ISSN 2201-2982
2017/4

# IELTS Research Reports
# Online Series

## Researching participants taking IELTS Academic Writing Task 2 (AWT2) in paper mode and in computer mode in terms of score equivalence, cognitive validity and other factors

Sathena Chan, Stephen Bax and Cyril Weir

IELTS™
BRITISH COUNCIL
idp IELTS AUSTRALIA
CAMBRIDGE ENGLISH Language Assessment
Part of the University of Cambridge

# Researching participants taking IELTS Academic Writing Task 2 (AWT2) in paper mode and in computer mode in terms of score equivalence, cognitive validity and other factors

This study investigates the extent to which 153 test-takers' cognitive processes, while completing IELTS Academic Writing in paper-based mode and in computer-based mode, compare with the real-world cognitive processes of students completing academic writing at university.

## Acknowledgements

We would like to thank our PhD students, Tanzeela Anbreen and Ekaterina Kandelaki, for their assistance in data collection, and the raters who participated in the study.

## Funding

This research was funded by the IELTS Partners: British Council, Cambridge English Language Assessment and IDP: IELTS Australia. Grant awarded 2013.

## Publishing details

Published by the IELTS Partners: British Council, Cambridge English Language Assessment and IDP: IELTS Australia © 2017.



## How to cite this article

Chan, S., Bax S. and Weir. C. 2017. Researching participants taking IELTS Academic Writing Task 2 (AWT2) in paper mode and in computer mode in terms of score equivalence, cognitive validity and other factors. *IELTS Research Reports Online Series, No. 4*. British Council, Cambridge English Language Assessment and IDP: IELTS Australia. Available at https://www.ielts.org/teaching-and-research/research-reports

# Introduction

This study by Sathena Chan, Stephen Bax, and Cyril Weir was conducted with support from the IELTS partners (British Council, IDP: IELTS Australia, and Cambridge English Language Assessment) as part of the IELTS joint-funded research program. Research funded by the British Council and IDP: IELTS Australia under this program complement those conducted or commissioned by Cambridge English Language Assessment, and together inform the ongoing validation and improvement of IELTS.

A significant body of research has been produced since the joint-funded research program started in 1995, with over 110 empirical studies receiving grant funding. After undergoing a process of peer review and revision, many of the studies have been published in academic journals, in several IELTS-focused volumes in the *Studies in Language Testing* series (http://www.cambridgeenglish.org/silt), and in *IELTS Research Reports*. Since 2012, in order to facilitate timely access, individual research reports have been made available on the IELTS website immediately after completing the peer review and revision process.

The study detailed in this report concerns the skill of writing; in particular, whether writing in a test written on paper and typed on computer are comparable to each other, as well as to writing in real-life academic contexts. It is a topic that is of relevance to IELTS for several reasons. One, it has been argued that most academic writing nowadays happens on computer, and thus, the IELTS Academic Writing test being handwritten is inauthentic, if not invalid. Two, IELTS is in fact beginning to be administered to some test-takers in computer format, alongside the paper version of the test. Thus, unlike other similar tests, which are only offered via computer, there is a need to demonstrate comparability across modality. Finally, it is also sometimes claimed that writing in IELTS Academic is not very similar to writing in academic contexts, so this study helps to address that charge.

So what did this study find? Scores for responses written on computer and on computer were not distinguishable. But beyond that, the study also investigated test-takers' cognitive operations in the writing process – for example, how they generated ideas, or how they monitored themselves while writing – to see whether they were engaging in the same type of activity across the two modes. That part of the study found that each aspect of the writing process was not distinguishable. In other words, it doesn't matter in which mode you take the test. You are doing the same thing, and you get scored in the same way.

In addition, the study drew from a separate study conducted by one of the authors (Chan, 2013), which focused on the extent to which different cognitive processes were engaged while writing in academic contexts. These turned out to be very similar to what candidates reported for the testing context. There was only one measure where some difference was observed: people engaged in monitoring while writing more often in the testing context, than in the non-testing context. That should not really be concerning. Because testing is, by definition, taking a sample of people's abilities, the more important thing is that each of the processes was engaged, rather than how often each one was engaged. In any event, the main finding here is that writing processes are comparable in real-world academic writing and in testing contexts.

The findings of this study should help reassure test-takers and test users about the validity of the IELTS Academic Writing test, and should give the IELTS partners greater confidence should they decide to expand provision of computer-based IELTS testing.

**Gad S. Lim**
**Principal Research Manager**
**Cambridge English Language Assessment**

**References:**

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

# Researching participants taking AWT2 in paper mode and in computer mode in terms of score equivalence, cognitive validity and other factors


## Abstract

Computer-based (CB) assessment is becoming more common in most university disciplines, and international language testing bodies now routinely use computers for many areas of English language assessment. Given that, in the near future, IELTS also will need to move towards offering CB options alongside traditional paper-based (PB) modes, the research reported here prepares for that possibility, building on research carried out some years ago which investigated the statistical comparability of the IELTS writing test between the two delivery modes, and offering a fresh look at the relevant issues.

By means of questionnaire and interviews, the current study investigates the extent to which 153 test-takers' cognitive processes, while completing IELTS Academic Writing in PB mode and in CB mode, compare with the real-world cognitive processes of students completing academic writing at university. A major contribution of our study is its use – for the first time in the academic 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.

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. Regarding analytic scores criteria, the differences in three areas (i.e. Task Achievement, Coherence and Cohesion, and Grammatical Range and Accuracy) were not significant, but the difference reported in Lexical Resources was significant, if slight. In summary, the difference of scores between the two modes is at an acceptable level.

With respect to the cognitive processes students employ in performing under the two conditions of the test, results of the Cognitive Process Questionnaire (CPQ) survey indicate a similar pattern between the cognitive processes involved in writing on a computer and writing with paper-and-pencil. There were no noticeable major differences in the general tendency of the mean of each questionnaire item reported on the two test modes. In summary, the cognitive processes were employed in a similar fashion under the two delivery conditions.

Based on the interview data (n=30), it appears that the participants reported using most of the processes in a similar way between the two modes. Nevertheless, a few potential differences indicated by the interview data might be worth further investigation in future studies. The Computer Familiarity Questionnaire survey shows that these students in general are familiar with computer usage and their overall reactions towards working with a computer are positive.

Multiple regression analysis, used to find out if computer familiarity had any effect on students' performances on the two modes, suggested that test-takers who do not have a suitable familiarity profile might perform slightly worse than those who do, in computer mode.

In summary, the research offered in this report offers a unique comparison with real-world academic writing, and presents a significant contribution to the research base which IELTS and comparable international testing bodies will need to consider, if they are to introduce CB test versions in future.

# Author biodata

## Sathena Chan

Dr Sathena Chan is a 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. She has been actively involved in different test development or validation projects for examination boards and educational organisations in the UK and worldwide. Her publications include the book, *Defining Integrated Reading-into-Writing Constructs* (2018, CUP), journal articles in *Assessing Writing* (2015) and *Language Testing in Asia* (2017), and research reports (e.g. Bax and Chan, 2016 on GEPT reading; Taylor and Chan, 2015 on test equivalence for GMC registration purposes). She also conducts professional training on integrated assessment literacy and statistical analyses for language testing.

## Stephen Bax

Professor Stephen Bax is a Professor of Modern Languages and Linguistics at the Open University. His research interests include the use of computers in language learning (CALL), the use of computers in language testing (CALT), and areas of discourse including Computer Mediated Discourse Analysis (CMDA). Most recently, he has been using eye tracking to research reading and reading tests. His 2013 article in *Language Testing*, the first article on eye tracking in the journal, won the 2014 TESOL Distinguished Research Award. His publications include the British Council research monograph *Researching English Bilingual Education in Thailand, Indonesia and South Korea*, the book, *Discourse and Genre* (2011, Macmillan), his 2013 book, *Researching Intertextual Reading* in the series *Contemporary Studies in Descriptive Linguistics*, as well as leading articles in the fields of teacher education, CALL and ICT, and areas of discourse. His 2003 article on CALL won an Elsevier prize.

## Cyril Weir

Professor 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. He is the author of a wide range of educational publications, including *Communicative Language Testing, Understanding and Developing Language Tests* and *Language Testing and Validation: An evidence-based approach*. He is also the co-author of *Examining Writing, Examining Reading, Evaluation in ELT, Reading in a Second Language, Measured Constructs*, and the co-editor of six *Studies in Language Testing* (SiLT) Volumes. He was elected as a Fellow of the Academy of Social Sciences in 2013 and awarded the OBE for services to English language assessment in 2015.

