Use of the CEFR in the development of an academic speaking test: The Test of English for Academic Purposes (TEAP)

Dan Joyce, Eiken Foundation of Japan
Fumiyo Nakatsuhara, CRELLA, University of Bedfordshire
Presentation Outline

① What is TEAP
② TEAP Speaking Test: Development
③ TEAP Speaking Test: Structure
④ Validation Studies
⑤ Study 1 and Study 2
⑥ Results
⑦ Summary/Conclusion
① What is TEAP?

*a new test of academic English for university entrance purposes in Japan*

Key terms:

- Benchmark
- university level
- EFL context of Japan
- model for revision
- four skills
- Japanese high school students
- Japanese Ministry of Education Course of Study for HS
② TEAP Speaking Test: Development

- Eiken
- Eiken Foundation of Japan
- Sophia University
- CRELLA
  Centre for Research in English Language Learning and Assessment
CRELLA’s role (Dr. Fumiyo Nakatsuhara)

- Literature review
- Language function surveys
- Exchanging ideas & info
- Developing draft test specs, draft rating scales, draft examiner frame etc.
  - Mini-trial
  - Trial: Study 1
  - Pilot: Study 2

\[ A \text{ priori Validation studies } \]
② TEAP Speaking: Test Development

Why use the CEFR?

- positive contribution to English–language learning and teaching in Japan by providing useful feedback to test takers beyond the usual pass/fail decisions associated with Japanese university entrance exams.
- facilitate stakeholders’ understanding of test scores and task requirements
- provide scores that indicate test takers’ approximate level in terms of a well-known external criterion
2 TEAP Speaking Test: Development

The role of the CEFR

- TEAP has used the CEFR as a reference point for defining relevant levels of proficiency.

- TEAP has used relevant descriptors from the scales of the CEFR as a springboard from which TEAP–specific descriptors were developed.
Takes account of the levels of English proficiency that we can legitimately expect high school students to display (A2–B1)

Looks forward to a higher level of proficiency beyond high school (B1–B2).

Acts as a bridge between high school and the TLU domain of the academic context of learning at Japanese universities.

As a minimum level of proficiency to access the language used in first-year university classrooms, TEAP focuses on B1–B2.

In order to provide meaningful feedback to as wide a range of test takers as possible, TEAP takes account of the A2 level of proficiency.
<table>
<thead>
<tr>
<th>Part</th>
<th>Task</th>
<th>Level</th>
<th>Language functions (Cognitive demands: grammatical encoding)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Interview</td>
<td>A2</td>
<td>- Providing specific personal information</td>
</tr>
<tr>
<td>2</td>
<td>Role play</td>
<td>A2/B1</td>
<td>- Initiating interaction&lt;br&gt;- Asking for information/opinions&lt;br&gt;- Commenting</td>
</tr>
<tr>
<td>3</td>
<td>Monologue</td>
<td>B1/B2</td>
<td>- Agreeing/disagreeing&lt;br&gt;- Justifying opinions&lt;br&gt;- Elaborating</td>
</tr>
<tr>
<td>4</td>
<td>Extended interview</td>
<td>B2</td>
<td>- Expressing opinions&lt;br&gt;- Justifying opinions&lt;br&gt;- Comparing&lt;br&gt;- Speculating&lt;br&gt;- Elaborating</td>
</tr>
</tbody>
</table>
1. **Level of tasks** were designed to increase across test (A2 $\rightarrow$ B2)

2. **Tasks were designed to reflect language functions** considered important by high school and university teachers

3. **Part 2: role play** designed to operationalise *asking for information, opinions*
   - Due to anticipated lack of familiarity with this test task in Japan, important part of the *a priori* validation studies
The analytic rating scales used by raters have five categories:

1. Pronunciation
2. Grammatical Range and Accuracy
3. Lexical Range and Accuracy
4. Fluency
5. Interactional Effectiveness

Score bands:
- B2
- B1
- A2
- Below A2
TEAP Speaking Test: Score report

Name: EIKEN HINAKO
Date of birth: 1995/04/08
Gender: Female
Test date: 2014/07/20
TEAP ID: TEAPEIKEN
Registration ID: 152-056-80015

Reading
Able to understand the content of the text, and can respond to questions about it. Can also express opinions and ideas accurately and coherently.

Listening
Able to follow conversations and understand the main points.

Writing
Able to write a clear and well-organized essay on a given topic.

Speaking
Able to speak fluently and with good interaction.
4 Validation Studies
Socio-cognitive framework for validating speaking tests (Weir, 2005, Taylor, ed. 2011)

- Test-taker and examiner feedback questionnaires
- Language functions of test-taker speech samples
- Linguistic and discourse features of test-taker speech samples
- Rating scores
- Rater feedback questionnaire & post-marking focus group discussion

- Collected various sources of empirical evidence that offered useful information to verify or modify the draft test materials and rating scales.
Study 1 & Study 2: Research Questions & Methodology
Research Questions

RQ1: To what extent does the test elicit intended language functions in each task? (Study 1)

RQ2: Is there any evidence from test-takers’ output language that validates the descriptors used to define the levels on each rating scale? (Study 1)

RQ3: What are the participating examiners’ and students’ perceptions of the testing procedures? (Studies 1)

RQ4: What are the participating raters’ perceptions of the testing and rating procedures? (Studies 1 and 2)

