Answer-changing Behavior in Multiple-choice Questions: Looking beyond the Impact of Changes

Zhi Li, Michelle Y. Chen, & Jayanti Banerjee
International Testing Commission (ITC), July 4th, 2018
Montréal, Québec
Overview of Answer Changes

• A common phenomenon in objective tests
  – Most test takers made some changes (Balance, 2006; Bath, 1967; Jacobs, 1972; Mathews, 1929)

• Effects of answer changes vs. common beliefs
  – First instinct fallacy vs. It-pays-to-switch (Foote & Belinky, 1972; Di Milla, 2007)

• Factors related to answer-changing behaviors
  – Test takers’ characteristics (proficiency, gender, personality)
  – Item characteristics (difficulty, discrimination, etc.)
Answer Changes in Language Tests

• Relatively few studies in language-related testing
  – The Graduate Record Examinations (Liu et al., 2015)

• No studies on listening tests
Listening Comprehension Tests

• Listening comprehension
  – As a complex process of meaning making
  – Goal setting, decoding aural/visual input, … monitoring comprehension
    (Taylor & Geranpayeh, 2011)

• Two types of listening performance tests
  – While-listening performance tests
  – Post-listening performance tests

• Three stages in a while-listening-performance tests
  – Question preview, Question responding, Answer review
Answer-changing Behaviors and Test Validation (I)

• Validity: “the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests” (Standards; AERA, APA, & NCME, 2014: P11)

• Response-process based validity evidence
  – One of the five major sources of validity evidence (AERA, APA, & NCME, 2014)
  – Contributes to the construct validity (Anderson et al., 1991; Cohen, 2006)
  – Often missing in validation studies (Zumbo & Chan, 2014)
    ➢ Response process is difficult to capture
Definition of Response Process

The *Standards* (AERA et al., 2014, p.15)

- “Cognitive process engaged in by test-takers”

A broader definition

- Response processes include test-takers’ cognitive processes, and processes related to their behaviors and emotions during a test (e.g., Hubley & Zumbo, 2017).
Answer-changing Behaviors and Test Validation (II)

- Answer-changing behaviors as part of response process
- They represent test takers’ behaviors
- They may reflect test takers’ strategies
  - e.g., make predictions, monitoring
- They can be recorded through *timestamped log data* in computer delivered tests
- The outcome of answer-changing behaviors is directly related to test performance/scores
Research Questions

With an eye towards test score validation, this study investigates:

1. Who made the changes?

2. When did the changes take place throughout the listening test?
   i. Was it dependent on the subskills measured?
   ii. Was it dependent on test takers’ proficiency levels?

3. What were the outcomes of the changes?
   i. Was it dependent on the subskills measured?
   ii. Was it dependent on test takers’ proficiency levels?
The Canadian Academic English Language (CAEL) Test, Computer Edition (CE) – An integrated and topic-based test of English for academic purposes – Reporting scale: 10-90 band score
Note: This is a screenshot of an example listening test.
The CAEL CE Listening Test

- We focused on the multiple choice questions in the three long listening testlets.

- Features of these listening testlets:
  - While-listening performance test
  - Academic topics: Two topics in arts & One topic in science
  - Subskills: Comprehending *local* information, Comprehending *global* information, & Making *inferences*
Participants

88 participants recruited for a pilot test
• Gender: 48 females and 40 males
• Major first language (L1) groups:
  – Chinese, Farsi, Arabic, Spanish, & Korean
• Proficiency levels (CAEL CE listening band score):
  – Low: Band score 20-40 (16)
  – Median: Band score 50-60 (41)
  – High: Band score 70-90 (31)
Data Collection & Analysis

• Data
  – Timestamped log data: answer-changing behaviors
  – Test performance and item score

• Preliminary Analysis
  – Mostly based on descriptive statistics to look for the patterns and possible relationships
### Overview of Answer-changing Behaviors

<table>
<thead>
<tr>
<th>Lecture topic</th>
<th># of TTs&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Total # of changes</th>
<th>Min, Max</th>
<th>Average # of changes per TT</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic 1</td>
<td>67</td>
<td>304</td>
<td>(1, 28)</td>
<td>4.5</td>
<td>5.0</td>
</tr>
<tr>
<td>Topic 2</td>
<td>61</td>
<td>208</td>
<td>(1, 36)</td>
<td>3.4</td>
<td>2.7</td>
</tr>
<tr>
<td>Topic 3</td>
<td>66</td>
<td>245</td>
<td>(1, 8)</td>
<td>3.7</td>
<td>3.5</td>
</tr>
<tr>
<td>TOTAL</td>
<td>87&lt;sup&gt;b&lt;/sup&gt;</td>
<td>757</td>
<td>(1, 47)</td>
<td>8.7</td>
<td>9.1</td>
</tr>
</tbody>
</table>

Note: TT = Test taker
<sup>a</sup> number of test takers who made at least one change (N = 88 TTs, k = 28 MCQs).
1. Who made the changes?

2. When did the changes take place throughout the listening test?
   i. Was it dependent on the subskills measured?
   ii. Was it dependent on test takers’ proficiency levels?

