Cognitive validity evidence in L2 reading tests using stimulated-recall interviews (SRI)

Nathaniel Owen
“Assessing general language proficiency: definitions and approaches”

• General language proficiency: reading
  – Critical perspectives (SRI)
  – Theoretical considerations (Khalifa and Weir, 2009)
  – Practical approaches: SRI

- Authenticity
- Test Setting
  - Timing
  - Conditions
  - Presentation Model
  - Delivery Model
  - Task and test content

- Participant
  - Nationality (L1)
  - Proficiency (θ)
  - Schematic knowledge
  - Affect
  - Fatigue

- Strategic metacognition
- Cognitive Processing
- Monitoring

- Observed test performance
- Stimulated-recall
- Interview
  - Timing/duration
  - Consistency
  - Nature
  - Number
  - Language of verbalisation

- Participant
  - Nationality (L1)
  - Proficiency (θ)
  - Schematic knowledge
  - Affect
  - Fatigue

- Verbalisations
  - Memory decay
  - Veridicality
  - Reactivity
  - Completeness
  - Metacognition

- Stimuli
### Theoretical considerations: A cognitive model of reading (Khalifa & Weir, 2009: 43)

<table>
<thead>
<tr>
<th>Metacognitive Activity</th>
<th>Processing Core</th>
<th>Monitor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monitor: goal checking</td>
<td>Creating an intertextual representation: construct an organised representation across texts</td>
<td>Text structure knowledge:</td>
</tr>
<tr>
<td>Remediation where necessary</td>
<td></td>
<td>Genre</td>
</tr>
<tr>
<td>Goal setter</td>
<td>Creating a text-level representation: construct an organised representation of a single text</td>
<td>Rhetorical tasks</td>
</tr>
<tr>
<td>Selecting appropriate type of reading:</td>
<td>Building a mental model: integrating new information; enriching the proposition</td>
<td>General knowledge of the world</td>
</tr>
<tr>
<td>Careful reading LOCAL: Understanding sentence GLOBAL: Comprehending main idea(s)</td>
<td>Inferencing</td>
<td>Topic knowledge</td>
</tr>
<tr>
<td></td>
<td>Comprehend overall text(s)</td>
<td>Meaning representation of text(s) so far</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expeditious reading LOCAL: Scan/search for specifics GLOBAL: Skim for gist</td>
<td>Establishing propositional meaning: at clause and sentence levels</td>
<td>Syntactic knowledge</td>
</tr>
<tr>
<td></td>
<td>Comprehend overall text(s)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lexicon Lemma: Meaning Word class</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lexicon Form: Orthography; Phonology Morphology</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Creating a text-level representation:
construct an organised representation of a single text

Building a mental model:
integrating new information; enriching the proposition

Inferencing:
At sentence/clause level
At word level

Establishing propositional meaning:
At clause level
At sentence level

Syntactic parsing

Lexical Access

Theoretical considerations: The processing core (Khalifa & Weir, 2009: 43)

• Model is
  – hypothetical
  – Hierarchical
  – Higher levels subsume lower levels

• Higher-level processes
  – Idea/concept unit of analysis
  – Amenable to participant verbalisation

• Lower-level processes
  – Word/clause unit of analysis
  – Automatised as participant becomes more proficient (Grabe, 2009)
Theoretical considerations

• Brunfaut & McCray (2015)
  – Used framework for SRI
  – No information re. how the framework applied to participant verbalisations
  – Test security; verbalisations will reveal content of text and items

• Bax (2013)
  – Experimental eye tracking distinguishes between lower metalinguistic levels (lexical access/syntactic parsing)
Research question 1

• Is the Khalifa and Weir model sufficient to account for all explanatory responses offered by the participants in the SRI?
  – Need to ensure meaningful relationship between stimuli and verbalisations
  – Evoking power of the stimuli
  – Ensure reliable coding
Practical approaches: The nature of the stimuli

- Memory decay
- Veridicality
- Reactivity
- Metacognition
- Completeness
- Objective
- Relatable
- Finely-grained

