation of the types of processing required in writing tasks in the real-world target setting (Glaser, 1991; Field, 2013).

Compared to score equivalence between PB and CB writing, there is rather insufficient research on writers' processes between the two test conditions, especially in the context of language testing. Shaw (2005) reports on earlier studies such as Hermann (1987) which found that the use of a computer interfered with students' composing process. Later studies, presumably as computer use become more commonplace, reported that regular use of word processors for writing over an extended period can lead to significant improvements in the students' writing skills (Owston & Wideman, 1997). Conversely, other research has suggested that the benefits of writing by hand may outweigh those of typing into a computer.

The work of child psychologists such as James (2012), however, throws interesting light on the value of hand writing as against typing in facilitating reading acquisition in young children finding that only writing a letter freehand fully activated the three areas of the brain essential for reading and writing. Berninger (2015) showed that children who wrote by hand, instead of typing on a keyboard, were better at generating composition ideas and experienced greater neural activity in the areas of the brain associated with reading and writing while doing this. Mueller (2014) found that “laptop note taking is less effective than longhand note taking for learning ... ... students who took notes on laptops performed worse on conceptual questions than students who took notes longhand... whereas taking more notes can be beneficial, laptop note takers’ tendency to transcribe lectures verbatim rather than processing information and reframing it in their own words is detrimental to learning (ibid, p.1)” Similar research on adults processing in the two modes would be a welcome addition to our knowledge base.

In a study comparing composing processes on an ESL placement writing test between the two conditions, Lee (2002) did not find significant difference in how test takers composed in the two tests. Cochran-Smith (1991) conducted a review of the literature comparing paper-based and computer-based writing in primary classrooms. The findings suggested that the mode CB writing itself does not lead to better overall quality of writing, but they noted that, under the CB conditions, students tend to make more revisions, and to produce longer, neater, more error-free texts. This finding is echoed in Grejda and Hannafin’s (1992) study in which students engaged in more mechanical revisions when composing under the CB conditions. However, in other studies, e.g. Haas (1989) it appears that when writing in PB conditions, writers were less hurried in generating text and this lead to better grammar and better mechanics. Due to the conflicting results regarding the underlying processes elicited by the two delivery test modes shown in the literature, researchers have, in time honoured fashion, stressed the need to investigate the issue further (Burke & Cizek, 2007).

2.3 Cognitive validity of IELTS and the theoretical model of Writing

Writing is an activity comprising a number of major phases of cognitive processing, e.g. planning, generating ideas, execution (translating ideas into words), organising, monitoring and revising (see Hayes and Flower, 1983; Kellogg, 1996; Shaw &Weir, 2007; Weigle, 2002). Writers may use multiple processes within a phase. For example, during the planning phase, a writer would typically ‘read the task prompt’, ‘set writing goals’ and ‘plan contents and structure’. It should be noted that while most writers compose following the general order of these cognitive phases, they often employ individual processes across different cognitive phases. For example, writers may evaluate and adjust their writing goals at the revision phase. While most researchers do not make the distinction between cognitive ‘phases’ and ‘processes’, for the purpose of examining the individual processes operationalised by IELTS between the two delivery modes, we consider the six cognitive phases and thirteen cognitive processes established in Chan (2013) (presented in Table 1) as the baseline of the target construct of IELTS Writing.

The cognitive processes elicited by IELTS Academic Writing Task 1 (AWT1) (in PB mode) was investigated in detail by Yu et al (2011). Using the think aloud approach for the main part of their study, they concluded the study by offering a model of cognitive processes consisting of three interrelated stages, specific to AWT1. They did not explicitly compare the test takers’ processes with those used in tasks in the target language use domain, although there seems to be an implicit assumption that the cognitive processes they examined under test conditions were in general of a kind relevant to what students are expected to do in the real-life academic writing situations. They were correct to report that Academic Writing Task 2 (AWT2) has received more research attention in general in support of their decision to look at AWT1. However, the cognitive processing of test takers taking AWT2 in CB mode has not previously been researched – an important gap in the research base if the IELTS writing test is to be computerised in future.

The research reported in this paper builds on the work of Yu et al (2011) by researching the cognitive processes of test takers completing AWT2 in PB mode but will extend it also to examine the processes used in CB mode, so as to investigate the cognitive equivalence of the two modes. The study will furthermore compare these cognitive processes with those reported by second language (L2) students in one dominant TLU domain of IELTS, i.e. academic writing at a UK university, to help to establish the cognitive validity of IELTS Writing in both modes.

Chan (2013) sampled two academic writing tasks, an essay and a report task, based on criteria established in the academic writing literature (e.g. Bridgeman & Carlson, 1983; Hale, Taylor, Bridgeman, Carson, Kroll & Kantor, 1996). The two sample real-life tasks were examined in terms of key contextual parameters, including, *purpose, time and length, topic domain, genre, interaction between input and response, language functions* and *intended reader*. The results showed that the sampled tasks resemble the characteristics of typical academic writing tasks as reported in previous comprehensive academic task survey studies (e.g. Bridgeman & Carlson, 1983; Carson, 2001; Horowitz, 1986a, 1986b; Johns, 1993; Leki & Carson, 1994. Chan (2013) then examined the cognitive processes of 200 L2 writers employed to complete the two sampled real-life tasks. As mentioned at the beginning of this Section, five key cognitive phases of composition were identified, namely, *conceptualisation*, *generating ideas*, *organising ideas*, *generating texts*, and *monitoring and revising* (see Table 1).

The results showed that students who scored higher on the real-life writing tasks reported employing most of the thirteen cognitive processes (in the five cognitive phases) more often than the low-scoring students. This suggests that the processes identified could be considered as appropriate cognitive parameters for evaluating academic writing tests. With the exclusion of a number of processes: *careful reading* and *scanning, skimming and search reading*, which relate specifically to the reading texts which served as input in the study, this list provides a useful baseline for the present study as to the cognitive processes which L2 writers in real academic contexts are likely to employ. We sought to determine the extent to which these processes are mirrored in the PB and CB versions of the IELTS AWT2 tests.

2.4 Impact of writers' computer familiarity on performance

Delivery mode has always been identified as one of the variables which might potentially have an impact on writers' performance (Weigle, 2002). Although the use of computers in academic writing has become very common, there is some residual concern that some test takers might be disadvantaged by unfamiliarity with computers. Most studies (e.g. Al-Amri 2008;

Russell, 1999; Shermis & Lombard 1998; Taylor, Jamieson, Eignor, & Kirsch, 1998; Taylor, Kirsch, Eignor & Jamieson, 1999) did not find that writers’ computer familiarity or anxiety has a significant impact on performance, at least not in a way directly observable by test scores.

On the contrary, studies (e.g. Russell, 1999) seemed to find that writers with a positive attitude towards the use of computer in writing tended to write more enthusiastically on computers, e.g. writing more extensively and revising more carefully in class. Weir et al. (2007) took careful account of three pertinent variables, namely computer familiarity, computer anxiety and computer attitudes, and found that the effect of these on performance was mostly negligible.

Although the impact of these variables seems to be far less powerful than might have previously been expected, Taylor et al (1998, 1999) stressed the importance of providing support e.g. a computer tutorial, to test takers as part of test preparation. Other researchers continue to press for more studies to investigate how these variables might affect writers’ performance before any final conclusions are drawn on the presence or absence of any impact (Hertz-Lazarowitz & Bar-Natan, 2002; McDonald, 2002).

2.5 Research Questions

1. Are there differences in the scores awarded to test takers’ writing performance according to delivery mode?
2. Are there differences in the cognitive processes test takers’ report as using according to delivery mode?
3. Are test takers’ performances in the computer-based mode impacted by specified affective variables associated with computer familiarity, usage and attitudes?

3. Research Methods

A mixed-methods design (Creswell & Plano Clark, 2011) was used. All participants completed two tests, one under the traditional PB mode and one in the experimental CB mode (see Section 3.3.1). Before the test event, all participants completed a Computer Familiarity Questionnaire (see Section 3.3.2). They also completed two Writing Process Questionnaires (see Section 3.3.3), each immediately after they had completed the PB and CB tests.

Embedded within the test study, qualitative data was collected in the form of an individual retrospective interview with participants (20%) where they described their writing processes under the two conditions. The results provide evaluation of both the outcomes (i.e. scores doubled rated by certified raters – see Section 3.2) and processing activated according to delivery mode. Figure 1 presents a summary of the research design in relation to data sources and analysis.

3.1 Participants

One-hundred and fifty-three test takers studying on undergraduate programmes at a British University participated in the study; 45.4% of them were male and 54.6% female. At the time of the study, all participants had a valid IELTS score, i.e. taken within 2 years. Their IELTS Writing Bands ranged from 4.5 to 8, see Table 2. Students who were required to attend pre-sessional English classes (i.e. those who had an IELTS overall scores 5.5 or below) were also recruited. They came from several major subject areas, including Business and Finance, Language and Communication, Science and Technology, and Social Sciences.

3.2 Raters

Four certificated, experienced IELTS raters (Raters A, B, C and D) participated in the study. All scripts were double marked using the confidential version of the rating scale. Rater A marked all the scripts whereas Raters B, C and D each double marked a sub-set of the scripts. The prerequisite checks of raters' reliability and severity are reported in Section 3.4.1.

3.3 Data sources

3.3.1 Test tasks and writing performances

All 153 participants completed two tests, one under the traditional PB mode and one in the experimental CB mode in a counter-balanced design. In CB mode participants composed the essay using Microsoft Word. All proofreading functions in the CB mode (e.g. grammar and spell check) were disabled. The research team selected eight versions from a pool of 20 retired IELTS Academic Writing Task 2. The eight were then examined by a panel of six experienced language testing practitioners. The two versions (Prompts 1 and 2) (see Appendix A), which were most comparable in terms of topic, domain, language functions, and expected output, were used in the study. Statistical analyses of the comparability of the two versions are presented in Section 3.4.1.

A total of 15 test sessions were conducted. Participants first completed ethics procedures, and then were divided at random into two groups. Each group, in a counter-balanced order, completed two AWT2 tests (Prompts 1 and 2) on paper and computer. The order of the version was also counterbalanced in alternate test sessions. Each test was 40 minutes long. The arrangement of the sessions is presented in Table 3. No breaks were provided.

