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

University of BRISTOL

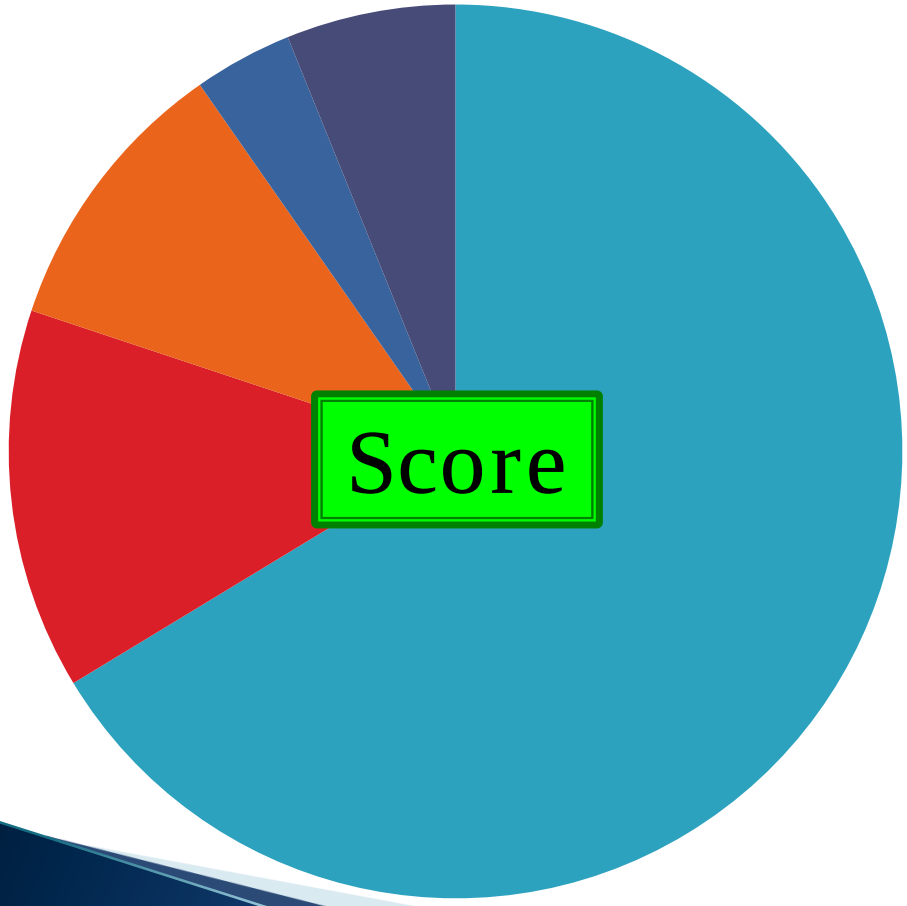
CÆR
Centre for Assessment & Evaluation Research

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

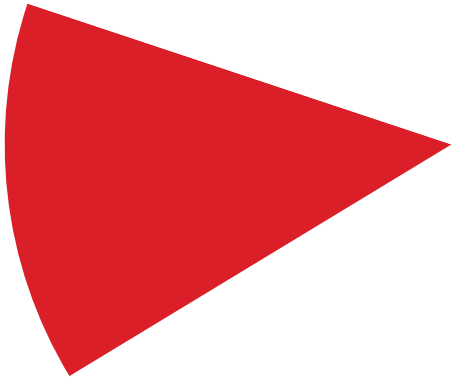
Hypothesized influences on ratings of L2 speech



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

Isaacs & Trofimovich (2011) →

Applied Psycholinguistics 32(1)



- 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

Research area	**Second language (L2) assessment**	**Psycholinguistics**
Research focus	*Rater* background characteristics	L2 learner *cognitive* variables
Variables examined (examples)	Gender First language Teaching experience	Musical ability Short-term memory Attention control

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)

1. Phonological memory

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

2. Attention control

Raters who allocate attention more efficiently could be:

- overly sensitive to additional shift costs imposed by L2 speech → **more severe ratings?**
- shift their attention effortlessly among different dimensions of speech → **more lenient ratings?**

15

3. Musical ability

Musical ability – ability to hear (internalize) music no longer present in the physical environment (Gordon, 1995)