Questions 20-26

Do the following statements agree with the information given in Reading Passage 2? In boxes 20-26 on your answer sheet, write

TRUE if the statement agrees with the information
FALSE if the statement contradicts the information
NOT GIVEN if there is no information on this

20. The FAA was created as a result of the introduction of the jet engine.


22. Beacons and flashing lights are still used by ATC today.

23. Some improvements were made in radio communication during World War II.

24. Class F airspace is airspace which is below 365m and not near airports.

25. All aircraft in Class E airspace must use IFR.

26. A pilot entering Class C airspace is flying over an average sized city.

Uncontrolled airspace is designated Class F, while controlled airspace below 5,490m above sea level and not in the vicinity of an airport is Class E. All airspace above 5,490m is designated Class A. The reason for the division of Class E and Class A airspace stems from the type of planes operating in them. Generally, Class E airspace is where one finds general aviation aircraft (few of which can climb above 5,490m anyway), and commercial turboprop aircraft. Above 5,490m is the realm of the heavy jets, since jet engines operate more efficiently at higher altitudes. The difference between Class E and A airspace is that in Class A, all operations are IFR, and pilots must be instrument-rated, that is, skilled and licensed in aircraft instrumentation. This is because ATC control of the entire space is essential. Three other types of airspace, Classes D, C and B, govern the vicinity of airports. These correspond roughly to small municipal, medium-sized metropolitan and major metropolitan airports respectively, and encompass an increasingly rigorous set of regulations. For example, all a VFR pilot has to do to enter Class C airspace is establish two-way radio contact with ATC. No explicit permission from ATC to enter is needed, although the pilot must continue to obey all regulations governing VFR flight. To enter Class B airspace, such as on approach to a major metropolitan airport, an explicit ATC clearance is required. The private pilot who cruises without permission into this airspace risks losing their license.
Background: Comparative study of reading components of IELTS and TOEFL iBT

- Participants:
  - Six postgraduate students (L1 Chinese)
  - Been in the UK for approximately 1 year
  - IELTS scores >6.0 overall and in reading

<table>
<thead>
<tr>
<th>Participant</th>
<th>Session 1 (Test 1)</th>
<th>Session 2 (Test 2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participant 1</td>
<td>IELTS 1 (Q1-13)</td>
<td>TOEFL 1 (Q1-13)</td>
</tr>
<tr>
<td>Participant 2</td>
<td>IELTS 2 (Q14-26)</td>
<td>TOEFL 2 (Q14-26)</td>
</tr>
<tr>
<td>Participant 3</td>
<td>IELTS 3 (Q27-40)</td>
<td>TOEFL 3 (Q27-38)</td>
</tr>
<tr>
<td>Participant 4</td>
<td>TOEFL 1 (Q1-13)</td>
<td>IELTS 1 (Q1-13)</td>
</tr>
<tr>
<td>Participant 5</td>
<td>TOEFL 2 (Q14-26)</td>
<td>IELTS 2 (Q14-26)</td>
</tr>
<tr>
<td>Participant 6</td>
<td>TOEFL 3 (Q27-38)</td>
<td>IELTS 3 (Q27-40)</td>
</tr>
</tbody>
</table>
Participants’ verbalisations

Since

Verbal statements relate directly to moments of action

On account of

• Video of specific moments played directly to participant (relatability of stimulus)
• Video replayed to participant immediately after test completion (memory)
• Participants are able to verbalise the thought processes that occurred to them at specific moments of the test (veridicality)

Unless

Claim (cognitive process)

• Verbalisations are metacognitive (verbalisations are reflective)
• Verbalisation reveals test-wise approach to item completion (construct-irrelevant action)
• Verbalisation inaccurate (contradicts action) (veridicality)

Qualifier

• Verbalisations are partially complete (only account for part of the test takers’ actions)

Toulmin (2003) diagram of logical argumentation for stimulated-recall interview data using video stimulus
Coding of ‘lexical access’ (P2)

Text reference: “...the high mountains of the Himalayas are only about 50 million years old. Lower mountains tend to be older, and are often the eroded **relics** of much higher mountain chains.”