3.3.2 Computer Familiarity Questionnaire

All participants completed a Computer Familiarity Questionnaire (see Appendix B) about their computer usage, comfort, perceived ability and interest in using computers (see Table 3). The questionnaire developed in Weir et al.'s (2007) study was deemed still generally fit for purpose by a focus group, but was slightly modified in a few respects to bring it up to date with current situation. For example, a new item (Q5) on participants' experience in taking writing tests in the two delivery modes was added. The version used in this study consists of fourteen Likert scale questions and one open-ended question about their preference of the delivery mode.

3.3.3 Writing Process Questionnaire

All participants completed two Writing Process Questionnaires, each immediately after they had completed the paper-based or computer-based tests (see Table 3). They were made aware that their responses to the Questionnaires would not have any bearing on the scores of the tests they had just completed. The Writing Process Questionnaire, was developed in Chan (2013) to examine the processes students use to complete real-life academic writing tasks. The questionnaire was developed based on models of writing in the literature (e.g. Hayes and Flower, 1983; Kellogg, 1996; Shaw and Weir, 2007) to aid the students to self-report the writing processes they used on a writing task. While for the most part the items in the 2013 questionnaire were deemed appropriate for this new related study by a focus group convened for this purpose, a few items, e.g. those about the processes of reading multiple sources, were deleted as they were seen as irrelevant to the IELTS AWT2 writing task. It was piloted with about 100 students. The validity of the questionnaire was then established with over 300 students (Chan, 2013; Chan, Weir & Wu, 2014). The internal consistency reliability of items, examined by correlational analysis, was satisfactory. The underlying structure of the questionnaire, i.e. distinct processes measured, was examined by Exploratory Factor Analysis. Based on the results, the final categories of the items are shown in Table 1. As a result, the new version contained a total of 40 Likert scale items (see Appendix C). The internal consistency reliability of items assigned to each cognitive phase, i.e conceptualisation, generating ideas, organising ideas, generating texts, monitoring and revising (low-level) and monitoring and revising (high-level), was examined again in this study. The figures ranged from $r(151) = .61$, $p < .01$ to $r(151) = .90$, $p < .01$, which indicates that the items of each cognitive phase were measuring a same construct. It should be noted that, while care has been taken to establish the reliability and validity of the questionnaire, it was designed to capture only limited aspects of cognitive processing as reported by the students in their retrospective accounts.

3.3.4 Interview

All participants were invited to participate in the interview. Thirty participants (20% of the total population) were randomly selected from those who expressed an interest. They were interviewed about their writing processes individually by the research team immediately after each test event. The average of their PB (M=5.80, SD=0.49) and CB (M: 5.80, SD: 0.55) bands were the same but the standard deviation of their CB band was slightly higher. Most of the interviewed participants had the same band under the two conditions. 16.7% had a difference of half a band, and 13.4% a difference of a band. Therefore, the interviewees' performances across the conditions were considered to be reasonably equivalent. All interviews were audio recorded, and the recordings were transcribed by two research assistants. 10% of the transcripts were double checked by a member of the research team to ensure accuracy (for data analysis, see Section 3.4.2).

3.4 Data Analysis

3.4.1 Score analysis

Test takers’ scores awarded under both the paper-based and computer-based conditions were compared by two sets of Multi-Facet Rasch Measurement (MFRM) analyses using

FACETS 3.71.2 (Linacre, 2013). The data were entered using the Rating Scale Model (RSM), which operates under the assumption that the rating scale associated with each category functions similarly.

Rasch logit scale and the Infit Mean Square index as a measure of fit (i.e. meeting the assumptions of the Rasch model) were used to analyse raters' reliability and severity. As mentioned in Section 3.2, Rater A rated all scripts whereas Raters B, C and D each rated a sub-set of the scripts as the second rater. The exact agreement between the first and second rater was 66.8%. As indicated by the Logit measure in Table 4, Rater B and D were more lenient than Rater A whereas Rater C was harsher than Rater A. Nevertheless, the difference in fair mean among the four raters was within 0.2, i.e. within half an IELTS band. Infit values for all the raters fall within the acceptable range between 0.7 and 1.3[1] (Bond and Fox, 2007). Therefore, the doubled marked scores reported in this study are considered reliable.

Table 5 reports the results of prerequisite analysis of the comparability of the test prompts used in the study. Judging by the observed mean and logit measure, Prompt 1 was significantly more difficult than Prompt 2 ($X^2=77.6, p<0.01$). However, while Prompt 1 was more difficult than Prompt 2, the differences in both the observed and fair mean scores of the two prompts were 0.25 or less. In other words, the differences were within half an IELTS band. After rounding, both the observed and fair mean scores of the two prompts would be the same, i.e. 5.5. In addition, as described in Section 3.3.1, the administration of versions was counter-balanced, any order effects being minimized. Therefore, we have confidence that test-version effect should not invalidate the findings of this study.

After we have confirmed that raters' reliability and severity, and the comparability of the task prompts was viewed as satisfactory, we analysed the data to answer RQ1. First, a 5-facet analysis with test takers' writing ability, delivery mode, essay topic, raters and rating category was conducted to examine the impact of each of the above facets on scores, and to compared test takers' overall scores between the two delivery modes.

Furthermore, to compare test takers' scores on each analytic rating category (i.e. Task Achievement, Coherence and Cohesion, Lexical Resources, and Grammatical Range and Accuracy) between the delivery modes, four 4-facet (i.e. test takers' writing ability, essay topic, raters and rating category) analyses were conducted. While delivery mode was not designated as a facet, the four analytic categories between the modes were treated as separate items, e.g. CB Task Achievement and PB Task Achievement. This allowed us to compare the four pairs of analytic scales between the delivery modes.

3.4.2 Cognitive equivalence between CB and PB mode

[1] Although Infit values in the range of 0.5 to 1.5 are considered ‘productive for measurement’ (Wright and Linacre 1994), a stricter range between 0.7 and 1.3 is preferred as the acceptable range of the Infit value in many studies (Bond and Fox, 2007). Given that IELTS is a high-stakes test, we refer to the latter as the acceptable range in this report.

Test takers' responses to the Writing Process Questionnaire under the paper-based and computer-based conditions were computed and analysed using SPSS (ver. 22) Descriptive statistics of individual questionnaire items were obtained. As the data of most items was not normally distributed, non-parametric Wilcoxon signed-tank tests were used to compare the results of the two modes, see Section 4.2.

To establish the extent to which the constructs measured by IELTS are relevant to the TLU domain, e.g. academic writing at a British university, the results in this study were compared descriptively to the findings reported in Chan (2013) with regards to undergraduates' cognitive processes on sampled academic writing tasks in real-life, see Section 2.3 for a review of the study. Since no inferential statistics were performed, the results should be interpreted with caution.

The purpose of the interview was to gain insights of the similarities and differences in test takers' processes under the two conditions. Based on the writing model presented in Table 1, the 30 transcripts were coded into one of the six key writing processes using NVivo v10 (see Appendix D for samples of coding). After that, the coded transcripts were classified as similar or different between the two test conditions. The data was second coded by a research assistant who was familiar with the writing model. The overall agreement rate was above 96%. Any discrepancies between their analyses were discussed until agreement was reached for every case.

3.4.3 Multiple Regression analysis of the impact of affective variables on CB performance

To reveal test takers' familiarity with computer and other relevant affective variables[2] in relation to their use of computer, descriptive statistics were calculated for the responses of participants who chose the options of definitely agree/always and mostly agree/often for each item in the Computer Familiarity Questionnaire (CFQ). The results of this study were compared descriptively to those obtained in Weir et al. (2007).

Furthermore, Multiple Regression analysis was used to examine if any of the affective variables influence test takers' CB test performance. After confirming that the data met the prerequisites for the analysis (including normality, homoscedasticity, linearity, no multicollinearity and no outliers), the items were submitted to Multiple Regression analysis. Stepwise method which includes or removes one independent variable at each step, based on the probability of F, was chosen.

The results from the above analyses of test-takers' scores, test takers' processes (questionnaire and interview data), test takers' computer familiarity and the impact of affective variables on computer-based performance were all taken into consideration to provide a more comprehensive examination of the equivalence of the writing test between the two delivery modes. The multiple sources of empirical data allowed us to examine the equivalence of the two delivery modes in relation to three key dimensions of test validity (Shaw & Weir, 2007), including test takers' characteristics, cognitive and evaluative (scoring) validity.

[2] Affective variables refer to those related to students' attitudes and familiarity with the computer delivery mode.

4. Results

4.1 Score equivalence between the PB and CB modes (RQ1)

After establishing raters' reliability and severity, and the comparability of the prompts in Section 3.4.1, we now report findings from the 5-facet MFRM analysis in relation to score equivalence between the paper-based and computer-based modes. After that, we report findings from the 4-facet MFRM analyses to compare individual analytic scores between the two modes.

To address RQ1, *are there differences in the scores awarded to test takers' writing performance according to delivery mode,* Table 6 presents the results of the delivery mode measurement. As indicated by the fixed chi-square statistics, test scores awarded under the paper-based and computer-based conditions were not statistically different in terms of difficulty ($X^2=1.8$, $p=0.18$). Test takers' performance under the PB and CB conditions in terms of both observed mean and fair mean scores were very close, with a difference of 0.12 in observed mean and 0.03 in fair mean. In addition, the lack of misfit data indicates that test scores obtained from the PB and CB delivery modes can be put on a common Rasch scale. The graphic representation of the placement of the two modes on a common Rasch scale is presented in Figure 2.

According to Tables 7-10, as indicated by the fixed chi-square statistics, differences in three of the four analytic scores (i.e. *Task Achievement, Coherence and Cohesion*, and *Grammatical Range and Accuracy*) between the two modes were not significant. However, the fair mean of the *Lexical Resources* was 0.07 ($X^2=8.2$, $p<0.01$) significantly higher under the PB than CB conditions (see Table 9). In real terms the difference in the *Lexical Resources* was very small and it did not contribute to a significant difference in test takers' overall scores between the two modes. Nevertheless, it is worth noting that the fair mean of *Lexical Resources* was below 6.0 in the CB mode but above 6.0 in the PB mode. It is therefore recommended that the test provider should monitor closely test takers' performance on *Lexical Resources* between the two modes. The interview data reported later provides some insight why test takers performed slightly better in *Lexical Resources* when they took the paper-based test.