3. Musical ability

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 nonmusicans

18

Aim of the 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

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

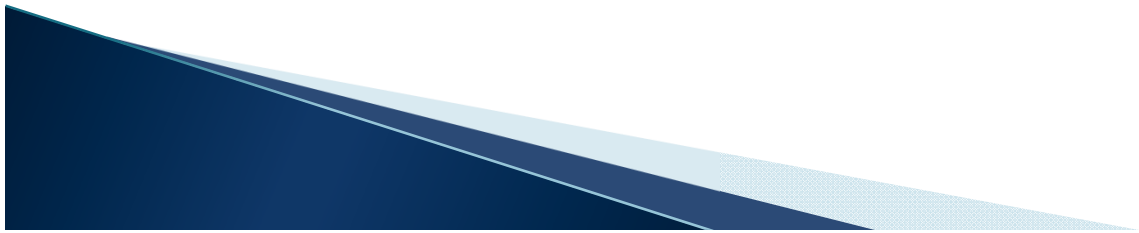
30 nonmusic majors

- 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

Raters scored 20-s speech samples (randomised)

1 2 3 4 5 6 7 8 9

Heavily accented **Not accented at all**

1 2 3 4 5 6 7 8 9

Very hard to understand **Very easy to understand**

1 2 3 4 5 6 7 8 9

Very dysfluent **Very Fluent**

Task: Phonological memory

Participant hears:

barch teeg nup guwk

barch teeg nup guwk

Same order? **YES**

Serial nonword recognition task (Gathercole et al., 2001)

Task: Phonological memory

Participant hears:

 mot **chan ped** kig

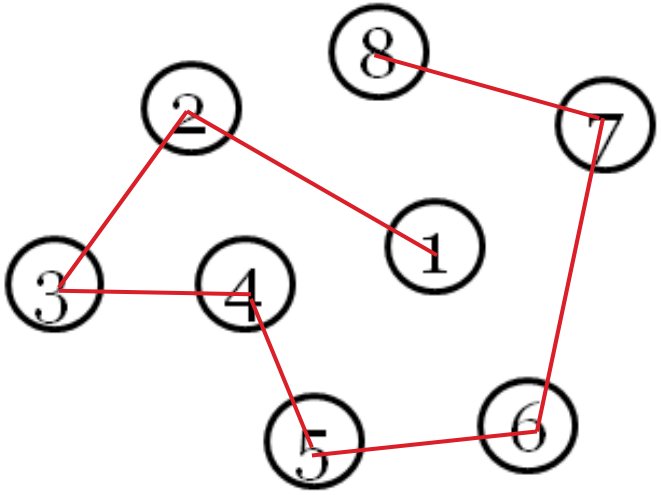
 mot **ped chan** kig

Same order? **NO**

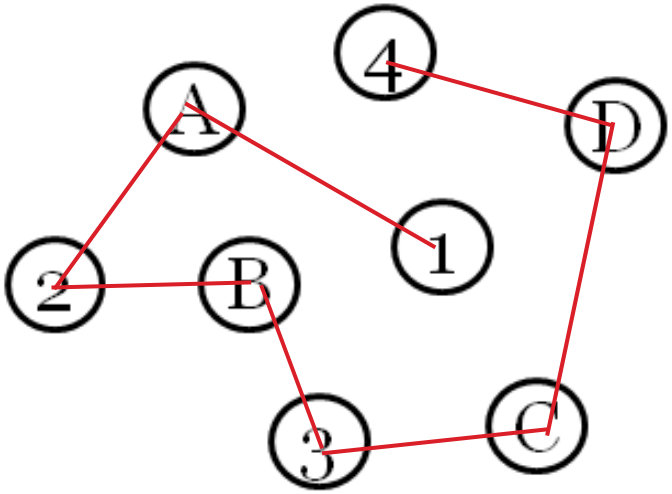
Phonological memory = number of sequences whose order was recognized correctly

Task: Attention control

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



0 Time A



0 Time B

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

Musical ability	*t*-test results
Melody	$t(58) = 5.67$, $\mathbf{p < .00001}$
Tempo	$t(58) = 3.79$, $\mathbf{p < .00001}$
Phrasing	$t(58) = 2.75$, $\mathbf{p < .01}$