30. The word **relics** (off-list) in the passage is closest in meaning to
   - Resemblances
   - Regions
   - *Remains
   - Restorations

**Observed behaviour of participant 3 for item 30:**
- Answers six seconds after answering item 29
- Participant selected option 3 for item 30 (correct) **without returning to the text.**
- Therefore either used her existing knowledge of the text held in working memory, OR she knew the answer based on her ‘lexical bank’.

**Participant verbalisation:**
“‘Relics’ means ‘remains’. I didn’t consider other options.” [Participant 3, test 2, item 30]

- She is immediately aware that the highlighted word (‘relics’) is synonymous with ‘remains’ (noun) in the context. Evidence that for this option, the participant only required ‘lexical access’ (P2).
Coding of ‘syntactic parsing’ (P3)

Text reference: “Causing participants in experiments to smile, for example, leads them to report more positive feelings and to rate cartoons (humorous drawings of people or situations).”

22. The word rate in the passage is closest in meaning to
- *judge
- reject
- draw
- Want

Observed behaviour of participant 5 for item 22:
- 00:42 Q9: Circles ‘rate’
- 00:43 Circles ‘rate’ in text (paragraph 4, line 3)
- 00:44 Q9: Circles ‘closest’
- 00:50 Underlines ‘humorous drawings of people or situations’
- 00:54 Q9: Considers options
- 01:00 Circles ‘report’ (paragraph 4, line 3)

Participant verbalisation:
“...because there’s ‘and’ here, so I was asked to answer this; ‘rate’, and here, ‘report’ is in the same position, so it should be the same function, so here it’s ‘report’, then it says it’s ‘report the feeling’, so here is should be ‘rate cartoons’... ‘report’ is like ‘describe’, like ‘describe feelings’”
Coding of ‘establishing propositional meaning at the clause level’ (P4c)

Text reference: “In classic research Paul Ekman took photographs of people exhibiting the emotions of anger disgust, fear, happiness, and sadness. He then asked people around the world to indicate what emotions were being depicted in them.”

17. The word them in the passage refers to

- emotions
- people
- photographs
- cultures

Observed behaviour of participant 2, item 17:
- Participant spends more than one minute focusing on above portion of the text.

Participant verbalisation:
- “I see in the sentence, ‘in classic research, Paul Ekman took photographs of people’, which means this phrase [photographs] is the main word in this phrase. So that is reassuring that I might make the right choice. They are not talking about people, they are talking about [photographs].”
Text reference: “Hills and mountains are often regarded as the epitome of permanence, successfully resisting the destructive forces of nature, but in fact they tend to be relatively short-lived in geological terms.”

35. Which of the sentences below best expresses the essential information in the highlighted sentence in the passage?

Incorrect choices change the meaning in important ways or leave out essential information.

• When they are relatively young, hills and mountains successfully resist the destructive forces of nature.
• *Although they seem permanent, hills and mountains exist for a relatively short period of geological time.
• Hills and mountains successfully resist the destructive forces of nature, but only for a short time.
• Hills and mountains resist the destructive forces of nature better than other types of landforms.
Coding of ‘establishing propositional meaning at the sentence level’ (P4s)

Observed behaviour of participant 6, item 35:
• 01:56 Q35: Begins reading stem and options
• 02:12 Q35: Underlines ‘best expresses’
• 02:15 Moves to text (paragraph 2). Reads highlighted sentence in context
• 03:25 Q35: Selects option 2

Participant verbalisation:
• “I think the option 2 is most suitable, is most similar with the meaning of the sentence, so I choose option 2. The sentence says ‘hills and mountains are often regarded as permanent, but...’. It’s [this] ‘but’, in fact they tend to be relatively short-lived, so this, option 2 is in the same format, ‘although’, they are permanent, ‘but’ ‘they exist for a relatively short geological time’”. [Participant 6, test 1, item 35]
Coding of ‘Inferring at the word’ level (P5w)

Text reference: “Under very cold conditions, rocks can be shattered by ice and frost. Glaciers may form in permanently cold areas, and these slowly moving masses of ice cut out valleys, carrying with them huge quantities of eroded rock debris.”

