Paper-based versus computer-based writing assessment: Investigating equivalence on integrated and independent tasks

Tineke Brunfaut, Luke Harding & Aaron O. Batty
EALTA, Bochum (Germany), May 2018

@TinekeBrunfaut @harding_luke @AaronOlafBatty

© Tineke Brunfaut, Luke Harding & Aaron O. Batty, 2018
Background

• Vast increases in computer accessibility, usage & familiarity
• Increases in computer-based language testing systems:
  – From inception
  – To replace paper-based test
  – Parallel offer of paper-based & computer-based mode
• Key motivations (Davey, 2011):
  – Target new constructs
  – Achieve more accurate and efficient scoring
  – Make test administration more accessible, efficient & cost-effective
  – Respond to market demands
  – Meet policy requirements
Background

- At face value, writing, of all the language skills, may be the most suitable to test in a computer-based environment.
- But L2 writing research shows that:
  - The writing medium may have an effect on writing process (e.g. Van Waes & Schellens, 2003)
  - Writing processes may in turn influence text quality (e.g. Breetvelt, van den Bergh & Rijlaarsdam, 2009)
So, are changes in delivery mode of writing tests valid and fair, in particular in contexts of paper-based replacements or parallel use of paper- & computer-based?

- Statistical equivalence?
- Experiential equivalence?
Prior research

*Paper-based vs computer-based L2 writing tests*

**Statistical equivalence (scores)**

- No uniform conclusions
  - Higher scores in paper-based mode
    (e.g., Chen, White, McCloskey, Soroui & Chun, 2011; Lei, Livingstone, Larkin & Bonett, 2004)
  - Higher scores in computer-based mode
    (e.g., Jin & Yan, 2017; Li, 2006)
  - Similar scores in both modes
    (e.g., Brelan & Muraki, 2005; Endres, 2012)
  - But also depending on, e.g.:
    • Type of writing task (e.g., Chen et al., 2011)
    • Test-taker characteristics (e.g., Brelan & Muraki, 2005; Endres, 2012)
    • Rating approach & criteria (e.g. Lee, 2004)
Prior research

*Paper-based vs computer-based L2 writing tests*

**Experiential equivalence (perceptions)**

- Mostly positive experiences in computer-based mode
  
  (e.g., Lee, 2004; Maycock & Green, 2005; Ling, 2017)

- But also depending on, e.g.:
  
  - Computer familiarity

---

**THUS**

- Mixed findings

- Complex; other factors

- Primarily independent writing tasks, BUT user-interface particularly in integrated writing tasks?
Our study
Brunfaut, Harding & Batty (2018)

Context

• Exam board wants to add an online alternative to a large-scale, paper-based ESL writing test
  – To make test delivery more efficient and accessible
  – To respond to changing stakeholder needs
• Integrated Skills of English (ISE) exam suite, Trinity College London
• Ongoing validation research to support parallel delivery modes
Research questions

1. Is there a **difference in test-takers’ scores** on the ISE writing test suite depending on delivery mode?

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

Instruments

Writing test

• Integrated Skills of English (ISE) exam suite, Trinity College London
  – ISE I (B1), ISE II (B2) and ISE III (C1)
  – Writing test: integrated + independent task

• Integrated reading-into-writing task (RIW)
  – Identify info from four texts that is relevant to the writing task
  – Paraphrase, summarize, synthesize the info
  – Combine the info to suit the writing purpose, e.g. suggest solutions
Methodology

Instruments

– Rating criteria:
  • Reading-for-writing (RfW)
  • Task fulfilment (TF)
  • Organisation and structure (O&S)
  • Language control (LC)
  • 0-4 scale
Methodology

**Instruments**

• Independent writing task (IW)
  – Produce a narrative, descriptive or instructional text in response to a writing prompt
  – Express facts/opinions (ISE I), evaluate & make suggestions (ISE II, III)
  – Rating criteria:
    • Task fulfilment (TF)
    • Organisation and structure (O&S)
    • Language control (LC)
    • 0-4 scale
Methodology

Instruments

Perception questionnaire

Format
• 12 four-point Likert scales, 2 MC questions, 2 open-ended questions

Focus
• **Impact** of delivery mode on test-takers’ **emotional state**
  nervousness, comfortability, frustration, confidence, boredom, happiness
• **Usability** of delivery mode
  understanding what to do; ease of writing with pen/keyboard; ease of revising/editing on paper/screen; clarity of layout; ease of navigation
• **Fairness** of delivery mode
  how well testing language ability; performance estimates; mode preference
Methodology

