Comparing the cognitive validity of a computer-based and paper-based EAP writing test

Dr. Sathena Chan, Prof. Cyril Weir
(CRELLA, University of Bedfordshire)
Prof. Stephen Bax
(Open University)

Research Background

- Writing on computer has become the norm in most university disciplines (Newman, Couturier & Scurry 2010)
- Responding to the need for computer-based assessment, most major standardised writing assessments offer their paper-based tests in computer-based mode, e.g. Cambridge English Exams and TOEFL iBT.
- Others have launched computer-based tests e.g. PTE Academic and Aptis.
- IELTS is currently a paper-based test.

Score Equivalence

- Early research had suggested that CB tests at the time tended to be more difficult than PB versions (e.g. Mazzeo & Harvey, 1988)
- More recent research shows that students' 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).
- But it is important to consider **rater's reliability and severity** across the modes.
  - Higher rater's reliability on CB rating (e.g. Wolfe & Manolo, 2005)
  - PB performance receiving lower scores
  - Illegible writing often penalised (e.g. Bae & Bachman, 2010)

Cognitive Validity

- The cognitive processes which a candidate draws on when completing the test writing task(s) are an accurate representation of the types of processing required in writing tasks in real life (Glaser, 1991; Shaw and Weir, 2007)
- There is a concern that "the different modes may be activating different executive processing within the candidate" (Weir et al, 2007: 9)
- The cognitive processes of participants completing paper-based IELTS Academic Task 1 has been investigated in detail (e.g. Yu et al, 2011).
- However, test takers' processes on AWT2, especially in CB mode, has not previously been researched – an important gap in the research base if the IELTS writing test is to be computerised in future.

Aims of the Project

To investigate comparability of IELTS Academic Writing essay task in paper-based (PB) and computer-based (CB) modes:

1) Score equivalence (take account of rater's severity)
2) Cognitive equivalence
   - *examine test takers' processes in both modes*
   - *furthermore compare these cognitive processes with those used by L2 students under genuine academic writing conditions*
3) To investigate the impact of test takers' familiarity with computer on performance

CRELLA University of Bedfordshire

Data Collection

- Participants: 153 first-year undergraduates (CEFR B1-C1)
- 2 prompts
- Raters: 4 certified IELTS raters approved by the British Council
- Test events (n=15): delivery mode x prompt countered-balanced

Group 1	Group 2	Mins
Computer Familiarity Questionnaire		5
Prompt 1 in **PB** mode	**Prompt 1** in **CB** mode	40
Writing Process Questionnaire		10
Prompt 2 in **CB** mode	**Prompt 2** in **PB** mode	40
Writing Process Questionnaire		10
20% were interviewed individually		20

Writing Process Questionnaire

- Developed based on models of writing (e.g. Hayes and Flower, 1980; Kellogg, 1996; Shaw & Weir, 2007)
- Adapted from Chan (2013), Chan, Wu & Weir (2014), Chan and Moore (2016)
- 40 items (4-point Likert scale)

Cognitive phases
Conceptualisation
Generating ideas
Organising ideas
Generating texts
Monitoring and revising at high-level
Monitoring and revising at low-level

Data Analysis

1) Score equivalence: Many-Facet Rasch Analysis (MFRM)
   - *One 5-Facet analysis (test takers' writing ability, delivery mode, prompts, raters and rating category)*
   - *Four further 4-facet analyses to examine individual analytic scoring categories between the modes (i.e. CB Task Achievement, PB Task Achievement are treated as separate items)*
2) Cognitive validity:
   - *Writing Process Questionnaire: **Wilcoxon signed-tank tests***
   - *Interviews: coded using Nvivo*
3) Impact of affective variables on test scores:
   - *Computer Familiarity Questionnaire: **Multiple Regression***

RESULTS

Raters' reliability and severity

- Rater A rated all scripts. Raters B, C & D each rated a sub-set of the scripts.
- The inter-rater reliability between the first rater and second rater was r=0.83 ($p<0.01$) using Spearman's rho test.

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

- Infit values for all the raters fall within the acceptable range. The difference in the Fair mean between the first rater and the second rater was within 0.12.

