Exploring rater behaviour in Aptis Writing performance assessment: an eye-tracking study

Dr. Nathaniel Owen
25th November 2018
Background

The Aptis test:
• English assessment tool from the British Council
• B2B business model
• A1 – C on the CEFR
• Mostly computer-based
• Five components:
  • Speaking
  • Reading
  • Writing
  • Listening
  • Core (grammar and vocabulary)
# Aptis Writing

<table>
<thead>
<tr>
<th>Test Design</th>
<th>Description</th>
<th>Focus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Part 1</td>
<td>Complete a form; basic personal information</td>
<td>Spelling; basic writing conventions (e.g. capital letters)</td>
</tr>
<tr>
<td>Part 2</td>
<td>Form filling – writing short texts</td>
<td>20-30 words. Writing sentences (on-topic) accurate spelling, punctuation, grammar.</td>
</tr>
<tr>
<td>Part 3</td>
<td>Social network style interaction. Respond to three questions.</td>
<td>Address all three questions, on-topic, accuracy, cohesiveness and coherence. 30-40 words.</td>
</tr>
<tr>
<td>Part 4</td>
<td>Two emails; one formal, one informal. Both in reaction to information about a change.</td>
<td>First email – 40-50 words. Second email – 120-150 words. Focus on register, range of vocabulary, grammar and cohesive devices.</td>
</tr>
</tbody>
</table>
You are doing a language course. Fill in the form. Write in sentences. Use 20-30 words. You have 7 minutes.

Aptis Language Club

Please write here about your free time and interests.

I am a big football fan and like to play and watch matches whenever possible. I'm also keen on politics and try to keep up to date with current affairs.
You are a member of a gardening club. You are talking to some other members in the gardening club chat room. Talk to them using sentences.

Use 30 - 40 words per answer. You have 10 minutes.

Aptis Gardening Club Chat

Sam: Hi! Welcome to the club. Can you tell me something about your garden?

Sure, My garden is not very big, but it’s really nice. There are some flowers, roses and tulips, and some bushes. Last year I installed a small fountain and added some rocks around the flowers.

Miguel: Welcome! Is gardening very popular in your country?

Yes, it is. Most of my friends have their own gardens and they really enjoy gardening! Every weekend we meet to discuss our new ideas to make our gardens more beautiful.

Michelle: What is your favourite season, and why do you like it?

My favourite season is spring. It’s warm, the sun is shining and the conditions for gardening are good. It’s the time when we can plant new flowers and trees. I really like spring!
Dear Members,

We are sorry to tell you that from next month membership fees will go up by a minimum of 15%. Also, due to maintenance the club will close from 4pm on Wednesdays for the next four weeks. Please feel free to email us at kjmn@goodhealth.com

Write an email to a friend. Write your feelings about the notice and suggest possible alternatives. Write 50 words. You have 10 minutes.

Hey Olga, what do you think about the fee increase at the sports club? I can’t believe they are increasing the fees and closing early. If you ask me the gym should offer a discount if they plan to close early. I just signed up and already a fee increase! Not happy!! Maria.

Also write an email to the Customer Service Team, explaining your feelings about the notice and suggesting possible alternatives. Write 120-150 words. You have 20 minutes.

Dear Sir or Madam,

I am writing to complain about the fee increase to take effect from next month and your plan to close early on Wednesdays.

I have been a member for only three months and feel it is unacceptable to increase the fee so soon after signing up while also closing early on Wednesdays, which is one of the days when I like to go to the sports club and meet my personal trainer.

I think you should rethink the fee increase if you are withdrawing services. One alternative is to implement the fee increase after you have completed the maintenance. Another alternative is to complete the maintenance after the gym is closed at 10pm.