3. What were the outcomes of the changes?
   i. Was it dependent on the subskills measured?
   ii. Was it dependent on test takers’ proficiency levels?
<table>
<thead>
<tr>
<th>Proficiency (Listening)</th>
<th># of TTs</th>
<th>Total # of changes</th>
<th>(Min, Max)</th>
<th>Average # of changes per TT</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>16</td>
<td>117</td>
<td>(1, 22)</td>
<td>7.3</td>
<td>5.9</td>
</tr>
<tr>
<td>Mid</td>
<td>41</td>
<td>448</td>
<td>(1, 47)</td>
<td>10.9</td>
<td>12.2</td>
</tr>
<tr>
<td>High</td>
<td>31</td>
<td>192</td>
<td>(1, 15)</td>
<td>6.2</td>
<td>4.2</td>
</tr>
</tbody>
</table>

TT = Test taker
Research Questions

1. Who made the changes?

2. When did the changes take place throughout the listening test?
   i. Was it dependent on the subskills measured?
   ii. Was it dependent on test takers’ proficiency levels?

3. What were the outcomes of the changes?
   i. Was it dependent on the subskills measured?
   ii. Was it dependent on test takers’ proficiency levels?
When Did the Changes Take Place

<table>
<thead>
<tr>
<th></th>
<th>Pre-listening</th>
<th>During-listening</th>
<th>Post-listening</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>43</td>
<td>370</td>
<td>344</td>
<td>757</td>
</tr>
</tbody>
</table>
Research Questions

1. Who made the changes?

2. When did the changes take place throughout the listening test?
   i. Was it dependent on the subskills measured?
   ii. Was it dependent on test takers’ proficiency levels?

3. What were the outcomes of the changes?
   i. Was it dependent on the subskills measured?
   ii. Was it dependent on test takers’ proficiency levels?
### How Were the Changes Related to the Types of Subskill

<table>
<thead>
<tr>
<th>Subskills</th>
<th>Pre-listening</th>
<th>During-listening</th>
<th>Post-listening</th>
<th>Total # of changes</th>
<th>Average # of changes</th>
<th>Average of TTs(^a) per item</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global (k=8)</td>
<td>12</td>
<td>128</td>
<td>135</td>
<td>275</td>
<td>1.9</td>
<td>17</td>
</tr>
<tr>
<td>Local (k=11)</td>
<td>15</td>
<td>144</td>
<td>69</td>
<td>228</td>
<td>1.4</td>
<td>15</td>
</tr>
<tr>
<td>Inference (k=9)</td>
<td>16</td>
<td>101</td>
<td>147</td>
<td>264</td>
<td>1.3</td>
<td>19</td>
</tr>
</tbody>
</table>

Note: \(a.\) number of test takers who made at least one change on one item (N = 88)
Research Questions

1. Who made the changes?

2. When did the changes take place throughout the listening test?
   i. Was it dependent on the subskills measured?
   ii. Was it dependent on test takers’ proficiency levels?

3. What were the outcomes of the changes?
   i. Was it dependent on the subskills measured?
   ii. Was it dependent on test takers’ proficiency levels?
When Did the Changes Take Place

<table>
<thead>
<tr>
<th>Proficiency (Listening)</th>
<th>Pre-listening</th>
<th>During-listening</th>
<th>Post-listening</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>6</td>
<td>65</td>
<td>46</td>
<td>117</td>
</tr>
<tr>
<td>Mid</td>
<td>24</td>
<td>227</td>
<td>197</td>
<td>448</td>
</tr>
<tr>
<td>High</td>
<td>13</td>
<td>78</td>
<td>101</td>
<td>192</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>43</strong></td>
<td><strong>370</strong></td>
<td><strong>344</strong></td>
<td><strong>757</strong></td>
</tr>
</tbody>
</table>
When Did the Changes Take Place

Average number of changes

- Pre-listening:
  - low: 0.4
  - mid: 0.6
  - high: 0.4
When Did the Changes Take Place

Average number of changes

<table>
<thead>
<tr>
<th></th>
<th>Pre-listening</th>
<th>During-listening</th>
<th>Post-listening</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average number of changes</td>
<td>0.4</td>
<td>4.1</td>
<td>2.9</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>5.5</td>
<td>4.8</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>2.5</td>
<td>3.3</td>
</tr>
</tbody>
</table>

- Pre-listening: 0.4, 0.6, 0.4
- During-listening: 4.1, 5.5, 2.5
- Post-listening: 2.9, 4.8, 3.3

- low
- mid
- high
Research Questions

1. Who made the changes?

2. When did the changes take place throughout the listening test?
   i. Was it dependent on the subskills measured?
   ii. Was it dependent on test takers’ proficiency levels?