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

- Prompt 2 scores were significantly higher than Prompt 1 scores, but the differences in both the observed and fair mean scores of the two prompts were within half an IELTS band.

CRELLA University of Bedfordshire

Score equivalence between CB and PB mode

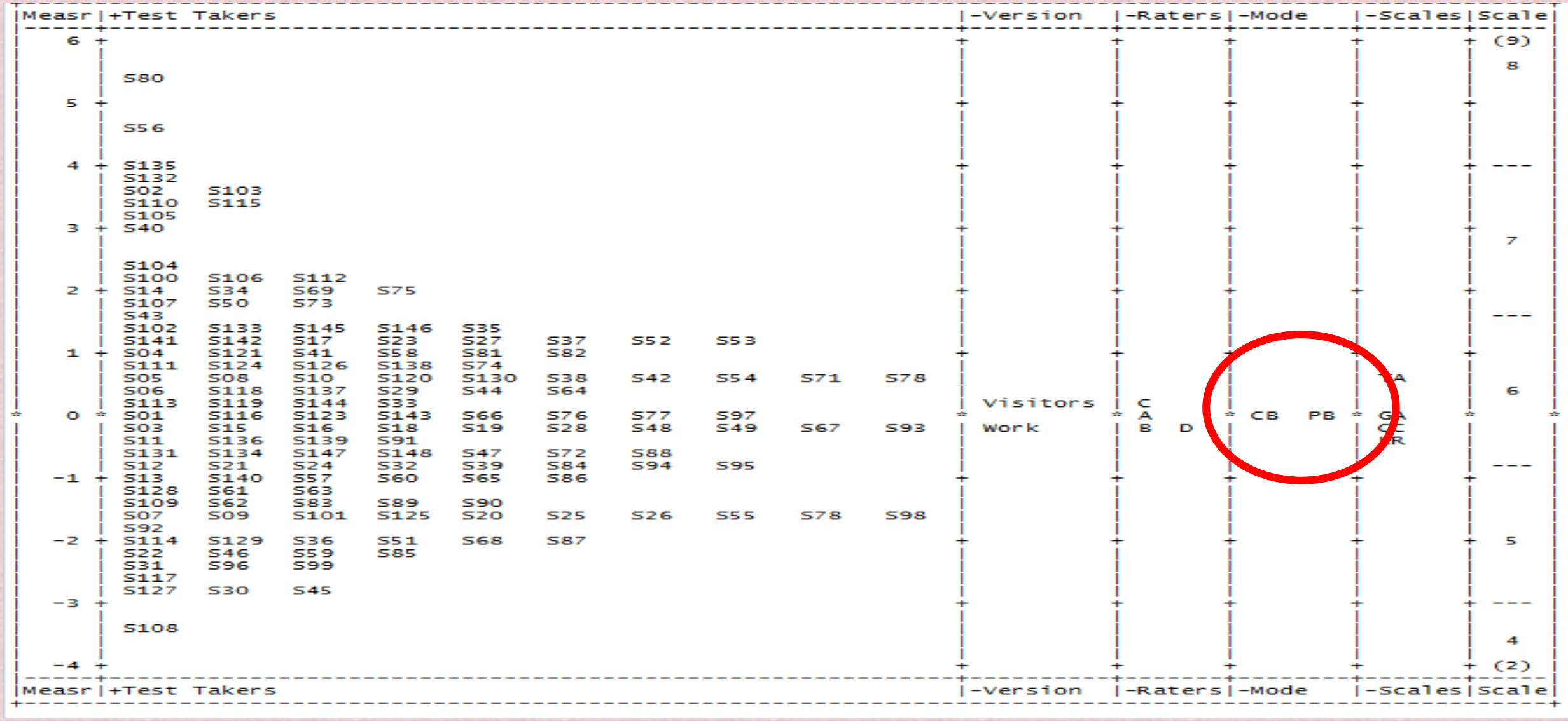
Test mode	N	Observed Mean	Fair Mean	Logit measure	Standard error	Infit mean square
CB	1104	5.75	5.83	.04	.05	.97
PB	1088	5.87	5.86	-.04	.05	1.02

(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

- Test scores obtained from the CB and PB mode were not statistically different.
- Both observed mean and fair mean scores of the two modes were very close.

