Building a validity argument for automated speaking assessment: Where are we now?

Jing Xu & Edmund Jones
Cambridge Assessment English

15th EALTA Conference, May 2018
Overview

• Literature review
• The prototype test (auto-marker)
• Research questions
• Methodology
• Results
• Conclusion
What is an Automated Speaking Test?

Traditional  Automated
Existing Test Products

High-stakes

VERSANT™

Low-stakes

TOEFL Go!

Liulishuo
Literature Review

Advantages of automated speaking assessment

• High consistency in scoring
• Near-instant score reporting
• On-demand testing
• Cost reduction

(Williamson, 2012; Xi, 2010)
Literature Review

Disadvantages of automated speaking assessment

• Lack of authenticity
  • Construct underrepresentation
    • Subversion and cheating behaviours
  • Potential negative washback

(Chun, 2006; Xi, 2010; Xi, Schmidgall, & Wang, 2016)
Structure of the Cambridge Prototype Speaking Test

- P1: Short answer questions
- P2: Read aloud
- P3: Individual long turn
- P4: Answerphone message with visual
- P5: Interview about a topic
Part 4 – Answerphone message with visual

You will have 1 minute to leave a message for an English-speaking friend about some visual information. First, you have 1 minute to look at the information and prepare what you are going to say. You will then have 1 minute to leave your message.

Your English-speaking friend wants to park at an airport for one week.

This table shows information about parking at the airport.

Look at the table and then leave a message for your friend about parking at the airport for one week.

<table>
<thead>
<tr>
<th>AIRPORT PARKING</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name of car park</td>
</tr>
<tr>
<td>Blue Zone</td>
</tr>
<tr>
<td>Green Zone</td>
</tr>
<tr>
<td>Silver Zone</td>
</tr>
<tr>
<td>Gold Zone</td>
</tr>
</tbody>
</table>
The Cambridge speech auto-marker

L2 Speech (Unscripted) → Speech Recognition → Feature Extraction → Scoring Model

Score Reporting
The Cambridge speech auto-marker

L2 Speech (Read aloud)

Speech Recognition

Feature Extraction

Scoring Model

Score Reporting
Human Assessment Criteria

- Speaking Ability
  - Pronunciation
    - Intelligibility
    - Stress
    - Intonation
  - Language Resource
    - Vocabulary
  - Fluency
    - Grammar
    - Hesitation
    - Rhythm
    - Coherence
  - Discourse Management
    - Content Appropriateness
  - Task Achievement

Automated Scoring Features

- Phoneme
- Stress pattern
- Intonation
- Lexical Diversity
- Complexity of language
- Formulaic sequences
- Grammatical Accuracy
- Grammatical Complexity
- Fluency Measures
- Sentence-level Coherence
- Topic Relevance

(Fulcher, 2003; Luoma, 2004)

### Research Questions

1. How well do the test tasks represent the TLU domain?

2. To what extent does the auto-marker agree with human examiners in grading?

3. Does speaking to a computer affect candidates’ test behaviours?

(Adapted from Kane, 2013)
Methodology

Participants

- 3537 ELLs worldwide
- 23 different countries
- 44 different L1s
- Mainly B1 and B2 speakers
Methodology

Procedure

Test responses → Auto-marker → Human raters
RQ1: How well do the test tasks represent the TLU domain?
Participants’ overall impressions of the speaking test

N=3192 participants responded to the questionnaire
<table>
<thead>
<tr>
<th>Perception</th>
<th>Disagree or strongly disagree</th>
<th>Neutral</th>
<th>Agree or strongly agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>The test instructions were clear and I understood well what I had to do in the test. (N=3142)</td>
<td>3.2%</td>
<td>11.2%</td>
<td>85.6%</td>
</tr>
<tr>
<td>The test allowed me to show my English speaking ability. (N=3151)</td>
<td>9%</td>
<td>21.4%</td>
<td>69.5%</td>
</tr>
<tr>
<td>The tasks in the test are similar to how I use English language in real life. (N=3121)</td>
<td>8.7%</td>
<td>26%</td>
<td>65.3%</td>
</tr>
<tr>
<td>The visuals I saw in the test were easy to understand. (N=3081)</td>
<td>5.8%</td>
<td>15.0%</td>
<td>79.2%</td>
</tr>
</tbody>
</table>
“I find the topics relevant, not too easy nor difficult. I think that these topics are related to what normally happens in daily life.”

