Testing the malleability of teachers' judgments of second language speech

Kym Taylor Reid, Mary Grantham O'Brien, Pavel Trofimovich, and Allison Bajt

This study examined whether a negative social bias can influence how teachers evaluate second language (L2) speech. Twenty-eight teachers of L2 German from Western Canada—14 native speakers (NSs) and 14 proficient non-native speakers (NNSs)—rated recordings of 24 adult L2 learners of German across five speech dimensions (accentedness, comprehensibility, vowel/consonant accuracy, intonation, flow) using 1,000-point scales. Immediately before rating, half of NS and NNS teachers heard critical comments about undergraduate German students' language skills, while the other half heard no biasing comments. Under negative bias, while the NNS teachers provided favorable evaluations across all five measures, NS teachers followed suit for only intonation and flow, downgrading L2 speakers' accentedness, comprehensibility, and vowel/consonant accuracy. Findings call into question the relative stability of L2 speech ratings and highlight the importance of social context and teacher status as native versus non-native speakers of the target language in assessments of L2 speaking performance.

Keywords

1. Introduction

People frequently make judgments about their interlocutors (e.g., D'Onofrio, 2018) and the language variety they speak (e.g., Preston, 1999). Although there is nothing inherent to a given language variety that makes it superior to another variety, listeners often prefer certain ways of speaking over others. For example, listeners judge speakers of some language varieties as more confident (Stewart, Ryan, & Giles, 1985), more intelligent (D'Onofrio, forthcoming), and friendlier (Watson & Clark, 2015) than those speaking other varieties. Some varieties are also frequently preferred to others as occurs, for instance, with English spoken in Alabama (Preston, 1999) or Liverpool English (Montgomery, 2007) that is consistently labeled as "incorrect." Furthermore, such listener-based preferences affect many (if not most) languages. Dailey-O'Cain (1999), for example, found that of 218 Germans tested across 46 regions, Westerners described the language of Easterners as less correct and less pleasant than that of German speakers in their own region (see also Boughton, 2006; van Bezooijen, 2002).

Although listeners understand a great deal of foreign-accented speech—that is, speech containing pronunciation features that both differ from an expected variety and mark an individual as a speaker of a specific variety (Derwing & Munro, 2015)—listeners are also quick to make judgments about second language (L2) speakers and the varieties that they speak on the basis of their accents (e.g., Cargile & Giles, 1997; Pantos & Perkins, 2013). The formal assessment of L2 learners often falls to teachers who, as members of their respective linguistic communities, might therefore be influenced by similar attitudinal biases. However, because of their experience with L2 speech and their training, teachers are generally considered to be expert judges of L2 performance (e.g., Anderson-Hsieh, Johnson, & Koehler, 1992; Lappin-Fortin & Rye, 2014; Rossiter, 2009). Consequently, the validity of the assumption that teachers are

immune to external biases shown by other language users is underexplored. This study's goal was to evaluate this assumption by examining the degree to which teachers' assessments of L2 German speech produced by classroom learners—specifically in terms of its accentedness, comprehensibility, vowel and consonant accuracy, intonation, and flow—can be manipulated as a result of teachers hearing critical comments about L2 students' speech.

2. Background Literature

2.1 Listener Biases and Their Origins

Some listener biases can be explained through reference to social identity, which is derived from a person's identification with a language group rather than with the language itself (Tajfel, 1972). One's language, in combination with other identity markers such as gender and age, can determine a member's status within a given social group (Norton & Toohey, 2011). Working with the assumption that listeners associate status traits such as intelligence and ambition (Lambert, Hodgson, Gardner, & Fillenbaum, 1960) and form solidarity with those who share the same in-group language (Ryan, Carranza, and Moffie, 1977), Lindemann (2003) tested 39 Michigan undergraduates, who distinguished English speakers from L2 speakers (all Korean), in audio clips. Although the undergraduates could not identify the accent they heard, in that they were unclear whether the accent was indeed Korean, Chinese, Japanese, Latino, or something else, all Korean-accented speakers were rated as less intelligent and of lower status than the English speakers. These accented speakers were likely identified as members of an out-group, suggesting that L2 speech is a marker of foreign (out-group) status and that a L2 speaker may be viewed negatively in comparison to the listener's identifying group.

In addition, in-groups and out-groups frequently differ in their relative dominance, such that one group and its language variety are often ascribed a higher status than other groups,

leading to negative evaluations of speakers of less dominant, preferred, or "standard" varieties (e.g., Ryan, Hewstone, & Giles, 1984). Reacting to these differences in group dominance, L2 speakers of English often downgrade fellow L2 speakers from their own group, who they perceive to be of lower status, relative to those who speak English natively (Chiba, Matsuura, & Yamamoto, 1995; Dalton-Puffer, Kaltenböck, & Smit, 1997; He & Miller, 2011; Xu, Wang, & Case, 2010).

Other listener biases towards L2 speech can be traced back to ideas which have little to do with the speech itself. In an early study by Rubin (1992), undergraduate students listened to a speech sample paired with an image of a speaker—supposedly a university instructor—who was either Caucasian or Asian. Regardless of the image that was presented, the speech sample was the same. Nonetheless, participants who saw the photograph of an Asian speaker rated the speech samples as more accented. Recent research on such biases, framed as reverse linguistic stereotyping, has shown that listeners often evaluate speech based on their beliefs, expectations, and stereotypical views rather than the speakers' actual performance (Hu & Su, 2015; Kang & Rubin, 2009). Such biases likely cause listeners to impose linguistic and evaluative features on the speech they hear, which may ultimately lead to L2 speakers being downgraded in evaluations.

Yet other sources of listener bias towards L2 speech might stem from people's moods. For example, participants who are in a negative mood tend to pay more attention to details, whereas those in a positive mood are more likely to process information globally (Beukeboom, & Semin, 2006; Vissers et al., 2010). This difference has been attributed (at least in part) to people in a happy mood being less likely to attend to errors because they are not motivated to put effort into processing information (Bodenhausen, Mussweiler, Gabriel, & Moreno, 2001). Such

findings may extend to listener evaluations of L2 speech. In a recent study by Reid, Trofimovich, and O'Brien (2019), some English listeners who were exposed to a positive anecdote about L2 speakers' language ability tended to upgrade L2 speakers in their ratings, whereas other listeners exposed to a negative anecdote before the rating tended to downgrade the same speakers in their assessments. Nonetheless, listener mood—as a sole explanation of listener bias—may not be straightforward. Although older listeners in Reid et al.'s study were indeed influenced by positive and negative biases, younger listeners performed in an unexpected way, providing higher ratings in response not only to a positive anecdote but also to a negative one, likely as a way of resisting an unfair (negative) characterization of L2 speakers.

2.2 Listener-Specific Factors

Although listener biases towards L2 speech might stem from different sources—including group identity, preconceived ideas, or mood states—such biases appear to depend on several listener-specific factors. One listener-specific factor shown to influence speech ratings is amount of exposure. For instance, English speakers with little prior exposure to Chinese-accented English were able to more accurately transcribe utterances when their experience with L2 speech included five different Chinese-accented speakers instead of just one (Bradlow & Bent, 2008). American listeners with international experience (which presumably enhanced their exposure to multiple varieties of L2 English) similarly showed a significant improvement in understanding L2 English speakers, compared to listeners with no international experience (Hansen Edwards, Zampini, & Cunningham, 2018). Listeners' familiarity with various speech varieties, as shown most clearly in research carried out within the world Englishes paradigm (for review, see Hansen Edwards et al., 2018), also contributes to favorable evaluations of speakers of those varieties, for instance, in terms of their intelligence or education (e.g., Chiba et al., 1995;

Hundt, Zipp, & Huber, 2015; Tan & Castelli, 2013). Furthermore, an increase in listener familiarity with L2 speech has also been linked to higher assessments in speaking tests (Carey, Mannell, & Dunn, 2011; Winke & Gass, 2013; Winke, Gass, & Myford, 2013).

Another listener-specific factor (which is directly relevant to this research) is listeners' status as L2 speakers themselves. With respect to listeners and speakers sharing the same linguistic background, for example, Foote and Trofimovich (2018) showed that Mandarin-speaking university students upgraded the comprehensibility ratings of other Mandarin speakers of L2 English in comparison to the Hindi and French speakers with whom the listeners did not share the same language background. Similarly, in Munro, Derwing, and Morton's (2006) study, Cantonese listeners found the speech from their own language background easier to understand than the speech by Japanese, Polish, and Spanish speakers. Yet, a shared listener–speaker background is not always an advantage. For example, Major, Fitzmaurice, Bunta, and Balasubramanian (2002) found that Spanish listeners performed better when listening to a Spanish speaker of L2 English, but that Japanese and Chinese listeners did not show an advantage when listening to L2 English speech of Japanese and Chinese speakers.