# Table of contents


**1 Introduction ..... 10**

**2 Aims of the project ..... 10**

**3 Context of the study ..... 11**

3.1 Rationale and previous research ..... 11

3.1.1 Computer-based academic writing in International Higher Education ..... 11

3.1.2 The use of CB writing in academic writing assessments ..... 12

3.1.3 Score equivalence ..... 13

3.1.4 Cognitive validity and cognitive equivalence ..... 13

3.1.5 Cognitive processing in IELTS Academic writing ..... 14

3.1.6 The potential impact of writers' computer familiarity on performance ..... 16

3.2 Research questions ..... 17

**4 Research design, data collection and analysis ..... 17**

4.1 General approach ..... 17

4.2 Participants ..... 17

4.3 Instruments ..... 18

4.3.1 Test tasks ..... 18

4.3.2 Computer Familiarity Questionnaire ..... 18

4.3.3 Writing Process Questionnaire ..... 19

4.3.4 Interview ..... 19

4.4 Data collection ..... 19

4.5 Data analysis ..... 20

**5 Results and discussion ..... 21**

5.1 Score equivalence between CB and PB mode (RQ1) ..... 21

5.2 Cognitive equivalence between CB and PB mode (RQ2) ..... 25

5.2.1 Evidence from questionnaire ..... 25

5.2.2 Evidence from interview data ..... 27

5.2.3 Conceptualisation ..... 27

5.2.4 Generating and organising ideas ..... 28

5.2.5 Generating texts ..... 28

5.2.6 Monitoring and revising (online and after writing/ at low- and high-levels) ..... 28

5.3 Relationship between affective variables and test performance in CB mode (RQ3) ..... 29

**6 Conclusions and recommendations ..... 31**

6.1 Summary ..... 31

6.2 Discussion ..... 32

**References ..... 34**

**Appendix 1: Test tasks ..... 37**

**Appendix 2: Computer Familiarity Questionnaire ..... 41**

**Appendix 3: Writing Process Questionnaire ..... 44**

**Appendix 4: Examples of interview coding ..... 47**

# List of tables

Table 1: Participants' previous IELTS scores ..... 17
Table 2: Structure of the Computer Familiarity Questionnaire ..... 18
Table 3: Structure of the Writing Process Questionnaire ..... 19
Table 4: Data collection procedures ..... 20
Table 5: Final totals of each data point ..... 20
Table 6: Rater measurement report ..... 21
Table 7: Version measurement report ..... 22
Table 8: Delivery mode measurement report ..... 22
Table 9: Analytic scales measurement report (Task Achievement) ..... 23
Table 10: Analytic scales measurement report (Coherence and Cohesion) ..... 24
Table 11: Analytic scales measurement report (Lexical Resources) ..... 24
Table 12: Analytic scales measurement report (Grammatical Range and Accuracy) ..... 24
Table 13: Mean of processes in each cognitive phase ..... 26
Table 14: Wilcoxon signed-ranks test on each cognitive phase (CB vs PB mode) ..... 27
Table 15: Descriptive statistics of the Computer Familiarity Questionnaire (CFQ) ..... 29
Table 16: Multiple regression analysis of CB scores on CFQ items ..... 30

# List of figures

Figure 1: FACETS variable map ..... 23
Figure 2: Differences in mean between the PB and CB Writing Process Questionnaire items ..... 25

# 1 Introduction

This document constitutes the final report on a research project entitled “Researching the cognitive processes of participants taking IELTS Academic Writing Task 2 (AWT2) in paper-based mode and in computer-based mode, in terms of score equivalence, cognitive validity, and other factors”, which was funded by the IELTS joint-funded research program 2013.

# 2 Aims of the project

This study relates to the first broad area of interest identified by the IELTS Joint Research Committee Call for Proposals 2013/14 namely “test development and validation issues”, specifically requesting “studies investigating the cognitive processes of IELTS test-takers”.

In line with the increasingly important 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). Many international language testing bodies now routinely use computers for many areas of Academic English written assessment. Although IELTS does not currently offer CB assessment, it seems foreseeable, given the cost and other perceived benefits of CB testing, that in the near future, IELTS will need to move towards offering CB options alongside traditional paper-based (PB) modes, if the test is to retain its reputation for cutting-edge language assessment at a competitive price.

In preparation for a possible move towards the CB assessment of IELTS writing some years ago, research was carried out to investigate differences between the CB and PB testing of IELTS writing (Weir, O’Sullivan, Yan & 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 the study raised. In addition, although some research has also been completed into the cognitive processes of writers completing IELTS writing tasks (e.g. Yu, Rea-Dickins & Kiely, 2011 on AWT1), no research has yet been conducted into the central issue of cognitive validity across the two writing modes. Shaw and Weir define cognitive validity as the extent to which the chosen task "represents the cognitive processing involved in writing contexts beyond the test itself" (2007, p.34). In other words, we are interested in the extent to which test-takers’ cognitive processes while completing IELTS Academic Writing in PB mode and in CB mode reflect the real-world cognitive processes of students completing academic writing at university. Such research is a necessary pre-requisite to the introduction of a CB version of the IELTS writing test.

This research study, therefore, aims to investigate the issue of cognitive validity as part of establishing equivalence across the two delivery modes, by examining the differences between the cognitive processes of test-takers completing the IELTS Academic Writing essay task (AWT2) in PB and CB modes. It is anticipated that the findings of this research will inform potential moves to offer computerised versions of the IELTS writing tasks, and computerised versions of other international tests, in future.

# 3 Context of the study

## 3.1 Rationale and previous research

### 3.1.1 Computer-based academic writing in International Higher Education

IELTS is now accepted by over 6000 university-level institutions in over 135 countries (including over 2000 in the US). Once accepted onto academic courses, millions of tertiary level students will, of necessity and with minimal further preparation, be required to plan, take notes, write, monitor, revise and submit extended pieces of written work via the computer, given that the essay is a staple of academic assessment (Newman, Couturier & Scurry 2010). In fact, many universities no longer permit the preparation and submission of academic essays which are not in electronic format, partly to allow for submission to online plagiarism checkers such as Turnitin (Turnitin, 2016). This strong momentum towards CB mode in academic assessment in universities worldwide, who indeed have been the major score users for IELTS in recent years, means that if the IELTS Academic Writing test is to retain credibility with this core constituency, it will need to plan for CB modes in the near future, so as to demonstrate that its assessment practice follows the real-world academic writing experience closely. It is essential that IELTS responds to this challenge in order to maintain its position in the field of academic language assessment.

A further important group with which IELTS needs to maintain its reputation is the test-takers themselves, particularly students applying for university entrance. Furneaux (2013) recently paraphrased an IELTS candidate's comment that he was "good in IELTS but rubbish at real academic writing". This perceived gap between aspects of the IELTS writing assessment and real-world academic writing, if not addressed, could potentially be harmful to IELTS' reputation in the long term. Modern students are often 'digital natives', more at ease with technology than in the past. However, this is not the case with all students, and indeed the situation has often been overstated (Jones, Ramanau, Cross & Healing, 2010; Bennett & Maton, 2010). Nonetheless, it is important that students taking the IELTS Writing test perceive the relevance of their test experience to their current and future experience of writing at university. This again suggests the need for IELTS to move towards CB writing assessment in the coming years.

Introducing a computer version of the IELTS Writing test naturally necessitates a research base from which such a CB mode could be reliably and confidently launched. Below, we review in greater detail the existing research into the IELTS CB mode, but it will be immediately apparent that the quantity and range of up-to-date research into the CB mode of delivery relevant to IELTS does not seem, at present, to provide a sufficiently robust research base from which to launch CB tests. This is the impetus behind the research proposal in this study, which aims to begin to supply that research base.

In other words, the time is ripe to consider such a move towards IELTS CB delivery. Not only are the potential advantages of CB delivery in terms of cost, security, efficiency and so on now established (Al-Amri, 2008; Puhan, Boughton, & Kim, 2007, citing Jodoin, 2003; Wise & Plake, 1990) but the changing nature of academic study and increased student familiarity with technology, as a generation of 'digital natives' enters university, means that a move towards technology is now a given in all areas of academic life. As Duderstadt, Atkins, & Van Houweling (2002: 24) note:

> *Soon those institutions skilled in the use of technology to improve learning will be seen as more dynamic and effective than their less engaged competitors… institutions and faculty that view themselves as excellent at teaching now need to excel at the use of technology if they are to remain leaders.*

The same is held true for language testing organisations.

### 3.1.2 The use of CB writing in academic writing assessments

In line with the shift towards computer-based academic writing in real-life, many international language testing organisations are moving towards the CB testing of writing, in some cases, abandoning PB mode altogether. Cambridge English Language Assessment has more than 350 test centres in 64 countries for its CB versions of KET, PET, FCE, CAE and CPE – in all of these, the writing component can be taken in CB versions. The TOEFL iBT in CB mode can already be taken in 1355 test centres in 149 countries – indeed, the PB format is now being phased out completely. A total of 96% of TOEFL test-takers worldwide take the TOEFL iBT® test, and that proportion is continuing to increase. Pearson offers the PTE Academic test, including academic writing, 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 2012:7). LTTC in Taiwan offers GEPT Advanced Writing Paper in both PB and CB modes. The British Council has launched a CB test – Aptis. Almost all major academic writing assessments offer some forms of CB essay tasks. In other words, the momentum towards CB mode among IELTS' competitors is compelling.