I look forward to hearing back from you if my recommendations are acceptable. If not, I will look to another sports club for my membership.
# Aptis writing – scoring

<table>
<thead>
<tr>
<th>Part 2</th>
<th>Part 3</th>
<th>Part 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 (B1)</td>
<td>5 (B2 or above)</td>
<td>6 (C2)</td>
</tr>
<tr>
<td>4 (A2.2)</td>
<td>4 (B1.2)</td>
<td>5 (C1)</td>
</tr>
<tr>
<td>3 (A2.1)</td>
<td>3 (B1.1)</td>
<td>4 (B2.2)</td>
</tr>
<tr>
<td>2 (A1.2)</td>
<td>2 (A2.2)</td>
<td>3 (B2.1)</td>
</tr>
<tr>
<td>1 (A1.1)</td>
<td>1 (A2.1)</td>
<td>2 (B1.2)</td>
</tr>
<tr>
<td>0 (A0)</td>
<td>0 (Below A2)</td>
<td>1 (B1.1)</td>
</tr>
<tr>
<td></td>
<td>0 (A1/A2)</td>
<td></td>
</tr>
</tbody>
</table>
Focus on writing

- O’Sullivan (2015) reported erratic level of agreement among raters for parts 3 and 4 during Angoff-scaling.
- Knoch et al (2015) reported improved level of agreement between online and f2f raters.
- …also reported a lack of scale steps for scores 1-3 for task 2.
- …but this may be due to small sample size in investigations to date.
- Greater validity evidence for the writing part of the test needed.
Investigating rater behaviour in L2 tests

- Raters differ in their interpretation of rating scales and therefore apply them differently to samples. Examples of rater effects include severity, halo and central tendency (Eckes, 2005).
- Raters differed significantly in their views on the importance of the various criteria (Eckes, 2008).
- Raters may not divide attention evenly among criteria (Eckes, 2008; Winke and Lim, 2015).
- Crucial to know how raters interpret scale content (Weigle, 2002) to determine reliability.
- Easier to determine rater effects in analytical scales with features marked individually (Hamp-Lyons, 1995)…
- …however, more inconsistencies in holistic scales by inexperienced raters (Barkaoui, 2011).
Recent research:

Ballard (2017) used eye-tracking to show that analytical scales susceptible to ordering effects (raters less able to recall left-most and right-most elements of a rubric).

Eye-tracking offers a ‘window to the mind’ of the participants, but must be used responsibly (Conklin and Pellicer-Sánchez, 2016; 2018).

Offers access to online processing without a secondary questionnaire.
Research questions

The research is motivated by the following main research question:

Q. (How) do Aptis raters engage with scoring rubrics when rating test-takers’ written responses?

Subsidiary research questions:

1. **Test tasks**: Can raters differentiate between scale steps for tasks 1-3?

2. **Rater variability**: what can eye-tracking reveal about how raters use rubrics when rating samples?
Research design

- Fifteen trained Aptis raters who have completed the two-day training session and CEFR familiarisation activities were recruited.
- Each rater was asked to rate responses to tasks 2-4 by 10 test takers (samples identified using Text Inspector).
- Each sample presented on-screen alongside task and relevant rubric.
- Participants’ eye-movements tracked as they rated each sample.
- Stimulated-recall interviews conducted three times for each rater (after test takers 1-3, 4-7 and 8-10) with eye-movement footage and scores as stimuli.
Eye-tracking

• Tobii Pro X3-120 mobile device.
Eye-tracking

Are you aware what your eyes are doing whilst you are reading this sentence?

Eye-tracking metrics (McCray and Brunfaut, 2015).
Eye-tracking

Question:
You are a member of the Teachers International Club. You received this e-mail from the club:

As you know, our club connects teachers from around the world and allows them to share their ideas and experiences in teaching. We would like to find out more about teaching in different parts of the world. We want you to write to us and tell us about teaching in your country. What attracts people to teaching? What are the advantages of the job? We will publish the most interesting responses on our website.

Write an e-mail to your friend. Tell your friend what you think about the questions from the club.

Write about 50 words. You have 10 minutes.

Write an e-mail to the manager of the club and answer the questions about the topic.

Write 120-150 words. You have 20 minutes.

1) Hi Ali,
How are you and how are your family? I hope all of them are okay. I received an email from the international club to share ideas about teaching in my country. As you know most of our student ask about the subject they want to read at the university and after graduation. It depends on the salary that they will have after graduation. So want to give me more ideas about teaching in our country.
Mohammad.