There is similarly little consensus regarding differences between native and L2 listener evaluations of L2 speech. For example, Rose (2017) showed that Japanese listeners judged the fluency of L2 English speakers more harshly than did English listeners across three tasks (picture description, topic narrative, and read aloud), even when the speaker was of the same Japanese background. This finding is similar to those showing that, compared to native listeners, L2 listeners are often more severe in their evaluations of L2 speech (Fayer & Krasinski, 1987; Kang, 2012; Rossiter, 2009). In other cases, however, L2 listeners may be more lenient than native listeners in rating L2 speech (Brown, 1995) or demonstrate no difference in rating (Derwing &

Munro, 2013). Thus, a common status as a fellow L2 speaker—and perhaps an accompanying feeling of solidarity—does not always lead to positive assessments of L2 speech for L2 listeners.

2.3 The Current Study

Human judges are considered to be the benchmark for determining L2 speakers' performance (e.g., Eskenazi, 2009), and teachers—by virtue of their educational training and instructional experience—have long been relied on for their rating expertise (e.g., Anderson-Hsieh et al., 1992; Bongaerts, van Summeren, Planken, & Schils, 1997; Rossiter, 2009). However, teachers are certainly not immune to biases when evaluating their students' language performance. For example, Ford (1984) showed that teachers in the southwestern United States evaluated the English writing samples of third and fourth grade native English speakers more negatively when the work was paired with speech samples of Spanish-accented speakers than speech samples of native English speakers. Similarly, English lecturers at a Slovenian university self-reported biased assessment of student writing (Sokolov, 2014), citing their knowledge of a student's L2 proficiency and their familiarity with the student as key factors impacting their assessments.

In addition to evaluating writing, teachers frequently assess their students' speaking skills. Teachers seem to be well aware of global characteristics of L2 speech, including comprehensibility (listeners' difficulty of understanding L2 speech), accentedness (listeners' perception of the extent to which speech differs from the local variety), and flow (listeners' evaluation of pacing and speed of an utterance), and exhibit sensitivity to specific linguistic dimensions of speech, such as segment accuracy (accurate articulation of consonants and vowels) and intonation (speech melody) (e.g., Anderson-Hsieh et al., 1992; Isaacs & Trofimovich, 2012; Isaacs & Thomson, 2013; Rossiter, 2009). What is largely unknown,

however, is whether teachers—as raters of their students' speaking skills—are susceptible to social and attitudinal influences from others (which we describe here as "social biases," whether positive or negative) in evaluating both global and specific dimensions of L2 speech. While amount of exposure to L2 speech might not be a critical factor in shaping teacher biases in assessment of their students (assuming that most teachers have ample experience with their students' speech), teachers' status as native speakers (NSs), that is, those who have been exposed to the target language in early childhood, versus nonnative speakers (NNSs), that is, those who learn the target language later in life, might be a key factor. As discussed previously, prior research has yielded inconclusive findings, implying that (at minimum) NS and NNS teachers might differ in assessments of their students' oral performance.

The main goal of this study was therefore to examine whether teachers—as evaluators of their students' L2 oral performance—might be sensitive to a social bias manipulation (as were many naïve listeners in Reid et al.'s 2019 study) and whether this sensitivity might differ for teachers who are NSs or NNSs of the target language. Focusing on teachers of L2 German, an underexplored population of instructors, we asked participants to evaluate speech samples produced by L2 German learners for five speech variables (accentedness, comprehensibility, vowel and consonant accuracy, intonation, and flow). Immediately before the rating task, half of the teachers heard a short personal opinion by the researcher criticizing L2 German skills of a hypothetical learner of German (negative bias group), while the remaining half heard no biasing opinion (baseline group). The negative bias manipulation, as implemented here, is consistent with similar attitudinal and affective manipulations, including affective priming (e.g., Fazio, Sanbonmatsu, Powell, & Kardes, 1986) 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. Our working assumption was that a brief biasing negative anecdote targeting the (poor) linguistic skills of a hypothetical L2 learner of German, introduced at the outset of a rating session, would influence the ratings provided by the teachers, relative to the ratings of those who heard no such anecdote. Although the impact of a negative bias (as operationalized here) may be more pronounced for the first few speaking performances evaluated by each teacher, our expectation was that the effects of such bias (if they exist) should be robust enough to be detected across all speaking performances being evaluated.

Because there was an equal number of teachers in each group who differed in their status (as either NSs or NNSs of German), it was also possible to examine whether teacher ratings were more or less susceptible to a biasing anecdote, depending on their NS–NNS status. We made no specific prediction regarding teacher status, based on inconclusive findings from prior work. On the one hand, NS and NNS teachers of German could provide similar ratings, regardless of the biasing orientation (e.g., Crowther, Trofimovich, & Isaacs, 2016; Derwing & Munro, 2013). On the other hand, NNS teachers might be less affected by a biasing anecdote, compared to NS teachers, feeling perhaps a certain degree of empathy with the fellow L2 speakers (e.g., Hansen, Rakić, & Steffens, 2014). Alternatively, NNS teachers might be more affected, given that L2 listeners sometimes provide harsher evaluations than native listeners (e.g., Kang, 2012; Rose, 2017; Rossiter, 2009). The following research question guided the study: Does a negative bias influence NS and NNS teachers' ratings of intermediate to advanced L2 German speech for comprehensibility, accentedness, consonant and vowel errors, intonation, and flow, relative to the ratings of the teachers not exposed to bias?

3. Method

3.1 Teacher Raters

The raters included 28 teachers of German working in a range of contexts and school boards, all residents of a western Canadian province. Half of the teachers were born to German-speaking parents in Germany (7), Switzerland (1), Romania (1), and Canada (5) (henceforth, NS teachers). The other half of the teachers were born to non-German-speaking parents in Canada (13) and Argentina (1) and self-identified as native speakers of English (8) and Spanish (1) or bilingual speakers of English and German (4) and Cantonese and English (1) (henceforth, NNS teachers). Thus, although both teacher groups included bilinguals or multilinguals, with German being one of the languages spoken, the groups differed in one crucial respect: NS teachers were born to German-speaking parents, while NNS teachers grew up in non-German households and were not exposed to German until at least school age. Thus, the assignment of teachers to the two groups was based on their status as early versus late learners of German, as opposed to, for instance, teachers' self-identification as a member of a social group.[1] The NS teachers reported either BA (13) or MA (1) as their highest degree, as well as having previously completed language pedagogy courses (11) and additional teacher training (12), including specialized courses in linguistics (12). The NNS teachers similarly reported either BA (12) or MA (2) degrees, with additional language pedagogy coursework (14) and teacher training (8), including instruction in linguistics (10). In terms of the range of L2 teaching experience, the NS teachers reported teaching German in elementary schools (10), junior and high schools (7), universities (1), and private language schools (3), with many engaged in teaching across multiple contexts. The NNS teachers similarly reported teaching German in elementary schools (10), junior and high schools (6), and universities (2), again with many engaged in instruction across several

levels.

The 14 NS and 14 NNS teachers were randomly assigned to two equal groups ($n = 7$). One group was then exposed to a negative social bias (described in detail below), while the other group received no bias (baseline) before the rating task. As shown in Table 1, the resulting four groups of teachers—NS and NNS teachers exposed to negative bias or no bias—did not differ in any background or language variables, including age, amount of daily German use in speaking, and daily German use with German NSs, as shown through nonparametric Kruskal-Wallis tests, $\chi(3) < 6.43$, $p > .09$. The groups also did not differ in their responses to a social attitudes questionnaire (see Appendix A) targeting the strength of teachers' pride for their ethnic group (4 questions, Cronbach's $\alpha = .87$), their perception of the role of language in their ethnic identity (3 questions, $\alpha = .78$), their perception of the role of German in society (3 questions, $\alpha = .83$), and their pride in teaching German (single question), $\chi(3) < 2.90$, $p > .41$.

Table 1 *Median and Range Values for German Teachers' Background Characteristics*

	Baseline		Negative bias	
Background variable	NS teachers	NNS teachers	NS teachers	NNS teachers
Gender (f–m)	7–0	4–3	4–3	7–0
Age (years)	44 (29–71)	37 (26–56)	54 (46–64)	40 (27–59)
Teaching German (years)	10 (2–40)	10 (1–25)	18 (13–25)	7 (1–27)
Daily German use (0–100%)	30 (20–60)	30 (10–80)	50 (30–90)	30 (10–60)
Daily German use with NSs (0–100%)	30 (0–100)	10 (10–70)	55 (30–90)	20 (10–90)
Pride for own ethnic group (1–9)[a]	8.3 (4.0–9.0)	7.3 (6.3–9.0)	7.8 (4.3–9.0)	9.0 (5.0–9.0)
Role of language in identity (1–9)[a]	8.0 (7.7–8.0)	8.3 (6.0–9.0)	9.0 (7.0–9.0)	8.3 (5.0–9.0)
Role of German (1–9)[a]	4.3 (1.3–8.3)	6.0 (6.0–7.7)	6.2 (2.3–7.0)	7.8 (4.3–9.0)
Pride in teaching German (1–9)	9.0 (8.0–9.0)	9.0 (9.0–9.0)	9.0 (8.0–9.0)	9.0 (8.0–9.0)

Note. [a]Mean of the question responses targeting this construct (see Appendix A).