RQ5: How well does the test function in terms of scoring validity, after incorporating modifications suggested in Study 1? (Study 2)
Participants
- 23 1st year university students
- 3 trained examiners & 3 trained raters

Data collection
- Speaking test sessions were video-recorded and transcribed
- Rating of the video-recorded performances by 3 raters, using the draft rating scales
- Examiner, student and rater feedback questionnaires
- Raters’ focus group discussion

Data Analysis
- Language function analysis (RQ1)
- Linguistic and discourse analysis of students’ speech samples (RQ2)
- Analysis of questionnaires and focus group discussion data (RQ3 & RQ4)

Modifying test materials, scales, etc.
Participants
- 120 3rd year high school students
- 5 trained examiners & 6 trained raters

Data collection
- Video-recorded speaking test performances were rated by 6 trained raters, using the modified rating scales
- Rater feedback questionnaires

Data Analysis
- Analysis of rater feedback questionnaire ($RQ4$)
- Analysis of rating scores ($RQ5$): FACETS analysis
6 Results
Transcripts were analysed and instances of functions from the list based on O’Sullivan et al (2002) were counted.
Part 1: Interview

- **Informational**
  - Expressing opinions
  - Giving info
  - Elaborating

- **Interactional**

- **Managing interaction**

**Mean**

- Giving information
- Elaborating
- Expressing opinions
Part 2: Role-play

- Informational
- Interactional
- Managing interaction

**Mean**

- Asking for info
- Asking for opinions
- Commenting
- Asking for permission

**Categories**

- Asking for opinions
- Commenting
- Asking for info
- Reciprocating
- Initiating
- Managing interaction
Part 3: Monologue

Informational

Interactional

Managing interaction

Elaborating

Justifying opinions

Agreeing, disagreeing
Part 4: Extended interview

Informational

- Elaborating
- Expressing opinions
- Justifying opinions
- Comparing
- Speculating

Interactional

Managing interaction

- Inferring
- Presenting
- Presenting points
- Justifying points
- Comparing points
- Speculating points
Main Findings

✓ Targeted functions were elicited by the relevant parts of the test as intended

[Suggestions for modifications]

Part 3 (Monologue):
Limit the examiner’s contribution only to non-verbalised response tokens (such as nodding, smiling)
Part 4 (Extended Interview):
Standardize the way that examiners end the test
(Following Brown’s (2006a) methodology)

• **List key assessment areas** specified in each rating category
• Identify linguistic and discourse features that could **quantify the key areas**
• **Analyse candidate output language** for these features
• **Compare the results** across different proficiency groups to see to what extent each of these features differs between adjacent levels of the rating scales.
Some of the Fluency measures

No. of unfilled pauses (utterance initial) per 50 words

Ratio of repair, false starts and repetition to AS units

Articulation rate

Note: No inferential statistics due to the small sample size
Main Findings

- All examined features broadly exhibited changes in the expected direction across the 3 levels.

  ➔ The rating scales are in general differentiating test-takers’ performance in a way congruent with the test designers’ intention.

- For some scales, the differences between levels were greater at one boundary than the other.

  ➔ In accordance with previous research (e.g. Brown, 2006a) indicating that specific aspects of performance are probably more relevant to differentiate particular levels.

  ➔ Worth following up in a larger-scale study.
Analysis of feedback questionnaires and raters’ focus group discussion (RQ3 & RQ4)
Main Findings

- Examiner training/post-interviewing questionnaires
  - The training session was useful
  - The test timings, instructions, questions and general test administration were appropriate

[Suggestions for modification]
- The test instructions could be clearer
- They need guidelines for what they should do when they feel a need for deviating from the interlocutor frame.

- Student feedback questionnaire
  - The role-play task especially was received positively, which confirmed the use of this innovative task in the Japanese context.
Main Findings

- Rater training/post-marking questionnaires and post-marking rater discussion + Test score analysis
  - The training session was useful and effective

[Suggestions for modifications]
- Some adjustments to the wording of descriptors in the Interactional Effectiveness scale (too easy in the draft scales)
- Raters need to be more explicitly instructed that an overall impression should not influence their individual analytic scores, especially on the Pronunciation scale.

All modifications suggested in Study 1 were discussed by the project team, and revised rating scales and test materials were prepared for Study 2.
Score Analysis & Rater Questionnaire Analysis in Study 2 (larger-scale pilot)
Score Analysis
- Multi-faceted Rasch analysis with 3 facets: examinees (N=120), raters (N=6) and rating categories (N=5)
- The scoring system generally worked well
  - No misfitting rater or rating category
  - All raters behaved with an acceptable level of consistency
  - For all rating categories, the rating scale steps progressed in the expected way

Rater questionnaire Analysis
- The revised rating scales worked better

Study 2 results demonstrated that changes made after Study 1 functioned in ways that test designers intended
7 Conclusion
Conclusion

- When coupled with thorough validation studies to guide its use, the CEFR can become a useful tool in test development.

- We can be confident that the TEAP Speaking Test is operationalising the test construct which the test was designed to measure.

But on-going validation studies are as important as *a priori* validation!
For a full validation report, see Nakatsuhara (forthcoming, online)

For more information about TEAP, see the following URL:

https://www.eiken.or.jp/teap/

Thank You!

d-joyce@eiken.or.jp
References:


