Volume 1·1 2019

Journal of Monolingual and Bilingual Speech

ISSN 2631-8407 (print)
ISSN 2631-8415 (online)

Exploring the stability of second language speech ratings through task practice in bilinguals' two languages

Kym Taylor Reid, Mary Grantham O'Brien, Pavel Trofimovich, and Aki Tsunemoto

Taylor Reid, K., O'Brien, M. G., Trofimovich, P., & Tsunemoto, A. (2020). Exploring the stability of second language speech ratings through task practice in bilinguals' two languages. *Journal of Monolingual and Bilingual Speech*, *2*, 315–329.
https://doi.org/10.1558/jmbs.17714

Introduction

When it comes to the assessment of second language (L2) speech, human ratings are deemed essential (Derwing & Munro, 2015). Teachers regularly evaluate L2 speakers in low-stakes assessments such as oral presentations and tests, and trained assessors evaluate L2 speakers in higher-stakes contexts such as standardized examinations. Naïve raters often rate speech for various dimensions, including accentedness and comprehensibility (Derwing & Munro, 2009). It is also common for untrained raters to provide evaluations of L2 speakers that extend beyond speech itself. For example, individuals with no training in speech assessment have been asked to judge L2 speakers' socioeconomic status (Deprez-Sims & Morris, 2010), educational achievement (Campbell-Kibler, 2007) and competence (Baquiran & Nicoladis, 2020). Such evaluations often have implications for future work and study opportunities, wages, and the quality of healthcare that a person receives (Halim et al., 2017; Timming, 2017).

Previous research investigating L2 speech ratings has demonstrated high levels of inter-rater reliability, indicating that raters agree with one another when they rate L2 speech along a variety of continua (Isaacs & Thomson, 2013). However, other research has called the stability of speech ratings into question. For example, visual information may affect speech ratings, such that raters' evaluations may be distorted by attributions of a speaker's group membership (Kang & Rubin, 2009). That is, raters sometimes impose accents on speech samples that are presented with images of speakers who look as though they might speak with a foreign accent. Along similar lines, hearing a positive or negative anecdote about an interaction with an L2 speaker may also affect speech ratings (Taylor Reid et al., 2019). Compared to raters who did not hear an anecdote about L2 speakers, positively-biased raters tended to rate speakers more favourably while negatively-biased raters (especially older individuals with likely more entrenched social attitudes) evaluated the same speakers more negatively.

Rater evaluations of speech can also be shifted through various interventions. For instance, in Staples et al. (2014), raters (native-speaking university students) completed several informal contact activities requiring them to cooperate with L2 speakers over eight weeks. The students who participated in the activities rated their L2 university instructors as being less accented, more comprehensible, and more effective teachers, compared to those who did not participate. In another study (Kang, 2008), English-speaking raters evaluated the speech of 11 international teaching assistants (ITAs) before and after a social-psychological intervention, which consisted of informal contact between small groups of raters and multiple ITAs. Raters were significantly more lenient in their speech assessments after spending time with the ITAs, suggesting that contact reduces bias by forging rater–speaker connections.

Increased rating reliability can also be achieved through rater training, although research on how rater training impacts the consistency and stability of speech ratings remains scarce. In a rare example focusing on L2 speaking, Davis (2016) asked 20 teachers of English to score TOEFL iBT speaking tests for the first time, showing that rater training – operationalized as familiarization with the scoring rubric and a review of exemplars with written score justification – resulted in increased inter-rater reliability and closer alignment with established reference scores. Raters seemed to adjust their ratings once familiar with scoring guidelines and the task model.

In addition to social contact and rater training, perspective taking has emerged as a way to stabilize ratings. Here, raters are asked to walk in the shoes of others and discover commonalities that they share with them, thereby mitigating prejudicial behaviour (Boland & Tenkasi, 1995). Weyant (2007) operationalized perspective taking as writing about a day in the life of a given person, after listening to a speech sample produced by either a native speaker or an L2 speaker. The participants who took the perspective of the L2 speaker assigned higher speech ratings to the speaker, compared to those who did not engage in perspective taking. Activating one's L2 speaking skills has also been linked to rater leniency. German raters who were called upon to use their L2 English before rating tended to assign higher ratings to L2 German speakers, compared to raters who did not use their L2 before the rating task (Hansen et al., 2014).