3. What were the outcomes of the changes?
   i. Was it dependent on the subskills measured?
   ii. Was it dependent on test takers’ proficiency levels?
RQ3: What Were the Outcomes of the Changes

<table>
<thead>
<tr>
<th>Outcomes</th>
<th>Topic1</th>
<th>Topic2</th>
<th>Topic3</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>W → W</td>
<td>67</td>
<td>56</td>
<td>64</td>
<td>187</td>
</tr>
<tr>
<td>R → W</td>
<td>32</td>
<td>31</td>
<td>25</td>
<td>88</td>
</tr>
<tr>
<td>W → R</td>
<td>56</td>
<td>43</td>
<td>56</td>
<td>155</td>
</tr>
<tr>
<td>R → R</td>
<td>9</td>
<td>16</td>
<td>13</td>
<td>38</td>
</tr>
</tbody>
</table>
RQ3: What Were the Outcomes of the Changes

<table>
<thead>
<tr>
<th>Outcomes</th>
<th>Topic1</th>
<th>Topic2</th>
<th>Topic3</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>W → W</td>
<td>67</td>
<td>56</td>
<td>64</td>
<td>187</td>
</tr>
<tr>
<td>R → W</td>
<td>32</td>
<td>31</td>
<td>25</td>
<td>88</td>
</tr>
<tr>
<td>W → R</td>
<td>56</td>
<td>43</td>
<td>56</td>
<td>155</td>
</tr>
<tr>
<td>R → R</td>
<td>9</td>
<td>16</td>
<td>13</td>
<td>38</td>
</tr>
</tbody>
</table>

Correct rate (WR) 0.34 0.29 0.35 **0.33**
Loss rate (RW) 0.20 0.21 0.16 **0.19**
### RQ3: What Were the Outcomes of the Changes

<table>
<thead>
<tr>
<th>Outcomes</th>
<th>Topic1</th>
<th>Topic2</th>
<th>Topic3</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>W → W</td>
<td>67</td>
<td>56</td>
<td>64</td>
<td>187</td>
</tr>
<tr>
<td>R → W</td>
<td>32</td>
<td>31</td>
<td>25</td>
<td>88</td>
</tr>
<tr>
<td>W → R</td>
<td>56</td>
<td>43</td>
<td>56</td>
<td>155</td>
</tr>
<tr>
<td>R → R</td>
<td>9</td>
<td>16</td>
<td>13</td>
<td>38</td>
</tr>
</tbody>
</table>

**Correct rate (WR)**

<table>
<thead>
<tr>
<th></th>
<th>Topic1</th>
<th>Topic2</th>
<th>Topic3</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>W → W</td>
<td>0.34</td>
<td>0.29</td>
<td>0.35</td>
<td>0.33</td>
</tr>
<tr>
<td>R → W</td>
<td>0.20</td>
<td>0.21</td>
<td>0.16</td>
<td>0.19</td>
</tr>
</tbody>
</table>

**Loss rate (RW)**

<table>
<thead>
<tr>
<th></th>
<th>Topic1</th>
<th>Topic2</th>
<th>Topic3</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>W → W</td>
<td>0.20</td>
<td>0.21</td>
<td>0.16</td>
<td>0.19</td>
</tr>
</tbody>
</table>
Research Questions

1. Who made the changes?

2. When did the changes take place throughout the listening test?
   i. Was it dependent on the subskills measured?
   ii. Was it dependent on test takers’ proficiency levels?

3. What were the outcomes of the changes?
   i. Was it dependent on the subskills measured?
   ii. Was it dependent on test takers’ proficiency levels?
### How Were the Changes Related to the Types of Subskill

<table>
<thead>
<tr>
<th>Subskills</th>
<th>Most changes happened at</th>
<th>Average correct rate (W→R)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global (k=8)</td>
<td>Post-listening, During-listening</td>
<td>0.41</td>
</tr>
<tr>
<td>Local (k=11)</td>
<td>During-listening</td>
<td>0.37</td>
</tr>
<tr>
<td>Inference (k=9)</td>
<td>Post-listening</td>
<td>0.44</td>
</tr>
</tbody>
</table>

Note: number of test takers who made at least one change on one item (N = 88)
Research Questions

1. Who made the changes?

2. When did the changes take place throughout the listening test?
   i. Was it dependent on the subskills measured?
   ii. Was it dependent on test takers’ proficiency levels?

3. What were the outcomes of the changes?
   i. Was it dependent on the subskills measured?
   ii. Was it dependent on test takers’ proficiency levels?
What Were the Outcomes of the Changes

Percentage of Outcomes of Answer Changes by Proficiency Groups

- WW
- RW
- RR
- WR

%: 0% 10% 20% 30% 40% 50%

low  mid  high
What Were the Outcomes of the Changes

Percentage of Outcomes of Answer Changes by Proficiency Groups

- WW
- RW
- RR
- WR

- low
- mid
- high
Answer-changing Behaviors and Test Validation

Understanding answer-changing behaviors

• Who
• Effectiveness of the changes

• When
  – Related to the target skills/constructs
  – Reflect test taking strategies? Metacognitive strategies?
Future Studies

Timestamped responding data + other data types
– The findings can be triangulated with an analysis of other behavioral data (e.g., eye-tracking) and/or think-aloud data

→ Better understanding of test-taking processes and their relationships with the measured construct
Thank You

research@paragontesting.ca
Selected References