3.2 Speech Materials

The target audio samples evaluated by the teachers included the speech of 24 L2 learners of German (12 women, 12 men), all native speakers of English living in Germany at the time of testing (M_{age} = 22.8 years, *range* = 20–31). The speakers were all adult learners of German with high intermediate to advanced German proficiency, as demonstrated by the results of an online proficiency test from the Goethe Institute (2004) that is aligned with the Common European Framework of Reference (CEFR), with 12 B2 (vantage level) speakers and 12 C1 (effective operational level) or C2 (mastery level) speakers. The speakers recorded brief narratives in response to an eight-panel picture story describing two passers-by who collided at a busy street

corner and exchanged similar-looking suitcases (Derwing, Rossiter, Munro, & Thomson, 2004).

The target samples included the first 20 seconds of each narrative with initial disfluencies,

including false starts and hesitations, removed, which aligns with prior research of L2 speech

assessment (e.g., Derwing & Munro, 2013). Thus, the samples were not modified in any way

except to remove pausing and hesitations (e.g., *uhm*) before the onset of speech. These speech

samples, collected as part of a previous study (O'Brien, 2014), were considered appropriate for

evaluation by L2 German teachers because they were collected in educational contexts that are

highly similar to those from which the teacher raters were drawn.

3.3 Rating Procedure

The teachers, who were tested individually in a quiet room in their workplace, provided

two sets of ratings in the same sequence: two global variables (accentedness, comprehensibility)

and three specific pronunciation variables (segmental errors, intonation, flow), all summarized in

Table 2. The ratings were carried out in the same session using 1,000-point sliding scales

programmed in the custom-built speech evaluation software *Z-Lab* (Yao, Saito, Trofimovich, &

Isaacs, 2013), with all information displayed in German for the teachers (see Appendix B for

screenshots of the rating interface). The endpoints of each scale were not labeled numerically,

and the scale contained no interval markings; however, the endpoints were identified as negative

(frowning face) and positive (smiling face) and corresponded to the ratings of 0 and 1,000,

respectively. The scales for accentedness and comprehensibility appeared on screen together, and

the teachers were required to listen to each sample once before providing their ratings, on the

assumption that accent and comprehensibility reflect initial, intuitive, perceptual judgments. The

scales for the remaining three variables (segmental errors, intonation, flow) also appeared

together, and the teachers were expected to replay each sample a second time before providing

their ratings for these three variables.

Table 2 *Summary of Rated Categories (English Translations) with Endpoint Descriptors*

Rated measure	Left endpoint	Right endpoint	Category summary
Accentedness	Heavily accented	No accent at all	How different a speaker sounds from a native German speaker
Comprehensibility	Hard to understand	Easy to understand	Ease or difficulty of raters' understanding of L2 speech
Segmental (vowel & consonant) errors	Frequent	Infrequent or absent	Errors in production of individual consonants and vowels within a word
Intonation	Unnatural	Natural	Appropriateness of pitch moves within speech, such as rising tones in yes/no questions
Flow	Disjointed, speech does not flow	Speech flows naturally and fluidly	Speakers overall pacing and speed of utterance delivery

At the beginning of each session, the teachers first examined the picture story described by the speakers. They were then instructed about each rating category using definitions and examples, and they were asked to evaluate three extra practice samples before proceeding to rate the 24 target speech samples (using headsets). The critical manipulation involved the researcher providing a biasing orientation in her native English to the teachers after they rated the practice samples. In the negative orientation group, the teachers heard a short (scripted and rehearsed but naturally delivered) personal opinion by the researcher about her recent experience interacting

with an undergraduate student majoring in German who, according to the researcher, could do little in the language and had a terrible accent and poor grammar. The researcher then lamented that a student majoring in German should be able to use the language, concluding that some German majors do not even bother to become fluent (see Appendix C for full script). The negative manipulation thus targeted a hypothetical learner of L2 German (i.e., not a specific speaker whose speech was included in the study) with the idea that the manipulation might color or otherwise influence teachers' assessments of all speech samples assigned to them and presented in a unique randomized order per teacher (for similar experimental logic, see Steele & Aronson, 1995). The teachers assigned to the baseline group received no social bias orientation, proceeding to rate the speaker files immediately after completing the practice samples.

At the end of the session, the teachers filled out a language background questionnaire, a social attitudes questionnaire, and the final debrief questionnaire, which asked them to judge the pleasantness of the rating session, the researcher's helpfulness, the rating task difficulty, and their confidence in rating, using 100-millimeter continuous scales (see Appendix D). The teachers were invited to comment about any part of their interaction with the researcher that might have affected their ratings. They were also debriefed after each session, and any final comments were recorded as field notes by the researcher.

3.4 Data Analysis

All ratings for the NS and NNS teachers were first checked for consistency (Cronbach's α) separately in the negative bias and baseline groups. As shown in Table 3, the teachers were consistent in their ratings, with reliability indexes exceeding the benchmark values of .70–.80 (Larson-Hall, 2009). Thus, individual scores were derived for each speaker, separately for NS and NNS teachers in the negative bias and baselined groups, by averaging across the relevant

ratings and across the seven teachers in each group for each measure.

Table 3 *Interrater Reliability Across Teacher Raters (Cronbach's α)*

Rated measure	Baseline		Negative bias	
	NS teachers	NNS teachers	NS teachers	NNS teachers
Accentedness	.95	.94	.97	.94
Comprehensibility	.86	.91	.93	.88
Segmental errors	.94	.93	.96	.95
Intonation	.90	.92	.90	.91
Flow	.94	.97	.96	.95

The debrief questionnaire responses were scored by measuring the distance (in millimeters) between the left endpoint of the scale and the teacher's mark (cross or checkmark) on the 100-millimeter scale. Comparisons of the debrief responses (using nonparametric Kruskal-Wallis tests) confirmed that the teacher groups did not differ in their responses, $\chi(3) <$ 6.57, $p > .09$. More specifically, the NS teachers found the experience pleasant ($Mdn_{baseline} =$ 100.0, $Mdn_{negative} = 100.0$) and the researcher helpful ($Mdn_{baseline} = 100.0$, $Mdn_{negative} = 100.0$); they also evaluated task difficulty similarly ($Mdn_{baseline} = 78.0$, $Mdn_{negative} = 80.0$) and were similarly confident in their ratings ($Mdn_{baseline} = 76.0$, $Mdn_{negative} = 85.0$). The NNS teachers also found the experience pleasant ($Mdn_{baseline} = 100.0$ $Mdn_{negative} = 97.0$) and the researcher helpful ($Mdn_{baseline} = 100.0$, $Mdn_{negative} = 98.0$), evaluating task difficulty at a comparable level ($Mdn_{baseline} = 82.0$, $Mdn_{negative} = 82.0$) and being similarly confident in their ratings ($Mdn_{baseline} =$ 80.0, $Mdn_{negative} = 79.0$).

The field notes were analyzed broadly for direct quotes from the teachers (which were

typically short phrases or single sentences) reflecting their awareness of bias. All teachers

appeared to be similarly unaware of the social bias, with 4–5 teachers (out of seven) per group

responding "no" or "not at all" to the question asking if they thought anything that the researcher

said could have influenced their ratings. The remaining teachers in each group commented on the

helpfulness of the researcher (e.g., "her explanations made the task as easy as it could be for

me," "helpful instructions, questions were clarified, doing a practice round was needed").

4. Results

To determine if the negative bias orientation influenced the NS and NNS teachers'

assessments of L2 speech, relative to the ratings of the teachers in the baseline group (all

summarized in Table 4), the ratings were submitted to two-way ANOVAs, one per rated

measure, with teacher status (NS, NNS) and bias (baseline, negative) as the targeted factors.