2) Dear manager,
I am a member of the teacher international club.

End of Response
## Research design

### Areas of interest (AoI)

<table>
<thead>
<tr>
<th>Test taker 1</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 B1 (or above)</td>
<td>Likely to be above A2 level.</td>
</tr>
</tbody>
</table>
| 4 A2.2 | • On-topic.  
• Uses simple grammatical structures to produce writing at the sentence level. Errors with basic structures common. Errors impede understanding of the response.  
• Mostly accurate punctuation and spelling.  
• Vocabulary is sufficient to respond to the question(s).  
• Some attempts at using simple connectors and cohesive devices to link sentences. |
| 3 A2.1 | • On-topic.  
• Uses simple grammatical structures to produce writing at the sentence level. Errors with basic structures common. Errors impede understanding in parts of the response.  
• Punctuation and spelling mistakes are noticeable.  
• Vocabulary is mostly sufficient to respond to the question(s) but inappropriate lexical choices are noticeable.  
• Response is a list of sentences with no use of connectors or cohesive devices to link sentences. |
| 2 A1.2 | • Not fully on-topic.  
• Grammatical structure is limited to words and phrases. Errors in basic patterns and simple grammar structures impede understanding.  
• Little or no use of accurate punctuation. Spelling mistakes common.  
• Vocabulary is limited to very basic words related to personal information and is not sufficient to respond to the question(s).  
• No use of cohesion. |
| 1 A1.1 | • Response limited to a few words.  
• Grammar and vocabulary errors so serious and frequent that meaning is unintelligible. |
| 0 A0 | No meaningful language or all responses are completely off-topic (e.g. memorised script, guessing). |

### Question:
You are a new member of the cooking club. Fill in the form. Write in sentences.

Use 20 - 30 words. You have 7 minutes.

**Aptis Cooking Club**

Please tell us about the days and times you can come and the food you enjoy cooking and eating.

I like prepare different types the foods Spanish and Italian. I can come every day, two hours.

**End of response.**
Data analysis

• Ratings were analysed using *Facets* to identify lenient/severe raters and to determine how raters engaged with rating scale.

• Fifteen raters judged ten test taker performances for each of the three task types for a total of 450 elements (15 x 10 x 3). Therefore, a three-facet model was adopted (raters, test-takers, tasks).

• Tasks 2 and 3 of the Aptis test are scored on a six-point scale (0-5); task 4 uses a seven-point scale (0-6).

• The research used a Rasch-Masters Partial Credit Model:

\[ \log_e \left( \frac{P_{nij}}{P_{ni(j-1)}} \right) = B_n - D_i - F_{ij} \]
Data analysis

- Raters fairly consistent (between 0 and 1 logits)
- Range of test taker ability represented (-4 to +3 logits)
- 9/10 test takers above task 2 in terms of ability/task difficulty; suggesting poor scale discrimination may be due to smaller sample sizes.
- Scales 1-3 for task 2 from -5 to +1 logits (scales 4-5 from +1 to +3 logits)
### Data analysis

<table>
<thead>
<tr>
<th>Total</th>
<th>Total Obsvd</th>
<th>Fair-M</th>
<th>Model</th>
<th>Infit</th>
<th>Outfit</th>
<th>Estim.</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score</td>
<td>Count</td>
<td>Average</td>
<td>Average</td>
<td>Measure</td>
<td>S.E.</td>
<td>MnSq ZStd</td>
<td>MnSq ZStd</td>
</tr>
<tr>
<td>-------</td>
<td>-------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>------</td>
<td>--------</td>
<td>--------</td>
</tr>
<tr>
<td>602</td>
<td>150</td>
<td>4.0</td>
<td>4.15</td>
<td>-2.14</td>
<td>.14</td>
<td>1.72</td>
<td>4.8</td>
</tr>
<tr>
<td>398</td>
<td>150</td>
<td>2.7</td>
<td>2.59</td>
<td>.35</td>
<td>.10</td>
<td>.44</td>
<td>-5.9</td>
</tr>
<tr>
<td>293</td>
<td>150</td>
<td>2.0</td>
<td>1.38</td>
<td>1.79</td>
<td>.09</td>
<td>.66</td>
<td>-2.8</td>
</tr>
<tr>
<td>431.0</td>
<td>150.0</td>
<td>2.9</td>
<td>2.71</td>
<td>.00</td>
<td>.11</td>
<td>.94</td>
<td>-1.3</td>
</tr>
<tr>
<td>128.3</td>
<td>0</td>
<td>.9</td>
<td>1.13</td>
<td>1.62</td>
<td>.02</td>
<td>.56</td>
<td>4.5</td>
</tr>
<tr>
<td>157.1</td>
<td>0</td>
<td>1.0</td>
<td>1.39</td>
<td>1.98</td>
<td>.03</td>
<td>.68</td>
<td>5.6</td>
</tr>
</tbody>
</table>