The paper-based IELTS Academic Writing test has established a strong reputation over many years for rigour and reliability, and is accepted by thousands of institutions worldwide for that reason. If IELTS is to move towards a CB version, perhaps initially alongside the PB version to ensure accessibility worldwide, and seeks at the same time to retain its current reputation, it must first ensure through appropriate research that the two modes deliver comparable results on all dimensions. As Weir et al (2007) noted in their research into PB and CB delivery of IELTS Writing tests, "this issue of equivalence cannot simply be ignored or assumed (ibid: 5)". Furthermore, Weir at el, drawing on work by McDonald (2002) and on Mead and Drasgow's (1993) often quoted meta-analysis of 159 correlations from published and non-published research in psychology, identify two fundamental types of equivalence which CB test needs to demonstrate as a minimum, namely, equivalence of scores on which test results are placed, and equivalence of the underlying construct that is being measured (as do, for example, Wolfe & Monolo, 2005).

We will now consider these two areas in turn, in terms of the CB and PB testing in general, and in terms of language testing in particular.

### 3.1.3 Score equivalence

In terms of score equivalence, a large body of research has concluded that, depending on appropriate design, the scores across the CB and PB modes can be considered comparable (e.g. Puhan, Boughton & Kim, 2007; Taylor, Jamieson, Eignor & Kirsch, 1998; Weir et al, 2007; Wise & Plake, 1989). Early research by Mazzeo and Harvey (1988) had 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 technology. More recent studies have concluded, however, that there is substantial comparability between scores in the two modes, perhaps in part owing to test-takers' increasing familiarity with computer use in education and daily life. This has also proved to be the case for research into language testing.

Taylor et al (1998), for example, 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 achievement tests yield very similar scores. More recently, 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". Puhan et al (2007) examined over 1000 participants in a test of writing in CB and PB modes, and also 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 CB versions was not significant.

The broad message from this is that, so long as the test design is carefully moderated, and assuming appropriate familiarity, attitude and anxiety levels on the part of test-takers, score equivalence is possible across the two modes in large scale tests of writing. Nevertheless, given that score equivalence in IELTS Writing across the two modes remains largely under-researched, it will form part of the research in this study. A questionnaire will also be used to examine the variable of computer familiarity and anxiety.

### 3.1.4 Cognitive validity and cognitive equivalence

Score equivalence is insufficient in itself to demonstrate the complete equivalence of CB and PB test modes. Weir (2005) points out that for a criterion-referenced test like IELTS, criterion-related decision consistency should be the major concern for test developers, i.e. the consistency in judgements of whether a set criterion has been met, rather than consistency of scoring in it. This in turn means that:

> *[i]n determining test equivalence we need to establish that the processing in CBA and P&P mode are similar in terms of theory-based validity.*
> (Weir et al. 2007:8, emphasis added).

This is because of the danger that "the different modes may be activating different executive processing within the candidate – therefore, making performance different in terms of interactional authenticity" (ibid: 9). The implication of this is that language test providers need to establish for both modes that the cognitive processes which a candidate draws on when completing the test writing task(s) are an accurate and comprehensive representation of the types of processing required in writing tasks in the real-world target setting (Glaser, 1991; Shaw & Weir 2007; Field, 2013). It must establish that this is comparatively equivalent for both CB and PB modes. In other words, it is important to ensure both cognitive equivalence, as well as score equivalence, if the two modes are to be used side by side.

An important part of this study is, therefore, the comparison between real-world cognitive processes, on the one hand, and those used on PB and CB in IELTS AWT2 test mode on the other. In this regard, the project uniquely draws on recent research by Chan (2013), building on a body of writing processing literature including Hayes and Flower (1980), Field (2004), and Shaw and Weir (2007), to investigate cognitive processing among L2 students in essay tasks in a real-life academic context. The parts of Chan's study related to academic essay writing forms the baseline with which cognitive processing in IELTS AWT2 writing can be compared. In broad terms, she investigated the processes involved in five phases of academic writing.

*Cognitive parameters for the analysis of academic writing (adapted from Chan, 2013)*

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

Chan's research investigated L2 students' writing in genuine academic settings, and as a result, identified a list of key cognitive processes which, in her view, should be at the heart of an academic reading-into-writing assessment. With the exclusion of two processes, (Careful reading (local/global) and Scanning, Skimming and Search reading, which relate specifically to the reading texts which served as input in her study), this list provides a useful baseline as to the cognitive processes which L2 writers in real academic contexts are likely to employ. The present project draws on this work to investigate the extent to which, and in what ways, Chan's conclusions regarding the key cognitive processes identified above are, in fact, mirrored in the PB and CB versions of the IELTS AWT2 test, and how the IELTS test in both modes might in future be adjusted in light of the comparison.

### 3.1.5 Cognitive processing in IELTS Academic writing

IELTS AWT2 requires participants "to write an essay in response to a point of view, argument or problem" in formal style. Participants are assessed on "their ability to present a solution to a problem; to present and justify an opinion; to compare and contrast evidence, opinions and implications; to evaluate and challenge ideas, evidence or an argument" (IELTS, 2013:5). IELTS AWT2 has been investigated in a number of studies (e.g. Mickan & Slater, 2003; Mickan, Slater & Gibson 2000), although not in terms of test-takers' cognitive processing.

The cognitive processing of participants completing IELTS Academic Task 1 (AWT1) (in PB mode) was investigated in detail by Yu et al (2011). Although they were correct to report that Academic Task 2 (AWT2) has received more research attention in general, the cognitive processing of participants on AWT2 has largely been neglected in research terms. In particular, the cognitive processing of participants taking AWT2 in CB mode has not previously been researched – this is an important gap in the research base if the IELTS Writing test is to be computerised in future.

Yu et al (2011) adopted a layered approach to data collection, using the think-aloud approach for the main part of their study, and concluded by offering a model of cognitive processes consisting of three interrelated stages, specific to AWT1. They did not explicitly compare the cognitive processes used by their test-takers with those used in similar real-world academic tasks in order to help establish the cognitive validity of the AWT1, although by implication, it seems to have been assumed that the cognitive processes they discovered under test conditions were, in general, of a kind relevant to genuine academic environments.

The research reported in our study builds on the work of Yu et al (2011) by researching the cognitive processes of test-takers completing AWT2 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. Furthermore, it will compare these cognitive processes with those used by second language (L2) students under genuine academic writing conditions at a UK university, to help to establish the test's cognitive validity in both modes. In this latter respect, as noted above, the project draws on recent data collected by Chan (2013), resulting from a detailed analysis of the key cognitive processes of L2 academic writers completing authentic writing tasks in a UK university context, data which act as an important baseline for the study. This has given us an opportunity to identify areas in which the IELTS AWT2 can be improved in terms of its cognitive validity vis-à-vis real-world academic writing, and at the same time, help to compare the cognitive processing of CB and PB modes. In this way, this research provides important evidence in these two key areas to inform decisions on whether or not the partners should launch a CB version of the IELTS writing component.

It could not be assumed that the cognitive processes of participants taking AWT2 in CB and PB modes will necessarily be the same. In a useful review of the ways in which the use of the computer can be either obstructive or advantageous in the writing process, Shaw (2005) reports on earlier studies, such as Hermann (1987) which found that the use of a computer could interfere with the composing process. However, later studies, presumably as computer use become more commonplace, began to find that "regular use of computers for writing over extended periods can lead to significant improvements to students' writing skills" (Shaw 2005:15). Most significantly in terms of cognitive processing, it is possible that "[in] the paper-based mode of composing, writers often expend considerable time and energy in intensive planning prior to writing in an attempt to obviate the need to rewrite or recopy text" (idem) whereas the CB mode may lead to planning during, as opposed to mainly before, the process of text production itself (Haas 1989).

To some extent, this picture may have changed in recent years, but in any case, we should expect to find that cognitive processing involved in planning and revising might be different in the two modes. Particularly in the revising process, it is possible that the CB mode will show not only different attention devoted to revision, but also different types of editing activity. It has been shown that L2 writers often revise more when composing in CB mode (Chadwick & Bruce 1989; Li & Cumming, 2001), while other studies have shown differences in both the quality of revisions and the time spent on revising in CB mode (Phinney & Khouri, 1993). In this study, therefore, careful attention is paid to areas of planning and revising, and also to other areas in which the use of the computer might engender different cognitive processing operations from PB writing.

In summary, before decisions can be made on the introduction of tests which allow for alternative output modes (in this case, PB and CB modes), it is important to gain a greater understanding of the underlying cognitive processes activated by these modes – in particular, we need to investigate the cognitive validity of test tasks employing these modes, and their cognitive equivalence. Ascertaining and ensuring this validity and equivalence is an important part of the research project proposed here. In order to do this, we are gathering data on the cognitive processing of participants on the CB and PB tasks by the means of a Writing Process Questionnaire (see Appendix 3) and retrospective interviews, as well as essential data on their aptitude with computers, issues to be addressed further below.

### 3.1.6 The potential impact of writers' computer familiarity on performance

Delivery mode has always been identified as one of the variables which potentially might have an impact on writers' performance (Shaw & Weir, 2007). Although the use of computers in academic writing has become very common, there is still some residual concern that a number of 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 (Hertz-Lazarowitz & Bar-Natan, 2002; McDonald,2002) on the presence or absence of any impact. This important dimension was, therefore, investigated in this study by means of a self-report questionnaire.

