How do examiners reach judgments?

Influences on rater judgments of second language speech

Talia Isaacs, University of Bristol, UK

Current Issues in Assessment Seminar

September 6, 2012
What’s in a score?

- L2 speech properties

Ideally, a score would perfectly reflect the qualities of a test-taker’s performance → extrapolated to underlying ability.
Hypothesized influences on ratings of L2 speech

- L2 speech properties
- Rater characteristics
- Speaking task
- Rating scale properties
- Residual
Rater characteristics

- Rater cognitive variables

Score
Background

- Economic globalisation & advancements in technology – brought people together
  - Greater student mobility, English lingua franca

- Postsecondary institutions seek to attract a diverse student body
  - Competition for human capital → global knowledge economy (Chiswick & Miller, 2007)
    - Attract top talent; counter funding shortfalls
Background

- Postsecondary institutions – responsible for providing valid language assessments

  • Most speaking components of tests used in academic setting scored by human raters

  • Rater judgments contribute to high-stakes decision-making → (e.g., admissions, granting Tier 4 visa)
Background

However, rater judgments do not always provide valid measures of speaking ability.

• Rater judgments might not reflect simply speakers’ performance but also individual differences among raters themselves
Previous research → Identified various sources of variability

<table>
<thead>
<tr>
<th>Research area</th>
<th>Second language (L2) assessment</th>
<th>Psycholinguistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Research focus</td>
<td><em>Rater</em> background characteristics</td>
<td>L2 learner <em>cognitive</em> variables</td>
</tr>
<tr>
<td>Variables examined (examples)</td>
<td>Gender</td>
<td>Musical ability</td>
</tr>
<tr>
<td></td>
<td>First language</td>
<td>Short-term memory</td>
</tr>
<tr>
<td></td>
<td>Teaching experience</td>
<td>Attention control</td>
</tr>
</tbody>
</table>

Research gap: Individual differences in rater cognitive abilities & effects on scoring
Aim of the study

- Relationship → individual differences in rater cognitive abilities & rater judgments of L2 speech

**Rater cognitive variables:**
- Musical ability
- Phonological memory
- Attention control

**Rated L2 speech measures:**
- Accentedness
- Comprehensibility
- Fluency
1. Phonological memory

Phonological memory span – an individual’s capacity to retain spoken sequences temporarily in the short-term memory system (Gathercole et al., 2001)
1. Phonological memory

Phonological memory is involved in:

- **Speech perception** (Jacquemot et al., 2006)

- **Perceptual learning of words** (Hervais-Adelman et al., 2008)

- **Subjective ratings of speech** (Gould et al., 2002)
Raters with larger phonological memory store → retain more speech

- **more severe** – overly sensitive to deviations?
- **more lenient** – listen to speech holistically?
2. Attention control

Attention-switching capacity – ability to both maintain focus on a single task & alternate attention between two simultaneous tasks (Arbuthnott & Frank, 2000)
2. Attention control

Attention control is related to:

- enhanced processing of relevant linguistic information & inhibited processing of irrelevant information (Eviatar 1998)

- allocating attention to speech signal
  - simultaneously processing form & meaning
Raters who allocate attention more efficiently could be:

• overly sensitive to additional shift costs imposed by L2 speech \(\rightarrow\) **more severe ratings**?

• shift their attention effortlessly among different dimensions of speech \(\rightarrow\) **more lenient ratings**?
3. Musical ability

Musical ability – ability to hear (internalize) music no longer present in the physical environment (Gordon, 1995)
Musicians’ extensive pitch processing experience positively transfers to:

• First language speech perception (Alexander et al., 2004)

• The perception & production of L2 speech (Slevc & Miyake, 2006)

• However, other studies have identified no relationship (Nakata, 2002)
3. Musical ability

Hypothesised: Raters with greater musical ability would judge L2 speech less favorably than less musical raters

Musical raters

• would be more sensitive to certain aspects of L2 speech (e.g., pitch fluctuations, voice quality)
  → downgrade their ratings relative to nonmusicians
Aim of the study

- Relationship $\rightarrow$ individual differences in rater cognitive abilities & raters’ judgments of second language (L2) speech

**Rater cognitive variables:**
- Musical ability
- Phonological memory
- Attention control

**Rated L2 speech measures:**
- Accentedness
- Comprehensibility
- Fluency
1. Accentedness

Accentedness – listeners’ judgments of how closely the pronunciation of an utterance approaches that of a native speaker (Munro & Derwing, 1999)
2. Comprehensibility

Listeners’ perceptions of how easily they understand L2 speech (Munro & Derwing, 1999)
3. Fluency

Fluency – listeners’ judgments of how smoothly & rapidly an utterance is spoken (Derwing et al., 2004):

- without undue pauses
- hesitations
- or dysfluencies
Speakers

40 adult French speakers from Quebec, Canada

Age of exposure to English: 8.7 years (0-17)

English (L2) use: 20% (0-70%)

French (L1) use: 80% (30-100%)

English proficiency (1-9): 6.1 (1-9)
Speaking prompt

8-frame picture narrative often used with adult learners from different proficiency levels in L2 pronunciation research (Derwing et al., 2004)
Raters

60 undergraduate student native English speakers (31 American, 29 Canadian)

30 music majors
- mean study of primary instrument = 9.5 years
- 80% formally trained in another instrument

30 nonmusic majors
- mean music study = 3.4 years
- 8 had no musical training
Raters

Music & nonmusic majors matched for language use background variables

French (L2) use: 8% ($sd = 9.4$)
English (L1) use: 92% ($sd = 10$)
French proficiency (1-9): 3.4 ($sd = 2.1$)
Raters scored 20-s speech samples (randomised)

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Heavily accented</strong></td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Very hard to understand</strong></td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Very dysfluent</strong></td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Very Fluent</strong></td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Task: Phonological memory

Participant hears:

- barch teeg nup guwkw
- barch teeg nup guwkw

Same order? YES

Serial nonword recognition task (Gathercole et al., 2001)
Task: Phonological memory

Participant hears:

- mot chan ped kig
- mot ped chan kig

Same order? NO

Phonological memory = number of sequences whose order was recognized correctly
Task: Attention control

Trail Making Test (US Army Individual Test Battery, 1944)

Attention control = Time B – Time A
Task: Musical ability

Three subtests from the Musical Aptitude Profile (Gordon, 1995)

- Melody
- Tempo
- Phrasing
Task 1: Musical ability

Melody

Participant hears:

Same basic song? YES
Task 2: Musical ability

Tempo

Participant hears:

Same tempo?

NO
Task 3: Musical ability

Phrasing

Participant hears:

Which one sounds better?

SECOND
Results
## Preliminary analysis 1

Musical Aptitude Profile subtests distinguished between music & nonmusic majors

<table>
<thead>
<tr>
<th>Musical ability</th>
<th>( t )-test results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Melody</td>
<td>( t(58) = 5.67, , p &lt; .00001 )</td>
</tr>
<tr>
<td>Tempo</td>
<td>( t(58) = 3.79, , p &lt; .00001 )</td>
</tr>
<tr>
<td>Phrasing</td>
<td>( t(58) = 2.75, , p &lt; .01 )</td>
</tr>
</tbody>
</table>

Pooled subtests (max 120)

- **Music**:  \( M=99.9, \, sd=4.5 \)
- **Nonmusic**:  \( M=88.1, \, sd=11.6 \)
Preliminary analysis 1

Phonological memory & attention control measures did not distinguish between music & nonmusic majors

Music & nonmusic majors differed solely in their musical ability
Preliminary analysis 2

Intraclass correlations computed separately for music & nonmusic major groups

- Coefficients of .98-.99 on the 3 rated speech measures

Raters from both groups overall internally consistent in their ratings
Phonological memory

Do speech ratings depend on phonological memory?

• Music & nonmusic majors’ ratings pooled
  • divided into high vs. low phonological memory groups (median split)

• analyses based on raters’ mean scores for each speaker
Phonological memory

No difference between the two groups for any measure.
Attention control

Do speech ratings depend on attention control?

• Divided raters into more vs. less efficient attention control groups (median split)
Attention control

No difference between the two groups for any measure.
Musical ability

Do speech ratings depend on musical ability?

• Music majors vs. nonmusic majors
Musical ability

**Accent**

- Music majors
- Non-music majors

**Comprehens.**

**Fluency**

**Rating scale**

*Accentedness rating: \( t \ (58) = 2.37, p = .02 \)
Reanalysed the data $\rightarrow$ Cross-classified multilevel models

MLWin, MCMC estimation (Browne, 2012; Rasbach et al., 2009)

Level 2: Speakers ($n = 40$); Raters ($n = 60$)
Level 1: Unique ratings ($n = 2400$)

***Identical results

$$y_{ij} \text{ (Accentedness)} = \beta_{0ijk} \text{ (Intercept)} + \beta_1 \text{ (Nonmusic major)} + \nu_{ok} \text{ (Speaker)} + u_{ojk} \text{ (Rater)} + e_{ij} \text{ (Residual)}$$