Model, Popln: RMSE .11 Adj (True) S.D. 1.62 Separation 14.35 Strata 19.46 Reliability 1.00
Model, Sample: RMSE .11 Adj (True) S.D. 1.98 Separation 17.58 Strata 23.78 Reliability 1.00
Model, Fixed (all same) chi-square: 544.7 d.f.: 2 significance (probability): .00
Model, Random (normal) chi-square: 2.0 d.f.: 1 significance (probability): .16

- Both infit and outfit mean square values for task 2 indicate misfit (>1.2), suggesting limited discriminatory power.
- Large infit mean square indicates unexpected patterns in the on-target observations; weak discrimination for individuals with writing ability equivalent to task 2?
- However: Knoch et al (2015) found that there is no point at which each of these levels becomes most probable on the logit scale. Current data indicates that the raters were able to discriminate across score boundaries 1-3:
Data analysis

Model = ?B,?,2B,R5 ; Task: Task 2

Measure relative to item difficulty

Category Probability
Data analysis – eye-tracking

How do raters engage with rubrics? Evidence from eye-tracking:

<table>
<thead>
<tr>
<th></th>
<th>Part 2</th>
<th>Part 3</th>
<th>Part 4</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Task</strong></td>
<td>4.12s</td>
<td>6.86s</td>
<td>5.58s</td>
</tr>
<tr>
<td><strong>Response</strong></td>
<td>15.07s</td>
<td>23.47s</td>
<td>34.02s</td>
</tr>
<tr>
<td><strong>Rubric</strong></td>
<td>2.32s</td>
<td>2.89s</td>
<td>3.01s</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Part 2</th>
<th>Part 3</th>
<th>Part 4</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Task</strong></td>
<td>15.63</td>
<td>25.90</td>
<td>20.11</td>
</tr>
<tr>
<td><strong>Response</strong></td>
<td>50.89</td>
<td>80.19</td>
<td>115.22</td>
</tr>
<tr>
<td><strong>Rubric</strong></td>
<td>8.62</td>
<td>10.82</td>
<td>11.07</td>
</tr>
</tbody>
</table>
Data analysis – eye-tracking

How do raters engage with rubrics? Evidence from eye-tracking:

<table>
<thead>
<tr>
<th>Time to first fixation (average across raters)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Task</td>
</tr>
<tr>
<td>Response</td>
</tr>
<tr>
<td>Rubric</td>
</tr>
</tbody>
</table>
Data analysis – eye-tracking

- Can eye-tracking reveal rater ‘types’?
- Logit measures used to divide raters into three groups (severe, lenient, intermediate).

### Number of rubric fixations

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lenient</td>
<td>150</td>
<td>21.3211</td>
<td>27.11939</td>
</tr>
<tr>
<td>Intermediate</td>
<td>150</td>
<td>34.9400</td>
<td>36.29169</td>
</tr>
<tr>
<td>Severe</td>
<td>150</td>
<td>38.8133</td>
<td>37.26614</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>450</td>
<td>31.6915</td>
<td>34.61781</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sample1-Sample2</th>
<th>Test Statistic</th>
<th>Std. Error</th>
<th>Std. Test Statistic</th>
<th>Sig.</th>
<th>Adj.Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lenient-Intermediate</td>
<td>-55.743</td>
<td>15.013</td>
<td>-3.713</td>
<td>.000</td>
<td>.001</td>
</tr>
<tr>
<td>Lenient-Severe</td>
<td>-75.957</td>
<td>15.013</td>
<td>-5.059</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>Intermediate-Severe</td>
<td>-20.213</td>
<td>15.013</td>
<td>-1.346</td>
<td>.178</td>
<td>.535</td>
</tr>
</tbody>
</table>
Data analysis – eye-tracking

- Logit measures used to divide raters into three groups (severe, lenient, intermediate). Is there an analytical/holistic distinction?