## 3.2 Research questions

The research questions for the study are as follows:

1. **Score equivalence:** Are there significant differences in the scores awarded by independent raters for performances in CB and PB mode in IELTS Academic Task 2 (AWT2)?
2. **Cognitive equivalence:** Do test participants use cognitive processes differently in completing IELTS Academic Writing Task 2 in CB mode and PB mode respectively?
3. **Are any of the independent variables** (computer familiarity, anxiety, etc.) effective in predicting test scores? (i.e. Is there any significant and meaningful link between results and computer familiarity or anxiety, etc.?)

# 4 Research design, data collection and analysis

## 4.1 General approach

To address these research questions, the study adopts a mixed-methods approach because a combination of the use of both qualitative and quantitative methods provides a better understanding of research matters than either approach alone (Johnson & Onwuegbuzie, 2004). Research tools utilised include two different questionnaires (Writing Processing Questionnaire and Computer Familiarity Questionnaire) (see Appendices 2 and 3), retrospective interview and score analysis.

## 4.2 Participants

### Test-takers

Students studying on undergraduate programs at a British university were recruited. Their English proficiency was estimated as ranging from B1 to C1 based on students' entrance profile. Table 1 presents their general language proficiency as demonstrated by IELTS scores. 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 included in the study, mainly because it was thought that writing mode, i.e. paper and computer in this case, would be likely to have more impact on writers at a lower proficiency level than those at a higher level. A total of 153 students participated in the study: 45.4% were male and 54.6% female. They came from a variety of major disciplines, including Business, Language and Communication and Computing.

**Table 1:** *Participants' previous IELTS scores*

| Overall IELTS | | Writing | |
|---|---|---|---|
| Band range | Percentage of participants | Band range | Percentage of participants |
| 4.5 | - | 4.5 | 2 |
| 5.0–5.5 | 34.6 | 5.0–5.5 | 35.3 |
| 6.0–6.5 | 54.2 | 6.0–6.5 | 52.3 |
| 7.0–7.5 | 10.5 | 7.0–7.5 | 9.2 |
| 8 | 0.7 | 8 | 1.2 |
| | 100 | | 100 |

### Raters

The scores reported were double rated by four certified IELTS raters approved by the British Council. Rater 1 marked all the scripts whereas Raters 2–4 each double marked a sub-set of the scripts. (Note: The overall scores were rounded down as in the operational IELTS test, i.e. 5.75 becomes 5.5, 5.25 becomes 5.0.) Inter-rater reliability, raters' severity and percentage of absolute agreement will be reported in Section 5.1.

## 4.3 Instruments

### 4.3.1 Test tasks

Two publicly available sample AWT2 tasks were used in a pre-pilot which involved 11 students. The feedback suggested that one of the tasks was more demanding than the other in terms of the topic domain. Another concern was the open access of the two sample tasks and the possibility that the participants might have seen them online previously. Therefore, upon the request of the research team, a set of retired AWT2 tasks was supplied by the test providers, eight of which were selected by the research team for further scrutiny. The eight tasks were presented to a panel of six expert language testing practitioners and researchers. They were asked to choose the two most comparable AWT2 tasks from the set in terms of topic, domain and language functions required. As a result, two AWT2 tasks were selected to be used in the main study of the research. The comparability of the two prompts will be discussed in Section 5.1.

In the study, each participant completed two tasks, one under the traditional PB mode and one in the experimental CB mode. In the CB mode, the participants composed the essay using Microsoft Word (see the tasks in Appendix 1). All proofing functions in the CB mode (e.g. grammar and spell check) were disabled. More information about the procedures of data collection is presented in Section 4.4.

### 4.3.2 Computer Familiarity Questionnaire

As suggested by previous research into comparability of CB and PB modes, participants' familiarity with computers might have an impact on their performance in a CB test. A Computer Familiarity Questionnaire was, therefore, deployed, adapted from Weir et al's (2007) study, with modifications to reflect the present research context. As a result, the Computer Familiarity Questionnaire (see Appendix 2) consists of 14 closed questions and one open-ended question. Q15, the open-ended question, asks the participants whether they would prefer to take the IELTS Academic Writing test on paper or computer. The structure of the questions is as shown in Table 2.

**Table 2:** *Structure of the Computer Familiarity Questionnaire*

| Categories | Questions |
|---|---|
| **Computer usage** | Q1, Q2, Q3, Q4, Q5 |
| **Comfort and perceived ability** | Q6, Q7, Q8, Q9, Q14 |
| **Interest in computers** | Q10, Q11, Q12, Q13 |

### 4.3.3 Writing Process Questionnaire

A Writing Process Questionnaire was then adapted from Chan's (2013) study on cognitive processes in academic writing. The questionnaire was modified to suit the task features of IELTS AWT2, with a total of 40 items (see Appendix 3). The internal consistency reliability of items representing each cognitive phase was checked to ensure that each category of items were measuring the same theoretic construct, i.e. the different processes in each cognitive phase in this study, see Table 3.

**Table 3:** *Structure of the Writing Process Questionnaire*

| Cognitive phases | Items | Internal consistency reliability |
|---|---|---|
| **Conceptualisation** | Q1, Q2, Q3, Q4, Q5, Q6, Q7, Q15, Q23, Q24 | 0.767 |
| **Generating ideas** | Q8, Q9, Q10, Q11, Q22 | 0.605 |
| **Organising ideas** | Q12, Q13, Q14, Q19, Q21 | 0.711 |
| **Generating texts** | Q16, Q17, Q18, Q20 | 0.722 |
| **Monitoring and revising at high-level** | Q25, Q26, Q27, Q28, Q29, Q33, Q34, Q35, Q36, Q37 | 0.895 |
| **Monitoring and revising at low-level** | Q30, Q31, Q32, Q38, Q39, Q40 | 0.793 |

### 4.3.4 Interview

Retrospective interviews were conducted to explore further the participants' cognitive processing on AWT2 in the two modes. Twenty percent (20%) of the participants (n=30) were interviewed individually by the research team immediately after each test event, with participants recruited on a voluntary basis. The average of their PB and CB bands were the same but the standard deviation of their CB band was slightly higher (PB - M=5.80, SD=0.49; CB - M: 5.80, SD: 0.55). Most of the interviewed participants had received the same band under the two conditions; 16.7% had a difference of 0.5 band, and 13.4% a difference of 1 band. The interviewees' performance across the conditions were, therefore, considered to be broadly equivalent. All interviews were voice recorded, and the recordings were transcribed by two research assistants. Ten percent (10%) of the transcripts were double-checked by a member of the research team to ensure accuracy. Information about analysis of the interview data is reported in Section 4.5.

## 4.4 Data collection

The research team visited classes of Academic Writing offered by the Department of Language and Communication to explain the overall research aim and recruit students. To encourage the students to participate in the study, we offered to send them written comments of their performance on the two tasks and arrange some open sessions for those who were interested to discuss their performance on the tests. Students who agreed to participate in the study were informed of the procedures of the test event and their assigned test time. Test events took place one week after the class visits.

With the support of the lecturers, about 15 test events were conducted during their normal class time. In each test event, participants first completed ethics procedures, and then were divided at random into two groups. Table 4 presents the data collection procedures. Each group, in a counter-balanced order, completed two AWT2 tasks (Prompt 1 and Prompt 2) on paper and computer. The order of the prompt was counterbalanced in each test event. Each test was 40 minutes long and the two tests were administered back to back (for details see Table 4).

**Table 4:** *Data collection procedures*

| Group A | Group B | Time (Mins) |
|---|---|---|
| All participants filled in a Computer Familiarity Questionnaire | | 5 |
| Completed AWT2 on paper | Completed AWT2 on computer | 40 |
| All participants filled in a Writing Process Questionnaire | | 10 |
| Completed AWT2 on computer | Completed AWT2 on paper | 40 |
| All participants filled in a Writing Process Questionnaire | | 10 |
| 20% of the group were interviewed individually | | 20 |

As previously mentioned, 153 students participated in the study. A total of 306 scripts (153 paper scripts and 153 computer scripts) were collected. The data from all questionnaires (including 153 Computer Familiarity Questionnaires and 306 Writing Process Questionnaires) were entered into a spreadsheet using SPSS v22. Scripts which received no mark due to illegible writing or inadequate writing sample were excluded. Questionnaires with 30% or above missing data were also excluded. The final totals of each data point is shown in Table 5.

**Table 5:** *Final totals of each data point*

| Data point | N |
|---|---|
| Paper-based scripts | 136 |
| Computer-based scripts | 138 |
| Computer Familiarity Questionnaire | 128 |
| PB Writing Process Questionnaire | 135 |
| CB Writing Process Questionnaire | 145 |

## 4.5 Data analysis

### RQ1: Score equivalence between CB and PB mode

To investigate score equivalence under the CB and PB conditions, two Multi-Facet Rasch Measurement (MFRM) analyses using FACETS 3.71.2 (Linacre, 2013) were conducted. First, a 5-facet analysis with test-takers' writing ability, testing mode, essay topic (prompts), raters and rating category was carried out to understand the impact of each of the above facets on scores and to investigate score equivalence across the conditions. This analysis compared test-takers' overall scores between the two modes. In order to investigate test-takers' performance on each analytic rating scale (i.e. Task Achievement, Coherence and Cohesion, Lexical Resources, and Grammatical Range and Accuracy) between the delivery modes, four further 4-facet analyses were conducted. In each of the 4-facet analyses, delivery mode was not designated as a facet. The individual analytic category between the modes, e.g. CB Task Achievement and PB Task Achievement, were treated as separate items. In other words, the four pairs of analytic scales between the delivery modes were compared one by one.