Accent – rating outcome
University major (rater attribute) – predictor
Reanalysed the data → Cross-classified multilevel models

**Fixed part** (Estimate, $SE$)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimate</th>
<th>$SE$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>4.7</td>
<td>.27</td>
</tr>
<tr>
<td>Nonmusic</td>
<td>.58</td>
<td>.27</td>
</tr>
</tbody>
</table>

Nonmusic majors’ accent ratings .58 higher than music majors’ ratings, 9-point scale

**Random part**

<table>
<thead>
<tr>
<th>Source of Variability</th>
<th>Attribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Differences between speakers</td>
<td>50%</td>
</tr>
<tr>
<td>Differences between raters</td>
<td>22%</td>
</tr>
</tbody>
</table>
Musical experience

<table>
<thead>
<tr>
<th>Measure</th>
<th>Music majors</th>
<th>Non-music majors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accent</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comprehens.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fluency</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*significant difference
Musical experience

Probe significant effect for accent further

• Grouped speakers into high, medium & low L2 speaking ability based on a combined measure of

  • accentedness ratings from an independent group of raters who had judged the same speech samples (Trofimovich et al., 2007)

  • speaking rate (syll/sec)
Musical experience

- Low ability: $t(58) = 2.61, p = .01$
- Medium ability: $t(58) = 2.25, p = .03$
Correlations among rated measures

<table>
<thead>
<tr>
<th>Measures</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Music Majors</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Accentedness</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Comprehensibility</td>
<td><strong>.47</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Fluency</td>
<td><strong>.58</strong></td>
<td><strong>.81</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Nonmusic Majors</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Accentedness</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Comprehensibility</td>
<td><strong>.78</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Fluency</td>
<td><strong>.83</strong></td>
<td><strong>.80</strong></td>
<td></td>
</tr>
</tbody>
</table>

**p < .001 (two tailed).
Implications for construct operationalisation in rating scales

Comprehensibility & accentedness often conflated in rating scale descriptors

- “Pronunciation is easily understood; Many features… are ‘nativelike’ ”

Cambridge ESOL Common Scale for Speaking
University of Cambridge ESOL Examinations (2008)

Goal: Describe comprehensible speech without resorting to native speaker standard
Musicians could help tease these dimensions apart
The present study

- Relationship → individual differences in rater cognitive abilities & raters’ judgments of second language (L2) speech

Rater cognitive variables:
- Musical ability
- Phonological memory
- Attention control

Rated L2 speech measures:
- Accentedness
- Comprehensibility
- Fluency
Conclusions

No relationship between ratings of L2 speech and two cognitive variables:

- phonological memory
- attention control

Listeners’ ratings of L2 speech do not appear to be influenced by individual differences in listeners’ phonological memory and attention control.
Conclusions

A relationship between accentedness ratings and musical experience:

- University-trained musicians rated accentedness more severely than nonmusic raters
- especially for L2 speakers of low “pronunciation” ability

Accentedness ratings are susceptible to effects of individual differences in listeners’ musical ability.
Implications

Taken together, the findings are reassuring

• Individual differences in phonological memory & attention control do not seem to threaten the validity of speaking assessments

• Why was there no effect?
Implications

- Musicians assigned lower mean scores solely for accent – intriguing from a research perspective
  - E.g., Which aspects of speech are musicians more sensitive to?

- However, implications for assessment limited
  - No indication, based on this study alone that raters
    - should be screened for musical ability
    - musically homogenous raters should be sought
Small effect size ($r = .3$) $\rightarrow$ differences in accent perception might not translate into differences in overall proficiency scoring

- Most applied linguists do NOT regard accent reduction as an appropriate goal for L2 communicative teaching or assessment (Levis, 2006)

- Accent not a criterion in IELTS or TOEFL
  - comprehensibility & fluency $\rightarrow$ nonsignificant
Future research

- Urgent need to examine the effects of musical ability/experience on intelligibility
  - Listeners’ actual understanding of L2 speech
    (measured by correctly transcribed words; Munro & Derwing, 1999)
Acknowledgments

Co-investigator: Pavel Trofimovich

- Tracey Derwing
- Harvey Goldstein
- Randall Halter
- Sarita Kennedy
- George Leckie
- Murray Munro
- Hyojin Song
- Sally Thomas
- Ron Thomson
- Carolyn Turner

talia.isaacs@bristol.ac.uk
References


References


References


References


References