34. The word them in the passage refers to

• cold areas
• *masses of ice
• valleys
• rock debris

Observed behaviour of participant 6, item 34:

• 00:10 Begins reading text (paragraph 6) carefully
• 01:00 Q34: Begins reading stem and options
• 01:06 Q34: Selects option 2

Participant verbalisation:

• “’them’ here is connected to the first part of the sentence, so it mentions that glaciers carry the masses of ice and then I think ‘them’ is continuing to talk about the masses of ice. I don’t know how glaciers can carry valleys, ‘rock debris’ is... because it’s carrying with rock debris, it’s not very correct, so I just chose ‘masses of ice’. I just put ‘masses of ice’ into the sentence, replacing [them] to read it again.”
As a general rule, the higher a mountain is, the more recently it was formed; for example, the high mountains of the Himalayas are only about 50 million years old. Lower mountains tend to be older, and are often the eroded relics of much higher mountain chains. About 400 million years ago, when the present-day continents of North America and Europe were joined, the Caledonian mountain chain was the same size as the modern Himalayas. Today, however, the relics of the Caledonian orogeny (mountain-building period) exist as the comparatively low mountains of Greenland, the northern Appalachians in the United States, the Scottish Highlands, and the Norwegian coastal plateau.

29. Which of the following can be inferred from paragraph 2 about the mountains of the Himalayas?

- Their current height is not an indication of their age.
- At present, they are much higher than the mountains of the Caledonian range.
- They were a uniform height about 400 million years ago.
- They are not as high as the Caledonian mountains were 400 million years ago.

“I think this sentence is correct from the information that I have found...as a general rule, the higher a mountain is, the more recently it was formed. So, because the age of the Himalayas is younger, so it’s higher than the Caledonian range” [Participant 6, 06:02:01:27-33].

Cedric Korthals, 06:02:01:27-33
Coding of ‘building a mental model’ (P6)

Paragraph 5: The weather, in its many forms, is the main agent of erosion. Rain washes away loose soil and penetrates cracks in the rocks. Carbon dioxide in the air reacts with the rainwater, forming a weak acid (carbonic acid) that may chemically attack the rocks. The rain seeps underground and the water may reappear later as springs. These springs are the sources of streams and rivers, which cut through the rocks and carry away debris from the mountains to the lowlands.

Participant verbalisation:
“‘Carbon dioxide’ is a kind of chemical, and it can erode rocks, but does the article talk about the origin of this chemical? I don’t recall the article talking about that… I read the sentence again when it talks about carbon dioxide, and certainly not comparing the two of them, because carbonic acid is simply the outcome after the interaction with the chemical and the rocks. What does it mean to say that rainwater penetrates soil? Because the previous sentence talks about ‘rain washes soil and penetrates cracks’, so this sentence still a continuation of that… thinking about whether it’s a continuation of that topic or it’s an initiation for a new idea, so I had to read the sentence again, and then I decided it’s a continuation of the idea, so I sort of confirmed with myself the choice” [Participant 3, test 2, item 32].
Coding of ‘creating a text-level representation’ (P7)

Directions: Three of the answer choices below are used in the passage to illustrate constructive processes, and two are used to illustrate destructive processes. Complete the table by matching appropriate answer choices to the processes they are used to illustrate. This question is worth 3 points.