4.2 Cognitive equivalence between CB and PB mode (RQ2)

Having established score equivalence between the two delivery modes, we now turn to the analysis of test takers' processes elicited by the test under the two conditions. We first report findings from the Writing Process Questionnaire, followed by the interview data.

4.2.1 Statistical evidence

As presented in Section 3.3.4, participants were asked to rate the extent to which, on a Likert scale of 1 to 4, they employed each of the 40 writing processes on the test immediately after completion of each of the paper-based and computer-based tests. The mean difference of test takers' response to all items between PB and CB modes are presented in Figure 3. The general tendency of the mean of each questionnaire item reported under the two test conditions was comparable. Most differences were 0.15 or below out of a 4-point scale. This indicates that, according to test takers' own perceptions, they employed the cognitive processes similarly under the two conditions. It should be noted that Item 30 showed the largest discrepancy (0.35) between the two delivered modes. Test takers checked the accuracy and range of the sentence structures of their writing more on the paper-based than computer-based test.

To examine the extent to which the constructs of IELTS Writing are relevant to the target language use domain, the findings from Chan's (2013) study of undergraduates' processes in completing academic writing tasks at a British university are provided as a baseline reference. The means in the six cognitive phases obtained in this study on PB and CP IELTS were largely comparable to those reported in the TLU domain, see Table 11. All differences between the test and TLU conditions appear to be very small (ranging 0.06 to 0.23). The most noticeable difference was obtained in the processes of monitoring and revising. The implications of this finding are discussed in Section 5.

After establishing the cognitive validity of IELTS Writing in relation to what writers do in a real-life academic context, we compared the results between the paper-based and computer-based modes. The findings in Table 11 show that the means of each cognitive phase obtained under the test conditions (both CB and PB) were between 3.17 and 3.40 (4=definitely agree; 3=agree; 2 disagree; 1=definitely disagree). The means in *conceptualisation*, *generating ideas*, *generating texts* and *organising ideas* were very close between the two modes, and the means in *monitoring and revising at low-level* were the same (see Table 11). Nevertheless, the participants reported doing slightly more *monitoring and revising at the high-level* under the computer-based than paper-based conditions. Given the practical difficulties of doing this in PB mode this is perhaps not too surprising and an obvious benefit of CB mode. The obtained differences were then subjected to Wilcoxon Signed Ranks Tests (Table 12). The results show that differences in test takers' reported use of the six writing processes between the PB and CB modes are not significant.

4.2.2 Descriptive evidence

As mentioned in Section 3.3.4, one-fifth of the participants (n=30) were interviewed. Drawing from the Writing Process Questionnaire (Appendix C) data, which was discussed in Section 4.2.1, and the interview data (see Appendix D), we now present a phase by phase description of the target cognitive processes elicited by the IELTS task under the two delivery conditions.

Conceptualisation

This is usually the initial phase when writers create a mental representation of the task and set macro-plans for their writing. There was not much difference in the way the participants reported how they approached the test under the two conditions. All interviewed participants began by reading the task prompt and instructions carefully, and planned what and how they were going to write to fulfill the task requirements. All participants were familiar with the test and did not have any difficulties understanding the instructions. Most participants planned mainly about the content and structure of their essay, though as many participants reported that the 'IELTS essay structure came automatically' to their planning.

However, a few differences in test takers' planning between the conditions emerged from the interview data. Participants tended to be more cautious with their planning under the paper-based than computer-based modes. Many reported strategies of producing a writing plan or listing the key ideas. It is interesting to note that most participants stayed very closely to this initial plan as they produced their essay on paper. One participant reported that he 'restricted his writing' to a neat four-paragraph essay structure, each containing a main idea, as previously taught. They were quite reluctant to make 'major changes' to their essay on paper. They believed the evident changes would lead to a lower mark due to untidiness. This concern reoccurs later in other phases. In contrast, participants were relaxed with their initial planning under the computer-based mode. They believed they did not need to start with a perfect plan because they felt more comfortable making changes to the plan or to the essay, processes which CB mode facilitates.

Generating and organising ideas

There was no noticeable difference reported by the participants about how they generated ideas for the essay between the two modes. Most of them appeared to generate ideas in an order following the structure of the essay. For example, one participant described how he generated a starting point for the introduction, one supporting idea, one opposing idea and a conclusion. About one-third of the participants explained that as they were familiar with what was required in IELTS, they just 'followed the flow' and 'ideas would come as they write'. About half of the interviewed participants mentioned that they drew upon their personal experience especially about the situations in their own country when generating ideas. The only difference was that under the PB condition, a few participants added points to their initial planning notes.

While test takers generated ideas using comparable processes, e.g. generating ideas following the structure of the essay or generating ideas from their knowledge of the topic, under the two conditions, we observed a few differences in how they organized ideas to achieve the writing purpose. According to students' verbal accounts of their processes in the retrospective interview, on the PB mode, they tended to organise their ideas at the whole text level according to the structure of their essay, i.e. the main purpose of each paragraph. Participants did this too under the CB mode but they tended to engage more in organising ideas at the levels of sentences and paragraphs. Some examples included 'prioritising ideas within a paragraph', 'distinguishing main ideas and support details', 'removing weaker or repetitive ideas', 'moving things around into a better order' and 'swapping order of sentences'. Such organizing processes sometimes overlapped with the online editing processes as they re-organised/edited the order of their clauses and sentences. It should be noted that the description represents the cognitive processes employed by the students based on self-report data (i.e. retrospective questionnaire and interview). Additional evidence from students' scripts should be analysed in future studies to triangulate the findings, i.e. to examine the extent to which the use of these processes results in any distinctive linguistics features in students' final products.

Generating texts

This is a phase when writers translate their mental ideas into words. On the paper-based mode, they execute this process via writing with a pen, whereas they type in the computer-based mode. A few participants were concerned that their typing speed or accuracy was not as good as their hand-writing, or vice versa. However, their concerns were not reflected in the score bands they received between the two modes.

Apart from the obvious difference in writing mode, the participants revealed some interesting differences in generating texts between the two. As mentioned previously, most participants in this study were reluctant to make changes to their essay under the paper-based condition. They reported that they were more careful when generating texts when composed on paper. Some described how they would think more carefully with their choice of words and sentence structures. On the contrary, in CB mode, they tended to focus more on 'getting the ideas out' during this phase, and they would make changes as they saw appropriate or at a later phase.

Monitoring and revising (online and post-writing/ at low- and high-levels)

More than one-third of the interviewed participants (35%) reported that they did not revise during writing in the PB condition as they felt it was inconvenient to make changes to existing hand-written texts. In comparison, only 25% reported that did not revise during writing in the CB mode. Similarly, slightly more participants reported engaging in post-writing monitoring and revising in the CB than PB modes (i.e. 70% vs 60% of the interviewed participants).

Broadly speaking, writers monitor and revise at two levels (Bereiter & Scardamalia, 1987). Monitoring and revising at the low level tend to be conventional, rule-governed and language-bound. In contrast, monitoring and revising processes at the high level tend to be driven by an awareness of the writing goal and hence meaning-bound. According to the interviews, participants tended to focus more on phrasing at the word level (e.g. to replace a previously used word to avoid repetition) and on correcting grammatical mistakes in the PB mode. In the CB mode, more participants reported making changes at the levels of clauses and sentences to improve coherence or argument. It should be noted that these findings only reflect the changes which the participants were aware of making, and do not necessarily reflect the actual changes they made. In future studies, textual analysis could be used to analyse the actual changes made by the writers to confirm the findings.

Based on the questionnaire and interview data, we have compared test takers' writing processes elicited by the test under the two conditions. We will now examine the possible impact of test takers' familiarity with computer on CB test performance.

4.3 Affective variables and their impact on CB test performance

The Computer Familiarity Questionnaire (CFQ) (see Appendix B) was administered to investigate affective variables of test takers' computer use in terms of *computer usage*, *comfort & perceived ability* and *interest in computers*. Based on the frequency data, a descriptive summary is provided in Table 13. Across the board findings, a clear majority of the participants reported using computers frequently at home and university for a variety of purposes, including surfing the Internet, electronic communications, study-related activities and, to a less extent, games and graphics.

When compared to the findings in Weir et al.'s (2007) study, participants nowadays, at least in the context of this study, appear to be more familiar and comfortable with using computers than eight years ago (see Table 13). There is a remarkable increase in the percentage of participants who have frequent access to computers and use them at home. Many more participants (i.e. an increase of 36.3%) frequently use computers and word-processing for study-related activities than in the previous study. Also, many more participants are comfortable in writing an essay (i.e. 23.1% more) and taking a test (i.e. 28.2%) on computer now, as compared to then. But interestingly, there is seemingly a slight decrease in participants' interest in computers while the computer has clearly become a necessity for study/work.

We next report which, if any, of these aspects of participants' familiarity with computers appear to have an impact on their performances on computer-based writing test. The Pearson correlation analysis established that there was a significant positive correlation, ranging from $r(120)=.176$, $p<.01$ to r(120)=.406, $p<.01$, between 10 CFQ items and students' CB test performance. Using the Stepwise method, a multiple regression analysis of CB score was performed on these 10 CFQ items. The analysis shows that only three items are useful to predict participants' performance on the computer-based task.

As shown in Table 14, frequency of using computers for word processing (CFQ4b) (β =.37, t=4.50, p < .01), access to computers at public library (CFQ1c) (β =.17, t = 2.08, p < .05), and forgetting time when using computer (CFQ13) (β = .17, t = 2.02, p < .05) significantly predicted test-takers' scores in the CB mode. These three variables (CFQ1c, CFQ4b and CFQ13) together explained 22.6% of the variance of the scores in the CB mode, indicating a low level of predictive power. In other words, participants in this study who had frequent access to computers at public places, who frequently used computers for word processing, and those who would forget the time when working with the computer performed significantly better, though the degree is slight, on the computer-based test.

5. Discussion

The findings of this study offer a useful addition to the equivalence debate by widening the normally accepted definition of equivalence solely based on scores to cover the cognitive processes elicited by the test under two delivery modes. The findings are also the first in the test equivalence literature to take account of reference data on writing processes in the real-world target context.