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 rated speech measures

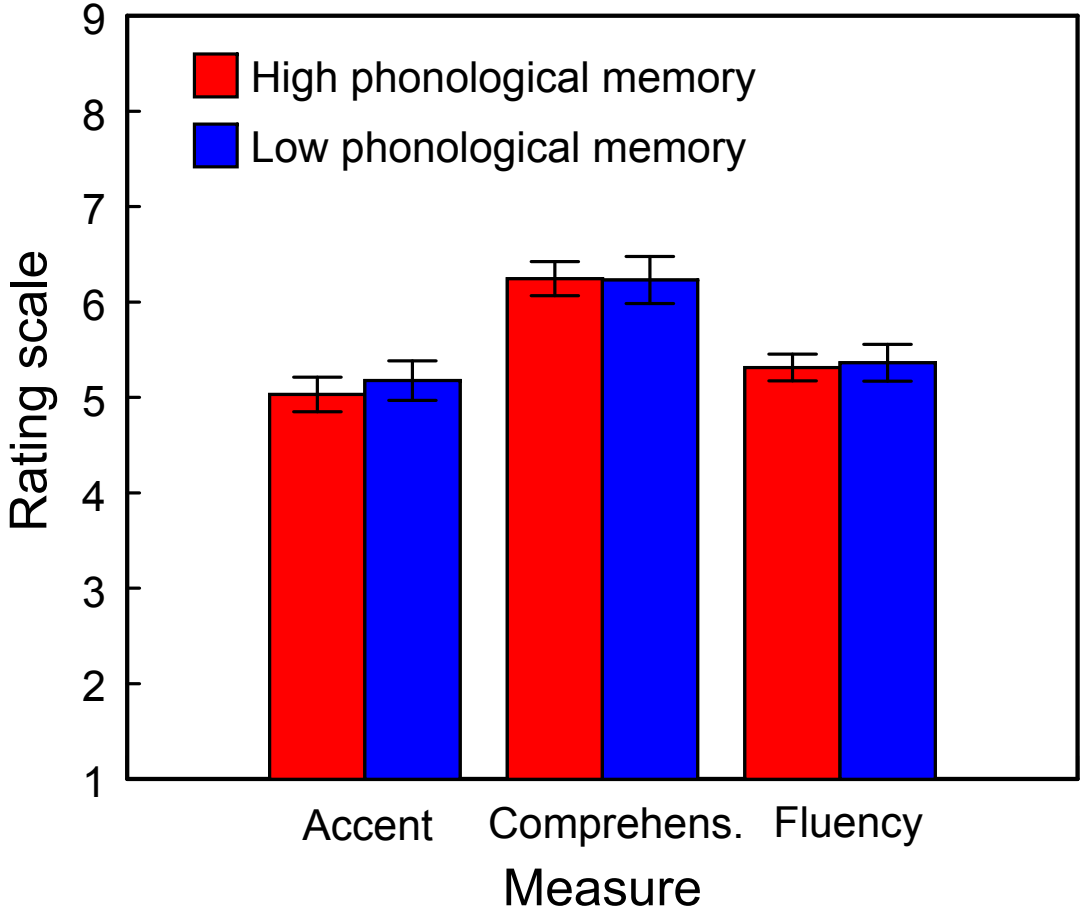
Raters from both groups overall internally consistent in their ratings

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.

