Comparing native and non-native raters of US Federal Government speaking tests

Rachel L. Brooks, PhD
Federal Bureau of Investigation
ALTE 2014
Native Speaker in Language Testing

This study

Zhang & Elder, 2011
Brown, 1995

Johnson & Lim, 2009
Hill, 1996

Ludwig, 1982
Kim, 2009
Shi, 2001
Kang, 2008

Fayer & Krasinski, 1987;
Galloway, 1980
Barnwell, 1989
Rossiter, 2009

Conservative

Aggressive

Rater Support

Sample Size
ILR Skill Level Descriptions

Current study

LT/SLA Research
### What is a native speaker?

**Paikeda, 1985**

1. A person who has a specified language as the mother tongue or first learned language
   - having at least a bachelor’s degree from a reputable college or university
2. A competent speaker of a specified language
   - who uses it idiomatically or in the usual way including structure, syntax, and grammar

**Davies, 2003**

1. Acquired L1/native language in childhood
2. Has intuitions (acceptability/productiveness) about his idiolectal grammar
3. Has intuitions about standard grammar
4. Is widely fluent, spontaneous, with huge vocabulary and communicative competence
5. Writes creatively

If L1/mother tongue is relevant, then a non-native rater with an L1 similar to the language tested might rate more accurately.

If competence/ability is relevant, then a non-native rater with a higher speaking proficiency might rate more accurately.
Rater Distribution (n=30)

Research Question 1

Research Question 2

Research Question 3

Linguistic Category Ratings

Final Ratings

Final Ratings

Final Ratings
Research Questions

1. Do native and non-native speaker raters assign comparable ratings on speaking tests?
2. Does speaking proficiency level affect a rater’s ability to reliably evaluate speaking proficiency?
3. Does the first language learned affect a rater’s ability to reliably evaluate speaking proficiency?
4. Do native and non-native raters assess the specific linguistic features of the speaking samples comparably?
## Raters/ Samples Evaluated

### Exams Rated

<table>
<thead>
<tr>
<th>ILR Level</th>
<th>NS</th>
<th>NNS</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>4/4+/5</td>
<td>5</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>3/3+</td>
<td>7</td>
<td>7</td>
<td>14</td>
</tr>
<tr>
<td>2/2+</td>
<td>1</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>Total</td>
<td>13</td>
<td>12</td>
<td>25</td>
</tr>
</tbody>
</table>

= 750 evaluations

### Raters

<table>
<thead>
<tr>
<th>Language</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>English (NS)</td>
<td>6</td>
</tr>
<tr>
<td>Arabic (NNS)</td>
<td>4</td>
</tr>
<tr>
<td>Farsi (NNS)</td>
<td>3</td>
</tr>
<tr>
<td>French (NNS)</td>
<td>3</td>
</tr>
<tr>
<td>German (NNS)</td>
<td>3</td>
</tr>
<tr>
<td>Mandarin (NNS)</td>
<td>4</td>
</tr>
<tr>
<td>Spanish (NNS)</td>
<td>4</td>
</tr>
<tr>
<td>Vietnamese (NNS)</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td>30</td>
</tr>
</tbody>
</table>
Inter-rater Reliability (Krippendorf’s alpha)

Research Question 1

- NS: 0.77

Research Question 2

- L5: 0.77
- L4: 0.58
- NNS: 0.59
- L3: 0.62
- L2: 0.62

Research Question 3

- En: 0.77
- Ar: 0.60
- Fa: 0.53
- Fr: 0.67
- Ge: 0.74
- Ma: 0.52
- Sp: 0.53
- Vi: 0.67
RQ 1: NS and NNS Group Mean Ratings

No Significant Differences
RQ 2: English Proficiency Level Group Mean Ratings

- **All Levels**
- **Level 2/2+**
- **Level 3/3+**
- **Level 4/4+/5**

- **ILR Level**
- **Exam Level**

- $p = 0.01$, partial $\eta^2 = 0.02$
- $p = 0.00$, partial $\eta^2 = 0.04$
RQ 3: First Language Mean Ratings

- **All Levels**: $p = 0.01$, partial $\eta^2 = 0.02$
- **Level 2/2+**: $p = 0.00$, partial $\eta^2 = 0.12$
- **Level 3/3+**: $p = 0.00$, partial $\eta^2 = 0.05$
- **Level 4/4+/5**: $p = 0.00$, partial $\eta^2 = 0.12$
RQ4: NS and NNS Raters: Mean Linguistic Category Ratings

Overall: \( p = 0.00, \text{ partial } \eta^2 = 0.04 \)
RQ4: English Proficiency:
Mean Linguistic Category Ratings

Overall:
\( p = 0.00, \) partial \( \eta^2 = 0.04 \)