Table 4 *Means (Standard Deviations) for Ratings Targeting 24 Speakers Varying in L2 Proficiency (0–1,000 Scale)*

Rated measure	Baseline		Negative bias	
	NS	NNS	NS	NNS
Accentedness	594 (302)	528 (319)	548 (321)	581 (258)
Comprehensibility	781 (172)	700 (229)	687 (222)	717 (167)
Segmental errors	591 (278)	535 (287)	560 (279)	583 (269)
Intonation	618 (217)	603 (235)	651 (204)	630 (197)
Flow	503 (285)	478 (306)	591 (249)	563 (266)

There were two sets of key findings. First, there were significant teacher status $\times$ bias

interaction effects for the ratings of accentedness, $F(1, 23) = 9.20$, $p = .006$, η_p^2 (effect size)

= .29, the ratings of comprehensibility, $F(1, 23) = 11.63$, $p = .002$, $\eta_p^2 = .34$, and the ratings of

segmental errors, $F(1, 23) = 7.89$, $p = .01$, $\eta_p^2 = .26$. These interaction effects (examined below)

implied that the NS and NNS teachers in the baseline and negative bias groups differed in their

assessments of these speech dimensions. Second, there was only a significant effect of bias for

the ratings of intonation, $F(1, 23) = 9.07$, $p = .006$, $\eta_p^2 = .28$, and the ratings of flow, $F(1, 23) =$

35.39, $p < .0001$, $\eta_p^2 = .61$. A significant effect of bias, in the absence of any other significant

effects, reflected a consistent influence of negative bias on all teachers, regardless of their NS or

NNS status. The teachers in the negative bias group provided higher ratings (i.e., evaluating

intonation and flow as being more natural) by an average of 30–87 points and with medium-to-

strong effects (Cohen's $d = 0.64$–1.41), according to Plonsky and Oswald's (2014) guidelines

and as compared to the teachers in the baseline group. These findings for intonation and flow are

illustrated using the ratings of flow in Figure 1.
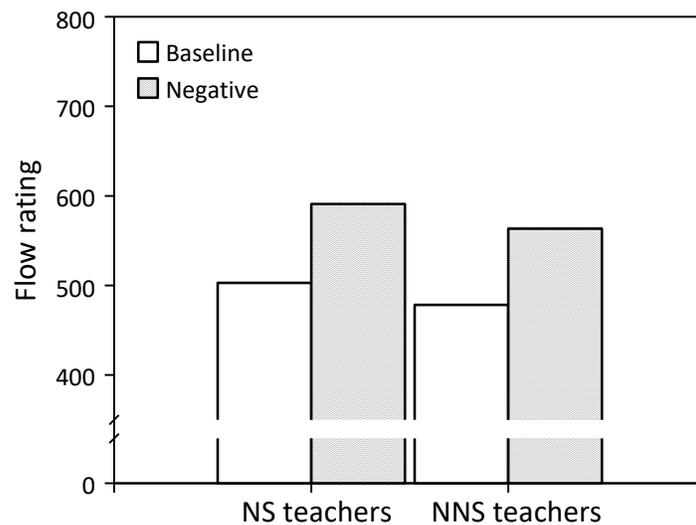
Figure 1. Mean ratings of flow for NS and NNS teachers evaluating the 24 L2 speakers

of German under negative bias or no bias (baseline).

To pinpoint the source of the significant interaction effects for the ratings of accentedness, comprehensibility, and segmental errors, these assessments were explored further using Bonferroni-adjusted pairwise comparisons. These comparisons (summarized in Table 5) revealed two rating patterns: First, in the baseline group (no bias), the NS teachers consistently rated the L2 speakers higher (i.e., less accented, more comprehensible, and speaking with fewer segmental errors) than the NNS teachers, by a mean of 56–82 points on the scale, which corresponded to medium-strength effect sizes ($d \geq 0.70$). Put differently, the NNS teachers were harsher in their assessments than the NS teachers when no social bias was imposed, suggesting that the NS and NNS teachers had different rating "baselines" for the speech they considered to be unaccented, comprehensible, and free from segmental errors.

Table 5 *Summary of Bonferroni Comparisons Across Teacher Groups*

Rated measure	Comparison	M_{diff}	P	D	95% CI
Accentedness	Baseline: NS > NNS	−66	.001*	0.81	[−101, −31]
	Negative: NS = NNS	+33	.231	0.24	[−22, 89]
	NS: Baseline > Negative	−47	.001*	0.72	[−73, −20]
	NNS: Baseline < Negative	+53	.042*	0.44	[2, 104]
Comprehensibility	Baseline: NS > NNS	−82	.002*	0.70	[−131, −33]
	Negative: NS = NNS	+31	.176	0.28	[−15, 76]
	NS: Baseline > Negative	−95	.001*	1.16	[−129, −61]
	NNS: Baseline = Negative	+17	.425	0.16	[−27, 63]
Segmental errors	Baseline: NS > NNS	−56	.001*	0.79	[−86, −26]
	Negative: NS = NNS	+23	.352	0.20	[−26, 71]
	NS: Baseline > Negative	−31	.034*	0.47	[−59, −3]
	NNS: Baseline < Negative	+48	.038*	0.45	[3, 93]
Intonation	Baseline < Negative	+30	.006*	0.64	[9, 51]
Flow	Baseline < Negative	+87	.001*	1.41	[56, 117]

Note. *p < .05 (Bonferroni-corrected).

Second, the NS and NNS teachers reacted differently to the negative bias. The NS teachers consistently downgraded the L2 speakers under negative bias for accentedness, comprehensibility, and segmental errors, by an average of 31–95 points on the scale. This effect was strong for comprehensibility ($d = 1.16$), medium for accentedness ($d = 0.72$) and small for segmental errors ($d = 0.47$). In contrast, the NNS teachers either upgraded the L2 speakers under negative bias for accentedness and segmental errors by an average of 48–53 points with

relatively small effects ($d = 0.45$) or remained unaffected by negative bias for comprehensibility. In the end, although the NS and NNS teachers seemed to have different rating baselines for these three speech dimensions (when no bias was imposed), they converged in their assessments by showing *opposite* reactions to the negative bias. These findings for the ratings of accentedness, comprehensibility, and segmental errors are illustrated using accentedness in Figure 2.
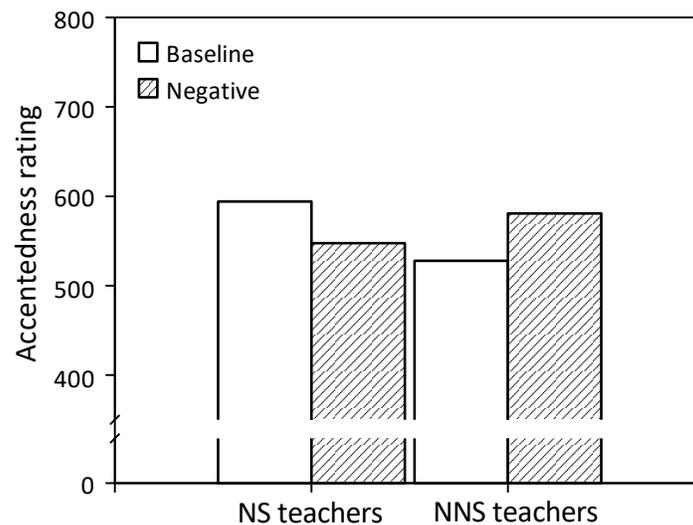

Figure 2. Mean ratings of accentedness for NS and NNS teachers evaluating the 24 L2 speakers of German under negative bias or no bias (baseline).

5. Discussion

This study's goal was to examine whether negative bias (introduced as a negative opinion about L2 German skills of a hypothetical learner of German) influenced NS and NNS teachers' evaluations of intermediate to advanced L2 German speech, relative to the ratings of the teachers not exposed to bias. For intonation and flow, the NS and NNS teachers rated L2 speech similarly when no bias was imposed, and both sets of teachers evaluated these dimensions more favorably, upgrading the speakers' performance under negative bias. However, the NS and NNS teachers

22

differed in their evaluations of L2 speakers' accentedness, comprehensibility, and segmental accuracy. When no negative bias was expressed, the NS teachers provided more lenient ratings for these dimensions, compared to those given by the NNS teachers. Yet, the two teacher groups differed in their response to negative bias: Whereas the NS teachers downgraded the performance of L2 speakers, the NNS teachers provided more favorable evaluations for the same L2 speakers.

5.1 L2 Speech Ratings and Listener Bias

With respect to the role of listener bias on teacher evaluations of L2 speech, the current findings revealed important interactions between bias and teacher status, notably, for the dimensions of accentedness, comprehensibility, and segmental (vowel and consonant) accuracy. The NS teachers "went along" with the negative bias and downgraded the L2 speakers in their evaluations of these three rated dimensions, relative to the ratings of teachers not exposed to a negative anecdote.[2] By contrast, the NNS teachers appeared to fight against the negative bias, upgrading their ratings (notably, for all five rated dimensions) of the same L2 speakers, compared to the assessments of the teachers who heard no negative anecdote. It is likely that the NNS teachers who heard negative comments about L2 students' German showed (enhanced) empathy with fellow L2 speakers. This behavior can be likened to that of the raters in the study by Hansen et al. (2014), for whom the requirement to use their L2 before evaluating the speech of other L2 speakers resulted in more favorable ratings, compared to those who did not use their L2 before assessment.

