Developing and Evaluating Performance-Based Assessments

Best Practices and Lessons Learned from an Online Chinese Course

Katharine B. Nielson and Megan C. Masters

Language Flagship: Results 2012
Friday October 26, 2012
New York City, NY
Outline

• Task-Based Language Assessment
• Research questions/methodology
• Results of empirical study
• Implications for classroom teachers
Task-Based Language Teaching


- Conduct a Needs Analysis (Long, 2005)
- Sequence course in terms of tasks (Robinson, 2001; Skehan, 1998)
- Promote learning by doing (Doughty & Long, 2003)
- Focus on form (Long, 1991; Long & Robinson, 1998)
- Use task as unit of analysis in assessments (Norris, 2002; Norris, 2009)
Performance-based, construct-based assessment, or combination?

- Performance-referenced assessment can be appropriate (Mislevy, et. al., 2002; Norris, 2002; Robinson & Ross, 1996)

- Performance-based assessment cannot stand alone and TBLT courses should include construct assessment (Bachman, 2002)
Develop rubrics

Specify criterial levels for each subtask, defining minimal evidence for task completion

Identify subtasks essential for task accomplishment

Identify target tasks

TBLA
Empirical Study

- What is the relationship between language performance and task accomplishment?
- How well do the rubrics (subtasks and success criteria) measure learner performance?
- How well does the rating scale work?
- Do rater differences affect scoring?
Research Setting

Yearlong, online, task-based Chinese course

35 Post-STARTALK, high school students

College-level intermediate course (CHIN 201)

3 college credits over 2 semesters
Methodology

Needs Analysis

Test Development

Test Administration

Rasch Analysis

Test Evaluation
## Sample Rubric

<table>
<thead>
<tr>
<th>Subtask</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Greets vendor</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Success criterion: Student responds appropriately to conversation partner’s greeting.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Requests food</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Success criterion: Conversation partner confirms choice of food.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clarifies food request</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Success criterion: Student responds appropriately to conversation partner’s clarification questions about food choice.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Requests drink</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Success criteria: Conversation partner confirms choice of drink.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# Sample Rubric

**Language Skills Evaluation:**
- If criteria met, in the “yes” column, indicate how well.
- 1 = student barely met success criteria
- 5 = student met success criteria perfectly

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Demonstrates Chinese language accuracy</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Success criteria: Student’s questions, comments, and responses are appropriate to the situation.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 - 2 - 3 - 4 - 5</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Demonstrates Chinese language fluency</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Success criteria: Student’s speech is clear and in Chinese (with the exception of proper names in English). Pauses and mis-starts do not detract from comprehensibility.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 - 2 - 3 - 4 - 5</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Pronounces Chinese language clearly</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Success criteria: Conversation partner is able to understand Student’s pronunciation and responds to questions and comments appropriately.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 - 2 - 3 - 4 - 5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Correlation Analysis

<table>
<thead>
<tr>
<th>Role-play Assessment</th>
<th>Mean percentage of subtasks completed correctly</th>
<th>Mean Likert rating for accuracy, fluency, and pronunciation</th>
<th>Pearson product-moment coefficient between percentage and Likert rating</th>
<th>Significant at p &lt; .05</th>
</tr>
</thead>
<tbody>
<tr>
<td>Follow street directions</td>
<td>89%</td>
<td>4.02</td>
<td>.547</td>
<td>yes</td>
</tr>
<tr>
<td>Order food from a vendor</td>
<td>98%</td>
<td>4.40</td>
<td>.53</td>
<td>yes</td>
</tr>
<tr>
<td>Order food at a restaurant</td>
<td>87%</td>
<td>3.94</td>
<td>.804</td>
<td>yes</td>
</tr>
<tr>
<td>Give street directions</td>
<td>84%</td>
<td>3.84</td>
<td>.561</td>
<td>yes</td>
</tr>
<tr>
<td>Buy something</td>
<td>82%</td>
<td>4.10</td>
<td>.905</td>
<td>yes</td>
</tr>
<tr>
<td>Arrange travel</td>
<td>87%</td>
<td>4.47</td>
<td>.810</td>
<td>yes</td>
</tr>
<tr>
<td>Final</td>
<td>83%</td>
<td>4.54</td>
<td>.819</td>
<td>yes</td>
</tr>
</tbody>
</table>
Multi-Faceted Rasch Analysis

- Person ability, item difficulty and rater severity converted to logit (log-odds) metrics
- Allows for direct comparisons of outcomes
- Consistency of person, item and rater calibration
- Visual examination of task item difficulty relative to person ability estimates
- Use of Likert scale
“Arranging a Trip”
Multi-faceted Output (n=19)

Raters do not exhibit substantial differences in severity

Majority of learners have ability estimates higher than most difficult tasks

Rubric does not adequately measure learners with ability estimates > 2.5 logits

Item redundancy
Probability Curves: Example

Uniform probability curves indicate equal-interval scale

Distinct portion of underlying construct of interest

Important for parametric analyses

F-thresholds
“Arranging a Trip” Probability Curves

Learners most likely to be given a rating of a 2 or 3

Learners least likely to be given a rating of 4

Absence of rating 1

Not representative of interval scale
### “Buying Something”

**Multi-faceted Output (n=22)**

- **Raters do not exhibit substantial differences in severity**
- **Majority of learners have ability estimates higher than most difficult tasks**
- **Rubric does not adequately measure learners with ability estimates > 3.5 logits**
- **Item redundancy**

<table>
<thead>
<tr>
<th>Measure</th>
<th>Examinees</th>
<th>Rater</th>
<th>Item</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>*</td>
<td>+</td>
<td></td>
<td>(5)</td>
</tr>
<tr>
<td>6</td>
<td>*</td>
<td>+</td>
<td></td>
<td>---</td>
</tr>
<tr>
<td>5</td>
<td>*</td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>*</td>
<td>+</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>*</td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>*</td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>*</td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>*</td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
"Buying Something" Probability Curves

Learners most likely to be given a rating of a 2 or 3

Learners least likely to be given a rating of 4

Idiosyncratic use of rating scale
“End-of-Course”
Multi-faceted Output (n=20)

Raters do not exhibit substantial differences in severity

Majority of learners have ability estimates higher than most difficult tasks

Rubric does not adequately measure learners with ability estimates > 2.0 logits

Item redundancy
“End-of-Course” Probability Curves

Uniform use of rating scale

Placement of learners proportional to range of learners’ ability estimates

Equal interval rating scale
Results of Rasch Analysis

• Commonalities among easy and difficult subtasks across modules
  – Clarifying information was difficult
  – Confirming information was easy

• Overall, assessment items were too easy for learners

• Likert scale could be reduced from 1-5 to 1-3
Conclusions

• More subtasks are needed for a nuanced picture of learner abilities
• Important to take rater severity into account when using criterion-referenced PBAs
• More and clearer criteria are needed
• Future iterations of Likert rating scale should be accompanied by category definitions
Implications for Classroom

Performance-based assessment can offer information in addition to standardized proficiency measures.

Rasch analysis can reveal issues with testing instruments and with rater severity estimates.

Teachers can use Rasch analysis to iteratively develop and validate their own tools.
Practical Considerations

• These assessments were developed for online instruction
  – More than one rater could be present
  – Fluent interlocutors were not limited by physical constraints
  – Tasks needed to be adapted
  – Technological constraints affected assessments
Questions?
Thank you!

• Special thanks to the Associate Directorate of Education and Training (ADET)
• …and to Dr. Der-lin Chao, Dr. Tamara Green and the many talented graduate students who collaborated with us on this project
References