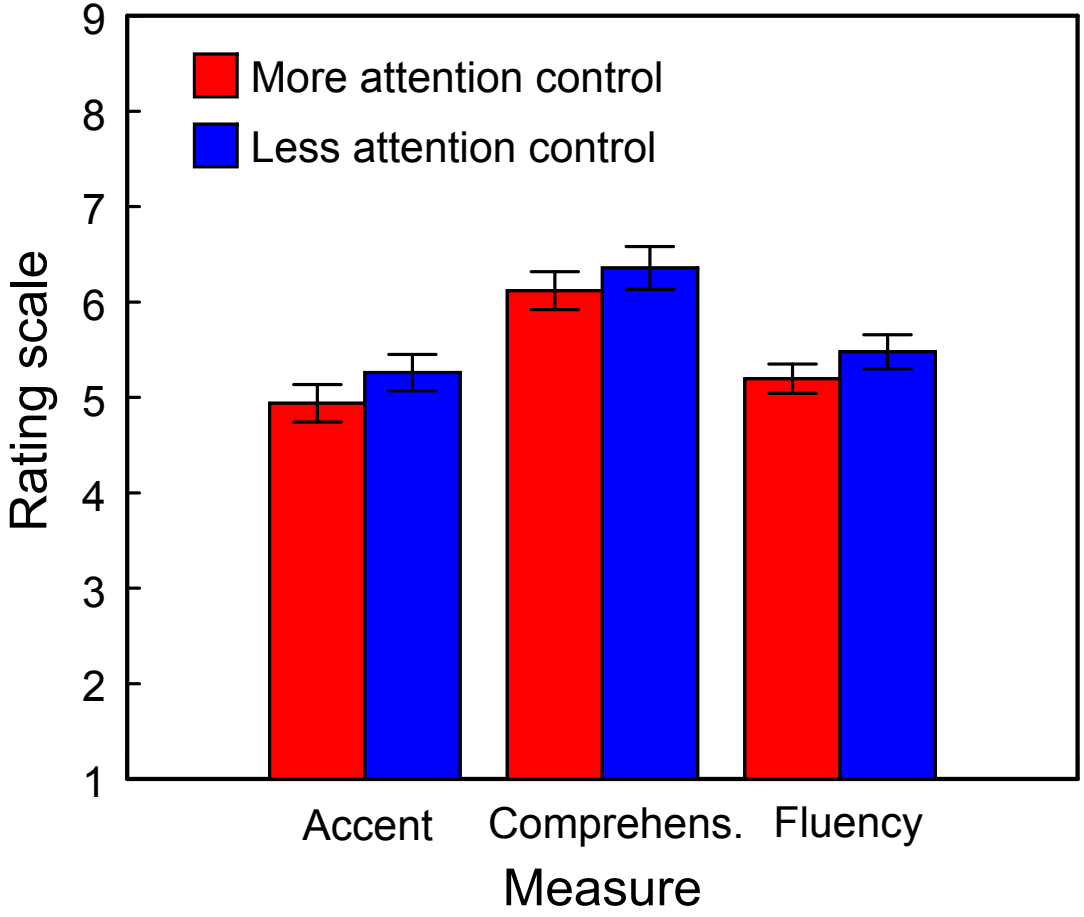
Attention control

Do speech ratings depend on attention control?

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

41

Attention control



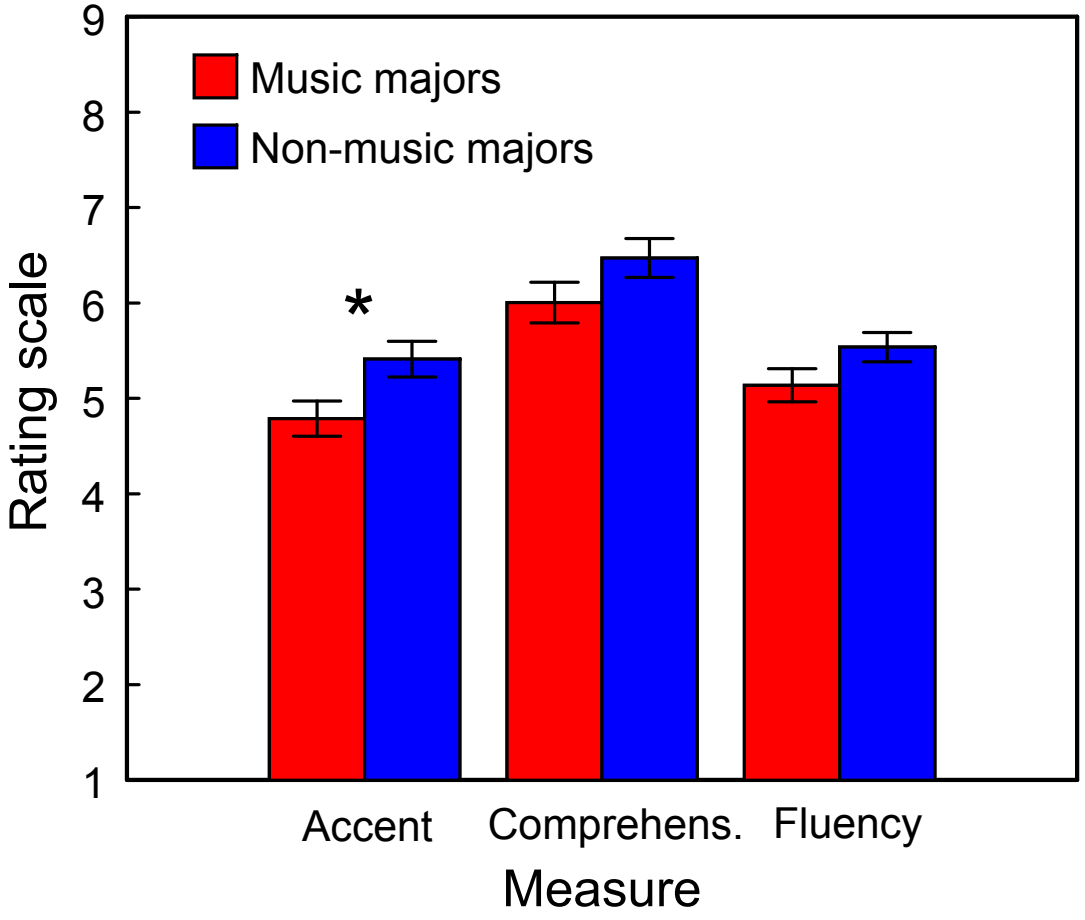
No difference between the two groups for any measure.

Musical ability

Do speech ratings depend on musical ability?

- Music majors vs. nonmusic majors

Musical ability



Accentedness rating: t (58) = 2.37, p = .02

Reanalysed the data → 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} (Accentedness) = β_{0ijk} (Intercept) + β_1 (Nonmusic major) + v_{ok} (Speaker) + u_{ojk} (Rater) + e_{ij} (Residual)

Accent – rating outcome

University major (rater attribute) – predictor

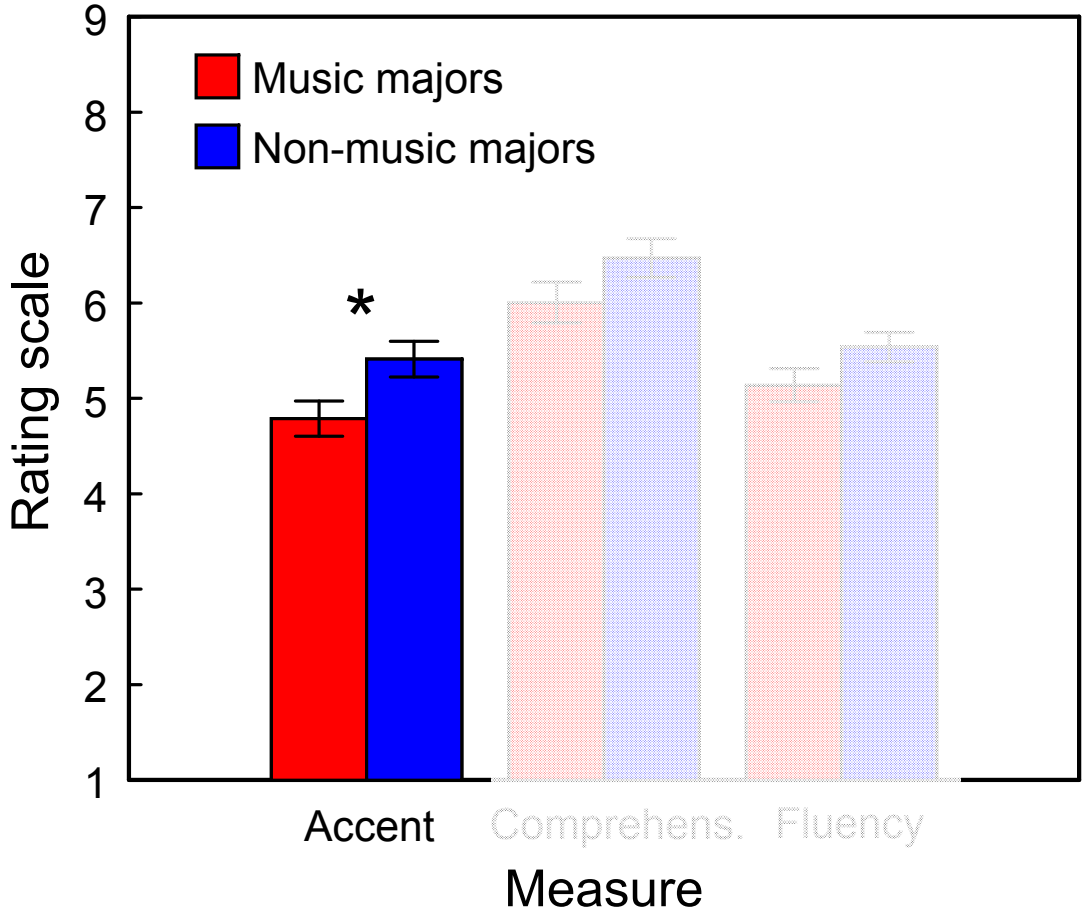
Reanalysed the data → Cross-classified multilevel models

Fixed part (Estimate, *SE*)	
Intercept	4.7 (*.27*)
Nonmusic	.58 (*.27*)

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

Random part	
Variability attributable to differences between speakers	50%
Variability attributable to differences between raters	22%

Musical experience

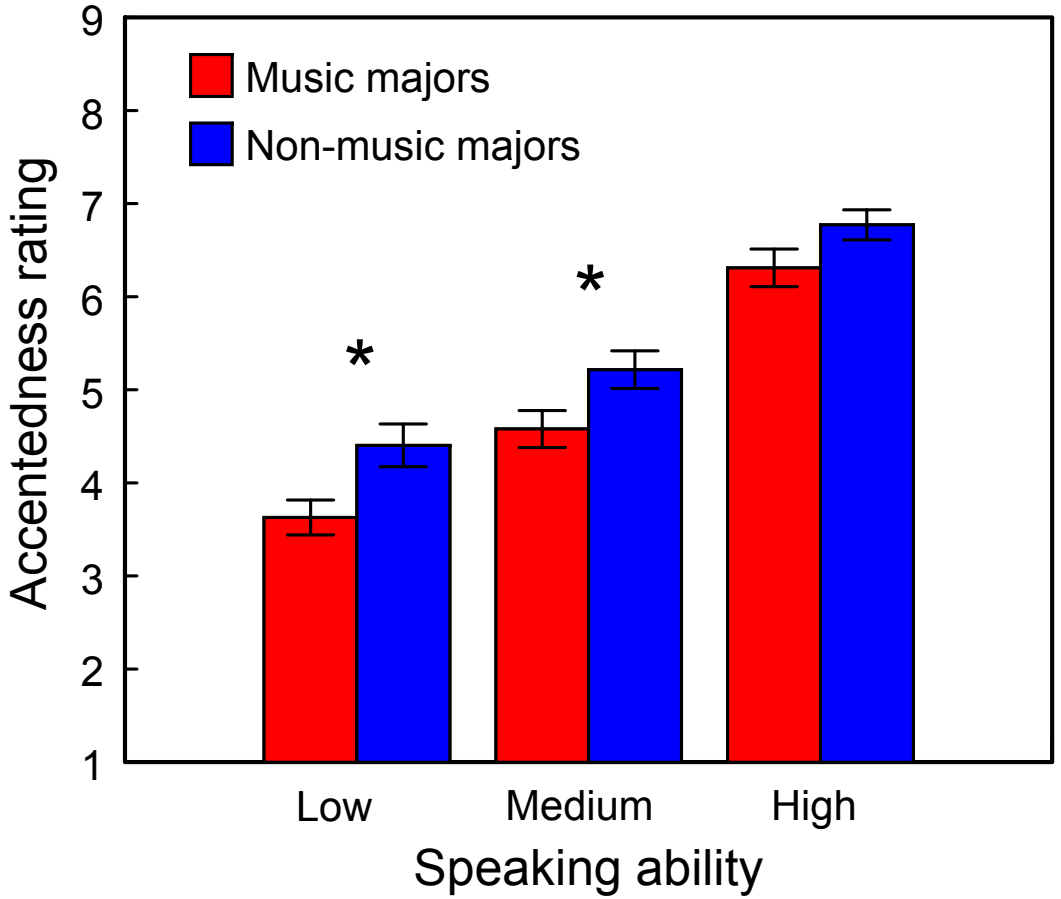


Musical experience

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

Measures	1	2	3
Music Majors			
1. Accentedness			
2. Comprehensibility	.47**		
3. Fluency	.58**	.81**	
Nonmusic Majors			
1. Accentedness			
2. Comprehensibility	.78**		
3. Fluency	.83**	.80**	