The present study

If speech ratings can be affected by social contact, rater training and perspective taking, then it is important to understand how other manipulations might impact rating stability. One such manipulation involves asking raters to complete the same speaking task as the speakers to be assessed. This strategy, whose goal is to familiarize raters with the speaking task while engaging them in perspective taking, is seldom found in assessment research, let alone studies of L2 speech. In a rare example, Zhang (2017) put pre-service music teachers in the shoes of their L2-speaking students by immersing the teachers in a 20-minute music class with all instructions and content in Mandarin. Verbal instruction was accompanied by aural, visual and kinaesthetic activities, which allowed the teachers to follow instruction. The teachers subsequently reported feeling anxious, confused and frustrated during the class, revealing an emotional response to perspective taking. With a similar goal in mind, Stewart (2010) embarked on action research in which she, a unilingual English speaker, enrolled in a Spanish course to see how the experience might influence her instruction of L2 English learners. She identified patterns and struggles similar to those she saw in her own students, which allowed her to develop new understanding of L2 speakers' challenges.

Though contextually different from the present study, these examples, combined with broader perspective-taking research and knowledge of the malleability of rater behaviour, motivate further investigations into the stability of L2 speech ratings. Therefore, this study's goal was to examine L2 speech ratings as a function of bilingual raters' performing the speaking task (i.e. the same task completed by those rated) in their stronger or weaker language. We asked 30 English–French bilinguals to evaluate French-accented L2 English speech for three dimensions. We focused on segmental errors (accuracy in articulation of consonants and vowels), intonation (natural rise and fall in pitch) and flow (overall pacing and speed of utterance delivery) because these are included in many speaking scale descriptors used by teachers in instructed settings (Saito et al., 2017) and trained raters in higher-stakes assessment contexts (e.g. IELTS, ACTFL).

Before providing the ratings, some raters performed the speaking task in their more dominant English or their less dominant French, while baseline raters performed no task. Completing the task (in either language) was expected to increase raters' task familiarity, leading to a potential change in their evaluations, compared to the assessments of baseline raters. However, completing the speaking task in French was expected to encourage raters to discover commonalities between them and L2 speakers, resulting in a potential difference in the extent to which performing the French and English task might shift raters' behaviour, relative to baseline raters' performance. The speaking task manipulation, as defined and implemented here, is similar to other experimental manipulations such as affective priming (e.g. Luoma-aho et al., 2019) and stereotype threat (e.g. Steele & Aronson, 1995) which expose people to visual or linguistic information that can potentially bias their responses or otherwise influence their performance. In our case, however, the speaking task manipulation was designed with a potential positive bias in mind. Due to lack of systematic prior work, we had no expectation as to differences in the impact of task performance on raters' assessments of the three target dimensions.

Method

Raters

Raters included 30 English–French bilingual residents of Montreal (22 females), all self-identified members of Quebec's anglophone community. They grew up in households with at least one native English-speaking parent and were exposed to a combination of English- and French-medium instruction through schooling. All raters reported English as their more dominant language. Using a 9-point scale (1 = 'beginner', 9 = 'nativelike'), they also self-assessed their English proficiency higher (*M* = 8.98, *SD* = 0.14) than their French proficiency (*M* = 6.23, *SD* = 1.29), *t*(29) = 11.15, *p* < 0.001. Of the 30 raters, 10 were tested by Taylor Reid (in review), providing baseline data. The remaining raters completed an identical testing sequence, except that they performed the speaking task in their stronger (English) or weaker (French) language (10 per group). The groups did not differ significantly in any background variables (see Table 1), including social attitudes (see Appendix A) towards raters' own ethnolinguistic group (5 questions, α = 0.93), their perception of the role of English in Quebec (5 questions, α = 0.68), their attitudes towards immigrants (5 questions, α = 0.77) and their views of other ethnolinguistic groups (5 questions, α = 0.69), as shown through between-group comparisons, *F*(2, 27) < 3.02, *p* > 0.07. This implied that the groups were comparable in several background characteristics, which minimized the likelihood that potential between-group differences in speech ratings would be attributable to these characteristics.