The teachers' debriefing comments revealed multiple examples of the NNS teachers who were exposed to negative bias expressing solidarity with L2 speakers, including their own students. For example, one NNS teacher pointed to the difficulties she had with her Spanish,

23

which she produces with a heavy English accent. Another NNS teacher indicated that she often works with her students on pronunciation because she is aware of her own accent in German. Yet another said that "[s]ome students don't have the ear for it." By contrast, the comments provided by the NS teachers in the negative bias condition showed their agreement with the negative anecdote. One NS teacher who heard the negative comments pointed to the need for some of the speakers to enroll in "remedial German," and another said that "learners really need to go to Germany to learn pronunciation." Finally, another NS teacher, who had commented on the debriefing questionnaire that she did not feel that the negative anecdote had affected her ratings, rhetorically asked "[s]hould learners like that even be able to study German?"

5.2 Sources of Teacher Bias

An interesting question arising from these findings concerns potential reasons for the teachers' rating behaviors. An enhanced empathy explanation for the more lenient ratings provided by the NNS teachers under negative bias is consistent with the assessments by the young (aged 18–40) English speakers from Montreal (Quebec) in Reid et al. (2019), where these participants (essentially untrained yet linguistically aware listeners) exposed to negative bias rated L2 English speech samples produced by native French speakers more favorably than did baseline participants of the same age. Young English speakers in French-speaking Quebec and NNS German teachers in a western (English-speaking) Canadian province share the status of being L2 speakers. Like the participants in Reid et al. (2019), the NNS teachers in this study, then, must regularly find themselves in contexts (especially in the workplace) where their status as L2 speakers of German is revealed when they speak. An enhanced awareness of one's own linguistic performance—particularly in the face of negative comments—may have enhanced the NNS teachers' solidarity with fellow L2 speakers, minimizing perceived teacher–student

distance and increasing the "self-other overlap" between themselves and the L2 German

students, which is in line with perspective-taking explanations of human behavior (Galinsky &

Moskowitz, 2000; Kawakami et al., 2012), and ultimately leads to more generous evaluations.

Or, perhaps the NNS teachers experienced a degree of cognitive dissonance (Visser & Cooper,

2007), realizing that going along with the negative bias manipulation and assessing L2 speakers

more harshly would be discrepant with their own status as L2 users. The NNS teachers then

acted in the opposite manner to reduce this perceived dissonance.

We could similarly speculate why the NS teachers were susceptible to the negative

manipulation in this study (at least for three of the five rated dimensions). A negative anecdote

targeting L2 speakers of German may have highlighted these NS teachers' status as in-group

members, casting L2 speakers in a negative light as out-group members, for example, as

imperfect speakers who would never reach high levels of linguistic performance (e.g., Tajfel,

1972). A negative anecdote may have also exacerbated the NS teachers' already existing,

preconceived ideas about the linguistic abilities of L2 speakers of German (e.g., Buckingham,

2014; Hu & Lindemann, 2009), especially in a context where German is not used outside

language classrooms. Alternatively, a negative anecdote may have made NS teachers more aware

of student errors or infidelities in their speech, relative to their expectations of what German

should sound like, and consistent with the idea that negative moods are associated with more

attention to detail (e.g., Beukeboom, & Semin, 2006). Regardless of the explanation, the NS

teachers were susceptible to the negative manipulation, in that they tended to evaluate L2

speakers more harshly.

It is important to note that the researcher who led the experiment was both a L2 speaker

of German and a fellow German teacher. Based on the concept of relational trust (Bryk &

Schneider, 2003), the teachers, in particular, might have been expected to align with the fellow teacher. For example, if she (as leader of the experiment) expressed negativity about the speakers, the other raters might have thought it appropriate to agree, even if only in the context of this particular rating task (Spillane, 2006; Spillane, Hallett, & Diamond, 2003). While the NS teachers indeed followed this pattern and agreed with the experimenter, the NNS teachers "pushed back" in the face of critical comments, which suggests that the negativity expressed by the person in charge has the potential to force bias in opposite directions, depending on who is listening. In sum, the NNS teachers' response may have been driven along the dimension of shared (L2) language status with the experimenter; in contrast, the NS teachers' response may have instead patterned along the dimensions of shared professional status with the experimenter.

5.3 Relationship Between Teacher Status and Negative Bias Varies Across Speech Dimensions

One noteworthy outcome of this research is that the relationship between listener bias and teachers' NS–NNS status (discussed in the previous two sections) appears to apply more readily to some rated speech dimensions than to others. The clearest evidence for how NS and NNS teachers reacted differently to a biasing anecdote was found for the ratings of accentedness, comprehensibility, and segmental accuracy. In contrast, the two teacher groups reacted similarly in their ratings of intonation and flow. Besides highlighting the nuanced nature of bias effects for teachers from different background profiles, these findings imply that different facets of a speaker's performance likely map onto different linguistic and experiential as well as social and attitudinal factors for the listener. For instance, in prior work, accentedness, comprehensibility, and segmental accuracy have been described as intuitive and conceptually simple for listeners to evaluate (as shown through high rating reliability values), compared to intonation and word

stress (Isaacs &Trofimovich, 2012). Therefore, it may well be that all teachers in this study—regardless of their NS–NNS status—had difficulty separating intonation and flow from the other measures because these dimensions are simply harder to evaluate, which minimized rating differences between NS and NNS teachers. One teacher, for example, noted that "it is difficult to rate intonation because the speakers are so unsure of themselves and stop speaking," which implies that intonation and flow were interrelated for this teacher. Two others brought up specifically their difficulty with rating intonation, and another mentioned the variability in speech rate that most certainly plays a role in ratings of flow. These insights are consistent with research showing that even experienced teachers have trouble describing specific pronunciation issues precisely (Isaacs & Trofimovich, 2012; Isaacs & Thomson, 2013).

Similarly, different facets of speaking performance might also carry varying degrees of social and attitudinal significance for the listener. For example, in a recent study conducted within the world Englishes paradigm that focused on listeners' attitudes towards speakers of various language varieties, Hansen Edwards, Zampini, and Cunningham (2019) showed that listener evaluations of speakers' fluency patterned with their ratings of speaker traits such as leadership, self-confidence, social status, and friendliness, whereas listener evaluations of speakers' accentedness and comprehensibility (i.e., speech traits) were evaluated differently from these other dimensions. Thus, it could be that NS and NNS teachers, especially when exposed to negative bias, might be reacting differently when evaluating speech traits as opposed to speaker traits. For instance, teachers' reactions may not differ along the NS–NNS status for such rated dimensions as fluency (and ostensibly also prosody as it captures some aspects of speech melody and thus flow) and speaker traits such as leadership and self-confidence, but may differ for such speech traits as accentedness, comprehensibility, and segmental accuracy. Needless to say, these

differences across rated dimensions of L2 speaking performance must be revisited in future work.

5.4 Uneven Baselines in the Absence of Negative Bias

Another outcome of this research is that the NS and NNS teachers appeared to demonstrate different "baselines" for what they considered to be the relevant speech ratings in the absence of negative bias. Put simply, the two teacher groups gave the L2 speakers similar ratings for intonation and flow, yet provided different assessments for accentedness, comprehensibility, and segmental accuracy, with NS teachers actually being more lenient than NNS teachers (cf. Figures 1 and 2, and see Table 2). These findings reflect the complex pattern of results previously reported for NS versus NNS listener evaluations of various dimensions of L2 speech. The findings for intonation and flow patterned with the results of prior work showing few differences between the assessments given by NS and NNS listeners (Crowther et al., 2016; Derwing & Munro, 2013). In contrast, the findings for accentedness, comprehensibility, and segmental errors aligned with prior findings demonstrating that NNS listeners provide harsher ratings compared to those provided by NS listeners (e.g., Fayer & Krasinski, 1987; Kang, 2012). Drawing definitive conclusions regarding potential sources of these NS–NNS listener differences in speech assessment (in the absence of any bias imposed) might be premature based on these results—until further, more systematic research is conducted—especially because prior studies, including this one, differ widely in the expertise and linguistic background of listeners, in the target language and proficiency of speakers, and most crucially in the methods and materials of speech assessment (see Winke, 2013). By way of summary, however, as counterintuitive as it sounds, the NS and NNS teachers—because they reacted to negative bias in the opposite ways—in the end were fully comparable in their assessments of accentedness, comprehensibility, and

segmental errors of the same L2 speakers. This raises an intriguing possibility, to be explored in future research, that making raters (including teachers) aware of desirable and perhaps even undesirable social biases may, in fact, serve rater training and rater calibration purposes.

6. Implications and Conclusion

Previous research has relied on teachers for their expert judgments of their students' performance (Anderson-Hsieh et al., 1992; Bongaerts et al., 1997; Rossiter, 2009). Nonetheless, teachers' judgments may be swayed by other teachers who hold leadership roles (Wheelan & Kaeser, 1997), and teachers' judgements may be affected by the real (or presumed) language ability of their students (Ford, 1984; Sokolov, 2014). The current results demonstrate an influence of social bias on teachers, a bias that differs in nontrivial ways in its influence on NS versus NNS teachers. The current findings also reveal that NS and NNS teachers differ in their baseline ratings performed in the absence of any bias.