In all analyses, the data were entered into the Rating Scale Model (RSM), which operates under the assumption that the rating scale associated with each item functions similarly.

### RQ2: Cognitive equivalence between CB and PB mode

The cognitive processes employed by the test-takers in CB and PB mode in IELTS Academic Task 2 (AWT2) were measured through the Writing Process Questionnaire. Descriptive statistics of individual questionnaire items from the CB and PB modes were obtained. As the data of most items was not normally distributed, non-parametric Wilcoxon signed-tank tests were, therefore, used to compare the results of the two modes. The results were also compared descriptively to the findings reported in an earlier study (Chan, 2013) with regards to students' cognitive processes on academic writing tasks in real-life.

In addition to questionnaire data, retrospective interviews were used to provide supplementary findings to the participants' writing processes between the two modes. The 30 transcripts were coded using NVivo v10 (see Appendix 4 for some examples).

### RQ3: Relationship between affective variables and test performance in CB and PB mode

To find out whether the affective variables have any effect on students' performances in the two modes, descriptive statistics were calculated for the scores of participants who chose the options of definitely agree/always and mostly agree/often for each item in the Computer Familiarity Questionnaires (CFQs). After confirming that the data met the pre-requisites for the analysis (including normality, homoscedasticity, linearity, no multicollinearity and no outliers), the items were submitted to Multiple Regression to examine their impact on the CB test performance. Stepwise method, which includes or removes one independent variable at each step, based on the probability of F, was chosen.

# 5 Results and discussion

## 5.1 Score equivalence between CB and PB mode (RQ1)

We first report findings regarding raters' reliability and severity, comparability of the prompts, and score equivalence between the two modes from the 5-facet MFRM analysis. After that, we report findings regarding individual analytic scores between the two modes from the 4-facet MFRM analyses.

Regarding raters' reliability and severity, Rasch logit scale and the Infit mean square index as a measure of fit (i.e. meeting the assumptions of the Rasch model) are reported in Table 6.

**Table 6:** *Rater measurement report*

| 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%

Before we discuss the severity of the raters, it is worth mentioning again that Rater A rated all scripts, whereas Raters B, C and D each rated a sub-set of the scripts as the second rater. As indicated by the Logit measure in Table 6, 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. In addition, Infit values for all the raters fall within the acceptable range. Although Infit values in the range of 0.5 to 1.5 are 'productive for measurement' (Wright and Linacre 1994), the range between 0.7 and 1.3 is usually taken as the acceptable range of the Infit value (Bond and Fox, 2007). Given that IELTS is a high-stakes test, we refer to the stricter acceptable range in this report. The exact agreement between the first and second rater was 66.8%.

Table 7 reports results regarding the comparability of the prompts used in the study. Judging by the observed mean and logit measure, Prompt 2 was significantly easier than Prompt 1 ($X2=77.6$, $p<0.01$). However, while Prompt 2 was significantly more difficult than Prompt 1, 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 previously, the administration of versions was counter-balanced, any order effects being minimised. Therefore, we have confidence that test-version effect should not invalidate the findings of this study.

**Table 7:** *Version measurement report*

| 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

Having showing that the raters' reliability and severity, and comparability of the prompts, were satisfactory, we now present the results for the delivery mode measurement, which is of most importance for addressing RQ1. As indicated by the fixed chi-square statistics in Table 8, test scores obtained from the CB and PB mode were not statistically different in terms of difficulty ($X2=1.8$, $p=0.18$). Test-takers' performance, in terms of observed mean and fair mean scores under the PB and CB conditions, were very close, with a difference of 0.12 in observed mean and 0.03 in fair mean. The lack of misfit data indicates that test scores obtained from the CB and PB 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 1.

**Table 8:** *Delivery mode measurement report*

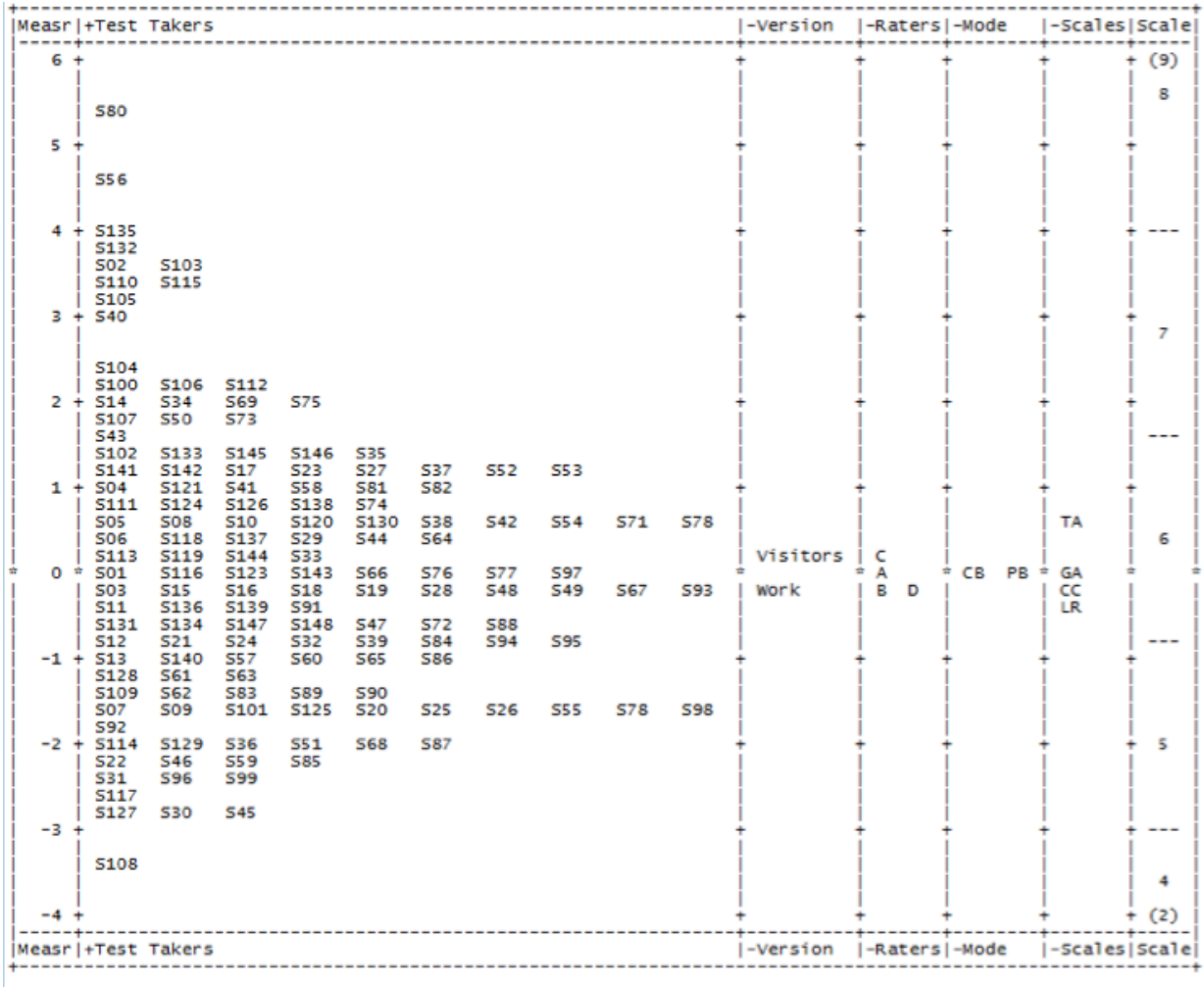
| Test mode | N | Observed mean | Fair mean | Logit measure | Standard error | Infit mean square |
|---|---|---|---|---|---|---|
| **Computer-based** | 1104 | 5.75 | 5.83 | .04 | .05 | .97 |
| **Paper-based** | 1088 | 5.87 | 5.86 | -.04 | .05 | 1.02 |

(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

**Figure 1:** *FACETS variable map*




In short, according to the 5-facet MFRM analysis, there was no significant difference in the overall PB and CB scores. Now, we report score equivalence findings in relation to the four analytic scores under both conditions based on the additional 4-facet MFRM analyses (see Tables 9–12).

**Table 9:** *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 10:** *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 11:** *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 12:** *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

According to Tables 9–12, as indicated by the fixed chi-square statistics, while differences in three of the four analytic scores (i.e. Task Achievement, Coherence and Cohesion, and Grammatical Range and Accuracy) were not significant, the difference reported in Lexical Resources was significant ($X2=8.2$, $p<0.01$). In terms of the fair mean of the Lexical Resources scores, test-takers scored 0.07 higher in the PB than CB conditions (see Table 11). This is consistent with the findings indicated by the interview that the participants were more cautious about their word choices in the PB mode (for details see Section 5.2). The difference was very small and, hence, as reported previously, did not contribute to a significant difference in the test-takers' overall scores between the two modes. Nevertheless, the fair mean Lexical Resources scores were below 6.0 in CB mode but above 6.0 in PB mode. It is, therefore, recommended that test providers should monitor closely test-takers' performance on Lexical Resources between the two modes.