CRELLA University of Bedfordshire

FACETS Variable map



Score equivalence: Analytic criteria

- No significant difference was obtained in the three pairs of analytic scores (Task Achievement, Cohesion and Coherence & Grammatical Range and Accuracy) between the two modes.
- But test takers scored significantly higher on Lexical Resources on PB than CB.

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

Test takers' processes: Descriptive

	Computer-based		Paper-based	
	M	SD	M	SD
Conceptualisation	3.25	0.40	3.27	0.42
Generating ideas	3.26	0.43	3.26	0.44
Organising ideas	3.25	0.49	3.24	0.48
Generating texts	3.40	0.47	3.39	0.51
Monitoring and revising (High-level)	3.22	0.50	3.17	0.50
Monitoring and revising (Low-level)	3.20	0.60	3.20	0.60

Test takers' processes

All differences were non-significant

Results from Wilcoxon Signed Ranks Tests

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		
Organising ideas	CB	3.20	66.5	-1.359	0.174
	PB	3.20	65.5		
Generating texts	CB	3.50	67.5	-1.631	0.103
	PB	3.50	66.5		
Monitoring and revising (High level)	CB	3.20	64.5	-1.649	0.990
	PB	3.20	65		
Monitoring and revising (Low level)	CB	3.17	66.5	0.000	1.000
	PB	3.17	66.5		

Difference in processes (Interview data)

(n=30)	Paper-based	Computer-based
Planning	• Planned carefully by writing down a plan or key ideas	• More relaxed about the initial plan • Did not need to start with a perfect plan
Organising	• Mainly at the whole text level in relation to the structure of their essay	• Also at the levels of sentences and paragraphs
Generating texts	• Were more carefully with their choice of words and sentence structures	• Focused more on 'getting the ideas out' during this phase
Monitoring and Revising	• Less willing to revise: inconvenient & lower band? • Changes were predominantly related to accuracy • Some phrasing at the word level	• Were more engaged to revise after writing • Also at the levels of clauses and sentences to improve coherence or argument

Students' familiarity with computer

- The vast majority of participants reported using computers frequently
  - at home (97.6%) (56.4%).
  - in university (82.7%) (84.3%).
- They used computers for a variety of purposes:
  - communications (94.5%) (89.9%)
  - word processing (92.9%) (68.0%)
  - study-related activities (96%) (59.7%)
- 96.1% (59.7%) have frequent access to computers at home; 89.8% (88.4%) at university; 78.6% in public library
- 90.6% (67.5%) are comfortable using a computer to write a paper.
- 81.2% (53.0%) are comfortable taking a test on computer.
- 78.0% (66.7%) would forget the time when working with computer.
- 71.4% (86.7%) consider working with a computer is really fun.

Impact of affective variables

- Significant positive correlations between 10 CFQ items and students' performance, ranging from $r(120)=.176$, $p<.01$ to $r(120)=.406$, $p<.01$.
- Multiple regression (stepwise): Three items explained 22.6% of the variance of the CB scores, indicating low level of predictive power.

	B	Std error	β	t	Sig.	
Frequently used PC for word processing	.297	.066	.374	4.496	.000	
Access to computers at public library	.093	.044	.174	2.083	.039	
Would forget the time	.107	.053	.166	2.020	.046	
R^2						.226
F						11.280

Main findings and Implications

1. There was no significant difference in the test scores across the PB and CB modes. But students scored *slightly* higher on Lexical Resources in PB than CB.
2. Results of the WPQ survey indicate a *similar pattern* between the cognitive processes involved in writing on a computer and writing with paper-and-pencil. However, a few potential differences indicated by the interview data might be worth further investigation in future studies.
3. Three of the computer familiarity variables have a small but significant impact on their performance in the computer mode. Test takers who do not have such a familiarity profile are likely to perform worse than those who do

   → Test providers might consider using these items to provide advice about the candidates' readiness for taking the test in the computer mode.

Acknowledgement

The project was funded by the IELTS Joint-funded research program.

CRELLA
University of Bedfordshire

Thank you!

sathena.chan@beds.ac.uk

cyril.weir@beds.ac.uk

CRELLA
University of Bedfordshire