Multi-faceted Rasch analysis in rating tasks in Japan’s university entrance examinations

Rie KOIZUMI (Juntendo University)
rkoiizumi@juntendo.ac.jp
PROMS (Pacific Rim Objective Measurement Society) 2016, Orient Hotel, Xi'an, China, August 1
Reforms in university entrance examinations in Japan (MEXT, 2014; Yomiuri Shimbun Kyouikubu, 2016)

- To keep up with the internationalization
- To increase those who have the ability to think, judge, and express ideas and perform actively
- Exam construct: knowledge, skills $\rightarrow +$ ability to think logically and express effectively
- L2 English test: reading, listening $\rightarrow +$ writing, speaking
Current types of university entrance examinations

- (1) General examinations
- (2) Recommendation-based examinations
- (3) Admissions Office (AO) examinations
- (4) Special selection examinations

→ Restructured? All should use both academic tests and other methods such as interviews and long essays.
Types of tests in general examinations

- (a) Only the National Center Test for University Admissions (Center Test)
  - Administered only once a year to about 550,000 examinees nationwide

- (b) Only a university-developed test
  - Limited quality control

- (c) Both
Reforms in the Center Test

- Format: multiple-choice (MC) with single answers → + MC with single and multiple answers + constructed-response
  - × Multiple administrations per year
  - × Not select applicants based on a one-point difference in test scores
- L2 English four-skill test, with the speaking section using voice recorders
- Analysis: CTT → + IRT
Concrete plan proposed for the Center Test reform

- For math and L1 Japanese:
  - 2020 to 2023: Elicit relatively short answers (40–80 Japanese characters)
  - From 2024: Administered digitally and elicit longer written responses (200–300 characters)
- Scored initially by humans, but later by artificial intelligence
40–80 Japanese characters

西安は観光するところも多く、滞在はとても充実しています。兵馬俑に行き、中国の歴史と文化を学びました。兵馬俑博物館では写真をとってよいことに驚きました。（76 Japanese characters）
Constructed responses require judgments using rating scales

- Rating involves many factors (McNamara, 1996; Fulcher, 2003)
- Human raters; automated system (e.g., characteristics, training)
- Rating scale (e.g., orientation, construct definition)
- Task (e.g., orientation, goals)
- Interlocutor
- Local performance condition
Multi-faceted Rasch analysis (MFRA)

- Useful in analyzing data affected by many factors (Barkaoui, 2013; Bond & Fox, 2015; Eckes, 2011; Engelhard, 2013; McNamara & Knoch, 2012)
- Facets: Test takers, tasks, raters, rating criteria, + α (Two or more facets)
- Obtain detailed information that cannot be derived from raw-score data analysis
- E.g., Do raters produce stable scores? Do rating scales work effectively? Do tasks work in an expected manner?
Argument-based approach to validity (Chapelle et al., 2008)

- **F. Utilization**
  - The degree of impact on learning and teaching

- **E. Extrapolation**
  - Relationship with other test scores

- **D. Explanation (reflection of the construct)**
  - Factor structure
  - Function analysis

- **C. Generalization (reliability)**
  - The test and raters provides stable estimates.

- **B. Evaluation (appropriate scoring)**
  - Statistical characteristics of tasks/a rating scale
  - Sufficient difficulty spread of tasks

- **A. Domain Definition**
  - Relevance and representativeness of the target domain
Issues in reforms

- Radical proposal in 2014
- Slow and limited progress

Background:
- Conflicting principles (Kuramoto & Koizumi, 2016)
  - Principle of education
  - Principle of measurement
- Methodological preferences
  - Some users of 2-parameter item response theory
  - Limited users of Rasch model
Focus on speaking tests testing L2 English ability 1/2

- Speaking tests for university entrance examinations
- (1) Four skill test
  - **TEAP** Eiken
  - **GTEC CBT** GTEC for Students
  - IELTS Cambridge English
  - TOEIC (LR + SW)
  - TOEFL iBT TOEFL Junior Comprehensive
Focus on speaking tests testing L2 English ability 2/2

- Separate speaking tests
  - As a supplement to the current exam (Mizohata, 2016)
  - SST
  - **TSST**
  - OPIc
  - Versant
  - TOEIC Speaking
TEAP (Test of English for Academic Purposes)

- Developed by the Eiken Foundation of Japan and Sophia University (http://www.eiken.or.jp/teap/)
- Construct: academic English proficiency required for learning and researching at universities
- 10-min face-to-face, one-on-one interviews
- CEFR about A2 to B2
- 0 to 30 points
- Rater: training session
TEAP tasks and criteria