5.1 Test equivalence of IELTS Academic Task 2 between the two delivery modes

The discussion over whether a different delivery mode would result in higher scores is clarified by our results, particularly in the context of high-stakes language tests like IELTS. The score data supports previous studies which found no significant difference in test takers' overall scores between the two modes (e.g. Neuman & Baydoun, 1998; Puhan et al. 2007; Taylor et al., 1998; Wise & Plake, 1989).

However, the literature suggests that score equivalence is likely to be dependent on several variables, including the prompt and rater effect (Burke and Cizek, 2007; McNamara, 2012). Therefore, in addition to ensuring the equivalence of task difficulty between the two prompts, counter-balancing the order of the prompts and order of mode in test delivery, we consider it necessary, in this kind of research, to use Multi-Facet Rasch analysis to investigate the effects of the above multiple facets on test scores and eliminate ambiguity (see Lynch & McNamara, 1996).

While participants' performance on IELTS Academic Task 2 can be considered equivalent between the two delivery modes, as no significant difference was observed between the two scores, it should be noted that participants in this study achieved slightly highly scores in *Lexical Resources* when they composed on paper. Although the difference in this single analytical criterion did not lead to a significant difference in students' overall scores, this might imply that some writing sub-constructs are being elicited slightly differently under the two modes.

5.2 Cognitive equivalence of IELTS Academic Task 2 between the two delivery modes

The quantitative questionnaire data show no significant differences in the use of key writing processes between the two delivery modes. This echoes the findings in the literature that test delivery mode does not necessary alter test takers' writing processes (Lee, 2002). Nevertheless, the qualitative interview data reveal some subtle differences in how test takers employed these processes between the two modes, which might have been overlooked if only a single quantitative data source had been used. Some of the differences in *planning*, *generating texts* and *monitoring and revising* (see Table 15 for a summary) might have an impact on test takers' performance, though such influence might not be reflected in the final scores. For example, an urge to produce 'perfect language' at the first attempt under the PB condition may have an important implication for writers at lower proficiency level as their executing process is disturbed by grammatical checks. However, as their proficiency in writing and/or L2 linguistic knowledge is still at a developmental stage, they are not likely to execute multiple processes successfully at the same time. The relationship between processes and performances was beyond the scope of this study, but should be investigated in future studies.

The findings also help to address the concern that writing in IELTS Academic is not very similar to writing in academic contexts. The results that all means under both the IELTS AWT2 PB and CB test conditions were slightly higher than the real-life figures yield some positive evidence supporting the cognitive validity of test. While it can be considered desirable for test performances to be equivalent to real-life writing activities, it should be noted that in the IELTS test, most test-takers are aware that they are being judged on language rather than on content (unlike in a 'real-life' university context). This might explain why the test-takers were doing more monitoring and revising than the undergraduates on their academic writing tasks. For this reason, there is a need to further investigate writers' monitoring and revising processes between test and real-life conditions in future studies.

5.3 Test takers who might be disadvantaged

While the multiple regression analysis shows that most affective variables investigated in this study did not have an impact on test takers' CB performance, three variables are found to have a mild but significant impact. This implies that test takers who do not have such an adequate computer familiarity profile are likely to perform worse than those who do in the CB mode. It is, therefore, recommended that the test provider might in future consider using these items to provide advice about the candidates' readiness for taking the test in the computer mode.

6. Conclusions

6.1 Limitations

Although this study has produced new insights into writing test equivalence between paper-based and computer-based conditions, it has several limitations.

First, the nature of a test equivalence study imposes difficulty in recruiting a large number of participants, as each needs to complete a test under two conditions. Nevertheless, we believe a sampled population of about 150 in this study is satisfactory for the findings to be generalisable to the wider test population. A further complication is that although the two versions of the task prompt used in this study may exhibit equivalence with one population, this may not necessarily hold true for another. In research designs such as the one used in this study, achieving complete equivalence of task may not be possible unless participants take both versions of the test in both conditions. As we considered inappropriate for participants to do the same version in both modes, we took the view that establishing acceptable boundaries of equivalence (e.g. counter-balancing the versions and conditions) within which we could have confidence was a suitable *modus operandi*.

6.2 Summary of findings

These limitations not withstanding, the most important conclusion from the study is that according to the 5-facet MFRM analysis, there were no significant differences in the scores awarded by two independent raters for candidates' performances on the tests taken under two conditions, one paper-and-pencil and the other computer. The difference between the fair means of the overall test scores in two modes was 0.03 for the whole group. Based on the 4-facet MFRM analyses, the differences in three analytic scores criteria (i.e. *Task Achievement*, *Coherence and Cohesion*, and *Grammatical Range and Accuracy*) were not significant, but the difference reported in *Lexical Resources* was significant.

With respect to the test takers' writing processes under the two conditions, results of the Writing Process Questionnaire indicate a *similar pattern* in the use of processes elicited by the PB and CB test. Most differences were 0.15 or below out of a 4-point scale. Secondly, the means of all items in each of six cognitive phases between the two modes were compared and tested by Wilcoxon Signed Ranks Test. All differences were non-significant. This indicates that the cognitive processes were employed in a similar fashion under the two delivery conditions.

This finding is confirmed by the interview data, where all test takers stated they composed the PB and CB tests in a comparable way. Nevertheless, a few differences in how test takers *planned*, *generated texts* and *monitored and revised* their texts emerged from the interviews. In terms of aspects of the revisions, some participants tended to focus more at the word level in the PB mode and more at the levels of clauses and sentences to improve coherence and argument in the CB mode. Drawing upon evidence from the questionnaire and interview data, the results on test takers' cognitive processes on the IELTS AWT2 in the CB mode of this study should be of great value to the test provider in specifying the cognitive parameters in the test specification.

The Computer Familiarity Questionnaire shows that participants in this study are familiar with computer usage, and their overall reactions towards working with a computer are positive. Most participants prefer to take the test under CB conditions. The results of Multiple Regression analysis indicate that three out of 15 of the computer familiarity variables (i.e. CFQ1c – access to computers at public library, CFQ4b – frequency of using computers for word processing, and CFQ13 - forgetting time) have a small but significant impact on their performance in the computer mode. This implies that test takers who do not have a suitable familiarity profile might perform slightly worse than those who do in computer mode.

6.3 Final thoughts

A difference of 0.25 in observed mean and 0.03 in fair mean between the test scores in the PB and CB modes were reported in this study. While no significant statistical difference was found in scores between the two modes, future research might investigate whether the test-takers themselves or test users would see the differences as 'non-significant'. Where there is a difference of one band, or even of half a band, it may turn out to be the difference between being accepted onto a programme or not, which might therefore have a 'significant' impact on a candidate's future. While the statistical test of significance is important and previous research has used this or similar measures, it is recommended that test developers need to bear in mind the human perception and consequences of even small differences such as a half band on IELTS between different modes and take steps accordingly, perhaps to the extent of issuing a "health warning" with results.

References

Al-Amri, S. (2008). Computer-based testing vs. paper-based testing: a comprehensive approach to examining the comparability of testing modes. *Essex Graduate Student Papers in Language & Linguistics, 10,* 22-44.

Berninger, V. (2012). *Evidence-Based, Developmentally Appropriate Writing Skills K–5: Teaching the Orthographic Loop of Working Memory to Write Letters So Developing Writers Can Spell Words and Express Ideas.* Presented at Handwriting in the 21st Century?: An Educational Summit, Washington, D.C., January 23, 2012.

Bond, T. G. and Fox, C. M. (2007). Applying the Rasch model. Fundamental measurement in the human sciences (2nd edition). University of Toledo.

Burke, J. N., & Cizek, G. J. (2006). Effects of composition mode and self-perceived computer skills on essay scores of sixth graders. *Assessing Writing*, *11*(3), 148–166.

Chan, S. (2013). *Establishing the construct validity of EAP reading-into-writing tests*. Unpublished PhD thesis. University of Bedfordshire, UK.

Chan, S. H. C., Wu, R. Y. F., & Weir, C. J. (2014). Examining the context and cognitive validity of the GEPT Advanced Writing Task 1: A comparison with real-life academic writing tasks. *LTTC-GEPT Research Reports*, RG-03, 1–89.

Cochran-Smith, M. (1991). Word processing and writing in elementary classrooms: Acritical review of related literature. Review of Educational Research, 61 (1), 107–155.

Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research (2nd ed.).* Thousand, Oaks, CA: Sage Publications.

Field, J. (2004). *Psycholinguistics: The Key Concepts*. London: Routledge.

Field, J. (2013). Cognitive validity. In Geranpayeh, A. & Taylor, L. (eds.) *Examining Listening.* Cambridge: Cambridge University Press.

Glaser, R. (1991). Expertise and assessment. In M. C. Wittrock and E. L. Baker (eds), *Testing and Cognition.* Prentice Hall, Englewood Cliffs, 17-30.

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

Grejda, G. F., & Hannafin, M. J. (1992). Effects of word processing on sixth graders' holistic writing and revisions. Journal of Educational Research, 85 (3), 144–149.

Haas, C. (1989). How the writing medium shapes the writing process: Effects of word processing on planning. *Research in the Teaching of English*, *23*, 181–207.

Hayes, J. R. and Flower, L. S. (1980). The dynamics of composing. In L.W. Gregg & E.R Steinberg (eds.) *Cognitive Processes in Writing*. Hillsdale, NJ: Lawrence Erlbaum Assoc.

Hermann, A (1987) *Research into writing and computers: Viewing the gestalt*. Paper presented at the Annual Meeting of the Modern Language Association, San Francisco, CA.

Hertz-Lazarowitz, R. & Bar-Natan, I. (2002). Writing development of Arab and Jewish students using cooperative learning (CL) and computer-mediated communication (CMC). *Computers & Education*, 39, 19-36.

James, K. H. (2012). The effects of handwriting experience on functional brain development in pre-literate children, *Trends in Neuroscience and Education*, *1*(1), 32-42.

Kellogg, R. T. (1996). A model of working memory in writing. In C. M. Levy, & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences and applications* (pp. 57–71). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Linacre, M. (2013). Facets computer program for many-facet Rasch measurement, version 3.71.2. Beaverton, Oregon: Winsteps.com.

Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing*, *15*(2), 158–180.

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

McDonald, A. (2002). The impact of individual differences on the equivalence of computer-based and paper-and-pencil educational assessments. *Computers & Education*, 39, 299-312.