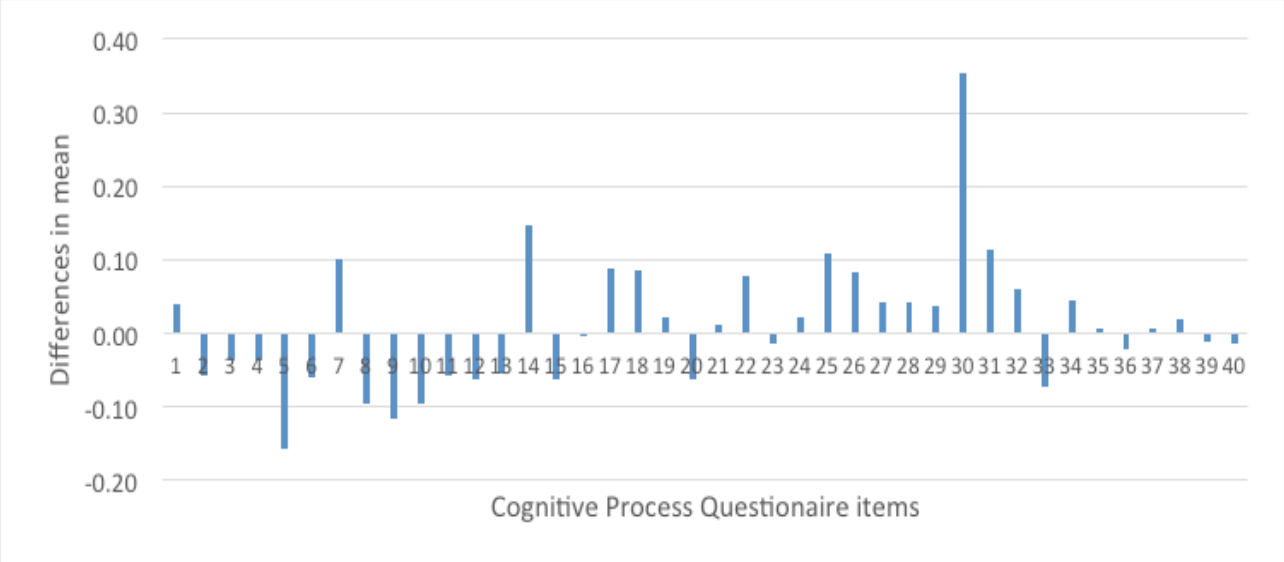
## 5.2 Cognitive equivalence between CB and PB mode (RQ2)

### 5.2.1 Evidence from questionnaire

To establish the cognitive validity of any writing test, it is essential to take into account the cognitive processes which writers actually employ when they write in the real-life situation. The findings of Chan's (2013) recent study, which consists of detailed analysis of the key cognitive processes of L2 academic writers completing authentic writing tasks in a UK university context, act as an important baseline of the real-life reference for this study. In other words, we aim to examine the extent to which these key cognitive processes identified in her study are, in fact, mirrored in the PB and CB versions of the IELTS AWT2 test.

Participants were asked to rate the extent to which, on a Likert scale of 1 to 4, they employed each cognitive process in completing the writing test immediately after completing the task under each condition (i.e. the paper and computer modes). The difference in means between the two modes are presented in Figure 2. It can be seen that there were no noticeable major differences in the general tendency of the mean of each questionnaire item reported on the two test modes. Most differences were 0.15 or below out of a 4-point scale. This indicates that the cognitive processes were employed to a comparable extent under the two delivery conditions. Only Item 30 showed a difference higher than 0.3. Participants reported checking the accuracy and range of the sentence structures more in PB than CB mode.

**Figure 2:** *Differences in mean between the PB and CB Writing Process Questionnaire items*




To gain a clearer picture of participants' reported processes in terms of each cognitive phase (for the categorisation of the questionnaire items, see Table 3), Table 13 compares the means of all items in the six cognitive phases between the two modes. The findings from Chan's (2013) study of undergraduate processing in completing academic writing tasks at a British university are also provided as a baseline real-life reference. It should be noted that no inferential statistics were performed to compare the results of this study and Chan's study. Therefore, the descriptive comparison provided below needs to be interpreted with some caution.

**Table 13:** *Mean of processes in each cognitive phase*

| | Computer-based | | | Paper-based | | | Real-life | | |
|---|---|---|---|---|---|---|---|---|---|
| | 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 |
| **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).

The findings show that the means of each cognitive phase obtained under the test conditions (both CB and PB modes) were between 3.17 and 3.40 (4=definitely agree; 3=agree; 2 disagree; 1=definitely disagree). This implies that participants in this study reported using all the key cognitive processes identified in Chan's study in both the PB and CB versions of the IELTS AWT2 test. In relation to the figures from Chan's (2013) study, the means obtained in this study were largely comparable to those reported under the real-life conditions. Based on the results that higher-scoring writers tended to report using these processes more than lower-scoring writers in authentic academic writing tasks, Chan (2013) argued that a valid academic writing assessment should target these key cognitive processes. The fact that all means under both the IELTS AWT2 test conditions were slightly higher than the real-life figures yields positive evidence supporting the cognitive validity of test. Although all differences between the test and real-life conditions, appear to be small (ranging 0.06 to 0.23), the most noticeable difference was obtained in the processes of monitoring and revising. 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, the test-takers are aware that they are being judged on language rather than on content (unlike in a 'real-life' university context), so it is perhaps understandable that the test-takers are doing more monitoring and revising. For this reason, there is a need to further investigate writers' monitoring and revising processes between test and real-life conditions in future studies.

Comparing the results between the CB and PB modes, a similar pattern was found in the way the participants reported the extent to which they employed the processes of the six cognitive phases. While the means in monitoring and revising at low-level were the same across conditions, the means in conceptualisation, generating ideas, generating texts and organising ideas were very close between the two modes (see Table 13). The participants reported doing slightly more monitoring and revising at the high-level under the computer-based condition than the paper-based condition. This will be further discussed based on the interview data. The obtained differences were further subjected to Wilcoxon signed-ranks tests. The Wilcoxon signed-ranks tests indicate that candidates' reported processes of the six cognitive phases in the PB mode are not significantly different from that in the CB mode (see Table 14).

**Table 14:** *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 | -0.065 | 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 | | |
| Monitoring and revising (low-level) | CB | 3.17 | 66.5 | 0.000 | 1.000 |
| | PB | 3.17 | 66.5 | | |

### 5.2.2 Evidence from interview data

One-fifth of the participants (n=30) were interviewed. Based on the interview data (see Appendix 4 for some coding examples), we include below a phase-by-phase description of the processes involved in writing the essay under the two different conditions. We focus the discussion on the similarities and differences reported between the two conditions.

### 5.2.3 Conceptualisation

This is the 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 started approaching the task. They began by reading the task prompt and instructions carefully and then planned what and how they were going to write to fulfil the task requirements. They did not report any difficulties understanding the tasks. The interviewed participants were primarily concerned about the content and structure of their essay at this stage of macro-planning. However, participants tended to be more cautious with their planning under the paper-based mode. Many reported writing down a plan or key ideas. Most of them 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 such evident changes would lead to a lower mark due to untidiness. This concern reoccurs in other phases. In contrast, participants appear to be more relaxed with their initial planning under the computer-based mode. They believed they did not need to start with a perfect plan before they wrote. They felt more comfortable making changes to the plan or to the essay under the CB mode.

### 5.2.4 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 according to 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 some concluding points. 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 participants then reported how they organised their ideas to achieve the writing purpose. The interview revealed some differences in how they did so between the PB and CB modes. 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. They did the same on 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 organising processes sometimes overlapped with the online editing processes as they re-organised the order of their clauses and sentences. These will be discussed in more detail below.

### 5.2.5 Generating texts

This is a phase when writers translate their mental ideas into words. On the PB mode, they execute this process via writing with a pen, whereas they type in the CB mode. Apart from this obvious difference, the participants revealed some interesting findings. As mentioned previously, most participants in this study were reluctant to make changes to their essay in the paper-based mode. They reported that they were more careful when generating texts in the PB than the CB mode. Some described how they would think more carefully with their choice of words and sentence structures. In contrast, in CB mode, they tended to focus more on 'getting the ideas out' during this phase and then they made changes as they saw appropriate or at a later phase. Such difference may have an important implication for writers at lower proficiency level, as the executing process is perhaps disturbed by an urge to produce 'perfect language' at the first attempt. 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. A few participants reported concerns that their typing speed or accuracy was not as good as their hand-writing. However, their concerns were not reflected in the score bands they received between the two modes.

### 5.2.6 Monitoring and revising (online and after writing/ at low- and high-levels)

Online monitoring and revising refers to the changes writers make to their text as writers write. They can make changes at the current position of their text or to the previously produced text. These changes can be at the levels of words, phrases, clauses, sentences and paragraphs, and on different aspects of writing quality, such as accuracy, coherence and line of argument. Thirty five percent (35%) of the interviewed participants reported that they did not do any online editing in the PB mode as they felt it was inconvenient to make changes to existing texts. In comparison, 25% of interviewed participants reported that they did not make any changes to the texts as they wrote in the CB mode. Participants also reported monitoring and making changes to the previously produced text after they had finished the whole draft.