Participants

- 283 English L2 learners
- 107 ISE I, 109 ISE II, 67 ISE III
- Approx. 50/50 M/F (ISE I & II), approx. 40/60 M/F (ISE III)
- 13-58 years old (mostly mid-20s)
- Mostly Italian, Spanish or Portuguese L1 speakers
  (a few Russian, Polish, Moldavian and French L1 speakers)
- Mostly European or South American background
Methodology

Data collection

1) Two ISE tests
   - One on paper, one on computer
   - Counterbalanced for:
     • delivery mode
     • test form
     • order

2) Perception questionnaire
Methodology

Analyses

Writing performances

• Rated by 7 trained ISE raters
• Many-facet Rasch measurement (MFRM)
  – Four-facet model: Test-takers, order, raters, rating criteria
  – Dummy facets: Test form, delivery mode, task
• Analysis of delivery mode effects on ISE writing scores through bias/interaction analyses of:
  1. mode and task
  2. mode and rating scale category
• Per ISE level
Methodology

Analyses

Perception questionnaire

• Selected-response questions
  – Descriptive statistics
  – Wilcoxon-signed rank tests

• Open-ended questions
  – Thematic analysis
## Key findings

### Performances

<table>
<thead>
<tr>
<th>Task</th>
<th>Scale</th>
<th>Mode &amp; Task</th>
<th>RIW</th>
<th>/</th>
<th>TF</th>
<th>O&amp;S</th>
<th>LC</th>
<th>MODE &amp; RATING SCALE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>PB</td>
<td>Meas</td>
<td>SE</td>
<td>Meas</td>
<td>SE</td>
<td>Contrast</td>
<td>SE</td>
</tr>
<tr>
<td>RIW</td>
<td></td>
<td>0.25 0.07</td>
<td>0.50</td>
<td>0.11</td>
<td>4.74</td>
<td>1077</td>
<td>0.000</td>
<td>0.14</td>
</tr>
<tr>
<td>IW</td>
<td>/</td>
<td>-0.09 0.08</td>
<td>0.09</td>
<td>0.08</td>
<td>1.45</td>
<td>807</td>
<td>0.148</td>
<td>0.04</td>
</tr>
<tr>
<td>RfW</td>
<td></td>
<td>0.52 0.15</td>
<td>0.48</td>
<td>0.20</td>
<td>2.33</td>
<td>267</td>
<td>0.021</td>
<td>0.14</td>
</tr>
<tr>
<td>RIW</td>
<td></td>
<td>-0.40 0.15</td>
<td>0.34</td>
<td>0.21</td>
<td>1.61</td>
<td>267</td>
<td>0.108</td>
<td>0.10</td>
</tr>
<tr>
<td>TF</td>
<td></td>
<td>-0.63 0.16</td>
<td>0.40</td>
<td>0.23</td>
<td>1.73</td>
<td>267</td>
<td>0.085</td>
<td>0.11</td>
</tr>
<tr>
<td>O&amp;S</td>
<td></td>
<td>-0.19 0.15</td>
<td>0.77</td>
<td>0.21</td>
<td>3.73</td>
<td>267</td>
<td>0.000</td>
<td>0.23</td>
</tr>
<tr>
<td>LC</td>
<td></td>
<td>0.62 0.14</td>
<td>0.06</td>
<td>0.20</td>
<td>0.30</td>
<td>267</td>
<td>0.768</td>
<td>0.02</td>
</tr>
<tr>
<td>TF</td>
<td></td>
<td>0.42 0.16</td>
<td>0.33</td>
<td>0.22</td>
<td>1.49</td>
<td>267</td>
<td>0.136</td>
<td>0.09</td>
</tr>
<tr>
<td>IW</td>
<td></td>
<td>0.92 0.14</td>
<td>0.16</td>
<td>0.20</td>
<td>0.80</td>
<td>267</td>
<td>0.424</td>
<td>0.05</td>
</tr>
<tr>
<td>LC</td>
<td></td>
<td>0.56 0.14</td>
<td>0.62</td>
<td>0.14</td>
<td>0.30</td>
<td>267</td>
<td>0.768</td>
<td>0.02</td>
</tr>
</tbody>
</table>

- Greater ease in the paper-based mode
- On the criteria Reading-for-Writing and Language Control
# Key findings