Table 1. Raters' background characteristics: means, standard deviations.

		Speaking task language	
Background variable	**Baseline**	**English**	**French**
Age (years)	25.40 (3.44)	22.70 (2.91)	23.40 (4.01)
Gender (f–m)	9–1	7–3	6–4
English use (%)	84.50 (18.02)	86.00 (13.50)	86.00 (10.75)
French use (%)	23.50 (12.92)	27.00 (14.94)	23.00 (17.67)
English proficiency (1–9 scale)[a]	8.93 (0.24)	9.00 (0.00)	9.00 (0.00)
French proficiency (1–9 scale)[a]	6.48 (1.23)	6.20 (1.30)	6.03 (1.43)
Pride in Anglophone group[b]	6.04 (2.89)	8.10 (1.08)	7.58 (1.27)
Role of English in Quebec[b]	5.64 (1.33)	6.52 (1.13)	5.38 (0.99)
Attitudes towards immigrants[b]	2.11 (0.74)	3.64 (1.70)	3.26 (1.69)
Feelings towards other groups[b]	6.78 (1.66)	7.42 (0.86)	7.04 (0.66)

Note. [a]1 = 'beginner', 9 = 'nativelike'. [b]Mean of five questions

Materials and procedure

Raters evaluated 40 English narratives by native speakers of Quebec French (27 females, age 18–61), all raised in French-speaking families and educated in French. The narratives, drawn from a speech corpus (Isaacs & Trofimovich, 2011), had been elicited through an 8-frame picture story describing two passers-by colliding on a street corner and accidentally exchanging identical-looking suitcases (Derwing et al., 2004). Consistent with prior work (Derwing & Munro, 2015), the initial 30 seconds of each narrative was included for rating, excluding initial false starts and hesitations. The recordings were evaluated for segmental errors, intonation and flow (see Appendix B for a screenshot of the interface) using 1,000-point sliding scales that included anchor descriptors but no numeric or interval markings (for scale validation, see Saito et al., 2017). Segmental errors were defined as errors in production of individual consonants and vowels within a word (frequent – infrequent or absent). Intonation referred to appropriateness of pitch moves within speech, such as rising tones in yes/no questions (unnatural – natural). Flow was described as speaker's overall pacing and speed of utterance delivery (disjointed, speech does not flow – speech flows naturally and fluidly).

Raters were first shown the same picture story described in the recordings and were asked to narrate the story in English (10 raters) or French (10 raters); 10 baseline raters inspected the picture story without recording their narrative. Then, raters received instructions about the rated dimensions, using definitions and examples, and evaluated three extra recordings to become familiar with the interface. The recordings were randomized for each rater, and raters could listen to each recording again after the initial playback. At the end of the session, raters were informally debriefed.

Results

Although raters did not differ in speaking task completion times in English (*M* = 39 seconds, *SD* = 15) versus French (*M* = 44 seconds, *SD* = 16), four of the 10 raters completing the French speaking task did not reference all 8 frames in their narratives, whereas all raters completing the English speaking task described all images. Additionally, five raters completing the French speaking task reported having difficulty or feeling nervous and embarrassed about their performance, whereas only one rater completing the English speaking task reported feeling uneasy. This implied that speaking task functioned as intended insofar as narrating the story for raters was generally more challenging in French than English. All ratings displayed high consistency within each group for segmental errors (α = 0.93–0.95), intonation (α = 0.91–0.94) and flow (α = 0.93–0.95); thus, individual scores were derived per rated dimension for each speaker, separately for each rater group (summarized in Table 2).

Table 2. Means (standard deviations) for speech ratings.