- Part 1: Interview
- Part 2: Role play (interviewing an examiner)
- Part 3: Monologue
- Part 4: Extended interview
- 5 criteria
  - 4 levels (0 to 3; Below A2, A2, B1, B2)
  - (a) Pronunciation, (b) Grammatical Range and Accuracy, (c) Lexical Range and Accuracy, (d) Fluency, (e) Interactional Effectiveness
TEAP analysis (Nakatsuahara, 2014)

- Analysis of the pilot study data
- 120 test takers
- 5 criteria
- 6 trained raters
- Facets
  - using a partial credit model
TEAP pilot study results

- Favorable overall
- High rater agreement (actual: 59.7%)
- 5.67 strata of test takers
- 2.8% underfitting test takers
- Appropriate rating scale functions
- Only a pilot study was analyzed using MFRM.
GTEC CBT (Global Test of English Communication Computer Based Testing)

- Developed by the Benesse Corporation and Center for Entrance Examination Standardization (http://www.benesse-gtec.com/cbt/en)
- Construct: English proficiency in four skills for academic purposes
- 20-min computer-based test
- CEFR about A2 to B2
- Maximum score: 350 points
GTEC CBT tasks

- **Part 1: Listening and responding (6 items)**
  - E.g., Where do people like to travel to in your country?
  - When is the best time of year to visit your country?

- **Part 2: Delivering and asking for information (3 items)**

- **Part 3: Expressing your opinion (3 items)**
  - Do you think technology has changed the way we live?
GTEC CBT criteria

- 23 criteria in total. Each item has 1 to 5.
- Full marks: 1 to 3 points
- Example of criteria
  - Part 1: Listening and responding
    - Respond to simple questions appropriately and clearly
  - Part 2: Delivering and asking for information
    - Based on the provided information, give the factual information and preference, and ask questions
  - Part 3: Expressing your opinion
    - State an opinion and provide reasons to support the opinion
GTEC CBT rater training, quality maintenance

- Practice session and certification
- Trial session
- Independent ratings by two raters.
- Divergent responses are assessed by the third experienced rater.

Analysis (Koizumi, Okabe, & Kashimada, 2016): 648 test takers, 23 criteria, 13 trained raters; Facets (Version 3.71.4), using partial credit model
### GTEC CBT Results 1/2

<table>
<thead>
<tr>
<th></th>
<th>Min〜Max</th>
<th>M (SD)</th>
<th>Strata</th>
<th>Reliability</th>
<th>% of under-fitting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test takers</td>
<td>-6.94〜 5.10</td>
<td>0.92 (1.44)</td>
<td>6.35</td>
<td>.95</td>
<td>8%</td>
</tr>
<tr>
<td>Criteria</td>
<td>-3.18〜 3.35</td>
<td>0.00 (1.54)</td>
<td>29.84</td>
<td>1.00</td>
<td>0%</td>
</tr>
<tr>
<td>Raters</td>
<td>-0.41〜 0.17</td>
<td>0.00 (0.17)</td>
<td>3.64</td>
<td>.86</td>
<td>0%</td>
</tr>
</tbody>
</table>
Favorable overall

High rater agreement (actual: 79.4% > expected: 63.4%)

Criteria and rater fit to the model

Bias analysis: Rater x criteria: 0%
Rater x test taker: 1.39%
Test taker x criteria: 0.01%

Off-topic responses received unexpectedly low scores.

Appropriate rating scale functions
TSST (Telephone Standard Speaking Test)

- Developed by ALC (https://tsst.alc.co.jp/tsst/e_contact.html)
- Based on ACTFL OPI
- Construct: how well a person can answer function-based questions spontaneously
- 15-min telephone-mediated test
- Prompts both in L1 Japanese and L2 English
- 1 to 9 levels (Novice to Advanced)
- CEFR A1 to B2
TSST

- 10 tasks (no preparation time; speaking for 45 sec)
- Questions selected randomly from a question pool
- From intermediate level to advanced level tasks
  - E.g.,
TSST rater training, quality maintenance

- Certification
- Training sessions annually
- Independent ratings by three raters
- One of them is an experienced rater.
- Analysis (Koizumi, 2016) 5406 test takers, 771 tasks, 32 trained raters
- Facets (Version 3.71.4), using a rating scale model
TSST Results 2/2

- Favorable overall
- Task and rater fit to the model
- 2 task separation is intentional (intermediate and advanced level tasks)
- High % of underfitting test takers
- Appropriate rating scale function
- Require further analysis
Toward reforms in Japanese university entrance exams

- Recent use of multi-faceted Rasch analysis in L2 speaking tests
  - Should be expanded to other types of tests
- Routine analysis and reporting to the public will benefit both test developers and test users.
- Need for cooperation with content experts and measurement professionals
Acknowledgments

- I appreciate Dr. Zhang and organizing committee members for inviting me to PROMS 2016.
- This work was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI, Grant-in-Aid for Scientific Research (C), Grant Number 26370737.
References