### Scoring rubric visit duration

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lenient</td>
<td>150</td>
<td>5.2049</td>
<td>7.27596</td>
</tr>
<tr>
<td>Intermediate</td>
<td>150</td>
<td>9.0643</td>
<td>9.60924</td>
</tr>
<tr>
<td>Severe</td>
<td>150</td>
<td>11.3199</td>
<td>11.32169</td>
</tr>
<tr>
<td>Total</td>
<td>450</td>
<td>8.5297</td>
<td>9.85575</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sample1-Sample2</th>
<th>Test Statistic</th>
<th>Std. Error</th>
<th>Std. Test Statistic</th>
<th>Sig.</th>
<th>Adj.Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lenient-Intermediate</td>
<td>-60.120</td>
<td>15.015</td>
<td>-4.004</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>Lenient-Severe</td>
<td>-89.800</td>
<td>15.015</td>
<td>-5.981</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>Intermediate-Severe</td>
<td>-29.680</td>
<td>15.015</td>
<td>-1.977</td>
<td>.048</td>
<td>.144</td>
</tr>
</tbody>
</table>
Data analysis – eye-tracking

• Logit measures used to divide raters into three groups (severe, lenient, intermediate).

Unexpected responses:

<table>
<thead>
<tr>
<th>Cat</th>
<th>Score</th>
<th>Exp.</th>
<th>Resd StRes</th>
<th>Nu Ra Nu Te N Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>.1</td>
<td>.9 3.5</td>
<td>9 9 8 8 4 Task 4</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>.2</td>
<td>1.8 4.4</td>
<td>15 15 8 8 4 Task 4</td>
</tr>
</tbody>
</table>

Rater 9, test taker 8, task 4:

“Only one response… it’s not on-topic. It’s either a 1 or a zero but it’s not really a zero because there are sentences and you know he may be A2, the benefit of the doubt.”

Rater 15, test taker 8, task 4:

“well they only did one so they weren’t going to get very high… it responds partially, I guess partially on-topic, but it’s very inaccurate basically, only one task done.”
Data analysis – eye-tracking

• Logit measures used to divide raters into three groups (severe, lenient, intermediate). Is there an analytical/holistic distinction?

Unexpected responses:

```
+-----------------------------------------------+---
| Cat | Score | Exp. | Resd StRes | Nu Ra Nu Te N Task |
+-----------------------------------------------+---
| 1   | 1     | .1   | .9 3.5     | 9 9 8 8 4 Task 4   |
| 2   | 2     | .2   | 1.8 4.4    | 15 15 8 8 4 Task 4 |
+-----------------------------------------------+---
```

Rater 7, test taker 8, task 4 [awarded a score of 0]:
“the only good bit of language was copied [from the task]”

Rater 5, test taker 8, task 4 [awarded a score of 0]:
“a three line response… I read it carefully, but I noticed that actually a lot of it was lifted from the input question. And yes, I mean basically they’d read the task, they’d kind of got an idea “email your friend”. And he put “my friend”, and he kind of lifted it and rewrote it slightly in a slightly different way.”
Data analysis – eye-tracking

- Lenient rater (rater 15, test taker 8, task 4)

Test taker 8

<table>
<thead>
<tr>
<th>Score</th>
<th>Description</th>
</tr>
</thead>
</table>
| 6     | Response shows the following features:  
        • Response on-topic and task fulfilled in terms of appropriateness of register. Two clearly different registers.  
        • Range of complex grammar constructions used accurately. Some minor errors occur but do not impede understanding.  
        • Range of vocabulary used to discuss the topics required by the task. Some awkward usage or slightly inappropriate lexical choices.  
        • A range of cohesive devices is used to clearly indicate the links between ideas. |
| 5     | Response shows the following features:  
        • Some complex grammar constructions used accurately. Errors do not lead to misunderstanding.  
        • Minor errors in punctuation and spelling occur but do not impede understanding.  
        • Sufficient range of vocabulary to discuss the topics required by the task. Inappropriate lexical choices do not lead to misunderstanding.  
        • A limited number of cohesive devices are used to indicate the links between ideas. |
| 4     | Response shows the following features:  
        • Some complex grammar constructions used accurately. Errors do not lead to misunderstanding.  
        • Minor errors in punctuation and spelling occur but do not impede understanding.  
        • Sufficient range of vocabulary to discuss the topics required by the task. Inappropriate lexical choices do not lead to misunderstanding.  
        • A limited number of cohesive devices are used to indicate the links between ideas. |
| 3     | Response shows the following features:  
        • Some complex grammar constructions used accurately. Errors do not lead to misunderstanding.  
        • Minor errors in punctuation and spelling occur but do not impede understanding.  
        • Sufficient range of vocabulary to discuss the topics required by the task. Inappropriate lexical choices do not lead to misunderstanding.  
        • A limited number of cohesive devices are used to indicate the links between ideas. |
| 2     | Response shows the following features:  
        • Control of simple grammatical structures. Errors occur when attempting complex structures.  
        • Punctuation and spelling is mostly accurate. Errors do not impede understanding.  
        • Limitations in vocabulary make it difficult to deal fully with the task. Errors impede understanding in parts of the text.  
        • Uses only simple cohesive devices. Links between ideas are not always clearly indicated. |
| 1     | Response shows the following features:  
        • Control of simple grammatical structures. Errors occur when attempting complex structures.  
        • Punctuation and spelling is mostly accurate. Errors do not impede understanding.  
        • Limitations in vocabulary make it difficult to deal fully with the task. Errors impede understanding in most of the text.  
        • Uses only simple cohesive devices. Links between ideas are not always clearly indicated. |
| 0     | Performance below B1, or no meaningful language or the responses are completely off-topic (e.g. memorised script, guessing). |

Question:
You are a member of a cooking club. You received this e-mail from the club’s manager.

Dear Member,

I am writing to inform you that next week’s cooking classes will be cancelled because your teacher is on holiday. Extra classes may be available when the teacher returns. I am sorry if this creates any difficulties for you. Unfortunately, we will not be able to offer you a refund at this stage.

The Manager:

Write an e-mail to your friend. Write about your feelings and what you think the club should do about the situation.

Write about 50 words. You have 10 minutes.

Write an e-mail to the manager of the club. Write about your feelings and what you think the club should do about the situation.

Write 120-150 words. You have 20 minutes.

1.)

My friend, the cooking class actually met up on Wednesday and we found out that the next week’s cooking classes will be cancelled because the teacher is on holiday. I’m sorry if this creates any difficulties for you and for you.

2.)

End of response.
## Data analysis – eye-tracking

- Severe rater (rater 12, test taker 8, task 4)