		Speaking task language	
Rated measure	**Baseline**	**English**	**French**
Segmental errors	414 (177)	429 (207)	403 (206)
Intonation	429 (161)	463 (182)	455 (205)
Flow	388 (179)	471 (217)	400 (231)

To examine whether engaging raters in a speaking task in English versus French impacted their ratings, relative to baseline raters' assessments, we computed two non-orthogonal (Bonferroni-corrected) contrasts per dimension (baseline versus English, baseline versus French). As summarized in Table 3, performing the speaking task in French before rating L2 speakers produced no statistically significant effect on raters' evaluations of any dimension. However, performing the speaking task in English resulted in raters significantly upgrading their evaluations (by up to +84 points on average) for two of the three dimensions (intonation and flow but not segmental errors). In terms of direction and magnitude, ratings generally tended to become more positive (lenient), relative to baseline raters' evaluations, with relatively small effect sizes (*d* < 0.50), except for ratings of flow after the English speaking task, where the effect size was very large (*d* > 1.00; Plonsky & Oswald, 2014).

Table 3. Summary of comparisons.

Rated measure	**Comparison**	***t*(39)**	***p***	***d***	M_{diff}	**95% CI**
Segmental errors	Baseline versus English	1.43	.482	0.23	+16	[–12, 43]
	Baseline versus French	1.09	.853	0.17	–11	[–36, 14]
Intonation	Baseline versus English	3.39	.005	0.54	+34	[9, 59]

	Baseline versus French	1.83	.225	0.29	+26	[–9, 61]
Flow	Baseline versus English	8.40	< 0.001	1.33	+84	[59, 109]
	Baseline versus French	1.24	.662	0.20	+13	[–13, 39]

To determine how rater evaluations related to their own language skills, we correlated raters' assessments with their estimates of daily English and French use, their self-ratings of English and French proficiency, and the time it took them to complete English or French speaking task (Spearman, two-tailed). Only one variable was significantly associated with ratings, and only for raters who performed the task in French, such that raters who provided greater estimates of their daily English use tended to assign lower (more severe) ratings of intonation (r_s = –0.68, p = 0.030) and flow (r_s = –0.70, p = 0.026) but not segmental errors (r_s = –0.29, p = 0.416).

Discussion

This study's results show that engaging raters in a speaking task before they evaluate L2 speech may affect rater assessments. That is, when bilingual raters completed the speaking task (i.e. the same task performed by speakers) in their stronger language (English), which was also the language of the task that they evaluated, they upgraded their ratings for intonation and flow (but not segmental errors), compared to raters who did not complete the speaking task. In contrast, bilinguals who completed the speaking task in their weaker language (French) did not differ from baseline raters, but their assessments were negatively associated with the amount of their daily English use.

Engaging raters in the same speaking task completed by L2 speakers may have functioned as a positive prime (Luoma-aho et al., 2019), encouraging perspective taking in similar ways to writing about the life of an L2 speaker (Weyant, 2007). Raters' leniency after performing the task in English implies that they may have enhanced their empathy, or shared understanding of the L2 speakers' challenges, through perspective taking. However, performing the task in French was clearly insufficient to encourage leniency (cf. Hansen et al., 2014). Assuming that perspective taking impacted both rater groups, any leniency activated among those who completed the French speaking task might have been neutralized by the anxiety of having to perform an unexpected task in the L2. Language anxiety, 'the feeling of tension and apprehension specifically associated with [L2] contexts' (MacIntyre & Gardner, 1994, p. 284), is a situation-specific occurrence that can spur negative emotions (Dewaele, 2002). Thus, the uncomfortable state induced by performing the speaking task in the L2 may have manifested itself as a lack of generosity while rating others. As one rater commented, 'I didn't think my French was that bad, but it was so hard to tell the story in my second language!' The negative associations between raters' self-reported daily English use and their ratings are certainly compatible with the idea that raters' greater dominance in English was tied to more discomfort in their L2 French and ultimately less generous L2 speech ratings.

The obtained difference for English versus French speaking task is also compatible with effects of mood induction (manipulation of affective states through exposure to happy versus sad music or video) on people's linguistic performance. Participants who are put into a negative mood tend to focus on linguistic detail (Beukeboom & Semin, 2006), so perhaps those performing the speaking task in French (who may have struggled, feeling anxious or embarrassed) scrutinized L2 speakers' performances more closely than those who completed the task in English. Put differently, any positivity induced through task familiarity and perspective taking may have been attenuated through an excessive focus on language detail (e.g. phonetic substitutions, non-native intonation patterns) brought on by negative emotions felt during the French task.