It is argued in the literature that the monitoring and revising phase is essential to successful academic writing, as writers read and evaluate their text carefully checking it against their writing goals and/or the task assessing criteria (e.g. Kellogg, 1996; Shaw & Weir, 2007). Consistent with the findings regarding online monitoring and revising, more interviewed participants reported engaging in post-writing monitoring and revising in the CB than PB modes (i.e. 70% vs 60% of the interviewed participants). This tendency of writers to revise more in the CB mode than the PB is consistent with what is reported in the literature (e.g. Cohen, 1994). In terms of aspects of the changes, 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 changes to correct 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 such findings.

## 5.3 Relationship between affective variables and test performance in CB mode (RQ3)

The Computer Familiarity Questionnaire (CFQ) was administered to investigate participants' familiarity with computers in terms of three aspects, i.e. Computer Usage, Comfort & Perceived Ability and Interest in Computers (see Appendix 2). The vast majority of participants reported using computers frequently at home and university, and they used computers for a variety of purposes, including surfing the Internet, electronic communications, study-related activities and, to a lesser extent, games and graphics. Based on the frequency data, a descriptive summary is provided in Table 15.

**Table 15:** *Descriptive statistics of the Computer Familiarity Questionnaire (CFQ)*

| 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 and 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 a paper |
| | 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.

When compared to relevant data in the literature (e.g. Weir et al, 2007), 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 15). In particular, there is a remarkable increase in the percentage of participants who have frequent access to computers and use them at home. Many more participants use computers for study-related activities and frequently use word processing than in Weir et al's 2007 study. Also, many more participants appear to be more comfortable using a computer to write a paper and take a test on computer now, as compared to then. But interestingly, there is seemingly a slight decrease in participants' interest in computers as addressed by Q11 and Q12, while the computer has clearly become a necessity for study/work.

We next investigated which, if any, of these aspects of participants' familiarity with computers have an impact on their performances on IELTS AWT2 in the computer mode. 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' performance on the CB task.

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 16, frequency of using computers for word processing (CFQ4b) ($\beta$ =.37, t=4.50, $p < .01$), access to computers at public library (CFQ1c) ($\beta$ =.17, t = 2.08, $p < .05$), and forgetting time when using computer (CFQ13) ($\beta$ = .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 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 mild, on the computer-based task.

**Table 16:** *Multiple regression analysis of CB scores on CFQ items*

| | B (unstandardised regression coefficient) | Standard error | β (standardised 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 | |
| **R2** | | | | | | .226 |
| **F** | | | | | | 11.280 |

As reported in the previous section, based on the results of MFRM, this study found no significant difference between participants' performance on the IELTS essay task between the two test delivery modes. However, the multiple regression analysis here indicates three of the computer familiarity intervening variables have a mild, but significant, impact on their performance in the computer mode. More importantly, this implies that test-takers who do not have such a familiarity profile are likely to perform worse than those who do, though there is no indication that they would perform better under the paper-based than computer-based condition. 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 and recommendations

## 6.1 Summary

This study set out to investigate whether meaningful differences would be observed, either in the scores achieved for written performance or in the cognitive processing of candidates, when a language test assessing academic writing was presented in two modes – pencil-and-paper and computer. Participants were asked to sit for a test consisting of two equivalent tasks in all respects, except that one was written directly on the computer, and the other written on paper. The resulting scripts were rated by trained and experienced approved IELTS examiners working independently. Before the test, candidates completed a Computer Familiarity Questionnaire (CFQ) which was intended to measure their familiarity with computers, and immediately after each task, they completed a Cognitive Process Questionnaire (CPQ) which was meant to gather evidence of the internal processing initiated during the two task performances.

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. Major supporting statistics for the conclusion include the fact that 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. In terms of the fair mean of the Lexical Resources scores, test-takers scored 0.07 higher in the PB than CB conditions but the difference was very small. In summary, the difference of scores between the two modes is at an acceptable level.

With respect to the cognitive processes students employ in performing under the two conditions of the test, results of the Writing Process Questionnaire indicate a similar pattern between the cognitive processes involved in writing on a computer and writing with paper-and-pencil. There were no noticeable major differences in the general tendency of the mean of each questionnaire item reported on the two test modes. 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.

In addition, 20% of the participants (n=30) were interviewed. Based on the interview data, it appears that the participants reported using most of the processes in a similar way between the two modes. Nevertheless, a few potential differences indicated by the interview data might be worth further investigation in future studies. For example, participants were more relaxed with their initial planning under the computer-based mode as they felt more comfortable making changes to the plan or to the essay in the CB mode. They tended to be more careful when generating texts in the PB than the CB mode. Some interviewed participants tended to engage more in organising ideas at the levels of sentences and paragraphs in the CB than the PB mode. 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 or argument in the CB mode.

The Computer Familiarity Questionnaire shows that, generally, these students are familiar with computer usage and their overall reactions towards working with a computer are positive. When compared to Weir et al (2007), participants nowadays, at least in the context of this study, appear to be more familiar and comfortable with using computers than eight years ago. In particular, there is a noted increase in the percentage of participants who have frequent access to, and use, computers at home. Many more participants use computers for study-related activities and frequently use word processing than in Weir et al's (2007) study. Also, many more participants appear to be more comfortable using a computer to write a paper and taking a test on computer now, as compared to then.

To find out if computer familiarity has any effect on students' performances on the two modes, multiple regression analysis was conducted. The results indicate that three 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, though there is no indication that they would perform better under the paper-based than computer-based conditions. It is, therefore, recommended that the test provider might consider using these items to provide advice about the candidates' readiness for taking the test in the computer mode.

## 6.2 Discussion

The findings of this study offer a useful addition to the equivalence debate by widening the normally accepted definition of equivalence to cover the cognitive processes initiated by the test tasks in different modes (cognitive validity), as well as score achievement. Our findings on equivalence are the more compelling, given that they are the first to take account of data on cognitive processes in writing in the real-world target academic situation.

The nature of this type of study, in particular, the difficulty in recruiting a large number of participants for two quite different task performances, means that the overall test population is relatively small. However, a population of about 150 is large enough for us to make relatively definitive statements about their performance. A further complication is that task difficulty is not a feature of the task itself but is affected by the interaction between test-takers and the task, and is therefore sample dependent, as are all reliability and correlation coefficients (Sawilowsky, 2000). What this means is that though two tasks 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 anchor groups take both forms of the test in each mode. In the study reported here, where it was inappropriate for a candidate to do the same test in both modes, we took the view that establishing acceptable boundaries of equivalence within which we could have confidence was a suitable modus operandi.

The score data support the findings of Neuman and Baydoun (1998) who found no significant difference in scores between the two modes. The discussion over which mode results in higher scores (Daiute, 1985, argued that computer mode would result in lower scores, while Russell and Haney, 1997, argued the opposite) is clarified by our results, where, as in Weir et al. (2007), it is clear that there was no significant difference in overall scores achieved by candidates between the two modes.

Nevertheless, 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-meaningful. 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 program 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.

Bennett, S. and Maton, K. (2010). Beyond the 'digital natives' debate: Towards a more nuanced understanding of students' technology experiences. *Journal of Computer Assisted Learning, 26(5),* 321–331.

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

Chadwick, S. and Bruce, N. (1989). The revision process in academic writing: From pen and paper to word processor. *Hong Kong Papers in Linguistics and Language Teaching, 12*, 1–27.

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

Daiute, C. (1985). *Writing and computers.* Addison-Wesley Longman.

Duderstadt, J. J., Atkins, D. E. and Van Houweling, D. E. (2002). *Higher education in the digital age: Technology issues and strategies for American colleges and universities.* Westport, CT: Praeger.

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

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

Furneaux, C. (2013). What are the academic writing requirements of Masters level study in the Humanities and how far can EAP proficiency tests, such as IELTS, replicate them? Paper presented at the *CRELLA Summer Research Seminar*, United Kingdom.

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

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 and 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. and 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.

IELTS (2013). *Information for participants*. Retrieved from http://www.ielts.org/pdf/Information%20for%20Participants_2013.pdf

Johnson, R. B. and Onwuegbuzie, A. J. (2004). Mixed Methods Research: A Research Paradigm Whose Time Has Come. E*ducational Researcher, 33(7)*, 14–26.

Jones, C., Ramanau, R., Cross, S. and Healing, G. (2012). Net Generation or Digital Natives: Is there a distinct new generation entering university? *Computers & Education, 54(3)*, 722–732.

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

Li, J. and Cumming, A. (2001). Word processing and second language writing: A longitudinal case study. *International Journal of English Studies, 1 (2),* 127–152.

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

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

Mickan, P. and Slater, S, (2003). Text analysis and the assessment of academic writing. In R. Tulloh (Ed), *IELTS Research Reports, Vol 4,* (pp. 59–88). IELTS Australia Pty Limited, Canberra,

Mickan, P, Slater, S, and Gibson, C (2000). Study of response validity of the IELTS writing subtest. *IELTS Research Reports, Vol 3*, (pp. 29–48). IELTS Australia Pty Limited, Canberra, .