These findings have important implications beyond the laboratory setting. Language teachers evaluate L2 learners on a regular basis. Sometimes they carry out high-stakes evaluations that may determine, for example, whether a L2 learner may study or work in a given target language setting. In such instances, it is essential that teachers be aware of both their own biases and the ways in which their assessments may be affected by comments provided by others. Importantly, too, employing multiple raters to carry out assessments in high-stakes settings may safeguard against the biases of individual raters.

Future research may probe the causes of teacher bias and determine whether training sessions designed to eliminate bias can be effective for all teachers equally or whether, for example, NNS teachers require different training from NS teachers. Furthermore, as the L2 learners in this study were of intermediate to advanced proficiency, an investigation of lower-

proficiency speakers might lend new perspectives as to the scope of teacher biases. It would also be valuable to explore whether the NS–NNS status of the researcher plays a role in the ratings assigned by participants. For now, a prudent take-home message arising from this research is that language teachers, as members of their respective sociolinguistic groups, are not immune to social biases that may affect the evaluations of their students.

Notes

1. An anonymous reviewer aptly pointed out that examining effects of negative bias on teachers' assessments might be more appropriate as a function of the social groups in which teachers claim membership, rather than based on teachers' NS–NNS status. In this study, the teachers self-identified as members of seven different groups, including German (12), German and Canadian (6), Anglophone Canadian (5), Anglophone Canadian and German (2), Latino (1), Chinese Canadian (1), and Swiss Canadian (1), which made it impossible to systematically investigate differences in ratings across multiple groups of unequal size.

2. Although the teachers assigned to different groups did not differ significantly in any of the background characteristics measured here (see Table 1), the NS teachers assigned to negative bias appeared to be slightly older and also to use German more, compared to the teachers in the remaining groups. As pointed out by an anonymous reviewer, these background variables may have contributed to NS teachers' harsher ratings under negative bias, which warrants future, systematic investigations of how teachers' background variables contribute to their speech assessments under negative bias.

7. References

Anderson-Hsieh, J., Johnson, R., & Koehler, K. (1992). The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentals, prosody, and syllable

structure. *Language Learning*, *42*, 529–555.

Beukeboom, C. J., & Semin, G. R. (2006). How mood turns on language. *Journal of Experimental Social Psychology*, *42*, 553–566. https://doi.org/10.1016/j.jesp.2005.09.005

Bodenhausen, G. V., Mussweiler, T., Gabriel, S., & Moreno, K. N. (2001). Affective influences on stereotyping and intergroup relations. In J. P. Forgas (Ed.), *Handbook of affect and social cognition* (pp. 319–343). Mahwah, NJ: Lawrence Erlbaum.

Bongaerts, T., van Summeren, C., Planken, B., & Schils, E. (1997). Age and ultimate attainment in the pronunciation of a foreign language. *Studies in Second Language Acquisition*, *19*, 447–465. https://doi.org/10.1017/S0272263197004026

Boughton, Z. (2006). When perception isn't reality: Accent identification and perceptual dialectology in French. *Journal of French Language Studies*, *16*, 277–304. https://doi.org/10.1017/S0959269506002535

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*, 707–729. https://doi.org/10.1016/j.cognition.2007.04.005

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing*, *12*, 1–15. https://doi.org/10.1177/026553229501200101

Bryk, A. S., & Schneider, B. (2003). Trust in schools: A core resource for school reform. *Educational Leadership*, *60*, 40–45.

Buckingham, L. (2014). Attitudes to English teachers' accents in the Arabian Gulf. *International Journal of Applied Linguistics*, *24*, 50–73. https://doi.org/10.1111/ijal.12058

Carey, M. D., Mannell, R. H., & Dunn, P. K. (2011). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing*, *28*,

201–219. https://doi.org/10.1177/0265532210393704

Cargile, A. C., & Giles, H. (1997). Understanding language attitudes: Exploring listener affect and identity. *Language & Communication*, *17*, 195–217. https://doi.org/10.1016/S0271-5309(97)00016-5

Chiba, R., Matsuura, H., & Yamamoto, A. (1995). Japanese attitudes toward English accents. *World Englishes*, *14*, 77–86. https://doi.org/10.1111/j.1467-971X.1995.tb00341.x

Crowther, D., Trofimovich, P., & Isaacs, T. (2016). Linguistic dimensions of second language accent and comprehensibility. *Journal of Second Language Pronunciation*, *2*, 160–182. https://doi.org/10.1075/jslp.2.2.02cro

D'Onofrio, A. (2018). Controlled and automatic perceptions of a sociolinguistic marker. *Language Variation and Change*, *30*, 261–285. https://doi.org/10.1017/S095439451800008X

D'Onofrio, A. (forthcoming). Sociolinguistic signs as cognitive representations. In L. Hall-Lew, E. Moore, & R. J. Podesva (Eds.), *Social meaning and linguistic variation*. Cambridge, UK: Cambridge University Press.

Dailey-O'Cain, J. (1999). The perception of post-unification German regional speech. In D. R. Preston (Ed.), *Handbook of perceptual dialectology* (Vol 1, pp. 227–242). Philadelphia, PA: John Benjamins.

Dalton-Puffer, C., Kaltenböck, G., & Smit, U. (1997). Learner attitudes and L2 pronunciation in Austria. *World Englishes*, *16*, 115–128. https://doi.org/10.1111/1467-971X.00052

Derwing, T. M., & Munro, M. (2013). The development of L2 oral language skills in two L1 groups: A 7-year study. *Language Learning*, *63*, 163–185. https://doi.org/10.1111/lang.12000

Derwing, T. M., & Munro, M. J. (2015). *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research*. Philadelphia, PA: 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*, 655–679. https://doi.org10.1111/j.1467-9922.2004.00282.x /

Eskenazi, M. (2009). An overview of spoken language technology for education. *Speech Communication*, *51*, 832–844. https://doi.org/10.1016/j.specom.2009.04.005

Fayer, J. M., & Krasinski, E. (1987). Native and nonnative judgments of intelligibility and irritation. *Language Learning*, *37*, 313–326. https://doi.org/10.1111/j.1467-1770.1987.tb00573.x

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology*, *50*, 229–238. https://doi.org/10.1037/0022-3514.50.2.229

Foote, J. A., & Trofimovich, P. (2018). Is it because of my language background? A study of language background influence on comprehensibility judgments. *Canadian Modern Language Review*, *74*, 253–278. https://doi.org/10.3138/cmlr.2017-0011

Ford, C. E. (1984). The influence of speech variety on teachers' evaluation of students with comparable academic ability. *TESOL Quarterly*, *18*, 25–40. https://doi.org/10.2307/3586333

Galinsky, A. D., & Moskowitz, G. B. (2000). Perspective-taking: Decreasing stereotype expression, stereotype accessibility, and in-group favoritism. *Journal of Personality and Social Psychology*, *78*, 708–724. https://doi.org/10.1037/0022-3514.78.4.708

Goethe Institut (2004). Einstufungstest [Placement test]. http://www.goethe.de/cgi-

bin/einstufungstest/einstufungstest.pl

Hansen Edwards, J. G., Zampini, M. L., & Cunningham, C. (2018). The accentedness, comprehensibility, and intelligibility of Asian Englishes. *World Englishes*, *37*, 538–557. https://doi.org/10.1111/weng.12344

Hansen Edwards, J. G., Zampini, M. L., & Cunningham, C. (2019). Listener judgments of speaker and speech traits of varieties of Asian English. *Journal of Multilingual and Multicultural Development*, *40*, 691–706. https://doi.org/10.1080/01434632.2018.1549057

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. https://doi.org/10.1177/0261927X13499761

He, D., & Miller, L. (2011). English teacher preference: The case of China's non-English-major students. *World Englishes*, *30*, 428–443. https://doi.org/10.1111/j.1467-971X.2011.01716.x

Hu, G., & Lindemann, S. (2009). Stereotypes of Cantonese English, apparent native/non-native status, and their effect on non-native English speakers' perception. *Journal of Multilingual and Multicultural Development*, *30*, 253–269. https://doi.org/10.1080/01434630802651677

Hu, G., & Su, J. (2015). The effect of native/non-native information on non-native listeners' comprehension. *Language Awareness*, *24*, 273–281. https://doi.org/10.1080/09658416.2015.1077853

Hundt, M., Zipp, L., & Huber, A. (2015). Attitudes in Fiji towards varieties of English. *World Englishes*, *34*, 688–707. https://doi.org/10.1111/weng.12160

Isaacs, T., & Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2

pronunciation: Revisiting research conventions. *Language Assessment Quarterly*, *10*,

135–159. https://doi.org/10.1080/15434303.2013.769545

Kang, O. (2012). Impact of rater characteristics and prosodic features of speaker accentedness on

ratings of international teaching assistants' oral performance. *Language Assessment

Quarterly*, *9*, 249–269. https://doi.org/10.1080/15434303.2011.642631

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, 441–456. https://doi.org/10.1177/0261927X09341950

Kawakami, K., Phills, C. E., Greenwald, A. G., Simard, D., Pontiero, J., Brnjas, A., & Dovidio,

J. F. (2012). In perfect harmony: Synchronizing the self to activated social categories.

Journal of Personality and Social Psychology, 102, 562–575.

https://doi.org/10.1037/a0025970

Lambert, W. E., Hodgson, R. C., Gardner, R. C., & Fillenbaum, S. (1960). Evaluational reactions

to spoken languages. *The Journal of Abnormal and Social Psychology*, 60, 44–51.

https://doi.org/10.1037/h0044430

Lappin-Fortin, K., & Rye, B. J. (2014). The use of pre-/posttest and self-assessment tools in a

French pronunciation course. *Foreign Language Annals*, *47*, 300–320.