Performing the task in English rather than French was also useful to raters because it provided them with a task model to follow in their evaluation of speakers' L2 English performance, which aligns with a positive role of task familiarity in rater calibration (Davis, 2016). Assessment of L2 speech, particularly by untrained raters, is often based on an abstract, idealized version of what English should sound like (Holliday, 2006). In this case, however, perhaps the raters who completed the English speaking task generated a more concrete, realistic model, thereby establishing their own imperfect narratives (e.g. complete with performance errors and native-speaker variations) as the baseline for comparison. In fact, several raters who completed the English speaking task criticized their performance in debrief comments, which is not atypical of native speakers (Campbell et al., 2001). In other words, the establishment of a realistic performance model coupled with self-criticism of their own narratives could have resulted in rating generosity for L2 speakers.

While intonation and flow patterned together in upgrades from the raters who performed the English speaking task, ratings of segmental errors remained relatively stable. A clue to this finding might have emerged in earlier research (Taylor Reid et al., in review), where the positive effects of perspective taking (as a way of mitigating externally-imposed social bias) were more pronounced for comprehensibility (raters' judgment of how easy or difficult it is for them to understand a speaker) than accentedness (raters' assessment of the extent to which a speaker sounds nativelike). Raters associate both intonation (Sereno et al., 2016) and flow (Derwing et al., 2004) with comprehensibility, whereas segmental errors are typically linked to accentedness (Saito et al., 2017). Because comprehensibility is influenced by topic familiarity (Gass & Varonis, 1984), which was activated when raters completed the narrative, it stands to reason that ratings of intonation and flow would be similarly enhanced. Segmental errors, however, similar to accentedness, would likely remain salient, given that even small segmental deviations reliably trigger the perception of a foreign accent (Pérez-Ramón et al., 2020).

Finally, the fact that the effect size for flow was far greater than that for intonation might link back to the idea that raters were using their own narratives as a model. Unlike monologues and conversations, a picture narrative is inherently more constrained in that it requires specific vocabulary for unfamiliar content (Derwing et al., 2004), and would naturally lead to pauses and hesitation, even for native speakers. Awareness of that struggle could have translated into especially generous ratings for others. Moreover, fluency or flow often stands as a proxy for L2 speakers' overall language ability, especially for laypersons, which would make the dimension of flow particularly appealing if raters chose to be generous in their speech ratings.

Implications, future work and conclusion

Future studies investigating human ratings of L2 speech should consider the impact of additional perspective-taking interventions on rater evaluations. These could include reading or listening to anecdotes about situations where L2 speakers experienced prejudice or enhanced empathy from their interlocutors on the basis of their speech. Similarly, raters could be asked to comment on situations in their own lives in which they experienced prejudice or enhanced empathy on the basis of their language performance or factors unrelated to their linguistic competence. If such interventions are effective in terms of encouraging more positive assessments of L2 speech in a laboratory setting, they could be utilized in the training of individuals tasked with assessing L2 speakers on a regular basis. For example, L2 teachers and examiners could be encouraged to complete tasks similar to those they are assessing before they begin their assessments. Along similar lines, one aspect of training provided to human resource personnel could involve roleplaying that teases apart linguistic issues associated with people's speech from other factors related to their professional competence. In even higher-stakes contexts where assessment of credibility is paramount – such as those involving legal interaction with L2 speakers (i.e. traffic stops, border crossings, courtroom proceedings) – efforts could move beyond conventional tactics aimed at increased understanding (e.g. diversity training) to those that might expand the mindset of the participant by combining perspective taking with other successful interventions such as intercultural communication opportunities, as can be achieved through virtual reality (see Salmanowitz, 2016). To conclude, our findings provide preliminary evidence of how bilingual raters' own experience with a speaking task might impact their assessments of others, highlighting the interplay of linguistic and social factors in L2 speech research.