**$p < .001$ (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

Implications

- Small effect size (r = .3) → 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 → 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

Social Sciences and Humanities Research Council of Canada

Fonds de recherche sur la société et la culture Québec

References

- Alexander, J. A., Wong, P. C. M., & Bradlow, A. R. (2005). Lexical tone perception in musicians and non-musicians. *Proceedings of Interspeech 2005, Eurospeech, 9th European Conference on Speech Communication and Technology*. Lisbon, Portugal.
- Arbuthnott, K., & Frank, J. (2000). Trail Making Test, Part B as a measure of executive control: Validation using a set-switching paradigm. *Journal of Clinical and Experimental Neuropsychology*, *22*, 518–528.
- Browne, W. J. (2012). *MCMC estimation in MLwiN version 2.25*. Centre for Multilevel Modelling, University of Bristol.
- Chiswick, B., & Miller, P. W. (2007). *The economics of language: International analyses*. London: Routledge.

References

- Derwing, T. M., Rossiter, M. J., Munro, M. J., & Thomson, R. I. (2004). Second language fluency: Judgments on different tasks. *Language Learning, 54*, 665–679.
- Eviatar, Z. (1998). Attention as a psychological entity and its effects on language and communication. In B. Stemmer & H. A. Whitaker (Eds.), *Handbook of neurolinguistics* (pp. 275–287). New York: Academic Press.
- Gathercole, S. E., Pickering, S. J., Hall, M., & Peaker, S. M. (2001). Dissociable lexical and phonological influences on serial recognition and serial recall. *Quarterly Journal of Experimental Psychology*, *54A*, 1–30.
- Gordon, E. E. (1995). *Manual: Musical Aptitude Profile*. Chicago: GIA Publications.

References

- Gould, O. N., Saum, C., & Belter, J. (2002). Recall and subjective reactions to speaking styles: Does age matter? *Experimental Aging Research*, *28*, 199–213.
- Hervais-Adelman, A., Davis, M. H., Johnsrude, I. S., & Carlyon, R. P. (2008). Perceptual learning of noise vocoded words: Effects of feedback and lexicality. *Journal of Experimental Psychology: Human Perception and Performance*, *34*, 460–474.
- Isaacs, T., & Trofimovich, P. (2011). Phonological memory, attention control, and musical ability: Effects of individual differences on rater judgments of second language speech. *Applied Psycholinguistics*, *32*, 113–140.
- Jacquemot, C., Dupoux, E., Decouche, O., & Bachoud-L´evi, A.-C. (2006).Misperception in sentences but not in words: Speech perception and the phonological buffer. *Cognitive Neuropsychology*,*23*, 949–971.

References

- Levis, J. M. (2006). Pronunciation and the assessment of spoken language. In R. Hughes (Ed.), *Spoken English, TESOL and applied linguistics: Challenges for theory and practice* (pp. 245–270). New York: Palgrave Macmillan.
- Munro, M. J., & Derwing, T. M. (1999). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning, 49*, 285–310.
- Nakata, H. (2002). Correlations between musical and Japanese phonetic aptitudes by native speakers of English. *Reading Working Papers in Linguistics, 6*, 1–23.
- Rasbash, J., Steele, F., Browne, W. J., & Goldstein, H. (2009). *A user's guide to MLwiN, version 2.10*. Centre for Multilevel Modelling, University of Bristol.

References

- Slevc, L. R., & Miyake, A. (2006). Individual differences in second-language proficiency: Does musical ability matter? *Psychological Science, 17*, 675–681.
- Trofimovich, P., Gatbonton, E., & Segalowitz, N. (2007). A dynamic look at L2 phonological learning: Seeking processing explanations for implicational phenomena. *Studies in Second Language Acquisition, 29*, 407–448.
- University of Cambridge ESOL Examinations. (2008). *Certificate of Proficiency in English: Handbook for teachers*. Cambridge: UCLES.
- US Army Individual Test Battery. (1944). *Manual of directions and scoring*. Washington, DC: Cambridge University Press.