## Performances

### ISE II (B2)

<table>
<thead>
<tr>
<th>Task</th>
<th>Scale</th>
<th>Meas</th>
<th>SE</th>
<th>Meas</th>
<th>SE</th>
<th>Contrast</th>
<th>SE</th>
<th>t</th>
<th>df</th>
<th>p</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>MODE &amp; TASK</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RIW</td>
<td>/</td>
<td>-0.06</td>
<td>0.07</td>
<td>0.06</td>
<td>0.07</td>
<td>0.12</td>
<td>0.1</td>
<td>1.14</td>
<td>1085</td>
<td>0.256</td>
<td>0.03</td>
</tr>
<tr>
<td>IW</td>
<td>/</td>
<td>-0.09</td>
<td>0.07</td>
<td>0.09</td>
<td>0.07</td>
<td>0.18</td>
<td>0.1</td>
<td>1.79</td>
<td>813</td>
<td>0.074</td>
<td>0.05</td>
</tr>
<tr>
<td>MODE &amp; RATING SCALE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RfW</td>
<td>RIW</td>
<td>0.18</td>
<td>0.15</td>
<td>0.21</td>
<td>0.15</td>
<td>0.03</td>
<td>0.21</td>
<td>0.15</td>
<td>269</td>
<td>0.880</td>
<td>0.01</td>
</tr>
<tr>
<td>TF</td>
<td>RIW</td>
<td>-0.47</td>
<td>0.14</td>
<td>-0.40</td>
<td>0.14</td>
<td>0.07</td>
<td>0.20</td>
<td>0.35</td>
<td>269</td>
<td>0.728</td>
<td>0.02</td>
</tr>
<tr>
<td>O&amp;S</td>
<td>RIW</td>
<td>-0.65</td>
<td>0.15</td>
<td>-0.61</td>
<td>0.15</td>
<td>0.04</td>
<td>0.21</td>
<td>0.20</td>
<td>269</td>
<td>0.840</td>
<td>0.01</td>
</tr>
<tr>
<td>LC</td>
<td>RIW</td>
<td>-0.27</td>
<td>0.14</td>
<td>0.03</td>
<td>0.14</td>
<td>0.31</td>
<td>0.20</td>
<td>1.54</td>
<td>269</td>
<td>0.126</td>
<td>0.09</td>
</tr>
<tr>
<td>TF</td>
<td>IW</td>
<td>0.15</td>
<td>0.12</td>
<td>0.30</td>
<td>0.12</td>
<td>0.15</td>
<td>0.16</td>
<td>0.90</td>
<td>269</td>
<td>0.369</td>
<td>0.05</td>
</tr>
<tr>
<td>O&amp;S</td>
<td>IW</td>
<td>0.11</td>
<td>0.14</td>
<td>0.28</td>
<td>0.14</td>
<td>0.17</td>
<td>0.19</td>
<td>0.85</td>
<td>269</td>
<td>0.394</td>
<td>0.05</td>
</tr>
<tr>
<td>LC</td>
<td>IW</td>
<td>0.44</td>
<td>0.13</td>
<td>0.69</td>
<td>0.13</td>
<td>0.25</td>
<td>0.18</td>
<td>1.36</td>
<td>269</td>
<td>0.176</td>
<td>0.08</td>
</tr>
</tbody>
</table>
## Key findings

### Performances

#### ISE III (C1)

<table>
<thead>
<tr>
<th>MODE &amp; TASK</th>
<th>Task</th>
<th>Scale</th>
<th>PB Meas</th>
<th>PB SE</th>
<th>CB Meas</th>
<th>CB SE</th>
<th>Contrast</th>
<th>CB SE</th>
<th>Rasch-W Welch</th>
</tr>
</thead>
<tbody>
<tr>
<td>MODE &amp; TASK</td>
<td>RIW</td>
<td>/</td>
<td>0.02</td>
<td>0.09</td>
<td>-0.02</td>
<td>0.09</td>
<td>-0.04</td>
<td>0.12</td>
<td>-0.3</td>
</tr>
<tr>
<td></td>
<td>IW</td>
<td>/</td>
<td>0.02</td>
<td>0.09</td>
<td>-0.02</td>
<td>0.09</td>
<td>-0.04</td>
<td>0.13</td>
<td>-0.29</td>
</tr>
</tbody>
</table>