“As I hear the questions I felt more comfortable and relax … The speaking text was developed gradually, so you feel good when notice that you start with a repetition, and then you answer easy questions, and then you have to think a little more to answer the final questions.”

“The speaking test could be a little more specific, like, explain me your study field, anything like that, this way we could talk about things we really know.”
RQ2: To what extent does the auto-marker agree with human examiners in grading?
Results

Reliability of human marking

Common set for inter-rater reliability:
for 60 candidates, all 5 parts were marked by all 5 examiners

- Reliability of a single human rater

  Intraclass correlation (Shrout & Fleiss, 1979)

<table>
<thead>
<tr>
<th>Part 1</th>
<th>Part 2</th>
<th>Part 3</th>
<th>Part 4</th>
<th>Part 5</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.84</td>
<td>0.87</td>
<td>0.90</td>
<td>0.88</td>
<td>0.91</td>
<td>0.91</td>
</tr>
</tbody>
</table>

- Inter-rater reliability

  Spearman’s rank correlation coefficient

<table>
<thead>
<tr>
<th>R1/R2</th>
<th>R1/R3</th>
<th>R1/R4</th>
<th>R1/R5</th>
<th>R2/R3</th>
<th>R2/R4</th>
<th>R2/R5</th>
<th>R3/R4</th>
<th>R3/R5</th>
<th>R4/R5</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.94</td>
<td>0.94</td>
<td>0.93</td>
<td>0.95</td>
<td>0.96</td>
<td>0.93</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
<td>0.94</td>
</tr>
</tbody>
</table>
Results

Automarker accuracy

- Spearman rank correlation between human and automarker scores:
  
  \[
  \text{Correlation} = 0.80
  \]

- Human–automarker agreement on CEFR levels:

<table>
<thead>
<tr>
<th></th>
<th>Exact agreement</th>
<th>Adjacent agreement</th>
<th>Misclassified</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>41.4%</td>
<td>89.3%</td>
<td>10.7%</td>
</tr>
</tbody>
</table>

N=2612
Results

Automarker accuracy

(Bland & Altman, 1986)
RQ3: Did speaking to a computer affect candidates’ test behaviours?
### Participants’ perceptions of the online speaking test

<table>
<thead>
<tr>
<th>Statement</th>
<th>Disagree or strongly disagree</th>
<th>Neutral</th>
<th>Agree or strongly agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>I felt comfortable speaking to a computer. (N=3127)</td>
<td>19.3%</td>
<td>26.9%</td>
<td>53.8%</td>
</tr>
<tr>
<td>Recording my voice in the test was easy. (N=3099)</td>
<td>18.0%</td>
<td>23.3%</td>
<td>58.7%</td>
</tr>
<tr>
<td>I had technical issues with the computer or internet. (N=3082)</td>
<td>40.3%</td>
<td>20.5%</td>
<td>39.2%</td>
</tr>
<tr>
<td>I felt nervous or worried. (N=3101)</td>
<td>31.7%</td>
<td>26.4%</td>
<td>41.9%</td>
</tr>
</tbody>
</table>
“I like the new experience to talk with a computer, I felt less pressure than talking with a person.”

“I felt free to talk to a computer just as if I was talking to a real person.”

“I got little nervous because I cannot see a face. This system is efficient but little lonely.”
Participants’ perceptions of an automarked speaking test

<table>
<thead>
<tr>
<th>Question</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Would you worry if you knew the speaking test was marked by a computer and not a human?</td>
<td>35.2%</td>
<td>64.8%</td>
</tr>
</tbody>
</table>
“I know technology has advanced a lot … However, I still do not feel confident that the computer will be able to analyse all the variation that is ok in our use of the language to be able to decide the level of proficiency.”

“Because I don't really know how it works.”
Conclusions

• From learners’ perspectives, the test tasks successfully represent characteristics of speaking in the TLU domain.

• The automarker requires further development and calibration (exact agreement: 41.4%; adjacent: 89.3%; slightly harsher).

• Some candidates feel anxious speaking to a computer and some do not. Most are comfortable with the idea of an automarker.
Directions for Future Research

- Improve automarker accuracy.
- Develop learning-oriented automated feedback.
- Create materials to explain how the automarker works.
- Develop an intelligent dialogic system.
References


Jing Xu: Xu.J@cambridgeenglish.org       Edmund Jones: Jones.E3@cambridgeenglish.org