Mead, A. & Drasgow, F. (1993). Equivalence of Computerized and Paper-and-Pencil Cognitive Ability Tests: A Meta-Analysis. *Psychological Bulletin*, *114(3)*, 449-458.

Mueller, P. A., & Oppenheimer, D. M. (2014). The pen is mightier than the keyboard: Advantages of longhand over laptop note taking. *Psychological Science*, *25*(6), 1159–1168.

Nakatsuhara, F., Inoue, C., Berry, V., & Galaczi, E. (2017). Exploring the Use of Video-Conferencing Technology in the Assessment of Spoken Language: A Mixed-Methods Study. *Language Assessment Quarterly*, *14*(1), 1–18.

Neuman, G., & Baydoun, R. (1998). Computerization of paper-and-pencil tests: When are they equivalent? *Applied Psychological Measurement*, 22(1), 71-83.

Newman, F., Couturier, L. & Scurry, J. (2010). The Future of Higher Education: Rhetoric, Reality, and the Risks of the Market. San Francisco: John Wiley & Sons.

Owston, R. & Wideman, H. (1997). Word processors and children's writing in a high-computer-access setting. *Journal of Research on Computing in Education*, 30 (2), 202–220.

Pearson (2012) "Into the fourth year of PTE Academic – our story so far." Retrieved from http://pearsonpte.com/media/Documents/fourthyear.pdf.

Puhan, P., Boughton, K., & Kim, S. (2007). Examining Differences in Examinee Performance in Paper and Pencil and Computerized Testing. *Journal of Technology, Learning, and Assessment*, *6*(3), 1-21.

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., & Plati, T. (2000). Effects of Computer Versus Paper Administrations of a State-Mandated Writing Assessment. *Technology and Assessment Study Collaborative,* 1–34.

Shaw, S. (2005). Evaluating the impact of word processed text on writing quality and rater behaviour, *Cambridge Research Notes*, *22,* 13-19.

Shaw, S. and Weir, C. (2007). *Examining Writing: Research and Practice in Assessing Second Language Writing. Studies in Language Testing*, Vol 26, Cambridge: Cambridge University Press.

Shermis, M., and Lombard, D. (1998). Effects of computer-based test administrations on test anxiety and performance. *Computers in Human Behavior*, *14(1),* 111 –123.

Taylor, C., Jamieson, J., Eignor, D., & Kirsch, I. (1998). *The relationship between computer familiarity and performance on computer-based TOEFL test tasks.* Research Reports 61. Princeton, NJ: Educational Testing Service.

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.

Weigle, S. C. (2002). *Assessing Writing*. Cambridge: Cambridge University Press.

Weir, C. J, O'Sullivan, B., Yan, J. and Bax, S. (2007). Does the computer make a difference? Reaction of participants to a computer-based versus a traditional handwritten form of the IELTS writing component: effects and impact. *IELTS Research Report*, Vol 7, (pp. 1–37). IELTS Australia, Canberra and British Council, London.

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. and Linacre, M. (1994). *Reasonable mean-square fit values.* Retrieved from http://www.rasch.org.

Wolfe, E. & Manolo, J. (2005). An investigation of the impact of composition medium on the quality of TOEFL writing scores. *ETS TOEFL Research Report 72*, 1-58.

Yu, G., Rea-Dickins, P. & Kiely, R. (2011). The cognitive processes of taking IELTS academic writing task one. *IELTS Reports, 11,* 373 – 449.

Acknowledgements

This research was funded by the IELTS Partners: British Council, Cambridge English Language Assessment and IDP (IELTS) Australia.

Sadly for all of us, Stephen Bax died while this paper was being prepared for publication. His incorrigible sense of humour and keen insights into our field will be sorely missed.

Table

Table 1. Cognitive parameters for academic writing tests (adapted from Chan, 2013)

Cognitive phases	Key processes
Conceptualisation	Task representation
	Macro-planning
Generating ideas	Careful reading (local/global)
	Scanning, Skimming and Search reading
	Connecting ideas and generating new representations
Organising ideas	Organising ideas in relation to input texts
	Organising ideas in relation to own texts
Generating texts	Translating ideas into linguistic forms
	Micro-planning
Monitoring and revising	Online monitoring and revising at low-level After writing monitoring and revising at low-level
	Online monitoring and revising at high-level After writing monitoring and revising at high-level.

Table 2. Participants' IELTS writing bands on entrance

Band range	Percentage of participants
4.5	2.0
5.0-5.5	35.3
6.0-6.5	52.3
7.0-7.5	9.2
8	1.2
	100

Table 3. Data collection procedures of a test session

Group A	Group B	Time (Mins)
All participants filled in a Computer Familiarity Questionnaire		5
Completed Prompt 1[1] on paper	Completed Prompt 1 on computer	40
All participants filled in a Writing Process Questionnaire		10
Completed Prompt 2 on computer	Completed Prompt 2 on paper	40
All participants filled in a second Writing Process Questionnaire		10
20% were interviewed individually		20

[1] The order of the versions (prompt 1 and prompt 2) was counterbalanced in alternate test sessions.

Table 4. Rater measurement report (5-facet analysis)

Rater	N	Observed Mean	Fair Mean	Logit measure	Standard error	Infit mean square
B	208	6.12	5.93	-.22	.11	1.04
D	372	5.84	5.91	-.17	.08	1.07
A	1096	5.80	5.81	.09	.05	.97
C	516	5.69	5.73	.29	.07	.99

Real, Populn: RMSE .08 Adj (True) S.D. .19 Separation 2.38 Strata 3.50 Reliability (not inter-rater) .85
Real, Sample: RMSE .08 Adj (True) S.D. .22 Separation 2.80 Strata 4.07 Reliability (not inter-rater) .89
Real, Fixed (all same) chi-square: 26.2 d.f.: 3 significance (probability): .00
Real, Random (normal) chi-square: 2.7 d.f.: 2 significance (probability): .26
Inter-Rater agreement opportunities: 1096 Exact agreements: 732 = 66.8% Expected: 483.2 = 44.1%

Table 5. Version Measurement Report (5-facet analysis)

Version	N	Observed Mean	Fair Mean	Logit measure	Standard error	Infit mean square
Prompt 1	1136	5.69	5.73	.30	.05	.91
Prompt 2	1056	5.94	5.96	-.30	.05	1.09

(Population): Separation 6.15; Strata 8.53; Reliability: 0.97
(Sample): Separation 8.75; Strata 12.00; Reliability: 0.99
Model, Fixed (all same) chi-square: 77.6 d.f.: 1; significance (probability): .00

Table 6. Delivery mode measurement report

Test mode	N	Observed Mean	Fair Mean	Logit measure	Standard error	Infit mean square
Computer-based	1104	5.75	5.83	.04	.05	.97
Paper-based	1088	5.87	5.86	-.04	.05	1.02

(Population): Separation .00; Strata .33; Reliability .00
(Sample): Separation .91; Strata 1.54; Reliability .45
Model, Fixed (all same) chi-square: 1.8; d.f.: 1; significance (probability): .18

Table 7. Analytic scales measurement report (Task Achievement)

Analytic scale	N	Observed Mean	Fair Mean	Logit measure	Standard error	Infit mean square
CB Task Achievement	276	5.51	5.63	.03	.09	.89
PB Task Achievement	272	5.63	5.65	-.03	.10	1.06

(Population): Separation .00; Strata .33; Reliability .00
(Sample): Separation .00; Strata .33; Reliability .00
Model, Fixed (all same) chi-square: .2; d.f.: 1; significance (probability): .70

Table 8. Analytic scales measurement report (Coherence and Cohesion)

Analytic scale	N	Observed Mean	Fair Mean	Logit measure	Standard error	Infit mean square
CB Coherence and Cohesion	276	5.88	5.93	-.13	.12	.83
PB Coherence and Cohesion	272	5.86	5.87	.13	.13	1.16

(Population): Separation .12; Strata .50; Reliability .02
(Sample): Separation 1.02; Strata 1.69; Reliability .51
Model, Fixed (all same) chi-square: 2.0; d.f.: 1; significance (probability): .15

Table 9. Analytic scales measurement report (Lexical Resources)

Analytic scale	N	Observed Mean	Fair Mean	Logit measure	Standard error	Infit mean square
CB Lexical Resources	276	5.89	5.97	.24	.12	.96
PB Lexical Resources	272	6.08	6.04	-.24	.12	.96

(Population): Separation 1.76; Strata 2.68; Reliability .76
(Sample): Separation 2.68; Strata 3.91; Reliability .88
Model, Fixed (all same) chi-square: 8.2; d.f.: 1; significance (probability): .00

Table 10. Analytic scales measurement report (Grammatical Range and Accuracy)

Analytic scale	N	Observed Mean	Fair Mean	Logit measure	Standard error	Infit mean square
CB Grammatical Range and Accuracy	276	5.71	5.76	.10	.12	1.07
PB Grammatical Range and Accuracy	272	5.90	5.82	-.10	.11	.87

(Population): Separation .00; Strata .33; Reliability .00
(Sample): Separation .64; Strata 1.18; Reliability .29
Model, Fixed (all same) chi-square: 1.4; d.f.: 1; significance (probability): .24

Table 11. Mean of processes in each cognitive phase

	Computer-based IELTS			Paper-based IELTS			TLU (Chan, 2013)		
	N	Mean	SD	N	Mean	SD	N	Mean	SD
Conceptualisation	129	3.25	0.40	130	3.27	0.42	143	3.17	0.49
Generating ideas	132	3.26	0.43	132	3.26	0.44	143	3.20	0.51
Generating texts	134	3.40	0.47	132	3.39	0.51	*N/A*	*N/A*	*N/A*
Organising ideas	132	3.25	0.49	130	3.24	0.48	143	3.13	0.52
Monitoring and revising (high-level)	128	3.22	0.50	129	3.17	0.50	143	3.00	0.59
Monitoring and revising (low-level)	132	3.20	0.60	132	3.20	0.60	143	2.97	0.62

Note. The processes of generating texts were not investigated in Chan (2013).