References

Baquiran C. L. C. & Nicoladis, E. (2020). A doctor's foreign accent affects perceptions of competence. *Health Communication, 35*, 726–30.

Beukeboom, C. J. & Semin, G. R. (2006). How mood turns on language. *Journal of Experimental Social Psychology, 42*, 553–66.

Boland, R. J. & Tenkasi, R. V. (1995). Perspective making and perspective taking in communities of knowing. *Organization Science, 6*, 350–72.

Campbell, K. S., Mothersbaugh, D. L., Brammer, C. & Taylor, T. (2001). Peer versus self assessment of oral business presentation performance. *Business Communication Quarterly, 64*, 23–40.

Campbell-Kibler, K. (2007). Accent, (ING), and the social logic of listener perceptions. *American Speech*, *82*, 32–64.

Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing*, *33*, 117–35.

Deprez-Sims A.-S. & Morris, S. B. (2010). Accents in the workplace: Their effects during a job interview. *International Journal of Psychology*, *45*, 417–26.

Derwing, T. M. & Munro, M. J. (2009). Comprehensibility as a factor in listener interaction preferences: Implications for the workplace. *Canadian Modern Language Review*, *66*, 181–202.

Derwing, T. M. & Munro, M. J. (2015). *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research.* Amsterdam: John Benjamins.

Derwing, T. M., Rossiter, M. J., Munro, M. J. & Thomson, R. I. (2004). Second language fluency: Judgments on different tasks. *Language Learning, 54*, 655–79.

Dewaele, J.-M. (2002). Psychological and sociodemographic correlates of communicative anxiety in L2 and L3 production. *International Journal of Bilingualism, 6*, 23–38.

Gass, S. & Varonis, E. M. (1984). The effect of familiarity on the comprehensibility of nonnative speech. *Language Learning*, *34*, 65–89.

Halim, M. L., Moy, K. H. & Yoshikawa, H. (2017). Perceived ethnic and language-based discrimination and Latina immigrant women's health. *Journal of Health Psychology*, *22*, 68–78.

Hansen, K., Rakić, T. & Steffens, M. C. (2014). When actions speak louder than words: Preventing discrimination of nonstandard speakers. *Journal of Language and Social Psychology, 33,* 68–77.

Holliday, A. (2006). Native-speakerism. *ELT Journal, 60,* 385–7.

Isaacs, T. & Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly, 10,* 135–59.

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*, 113–40.

Kang, O. & Rubin, D. L. (2009). Reverse linguistic stereotyping: Measuring the effect of listener expectations on speech evaluation. *Journal of Language and Social Psychology*, *28*, 441–56.

Kang, O. (2008). Ratings of L2 oral performance in English: Relative impact of rater characteristics and acoustic measures of accentedness. Doctoral dissertation, University of Georgia, Athens, GA. Retrieved from https://getd.libs.uga.edu/pdfs/kang_okim_200805_phd.pdf.

Luoma-aho, V., Pirttimäki, T., Maity, D., Munnukka, J. & Reinikainen, H. (2019). Primed authenticity: How priming impacts authenticity perception of social media influencers. *International Journal of Strategic Communication*, *13*, 352–65.

MacIntyre, P. D. & Gardner, R. C. (1994). The subtle effects of language anxiety on cognitive processing in the second language. *Language Learning*, *44*, 283–305.

Pérez-Ramón, R., Cooke, M. & García Lecumberri, M. L. (2020). Is segmental foreign accent perceived categorically? *Speech Communication*, *117*, 28–37.

Plonsky, L. & Oswald, F. L. (2014). How big is ‘big?’ Interpreting effect sizes in L2 research? *Language Learning, 64*, 878–912.

Saito, K., Trofimovich, P. & Isaacs, T. (2017). Using listener judgements to investigate linguistic influences on L2 comprehensibility and accentedness: A validation and generalization study. *Applied Linguistics*, *38*, 439–62.

Salmanowitz, N. (2016). Unconventional methods for a traditional setting: The use of virtual reality to reduce implicit racial bias in the courtroom. *UNHL Review, 15*, 117–60.