Lindemann, S. (2003). Koreans, Chinese or Indians? Attitudes and ideologies about non-native

English speakers in the United States. *Journal of Sociolinguistics*, *7*, 348–364.

https://doi.org/10.1111/1467-9481.00228

Major, R. C., Fitzmaurice, S. F., Bunta, F., & Balasubramanian, C. (2002). The effects of non-

native accents on listening comprehension: Implications for ESL assessment. *TESOL*

Quarterly, *36*, 173–190.

Montgomery, C. (2007). *Northern English dialects: A perceptual approach* (Unpublished PhD
dissertation). University of Sheffield, Sheffield, United Kingdom.

Munro, M. J., Derwing, T. M., & Morton, S. L. (2006). The mutual intelligibility of L2 speech.
Studies in Second Language Acquisition, *28*, 111–131.
https://doi.org/10.1017/S0272263106060049

Norton, B., & Toohey, K. (2011). Identity, language learning, and social change. *Language
Teaching*, *44*, 412–446. https://doi.org/10.1017/S0261444811000309

O'Brien, M. G. (2014). L2 learners' assessments of accentedness, fluency, and comprehensibility
of native and nonnative German speech. *Language Learning*, *64*, 715–748.
https://doi.org/10.1111/lang.12082

Pantos, A. J., & Perkins, A. W. (2013). Measuring implicit and explicit attitudes toward foreign
accented speech. *Journal of Language and Social Psychology*, *32*, 3–20.
https://doi.org/10.1177/0261927X12463005

Preston, D. (1999). *Handbook of perceptual dialectology 1*. Philadelphia, PA: John Benjamins.
https://doi.org/10.1075/z.hpd1

Reid, K. T., 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*, 419–442.
https://doi.org/10.1017/S0272263118000244

Rose, R. L. (2017). Differences in second language speech fluency ratings: Native versus
nonnative listeners. In L. Degand et al. (Eds.), *Proceedings of the International
Conference "Fluency & Disfluency Across Languages and Language Varieties"* (pp.

101–103). Louvain-la-Neuve, Belgium: Université catholique de Louvain.

Rossiter, M. J. (2009). Perceptions of L2 fluency by native and non-native speakers of English. *Canadian Modern Language Review*, *65*, 395–412. https://doi.org/10.3138/cmlr.65.3.395

Rubin, D. L. (1992). Nonlanguage factors affecting undergraduates' judgments of nonnative English-speaking teaching assistants. *Research in Higher Education*, *33*, 51–531. https://doi.org/10.1007/BF00973770

Ryan, E. B., Carranza, M. A., & Moffie, R. W. (1977). Reactions toward varying degrees of accentedness in the speech of Spanish-English bilinguals. *Language and Speech*, *20*, 267–273.

Ryan, E. B., Hewstone, M., & Giles, H. (1984). Language and intergroup attitudes. In J. R. Eiser (Ed.), *Attitudinal judgment* (pp. 135–158). New York: Springer.

Sokolov, C. (2014). Self-evaluation of rater bias in written composition assessment. *Linguistica*, *54*, 261–275. https://doi.org/10.4312/linguistica.54.1.261-275

Spillane, J. P. (2006). *Distributed leadership*. San Francisco, CA: Jossey-Bass. https://doi.org/10.1093/obo/9780199756810-0123

Spillane, J. P., Hallett, T., & Diamond, J. B. (2003). Forms of capital and the construction of leadership: Instructional leadership in urban elementary schools. *Sociology of Education*, *76*, 1–17. https://doi.org/10.2307/3090258

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. https://doi.org/10.1037/0022-3514.69.5.797

Stewart, M. A., Ryan, E. G., & Giles, H. (1985). Accent and social class effects on status and solidarity evaluations. *Personality and Social Psychology Bulletin*, *11*, 98–105.

https://doi.org/10.1177/0146167285111009

Tajfel, H. (1972). Experiments in a vacuum. In J. Israel & H. T. Triandis (Eds.), *The context of social psychology: A critical assessment* (pp. 69–119). London, UK: Academic Press.

Tan, Y.-Y., & Castelli, C. (2013). Intelligibility and attitudes: How American and Singapore English are perceived around the world. *English World-Wide*, *34*, 177–201. https://doi.org/10.1075/eww.34.2.03tan

Isaacs, T., & Trofimovich, P. (2012). Deconstructing comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings. *Studies in Second Language Acquisition*, *34*, 475–505. https://doi.org/10.1017/S0272263112000150

van Bezooijen, R. (2002). Aesthetic evaluation of Dutch: Comparisons across dialects, accents and languages. In D. Long & D. Preston (Eds.), *Handbook of perceptual dialectology* (Vol. 2, pp. 13–30). Amsterdam, The Netherlands: John Benjamins. https://doi.org/10.1075/z.hpd2.07bez

Visser, P. S., & Cooper, J. (2007). Attitude change. *The Sage handbook of social psychology* (pp. 197–218). Thousand Oaks, CA: Sage. https://doi.org/10.4135/9781848608221.n9

Vissers, C. T. W. M., Virgillito, D., Fitzgerald, D. A., Speckens, A. E. M., Tendolkar, I., Van Oostrom, I., & Chwilla, D. J. (2010). The influence of mood on the processing of syntactic anomalies: Evidence from P600. *Neuropsychologia*, *48*, 3521–3531.

Watson, K., & Clark, L. (2015). Exploring listeners' real-time reactions to regional accents. *Language Awareness*, *24*, 38–59, https://doi.org/10.1080/09658416.2014.882346

Wheelan, S., & Kaeser, R. M. (1997). The influence of task type and designated leaders on developmental patterns in groups. *Small Group Research*, *28*, 94–121. https://doi.org/10.1177/1046496497281004

Winke, P. (2013). Rating oral language. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1–7). Hoboken, NJ: Blackwell. https://doi.org/10.1002/9781405198431.wbeal0993

Winke, P., & Gass, S. (2013). The influence of second language experience and accent familiarity on oral proficiency rating: A qualitative investigation. *TESOL Quarterly*, *47*, 762–789. https://doi.org/10.1002/tesq.73

Winke, P., Gass, S., & Myford, C. (2013). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing*, *30*, 231–252. https://doi.org/10.1177/0265532212456968

Xu, W., Wang, Y., & Case, R. E. (2010). Chinese attitudes towards varieties of English: A pre-Olympic examination. *Language Awareness*, *19*, 249–260. https://doi.org/10.1080/09658416.2010.508528

Yao, Z., Saito, K., Trofimovich, P., & Isaacs, T. (2013). Z-Lab [Computer software]. https://github.com/ZeshanYao/Z-Lab

Correspondence concerning this article should be addressed to Kym Taylor Reid, Department of Education, Concordia University (FG 5.150), 1455 de Maisonneuve Blvd W., Montreal, Quebec, Canada H3G 1M8. Email: kym.taylor@concordia.ca

The second, third, and fourth authors' addresses are included below:

Mary O'Brien, School of Languages, Linguistics, Literatures and Cultures, University of Calgary, C216 Craigie Hall, 2500 University Drive NW, Calgary, Alberta, CANADA T2N 1N4

Email: mgobrien@ucalgary.ca

Pavel Trofimovich, Department of Education, Concordia University (FG 5.150), 1455 de

Maisonneuve Blvd W., Montreal, Quebec, Canada H3G 1M8

Email: pavel.trofimovich@concordia.ca

Allison Bajt, School of Languages, Linguistics, Literatures and Cultures, University of Calgary,