Table 12. Wilcoxon Signed Ranks Test on each cognitive phase (CB vs PB mode)

Cognitive phase	Delivery mode	Median	Mean rank	Z	Sig. (2-tailed)
Conceptualisation	CB	3.20	65.0	-0.065	0.948
	PB	3.30	65.5		
Generating ideas	CB	3.20	66.5	0.000	1.000
	PB	3.20	66.5		
Generating texts	CB	3.50	67.5	-1.631	0.103
	PB	3.50	66.5		
Organising ideas	CB	3.20	66.5	-1.359	0.174
	PB	3.20	65.5		
Monitoring and revising (high-level)	CB	3.20	64.5	-1.649	0.99
	PB	3.20	65.0		
Monitoring and revising (low-level)	CB	3.17	66.5	0.000	1.000
	PB	3.17	66.5		

Table 13. Descriptive statistics of the Computer Familiarity Questionnaire

Categories	Items	N	Percentage
Computer Usage	Q1	128	96.1% (59.7%) have frequent access to computers at home; 89.8% (88.4%) at university; 78.6% in public places
	Q2	127	97.6% (56.4%) use computers frequently at home; 82.7% (84.3%) at university; 40.7% in public places
	Q3	126	87.3% (95.7%) frequently use computers for surfing the Internet; 94.5% (89.9%) electronic communication; 96%(59.7%) for study-related activities; 66.7% for other purposes
	Q4	125	92.9% (68.0%) frequently use word processing; 55.6% spreadsheets; 57.9% data analysis; 31.7% graphics; 28.0% games; 64.3% other purposes
	Q5	127	86.6% frequently take a test on paper; 64.2% on computer
Comfort & Perceived Ability	Q6	128	81.2% (79.0%) are comfortable using a computer in general
	Q7	128	90.6% (67.5%) are comfortable using a computer to write an essay
	Q8	128	81.2% (53.0%) are comfortable taking a test on computer; 94.2% on paper
	Q9	128	89.2% (71.1%) are comfortable typing with keyboard
	Q14	126	60.3% (49.0%) are good or excellent at using a computer
Interest in Computers	Q10	127	87.4% (84.8%) consider very important to work with a computer
	Q11	126	71.4% (86.7%) consider playing or working with a computer is really fun.
	Q12	127	63.0% (67.6%) use a computer because they are very interested in this.
	Q13	127	78.0% (66.7%) would forget the time when working with the computer

Note: Figures of equivalent CFQ items from Weir et al. (2007) are provided in brackets for reference. New CFQ items added in this study do not have any comparative figures.

Table 14. Multiple regression analysis of CB scores on CFQ items

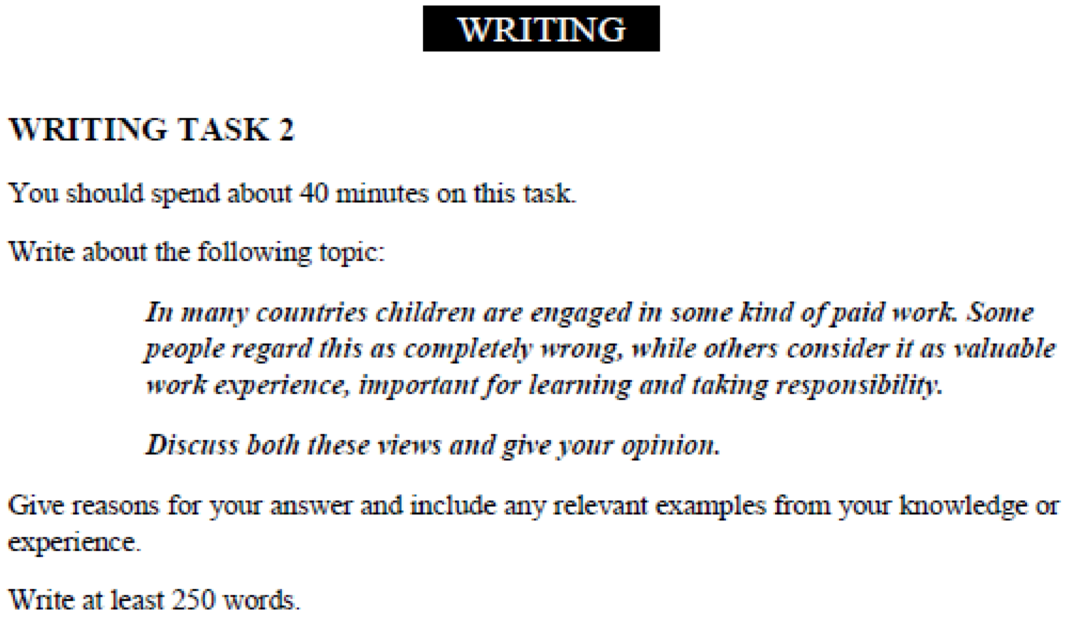
	B (unstandardized regression coefficient)	Standard error	β (Standardized regression coefficient)	t	Sig.	
CFQ 4b	.297	.066	.374	4.496	.000	
CFQ 1c	.093	.044	.174	2.083	.039	
CFQ 13	.107	.053	.166	2.020	.046	
R^2						.226
F						11.280

Table 15. Summary table of findings emerged from the interview data

	Differences observed between the two modes
Conceptualisation	• More detailed planning under the PB mode, and followed the plan closely to avoid major changes • Most did not start writing with a ‘perfect’ plan under the CB mode
Generating ideas	• Some referred to planning notes under the PB mode
Generating texts	• Handwriting vs. typing • More careful about choice of words and sentence structure when composed on paper • Focused more on expressing the ideas under the CB mode
Organising ideas	• Engaged more in organising ideas at the levels of sentences and paragraphs under the CB mode
Monitoring and revising (high-level)	• More test takers revised texts during writing under the CB than PB mode • More engaged in post-writing revising under the CB than PB mode • Made more changes at the levels of clauses and sentences to improve coherence under the CB mode
Monitoring and revising (low-level)	• Focused on phrasing at the word level and correcting grammatical mistakes under the PB mode

Appendix A Test Tasks

IELTS AWT2 Prompt 1



WRITING

WRITING TASK 2

You should spend about 40 minutes on this task.

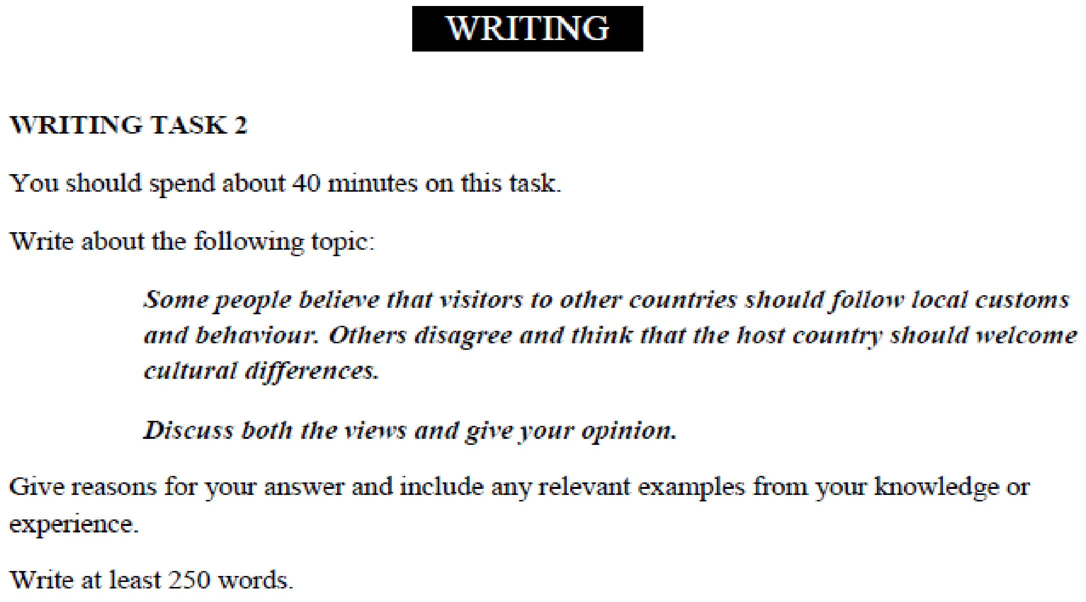
Write about the following topic:

> ***In many countries children are engaged in some kind of paid work. Some people regard this as completely wrong, while others consider it as valuable work experience, important for learning and taking responsibility.***
>
> ***Discuss both these views and give your opinion.***

Give reasons for your answer and include any relevant examples from your knowledge or experience.

Write at least 250 words.

IELTS AWT2 Prompt 2



WRITING

WRITING TASK 2

You should spend about 40 minutes on this task.

Write about the following topic:

> ***Some people believe that visitors to other countries should follow local customs and behaviour. Others disagree and think that the host country should welcome cultural differences.***
>
> ***Discuss both the views and give your opinion.***

Give reasons for your answer and include any relevant examples from your knowledge or experience.

Write at least 250 words.

Appendix B – Computer Familiarity Questionnaire Items

		Always	Very often	Often	Once a while	Never
1	How often is there a computer available to you to use at each of the following places?					
	a) Home	5	4	3	2	1
	b) University	5	4	3	2	1
	c) Public place	5	4	3	2	1
	d) Others, if any (please specify)					
2	How often do you use a computer at each of the following places?					
	a) Home	5	4	3	2	1
	b) University	5	4	3	2	1
	c) Public Library	5	4	3	2	1
	d) Others, if any (please specify)					
3	How often do you use a computer for					
	a) surfing the Internet?	5	4	3	2	1
	b) electronic communications, e.g. emails?	5	4	3	2	1
	c) study-related activities?	5	4	3	2	1
	d) others, if any (please specify)					
4	How often do you use a computer software for?					
	a) games?	5	4	3	2	1
	b) word processing?	5	4	3	2	1
	c) spreadsheets?	5	4	3	2	1
	d) painting or graphics?	5	4	3	2	1
	e) data or text analysis?	5	4	3	2	1
	f) Others, if any (please specify)					
5	How often do you take a test on					
	a) paper and pencil?	5	4	3	2	1
	b) computer?	5	4	3	2	1
		Very comfortable	Quite comfortable	Comfortable	Quite uncomfortable	Very uncomfortable
6	How comfortable are you with using a computer in general?	5	4	3	2	1
7	How comfortable are you with using a computer to write an essay?	5	4	3	2	1
8	How comfortable are you with taking a test on					
	a) computer?	5	4	3	2	1
	b) paper and pencil?	5	4	3	2	1
9	How do you feel about using the keyboard (typing)?	5	4	3	2	1

	To what extent do you agree with the following statements?	Strongly agree Mostly agree Neutral Mostly disagree Strongly disagree
10	It is very important to me to work with a computer.	5 4 3 2 1
11	To play or work with a computer is really fun.	5 4 3 2 1
12	I use a computer because I am very interested in this.	5 4 3 2 1
13	I forget the time, when I am working with the computer.	5 4 3 2 1
		Excellent Good Fair Poor Very poor
14	If you compare yourself with other students, how would you rate your ability to use a computer?	5 4 3 2 1
15	If you are allowed to choose, do you prefer to take the IELTS Academic Writing test on paper or computer? Why?	