Sereno, J., Lammers, L. & Jongman, A. (2016). The relative contribution of segments and intonation to the perception of foreign-accented speech. *Applied Psycholinguistics*, *37*, 303–22.

Staples, S., Kang, O. & Wittner, E. (2014). Considering interlocutors in university discourse communities: Impacting US undergraduates’ perceptions of ITAs through a structured contact program. *English for Specific Purposes*, *35*, 54–65.

Steele, C. M. & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, *69*, 797–811.

Stewart, M. A. (2010). Walking in my students’ shoes: An ESL teacher brings theory to life in order to transform her classroom. *Networks*, *12*, 1–6.

Taylor Reid, K., Trofimovich, P. & O’Brien, M. G. (2019). Social attitudes and speech ratings: Effects of positive and negative bias on multiage listeners’ judgments of second language speech. *Studies in Second Language Acquisition*, *41*, 419–42.

Taylor Reid, K., Trofimovich, P., O’Brien, M. G. and Tsunemoto, A. (under review). Using task practice to reduce social influences on listener evaluations of second language accent and comprehensibility. Manuscript submitted for publication.

Timming, A. R. (2017). The effect of foreign accent on employability: A study of the aural dimensions of aesthetic labour in customer-facing and non-customer-facing jobs. *Work, Employment and Society*, *31*, 409–28.

Weyant, J. M. (2007). Perspective taking as a means of reducing negative stereotyping of individuals who speak English as a second language. *Journal of Applied Social Psychology*, *37*, 703–16.

Zhang, Y. (2017). Walking a mile in their shoes: Developing pre-service music teachers’ empathy for ELL students. *International Journal of Music Education*, *35,* 425–34.

Appendix

Social Attitudes Questionnaire

Indicate the degree to which each of these statements accurately reflects how you feel.

Disagree **Agree**

PRIDE FOR ETHNIC GROUP

1. I am proud to be a member of my ethnic group. 1 2 3 4 5 6 7 8 9
2. I am proud to let people know that I belong to my ethnic group. 1 2 3 4 5 6 7 8 9
3. I am proud of the achievements of my ethnic group. 1 2 3 4 5 6 7 8 9
4. I feel proud to see symbols of my ethnic group (such as a flag) displayed around me. 1 2 3 4 5 6 7 8 9
5. I am proud to be able to speak the language of my ethnic group. 1 2 3 4 5 6 7 8 9

ENGLISH IN QUEBEC

6. Anglophone Quebecers do not have considerable economic power in Quebec. 1 2 3 4 5 6 7 8 9
7. Anglophone Quebecers do not have considerable political power in Quebec. 1 2 3 4 5 6 7 8 9
8. In my daily life (for example, in a restaurant, shop, doctor's office), I should have the right to speak English in Quebec. 1 2 3 4 5 6 7 8 9
9. I should have the freedom to choose if I want my children to be educated in English in Quebec. 1 2 3 4 5 6 7 8 9
10. Anglophone Quebecers' contribution to Quebec is not recognized or valued. 1 2 3 4 5 6 7 8 9

ATTITUDES TOWARDS IMMIGRANTS

11. The influx of immigrants is lowering the standard of living of people in Quebec. 1 2 3 4 5 6 7 8 9
12. Laws, customs, and traditions that are specific to immigrant groups should not be imposed on the Quebec society as a whole. 1 2 3 4 5 6 7 8 9
13. Immigrants should adopt the Quebec way of life and values to replace their traditional way or life and values. 1 2 3 4 5 6 7 8 9
14. Immigrants are bringing conflicts in their home countries into Quebec. 1 2 3 4 5 6 7 8 9
15. Immigrants benefit a lot from being in Quebec so they should be loyal to Québec. 1 2 3 4 5 6 7 8 9