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

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

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

Phinney, M. and Khouri, S. (1993). Computers, revision, and ESL writers: The role of experience. *Journal of Second Language Writing, 2,* 257–277.

Puhan, P., Boughton, K. and 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. and Haney, W. (1997). Testing writing on computers: an experiment comparing student performance on tests conducted via computer and via paper-and-pencil. *Education Policy Analysis Archives, 5(3),* 1–20.

Sawilowsky, S. S. (2000). Psychometrics versus datametrics: comment on Vacha-Haase's "reliability generalization" method and some EPM editorial policies. *Educational and Psychological Measurement, 60(2),* 157–173.

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

Turnitin. (2016). *User guides*. Retrieved from: http://guides.turnitin.com/01_Manuals_and_Guides

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

Weir, C. J. (2005). *Language Testing and Validation: An Evidence-Based Approach.* Basingstoke: Palgrave Macmillan.

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.

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

Wolfe, E. and 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. and Kiely, R. (2011). The cognitive processes of taking IELTS Academic Writing Task 1. *IELTS Reports, Vol 11,* (pp. 373–449). IDP: IELTS Australia, Canberra, and British Council, London.

# Appendix 1: Test tasks

IELTS AWT2 Prompt 1 – Paper-based

**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.

Do not write below this line

**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.

**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.

Do not write below this line

**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 2: Computer Familiarity Questionnaire

## Computer Familiarity Questionnaire

Name:_____ (Given name) _____ (Surname)

Student number:_____ Gender: M/F (Please circle)

IELTS result (latest): Overall band:_____ Writing band:_____

There are 15 questions about your experiences and opinions of working with computer. Please **CIRCLE** the number to indicate your answer to each question/statement.

| | | 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 library | 5 | 4 | 3 | 2 | 1 |
| | d) Others (please specify): _____ | 5 | 4 | 3 | 2 | 1 |
| 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 (please specify): _____ | 5 | 4 | 3 | 2 | 1 |
| 3) | How often do you use computer for | | | | | |
| | a) surfing the Internet ? | 5 | 4 | 3 | 2 | 1 |
| | b) electronic communication, e.g. emails? | 5 | 4 | 3 | 2 | 1 |
| | c) study-related activities? | 5 | 4 | 3 | 2 | 1 |
| | d) Others (please specify): _____ | 5 | 4 | 3 | 2 | 1 |

Page **1** of **3**

| | | Always | Very often | Often | Once a while | Never |
|---|---|---|---|---|---|---|
| 4) | How often do you use the 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 |
| | e) Others (please specify): _____? | 5 | 4 | 3 | 2 | 1 |
| 5) | How often do you take a test on | | | | | |
| | a) paper? | 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 a paper? | 5 | 4 | 3 | 2 | 1 |
| 8) | How comfortable are you with taking a test on | | | | | |
| | a) computer? | 5 | 4 | 3 | 2 | 1 |
| | b) paper? | 5 | 4 | 3 | 2 | 1 |
| 9) | How do you feel about using the keyboard (typing)? | 5 | 4 | 3 | 2 | 1 |

| | | 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?

Thank you for your opinions!

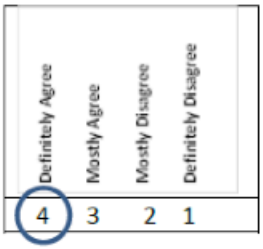
# Appendix 3: Writing Process Questionnaire

Student number:_____

There are 40 statements about how you might have completed the IELTS essay task. Please answer **ALL** the questions, thinking about different phases of the task completion.

- while reading the task instructions
- before starting to write
- while writing the first draft
- after writing the first draft

Please **CIRCLE** the number to indicate the extent of your agreement or disagreement to each statement. See an example below:




| **While READING the task instructions...** | | Definitely Agree | Mostly Agree | Mostly Disagree | Definitely Disagree |
|---|---|---|---|---|---|
| 1 | I read the **whole** task instructions carefully. | 4 | 3 | 2 | 1 |
| 2 | I thought about **how well I understood** the task requirements. | 4 | 3 | 2 | 1 |
| 3 | I thought about what I knew about the **topic**. | 4 | 3 | 2 | 1 |
| 4 | I thought about what I knew about the **genre.** | 4 | 3 | 2 | 1 |
| 5 | I thought about the **purpose** of the task. | 4 | 3 | 2 | 1 |
| 6 | I thought about **what I might need to write** to make my essay relevant and adequate to the task. | 4 | 3 | 2 | 1 |
| 7 | I thought about the **intender reader** of my essay and their **expectations**. | 4 | 3 | 2 | 1 |

| **BEFORE starting to write ...** | | Definitely Agree | Mostly Agree | Mostly Disagree | Definitely Disagree |
|---|---|---|---|---|---|
| 8 | I thought about or jotted down **ideas** which are **relevant** to the task/topic. | 4 | 3 | 2 | 1 |
| 9 | I **prioritised** these ideas based on the task requirements. | 4 | 3 | 2 | 1 |
| 10 | I **linked** these **ideas** to what I know already about the topic from **memory**. | 4 | 3 | 2 | 1 |
| 11 | I worked out how these ideas **relate** to each other, e.g. main ideas or examples. | 4 | 3 | 2 | 1 |
| 12 | I thought about the **structure** of my essay. | 4 | 3 | 2 | 1 |
| 13 | I **recombined** or **reordered** some of the ideas to fit the structure of my essay. | 4 | 3 | 2 | 1 |
| 14 | I **removed** some ideas I planned to write because they did not fit the structure of my essay. | 4 | 3 | 2 | 1 |
| 15 | I **re-read** the task instructions. | 4 | 3 | 2 | 1 |
| **WHILE writing the first draft ...** | | Definitely Agree | Mostly Agree | Mostly Disagree | Definitely Disagree |
| 16 | I thought about the **appropriate words** to express my ideas. | 4 | 3 | 2 | 1 |
| 17 | I thought about the **correct sentence structures** to express my ideas. | 4 | 3 | 2 | 1 |
| 18 | I thought about the **correct grammar** to express my ideas. | 4 | 3 | 2 | 1 |
| 19 | I thought about how to **connect** my ideas **smoothly** in the whole essay | 4 | 3 | 2 | 1 |
| 20 | I thought about how to make my ideas **persuasive** to the reader. | 4 | 3 | 2 | 1 |
| 21 | I **organised** my sentences and paragraphs in a **logical** order. | 4 | 3 | 2 | 1 |
| 22 | I developed new ideas or a **better understanding** of the topic. | 4 | 3 | 2 | 1 |
| 23 | I **re-read** the task instructions. | 4 | 3 | 2 | 1 |
| 24 | I **changed** my writing **plan** (e.g. structure and content). | 4 | 3 | 2 | 1 |
| 25 | I checked that the content was **relevant** and revised accordingly. | 4 | 3 | 2 | 1 |
| 26 | I checked that my essay was **well-organised** and revised accordingly. | 4 | 3 | 2 | 1 |
| 27 | I checked that my essay was **coherent** and revised accordingly. | 4 | 3 | 2 | 1 |
| 28 | I checked that I included **my own viewpoint** on the topic and revised accordingly. | 4 | 3 | 2 | 1 |
| 29 | I checked the possible effect of my essay on the **intended reader** and revised accordingly. | 4 | 3 | 2 | 1 |

| | | | | | |
|---|---|---|---|---|---|
| 30 | I checked the accuracy and range of the **sentence structures** and revised accordingly. | 4 | 3 | 2 | 1 |
| 31 | I checked the **grammar** (e.g. part of speech and tenses) and revised accordingly | 4 | 3 | 2 | 1 |
| 32 | I checked the appropriateness and range of **vocabulary** and revised accordingly. | 4 | 3 | 2 | 1 |
| **AFTER writing the first draft ...** | | Definitely Agree | Mostly Agree | Mostly Disagree | Definitely Disagree |
| 33 | I checked that the content was **relevant** and revised accordingly. | 4 | 3 | 2 | 1 |
| 34 | I checked that my essay was **well-organised** and revised accordingly. | 4 | 3 | 2 | 1 |
| 35 | I checked that my essay was **coherent** and revised accordingly. | 4 | 3 | 2 | 1 |
| 36 | I checked that I included **my own viewpoint** on the topic and revised accordingly. | 4 | 3 | 2 | 1 |
| 37 | I checked the possible effect of my essay on the **intended reader** and revised accordingly. | 4 | 3 | 2 | 1 |
| 38 | I checked the accuracy and range of the **sentence structures** and revised accordingly. | 4 | 3 | 2 | 1 |
| 39 | I checked the **grammar** (e.g. part of speech and tenses) and revised accordingly | 4 | 3 | 2 | 1 |
| 40 | I checked the appropriateness and range of **vocabulary** and revised accordingly. | 4 | 3 | 2 | 1 |

If some of your writing processes are not included in the above 40 statements, please write them down in the box below.

| During the test, I also ..... |
|---|
| |

Thank you very much for your participation!

# Appendix 4: 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.<br>• 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.<br>• 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. |
| **Online monitoring and revising** | • 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. |
| **Editing (after writing monitoring and revising)** | • I read the essay again made some changes according to what the intended reader needs to know |