<table>
<thead>
<tr>
<th>Answer Choices</th>
<th>Constructive Processes</th>
<th>Destructive Processes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Collision of Earth’s crustal plates</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Separation of continents</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Wind-driven sand</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Formation of grass roots in soil</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Earthquakes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Volcanic activity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Weather processes</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Participant verbalisation:
“I remember when I was reading the article at the very beginning, I noted down a few key words, and in one paragraph in particular, it talks about three different constructive processes, and I have impressions of them being crustal plates, earthquakes and volcanic activity, so that’s why I very quickly picked out 1, 5 and 6. I did go back to the text when I was deciding on number 3. For number 7 I decided that very quickly because I remember one of the main destructive forces mentioned was weather processes. I eliminated number 4 because the ‘formation of grass roots’ is actually part of the forces that protect rocks from erosion. But then I was thinking about separation of continents and thinking whether in the texts it talks about separation of continents within the discussion about destructive processes. This phrase, to remind myself as cue that from this moment onwards, the author begins to talk about destructive forces.” [Participant 3, item 38].
Outcomes

• ‘Specification matrix’ (Buck, 2001)
• Scale weighting of frequency of cognitive processes divided by the number of items per item type (adapted from Cohen and Upton, 2006)

<table>
<thead>
<tr>
<th>Processing level</th>
<th>TOEFL</th>
<th>(BC-v)</th>
<th>(BC-nf)</th>
<th>(BC-ss)</th>
<th>(BC-pr)</th>
<th>(l-rp)</th>
<th>(l)</th>
<th>(l-it)</th>
<th>(R2L-ps)</th>
<th>(R2L-st)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOW</td>
<td>P2</td>
<td>VH</td>
<td>VH</td>
<td>VH</td>
<td>M</td>
<td>VH</td>
<td>VH</td>
<td>H</td>
<td>VH</td>
<td></td>
</tr>
<tr>
<td></td>
<td>P3</td>
<td>M</td>
<td>M</td>
<td>H</td>
<td>H</td>
<td>M</td>
<td>H</td>
<td>H</td>
<td>M</td>
<td></td>
</tr>
<tr>
<td></td>
<td>P4c</td>
<td>H</td>
<td>VH</td>
<td>VH</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td></td>
</tr>
<tr>
<td></td>
<td>P4s</td>
<td>L</td>
<td>H</td>
<td>VH</td>
<td>VH</td>
<td>H</td>
<td>VH</td>
<td>H</td>
<td>H</td>
<td></td>
</tr>
<tr>
<td></td>
<td>P5w</td>
<td>L</td>
<td>S</td>
<td>S</td>
<td>M</td>
<td>L</td>
<td>S</td>
<td>L</td>
<td>L</td>
<td></td>
</tr>
<tr>
<td></td>
<td>P5s/c</td>
<td>S</td>
<td>S</td>
<td>L</td>
<td>H</td>
<td>L</td>
<td>H</td>
<td>L</td>
<td>S</td>
<td></td>
</tr>
<tr>
<td></td>
<td>P6</td>
<td>S</td>
<td>M</td>
<td>S</td>
<td>S</td>
<td>H</td>
<td>S</td>
<td>VH</td>
<td>VH</td>
<td></td>
</tr>
<tr>
<td></td>
<td>P7</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>H</td>
<td></td>
</tr>
</tbody>
</table>

Very high (VH) frequency ≥ 2.00
High (H) frequency ≥ 1.00
Moderate (M) frequency ≥ 0.50
Low (L) frequency ≥ 0.30
Sporadic (S) frequency ≤ 0.29
Discussion

• Methodology not a ‘magic bullet’. Claims regarding participant utterances remain *inferential* (Gass & Mackey, 2000)
• (One of) the first to provide substantive evidence for each of the levels of Khalifa and Weir’s (2009) processing core
• (Some) evidence for hierarchical nature of the model
• Methodological utility: supplement to ET studies (Bax, 2013; Brunfaut & McCray); test development and validation (test-wiseness and local independence)
• **Caveats**
  – Metacognition vs cognition in SRI
  – The ‘Negative response dilemma’
  – Number of participants and cognitive load on the researcher
  – Participant requirements (high-level) and Hawthorne effect
  – Unnatural test completion: artificial metacognitive decision-making
Thank you!

Q&A