PERSONAL RELATIONS WITH OTHER GROUPS

16. I feel accepted and respected by other ethnic groups in Quebec. 1 2 3 4 5 6 7 8 9

17. Members of other ethnic groups do not mind me living in close proximity to them. 1 2 3 4 5 6 7 8 9

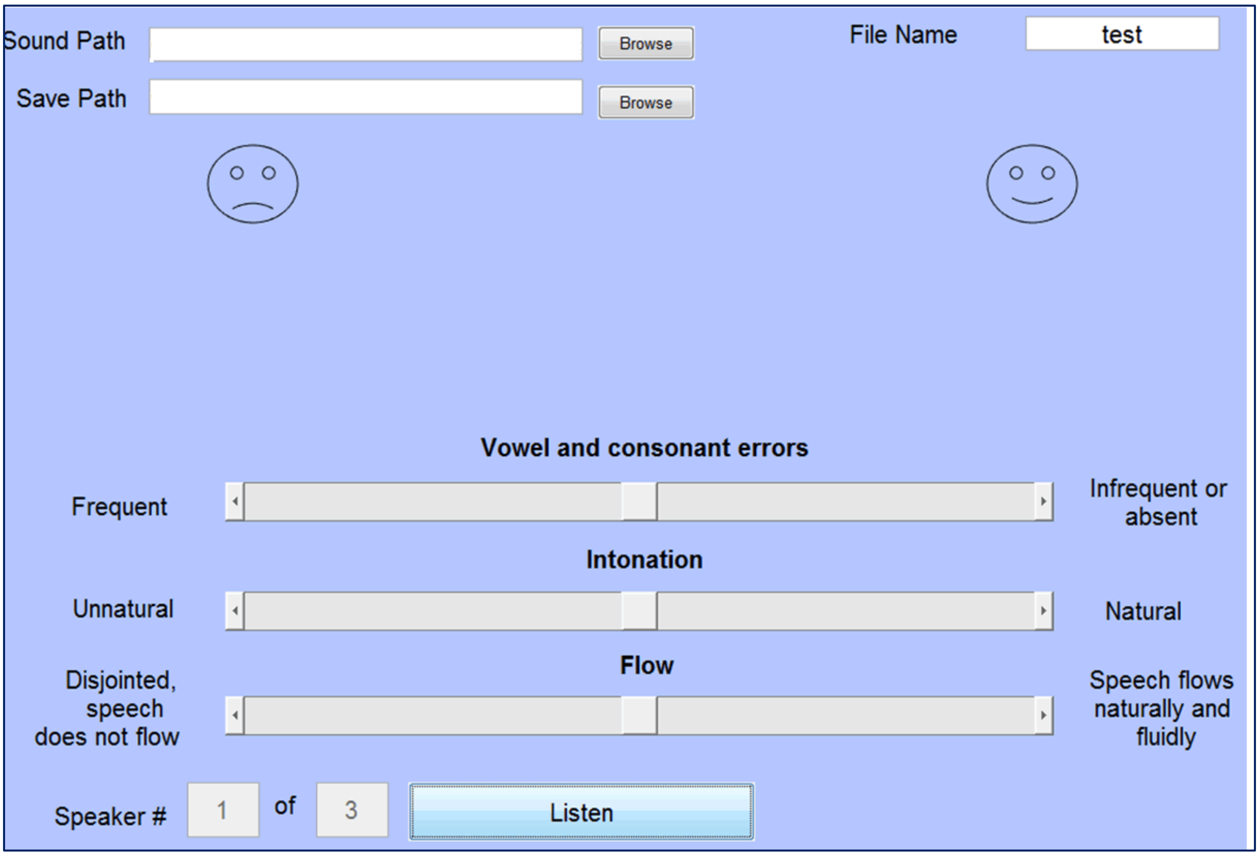
18. Members of other ethnic groups would not object to my children marrying their children. 1 2 3 4 5 6 7 8 9

19. Children who grow up in an ethnically diverse Quebec are more prepared to live in today’s world. 1 2 3 4 5 6 7 8 9

20. Children growing up in an ethnically diverse Quebec are more tolerant of other groups. 1 2 3 4 5 6 7 8 9

Note. Category labels were not presented to participants. Questionnaire is based on materials from Authors (xxxx).

Screenshot of the Rating Interface



File Name: test

Speaker # 1 of 3