### Linguistic Category

- **Functions**
  - L2
  - L3
  - L4
  - L5

- **Organization**
  - L2
  - L3
  - L4
  - L5

- **Structures**
  - L2
  - L3
  - L4
  - L5

- **Vocabulary**
  - L2
  - L3
  - L4
  - L5

- **Fluency**
  - L2
  - L3
  - L4
  - L5

- **Pronunciation**
  - L2
  - L3
  - L4
  - L5

- **Social/Cultural Appropriateness**
  - L2
  - L3
  - L4
  - L5

### Statistics

- **Functions**
  - L2: \( p = 0.00, \) partial \( \eta^2 = 0.03 \)
  - L3: \( p = 0.00, \) partial \( \eta^2 = 0.02 \)
  - L4: \( p = 0.01, \) partial \( \eta^2 = 0.02 \)
  - L5: \( p = 0.05, \) partial \( \eta^2 = 0.01 \)

- **Organization**
  - L2: \( p = 0.00, \) partial \( \eta^2 = 0.02 \)
  - L3: \( p = 0.00, \) partial \( \eta^2 = 0.02 \)
  - L4: \( p = 0.01, \) partial \( \eta^2 = 0.02 \)
  - L5: \( p = 0.05, \) partial \( \eta^2 = 0.01 \)

- **Structures**
  - L2: \( p = 0.00, \) partial \( \eta^2 = 0.03 \)
  - L3: \( p = 0.00, \) partial \( \eta^2 = 0.02 \)
  - L4: \( p = 0.01, \) partial \( \eta^2 = 0.02 \)
  - L5: \( p = 0.05, \) partial \( \eta^2 = 0.01 \)

- **Vocabulary**
  - L2: \( p = 0.00, \) partial \( \eta^2 = 0.03 \)
  - L3: \( p = 0.00, \) partial \( \eta^2 = 0.02 \)
  - L4: \( p = 0.01, \) partial \( \eta^2 = 0.02 \)
  - L5: \( p = 0.05, \) partial \( \eta^2 = 0.01 \)

- **Fluency**
  - L2: \( p = 0.00, \) partial \( \eta^2 = 0.03 \)
  - L3: \( p = 0.00, \) partial \( \eta^2 = 0.02 \)
  - L4: \( p = 0.01, \) partial \( \eta^2 = 0.02 \)
  - L5: \( p = 0.05, \) partial \( \eta^2 = 0.01 \)

- **Pronunciation**
  - L2: \( p = 0.00, \) partial \( \eta^2 = 0.03 \)
  - L3: \( p = 0.00, \) partial \( \eta^2 = 0.02 \)
  - L4: \( p = 0.01, \) partial \( \eta^2 = 0.02 \)
  - L5: \( p = 0.05, \) partial \( \eta^2 = 0.01 \)

- **Social/Cultural Appropriateness**
  - L2: \( p = 0.00, \) partial \( \eta^2 = 0.03 \)
  - L3: \( p = 0.00, \) partial \( \eta^2 = 0.02 \)
  - L4: \( p = 0.01, \) partial \( \eta^2 = 0.02 \)
  - L5: \( p = 0.05, \) partial \( \eta^2 = 0.01 \)
Conclusions

1. No significant difference between NS and NNS raters
   ▫ Any differences can be overcome by training
   ▫ FBI SPT raters are not typical people
   ▫ Inter-rater reliability impact?

2. Proficiency should be considered over NS
   ▫ Level 2+ raters should be excluded

3. L1 has an impact on rating
   ▫ But not compared to English raters
   ▫ Language distance matters

4. Ratings of specific features show more group differences
   ▫ Rater proficiency and L1 groups
   ▫ Differences never occur in “structures”
The native speaker

• (Re)defined
  ▫ Need for clear definition
  ▫ Native speaker assumptions
  ▫ Native speaker is a social construct, not a measurement construct
    • It is associated with acquisition method, culture, identity, confidence
  ▫ Call the ideal speaker something else, specify what it is

• Justification
  ▫ Appropriateness for use: is it fair?
    • Decisions: standard variety, correctness
    • Consequences of misuse go beyond test itself
Qualifying speaking raters by speaking proficiency ability

• It takes one to know one?
  ▫ Much of impact seen in Level 3 tests
    • Level 2 is below rater’s proficiency level
    • Level 4-5 is limited by ceiling effect

• Competence vs. performance

• Training: the great equalizer
Holistic versus analytic rating

- Trend:
  - No differences are found in overall ratings
  - Differences found in linguistic features, except Structures

- What construct are raters using to rate?
- Does construct matter if final ratings are not significantly different?
Limitations and future research

- Current study deals only with rating, not test administration
- Replicate with NS raters < ILR 5
- Replicate in a language other than English
- Analyze the rater comments
- Further investigate rater competencies: linguistic, cognitive, cultural, and evaluative competencies
Thank you

Rachel Lunde Brooks
rachel.brooks@ic.fbi.gov