| MODE & RATING SCALE | CATEGORY | RIW  | RfW | 0.21 | 0.17 | 0.00 | 0.17 | -0.21 | 0.24 | -0.86 |
| MODE & RATING SCALE |          | TF   | -0.42| 0.18| -0.37| 0.18| 0.05 | 0.25 | 0.21 | 0.833 |
| MODE & RATING SCALE |          | O&S | -0.48| 0.18| -0.52| 0.18| -0.05 | 0.26 | -0.19 | 0.853 |
| MODE & RATING SCALE |          | LC  | -0.19| 0.17| -0.13| 0.16| 0.05 | 0.23 | 0.23 | 0.817 |

| MODE & RATING SCALE | CATEGORY | TF   | 0.24 | 0.16| 0.12 | 0.16| -0.12 | 0.22 | -0.53 | 0.598 |
| MODE & RATING SCALE |          | IW  | 0.29 | 0.17| 0.19 | 0.18| -0.10 | 0.25 | -0.42 | 0.675 |
| MODE & RATING SCALE |          | LC  | 0.48 | 0.16| 0.58 | 0.16| 0.10  | 0.23 | 0.45  | 0.653 |
Key findings

Perceptions

Overall

• Generally positive about the test in both modes of delivery
• More positive about online mode, particularly for higher proficiency levels

Impact on emotional state

• Generally somewhat mixed
• ISE I (B1) mostly similar between modes, but happier online
• ISE II (B2) higher perceived positive emotional state online (more comfortable, confident, happy)
• ISE III (C1) higher perceived positive emotional state online (more comfortable, confident, happy; less nervous, frustrated)
Key findings

Perceptions

Usability

- Generally positive in both delivery modes
- ISE I (B1) higher in online mode
  (easy to revise and edit, to navigate)
- ISE II (B2) higher in online mode
  (easy to type on keyboard, revise and edit, navigate)
- ISE III (C1) higher in online mode
  (easy to type on keyboard, revise and edit, clear layout)

Fairness

- Generally perceived as fair in both modes
- ISE I (B1) online perceived to test writing ability better, more prefer online mode, more perceived higher score online
- ISE II (B2) more prefer online mode, more perceived higher score online
- ISE III (C1) online perceived to test writing ability better, more prefer online mode, more perceived higher score online
Key findings
Perceptions

Qualitative comments:

- **More favourable** usability evaluations of **online** mode:
  - advantages for editing, speed and neatness of typing, keeping track of word count

  **Computer-based mode:** “I prefer the CB: it's fast and clean. You can change words in any time without putting crosses over the mistakes.”

  **Paper-based mode:** “It is difficult to edit what you have written, if you make mistakes, you don't have the possibility to correct them in a 'nice way’.”
Key findings
Perceptions

Qualitative comments:
• Specific to **reading-into-writing**, in both modes but more salient in **online** mode

Paper-based mode: “We need to have a look in the texts and for that, it's hard to find the right pages”

Computer-based mode: “It was uncomfortable to scroll down and up the page every time I had to source some information in the texts while I was writing the article about the previous texts.”

Computer-based mode: “During the CB writing test we could not underline parts of the text as we could do during the reading part. I would have personally used that function again.”
In sum

ISE II & III (B2 & C1)
• No discernible effect of delivery mode on scores at task level and at the individual rating category level
• General preference for computer-based mode

ISE I (B1)
• Preference for computer-based mode less prominent
• Paper-based mode easier for the Reading-into-Writing task (clear, though relatively small effect)
• On the rating scale categories of Reading for Writing & Language Control
In sum

• Cognitive load for integrated tasks at ISE I (B1) in the unfamiliar online mode poses costs in terms of linguistic accuracy and test-takers’ ability to draw effectively on source materials in their response?
  
  Cf. Noyes, Garland & Robbins (2004); Skehan’s “trade-off” hypothesis
Implications

1. Supportive evidence for the context validity of delivery of ISE writing exams in both modes
   - Esp. at higher levels of proficiency (ISE II & III)
   - Not for reading-into-writing task at ISE I
   → Further research needed on the interaction between task complexity, proficiency level, and delivery mode

2. Test-taker experience: computer-based writing preference
   - Esp. at higher levels of proficiency (ISE II & III)
   - Some usability issues
   → User-interface improvements made
   - Shift towards typing/screen writing as the norm
   → Continuation of research on delivery mode effects as digital usage and literacy increase
This research was funded by, and carried out under Trinity College London’s funded research programme. Any opinions, findings, conclusions, or recommendations expressed are those held by the authors and do not necessarily reflect the views of Trinity, its examiners, service providers, or registered examination centres.


Thank you!

@TinekeBrunfaut  @harding_luke  @AaronOlafBatty