### Test taker 8

<table>
<thead>
<tr>
<th>6</th>
<th>C2</th>
<th>Likely to be above C1 level.</th>
</tr>
</thead>
</table>
| 5 | C1 | Response shows the following features:  
  - Response on-topic and task fulfilled in terms of appropriateness of register. Two clearly different registers.  
  - Range of complex grammar constructions used accurately. Some minor errors occur but do not impede understanding.  
  - Range of vocabulary used to discuss the topics required by the task. Some awkward usage or slightly inappropriate lexical choices.  
  - A range of cohesive devices is used to clearly indicate the links between ideas. |
| 4 | B2.2 | Response on-topic and task fulfilled in terms of appropriateness of register: appropriate register used consistently in both responses. Response shows the following features:  
  - Some complex grammar constructions used accurately. Errors do not lead to misunderstanding.  
  - Minor errors in punctuation and spelling occur but do not impede understanding.  
  - Sufficient range of vocabulary to discuss the topics required by the task. Inappropriate lexical choices do not lead to misunderstanding.  
  - A limited number of cohesive devices are used to indicate the links between ideas. |
| 3 | B2.1 | Response partially on-topic and task partially fulfilled in terms of appropriateness of register: appropriate register used consistently in one response. Response shows the following features:  
  - Some complex grammar constructions used accurately. Errors do not lead to misunderstanding.  
  - Minor errors in punctuation and spelling occur but do not impede understanding.  
  - Sufficient range of vocabulary to discuss the topics required by the task. Inappropriate lexical choices do not lead to misunderstanding.  
  - A limited number of cohesive devices are used to indicate the links between ideas. |
| 2 | B1.2 | Response partially on-topic and task not fulfilled in terms of appropriateness of register: appropriate register not used consistently in either response. Response shows the following features:  
  - Control of simple grammatical structures. Errors occur when attempting complex structures.  
  - Punctuation and spelling is mostly accurate. Errors do not impede understanding.  
  - Limitations in vocabulary make it difficult to deal fully with the task. Errors impede understanding in most parts of the text.  
  - Links between ideas are not always clearly indicated. |
| 1 | B1.1 | Response off-topic and task not fulfilled in terms of appropriateness of register. No evidence of awareness of register. Response shows the following features:  
  - Control of simple grammatical structures. Errors occur when attempting complex structures.  
  - Punctuation and spelling is mostly accurate. Errors do not impede understanding.  
  - Limitations in vocabulary make it difficult to deal fully with the task. Errors impede understanding in most of the text.  
  - Unusually simple cohesive devices. Links between ideas are not always clearly indicated. |
| A1/A2 | | Performance below the level of meaningful language or the responses are completely off-topic (e.g. memorised script, guessing). |

**Question:**

You are a member of a cooking club. You received this e-mail from the club's manager.

Dear Member,

I am writing to inform you that next week's cooking classes will be cancelled because your teacher is on holiday. Extra classes may be available when the teacher returns. I am sorry if this creates any difficulties for you. Unfortunately, we will not be able to offer you a refund at this stage.

The Manager

Write an e-mail to your friend. Write about your feelings and what you think the club should do about the situation.

Write about 50 words. You have 10 minutes.

Write an e-mail to the manager of the club. Write about your feelings and what you think the club should do about the situation.

Write 120-150 words. You have 20 minutes.

1.)

My situation is not ideal. Unfortunately, I am out of town next week and I am not able to attend the cooking class. Since it is not convenient for me, I do not think the club should offer refunds. I would appreciate it if they could arrange for substitute classes. This could be a great opportunity for me to learn new skills and meet other club members.

2.)

End of response.
Data analysis – eye-tracking

• Severe rater (rater 12, test taker 8, task 4)
Conclusions and further work:

- **Test tasks:** Can raters differentiate between scale steps for tasks 1-3?
  - Yes (Knoch et al. 2015 not replicated). However, task 2 still lacks discriminatory power. Wide range of ability across lower bands.

- **Rater variability:** do more lenient/severe raters use rubrics in different ways when rating the same samples?
  - Yes. Eye-tracking evidence emerging of how lenient and severe rater types behave differently.
  - Hypothesis: Severe/lenient distinction
    - Analytic raters spend longer examining the rubric to determine how the scripts correspond to specific areas.
    - Holistic raters respond more on basis of an overall impression of the script and spend less time focusing on the rubric.

- **Caveat:** some evidence of method effect:
  - “If I was doing it on my own at home, there is a difference. That’s the tricky thing, isn’t it?” (Rater 15)
Conclusions and further work:

- What next?
  - **Test tasks**: do lenient/severe distinctions hold across tasks?
  - Association between score awarded and band first fixated?
  - Analysis of Factorial ANOVA/weighted data
  - Saccades/regressions still being analysed
  - Qualitative data coding ongoing

- Outcomes:
  - Findings will inform rater training re. analytical/holistic rating. Are parts of the rubrics for some tasks being overlooked in favour of others?
  - Significant benefits of eye-tracking methodology (cross-referenced with Rasch measurement)
THANK YOU