Appendix C. Writing Process Questionnaire items

While READING the task instructions ...	
1	I read the **whole** task instructions carefully.
2	I thought about **how well I understood** the task requirements.
3	I thought about what I knew about the **topic**.
4	I thought about what I knew about the **genre.**
5	I thought about the **purpose** of the task.
6	I thought about **what I might need to write** to make my essay relevant and adequate to the task.
7	I thought about the **intender reader** of my essay and their **expectations**.
BEFORE starting to write ...	
8	I thought about or jotted down **ideas** which are **relevant** to the task/topic.
9	I **prioritised** these ideas based on the task requirements.
10	I **linked** these **ideas** to what I know already about the topic from **memory**.
11	I worked out how these ideas **relate** to each other, e.g. main ideas or examples.
12	I thought about the **structure** of my essay.
13	I **recombined** or **reordered** some of the ideas to fit the structure of my essay.
14	I **removed** some ideas I planned to write because they did not fit the structure of my essay.
15	I **re-read** the task instructions.
WHILE writing the first draft ...	
16	I thought about the **appropriate words** to express my ideas.
17	I thought about the **correct sentence structures** to express my ideas.
18	I thought about the **correct grammar** to express my ideas.
19	I thought about how to **connect** my ideas **smoothly** in the whole essay
20	I thought about how to make my ideas **persuasive** to the reader.
21	I **organised** my sentences and paragraphs in a **logical** order.
22	I developed new ideas or a **better understanding** of the topic.
23	I **re-read** the task instructions.
24	I **changed** my writing **plan** (e.g. structure and content).
25	I checked that the content was **relevant** and revised accordingly**.**
26	I checked that my essay was **well-organised** and revised accordingly**.**
27	I checked that my essay was **coherent** and revised accordingly**.**
28	I checked that I included **my own viewpoint** on the topic and revised accordingly**.**
29	I checked the possible effect of my essay on the **intended reader** and revised accordingly**.**
30	I checked the accuracy and range of the **sentence structures** and revised accordingly**.**
31	I checked the **grammar** (e.g. part of speech and tenses) and revised accordingly
32	I checked the appropriateness and range of **vocabulary** and revised accordingly**.**
AFTER writing the first draft ...	
33	I checked that the content was **relevant** and revised accordingly**.**
34	I checked that my essay was **well-organised** and revised accordingly**.**

35	I checked that my essay was **coherent** and revised accordingly**.**
36	I checked that I included **my own viewpoint** on the topic and revised accordingly**.**
37	I checked the possible effect of my essay on the **intended reader** and revised accordingly**.**
38	I checked the accuracy and range of the **sentence structures** and revised accordingly**.**
39	I checked the **grammar** (e.g. part of speech and tenses) and revised accordingly
40	I checked the appropriateness and range of **vocabulary** and revised accordingly**.**

Appendix D. Examples of interview coding

Categories	Examples
Task representation	• I read through the instructions and the question and thought about how to approach the task.
Macro-planning	• I spent about 10 minutes for planning. I thought about some key-points and ideas to put in the essay.
Generating ideas	• The ideas just came. When I started to write more ideas came. • I thought about my experience related to the topic like the situation in my home country.
Organising ideas	• I organised the ideas according to the structure of my essay: introduction, main body and the conclusion.
Generating texts	• I just wrote down all my ideas as quickly as possible without much planning. • I first wrote the introduction. After that I wrote about the first supporting argument, but I left it there for a while because I wanted to write down some idea about the second supporting argument.
Monitoring and revising (during writing)	• I made some changes while I was writing the essay. Sometimes I made changes to a sentence to make it flow better or sometimes I just changed a particular word.
Monitoring and revising (after writing)	• I read the essay again made some changes according to what the intended reader needs to know

Figure

Figure 1. Summary of research design

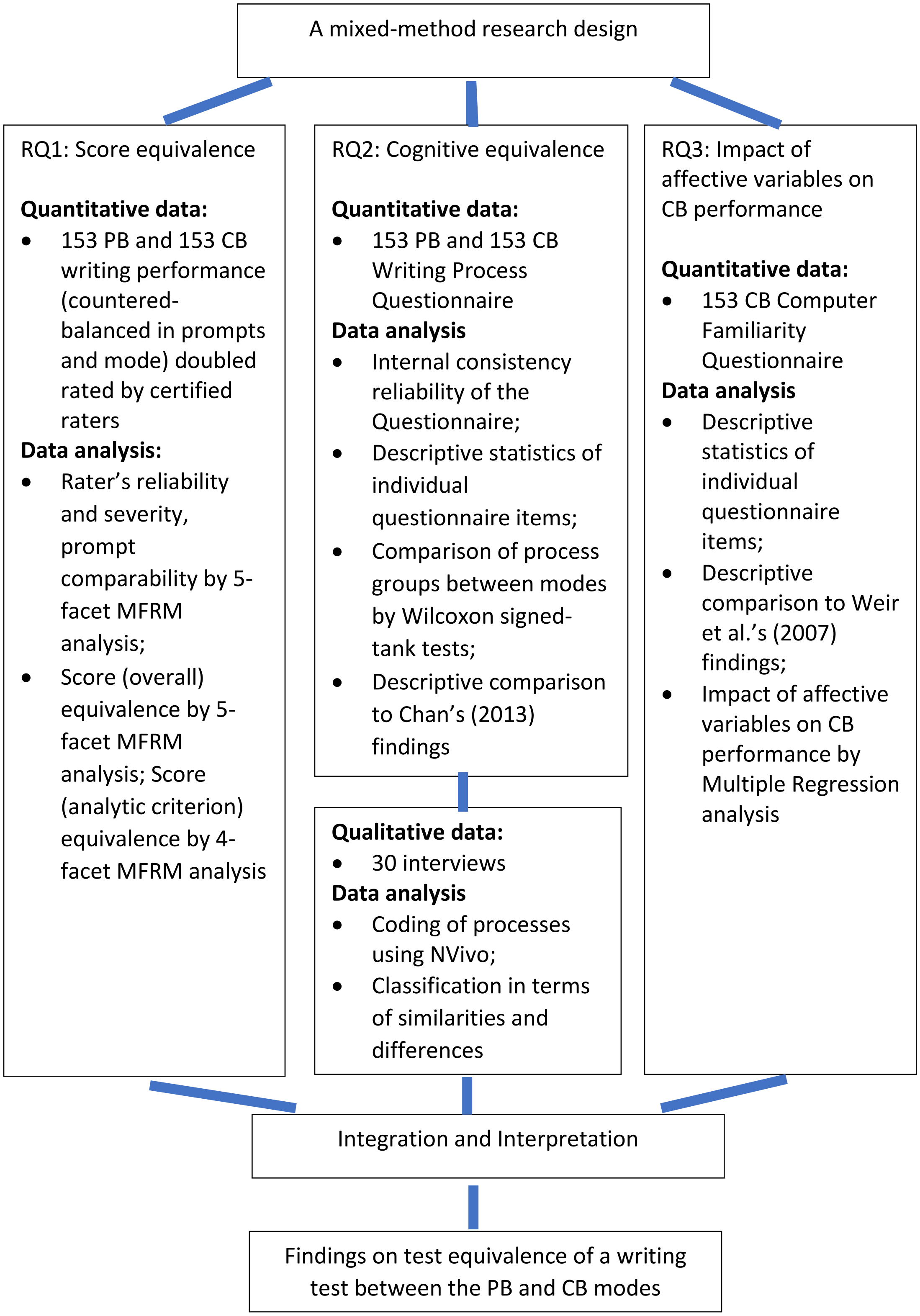
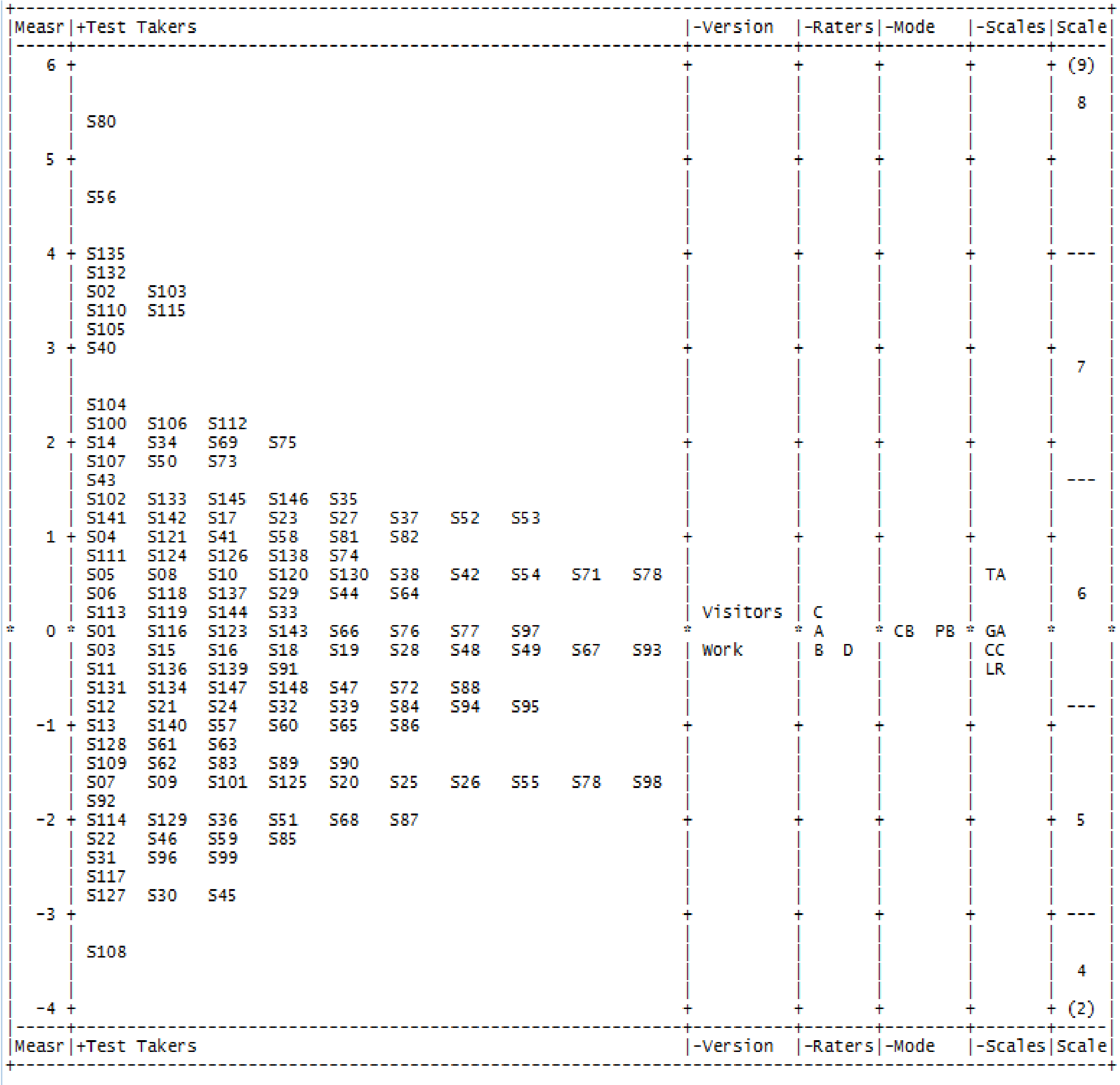
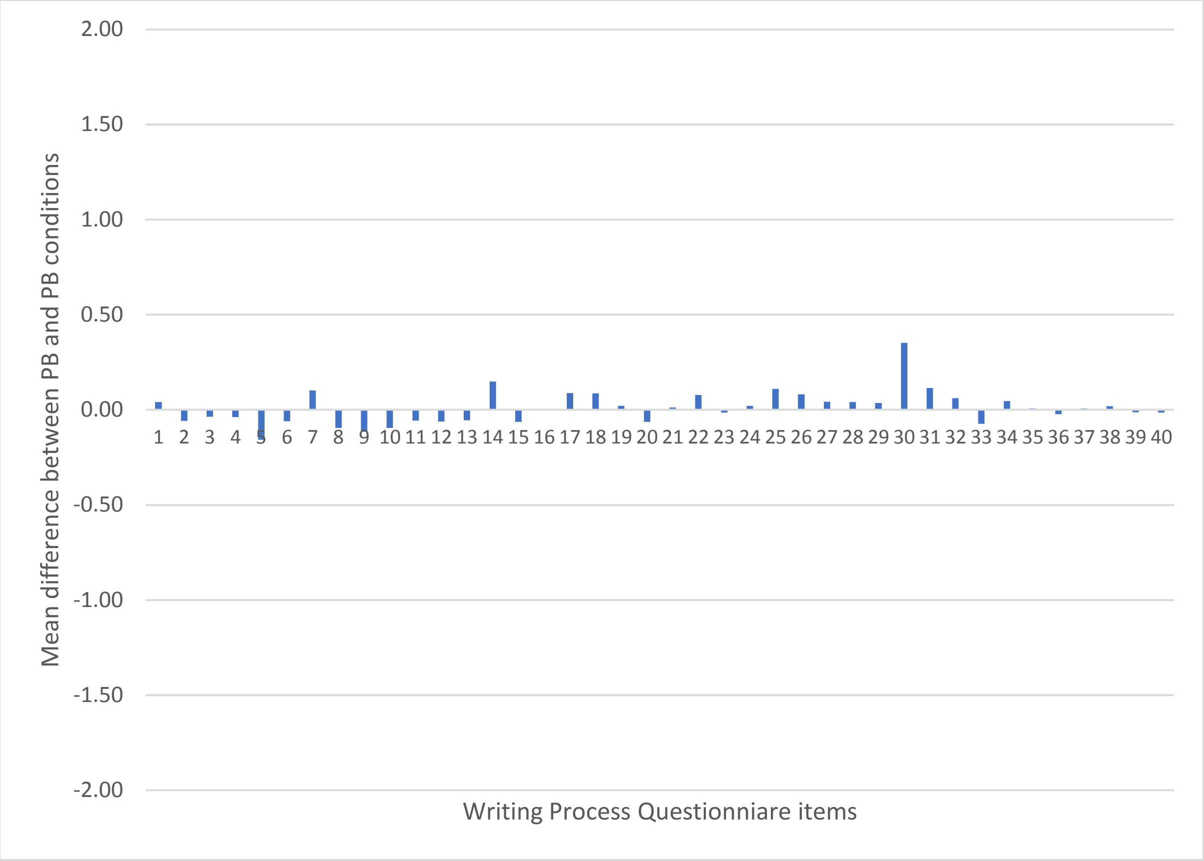