C216 Craigie Hall, 2500 University Drive NW, Calgary, Alberta, CANADA T2N 1N4

Email: ambajt@ucalgary.ca

Appendix A

Social Attitudes Questionnaire

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

		Disagree							**Agree**	
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
6.	The ability to speak my ethnic language is important in defining my personal identity.	1	2	3	4	5	6	7	8	9
7.	The ability to speak German is important in defining my personal identity.	1	2	3	4	5	6	7	8	9
8.	Speaking German also means embracing German culture.	1	2	3	4	5	6	7	8	9
9.	People who speak German are more prepared to live in today's world.	1	2	3	4	5	6	7	8	9
10.	People who speak German are more tolerant of other groups.	1	2	3	4	5	6	7	8	9
11.	I take pride in teaching German.	1	2	3	4	5	6	7	8	9

Appendix B

Screenshots of the Rating Interface
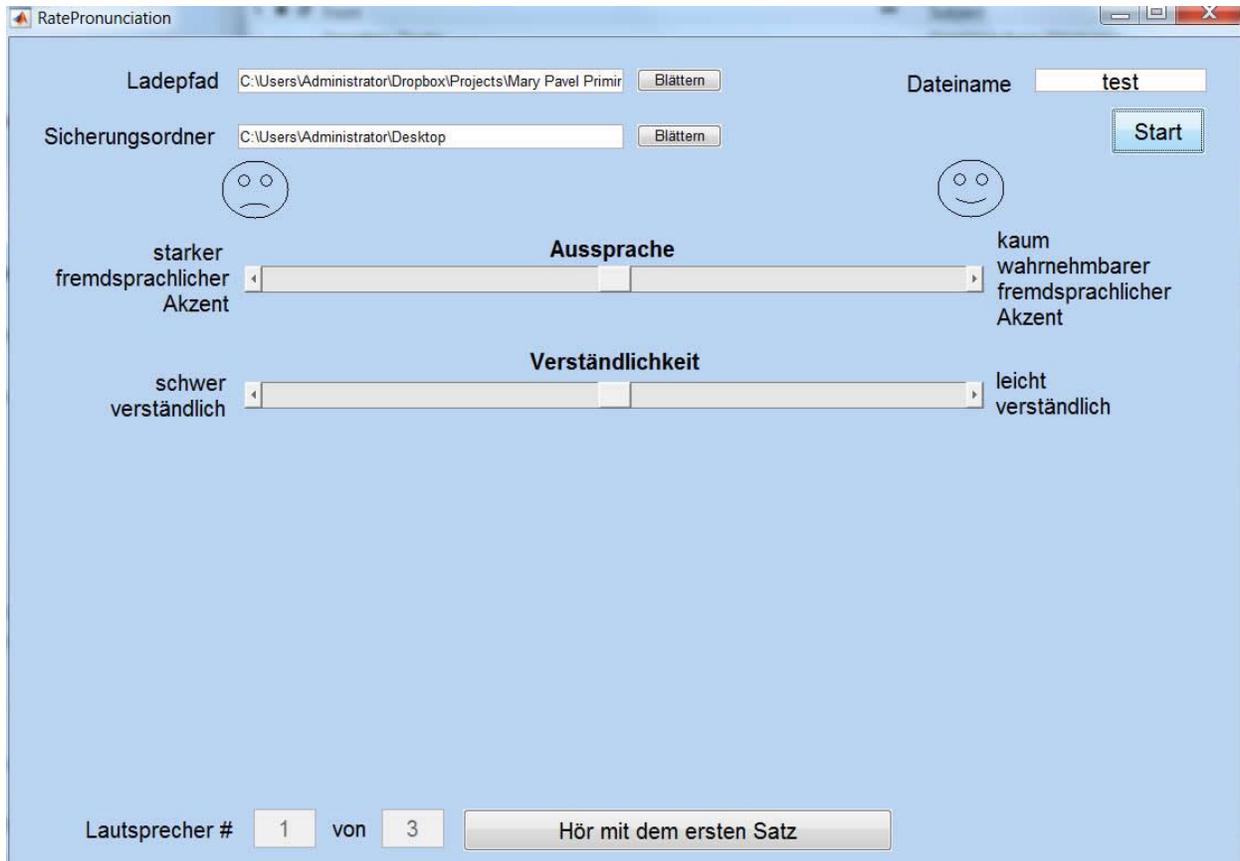
Screenshot of the interface for the rating of accentedness and comprehensibility.
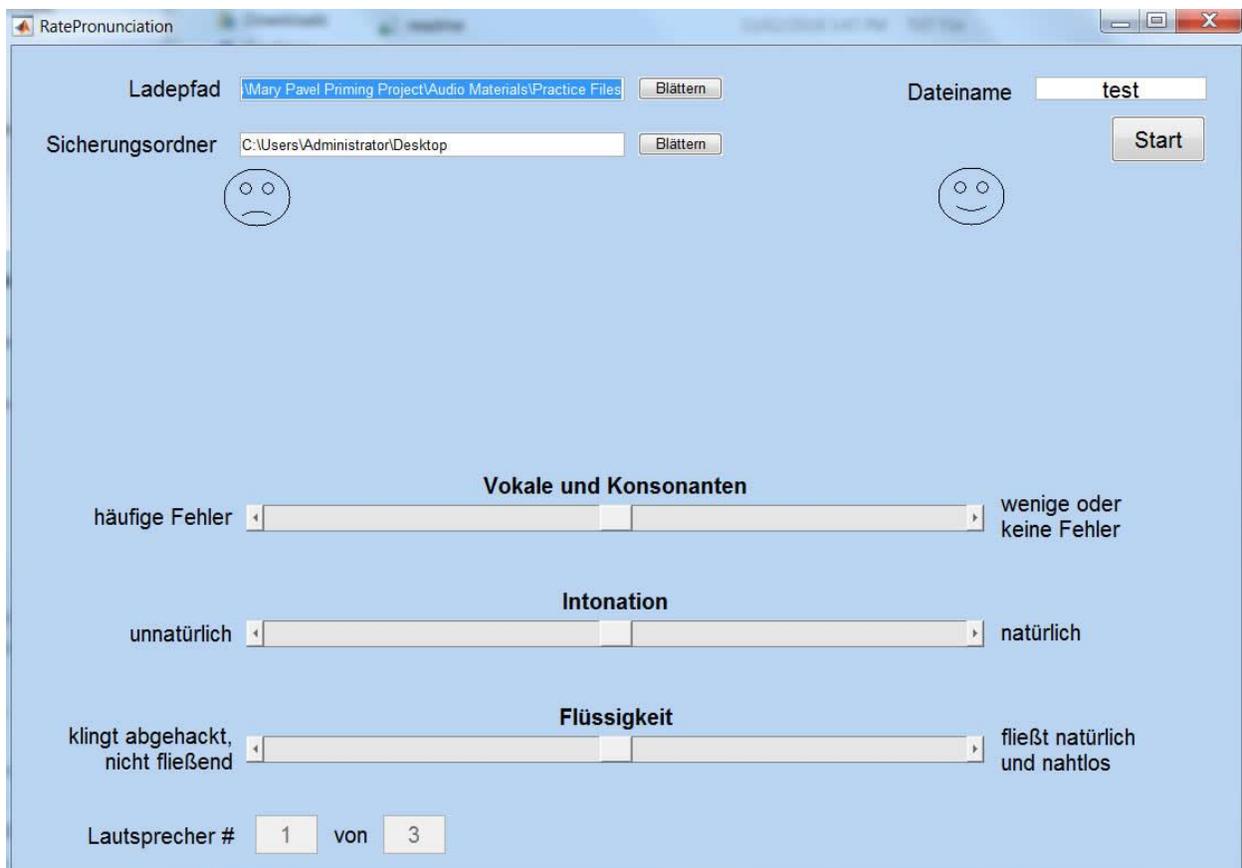
Screenshot of the interface for the rating of segmental errors (vowels and consonants), intonation, and flow.

Appendix C

Script for the Negative Social Bias Manipulation

Negative

I'm sorry. I just have to vent about something. You also teach German, so you understand my situation. I was just chatting with an undergraduate student majoring in German, and I honestly couldn't believe how little she could do in the language. I mean, you'd think that when you've nearly completed a degree in a language, you'd be able to do more with it. I could barely understand her. Her accent was awful, and her grammar didn't even make sense. You know, if you major in German you should be able to use it with anyone, especially fellow German learners. I can't believe that some German majors don't even bother to become fluent. It's ridiculous.

Appendix D

Final Debrief Questionnaire

Please rate your experience in today's session by putting an X in the appropriate spot on the scale.

1. How pleasant was your experience in this rating session?

Very unpleasant ●———————————————————● Very pleasant

2. How helpful was the researcher during the session?

Not helpful at all ●———————————————————● Very helpful

3. How difficult was the rating task for you?

Very difficult ●———————————————————● Very easy

4. How confident are you in your ratings?

Not confident at all ●———————————————————● Fully confident

5. Did any part of your interaction with the researcher affect your ratings?

_____

_____

_____

_____