A mixed-method research design
RQ1: Score equivalence
Quantitative data:
153 PB and 153 CB writing performance (countered-balanced in prompts and mode) doubled rated by certified raters
Data analysis:
Rater's reliability and severity, prompt comparability by 5-facet MFRM analysis;
Score (overall) equivalence by 5-facet MFRM analysis; Score (analytic criterion) equivalence by 4-facet MFRM analysis
RQ2: Cognitive equivalence
Quantitative data:
153 PB and 153 CB Writing Process Questionnaire
Data analysis
Internal consistency reliability of the Questionnaire;
Descriptive statistics of individual questionnaire items;
Comparison of process groups between modes by Wilcoxon signed-tank tests;
Descriptive comparison to Chan's (2013) findings
Qualitative data:
30 interviews
Data analysis
Coding of processes using NVivo;
Classification in terms of similarities and differences
RQ3: Impact of affective variables on CB performance
Quantitative data:
153 CB Computer Familiarity Questionnaire
Data analysis
Descriptive statistics of individual questionnaire items;
Descriptive comparison to Weir et al.'s (2007) findings;
Impact of affective variables on CB performance by Multiple Regression analysis
Integration and Interpretation
Findings on test equivalence of a writing test between the PB and CB modes

Figure 2. FACETS Variable Map (5-facet analysis)



```
+------------------------------------------------------------------------------------------------------------+
|Measr|+Test Takers                                                |-Version  |-Raters|-Mode   |-Scales|Scale|
|-----+------------------------------------------------------------+----------+-------+--------+-------+-----|
|   6 +                                                            +          +       +        +       + (9) |
|     |                                                            |          |       |        |       |     |
|     |                                                            |          |       |        |       |  8  |
|     | S80                                                        |          |       |        |       |     |
|     |                                                            |          |       |        |       |     |
|   5 +                                                            +          +       +        +       +     |
|     |                                                            |          |       |        |       |     |
|     | S56                                                        |          |       |        |       |     |
|     |                                                            |          |       |        |       |     |
|     |                                                            |          |       |        |       |     |
|   4 + S135                                                       +          +       +        +       + --- |
|     | S132                                                       |          |       |        |       |     |
|     | S02   S103                                                 |          |       |        |       |     |
|     | S110  S115                                                 |          |       |        |       |     |
|     | S105                                                       |          |       |        |       |     |
|   3 + S40                                                        +          +       +        +       +     |
|     |                                                            |          |       |        |       |  7  |
|     |                                                            |          |       |        |       |     |
|     | S104                                                       |          |       |        |       |     |
|     | S100  S106  S112                                           |          |       |        |       |     |
|   2 + S14   S34   S69   S75                                      +          +       +        +       +     |
|     | S107  S50   S73                                            |          |       |        |       |     |
|     | S43                                                        |          |       |        |       | --- |
|     | S102  S133  S145  S146  S35                                |          |       |        |       |     |
|     | S141  S142  S17   S23   S27   S37   S52   S53              |          |       |        |       |     |
|   1 + S04   S121  S41   S58   S81   S82                          +          +       +        +       +     |
|     | S111  S124  S126  S138  S74                                |          |       |        |       |     |
|     | S05   S08   S10   S120  S130  S38   S42   S54   S71   S78  |          |       |        | TA    |     |
|     | S06   S118  S137  S29   S44   S64                          |          |       |        |       |  6  |
|     | S113  S119  S144  S33                                      | Visitors | C     |        |       |     |
*   0 * S01   S116  S123  S143  S66   S76   S77   S97              *          * A     * CB  PB * GA    *     *
|     | S03   S15   S16   S18   S19   S28   S48   S49   S67   S93  | Work     | B  D  |        | CC    |     |
|     | S11   S136  S139  S91                                      |          |       |        | LR    |     |
|     | S131  S134  S147  S148  S47   S72   S88                    |          |       |        |       |     |
|     | S12   S21   S24   S32   S39   S84   S94   S95              |          |       |        |       | --- |
|  -1 + S13   S140  S57   S60   S65   S86                          +          +       +        +       +     |
|     | S128  S61   S63                                            |          |       |        |       |     |
|     | S109  S62   S83   S89   S90                                |          |       |        |       |     |
|     | S07   S09   S101  S125  S20   S25   S26   S55   S78   S98  |          |       |        |       |     |
|     | S92                                                        |          |       |        |       |     |
|  -2 + S114  S129  S36   S51   S68   S87                          +          +       +        +       +  5  |
|     | S22   S46   S59   S85                                      |          |       |        |       |     |
|     | S31   S96   S99                                            |          |       |        |       |     |
|     | S117                                                       |          |       |        |       |     |
|     | S127  S30   S45                                            |          |       |        |       |     |
|  -3 +                                                            +          +       +        +       + --- |
|     |                                                            |          |       |        |       |     |
|     | S108                                                       |          |       |        |       |     |
|     |                                                            |          |       |        |       |  4  |
|     |                                                            |          |       |        |       |     |
|  -4 +                                                            +          +       +        +       + (2) |
|-----+------------------------------------------------------------+----------+-------+--------+-------+-----|
|Measr|+Test Takers                                                |-Version  |-Raters|-Mode   |-Scales|Scale|
+------------------------------------------------------------------------------------------------------------+
```

Figure 3. Mean difference of the individual processes between PB and CB mode


Mean difference between PB and PB conditions
Writing Process Questionniare items