Reactions to second language speech: Influences of discrete speech characteristics, rater experience, and speaker first language background

Talia Isaacs (University College London) & Ron I. Thomson (Brock University)

Isaacs, T., & Thomson, T. (2020, in press). Reactions to second language speech: Influences of discrete speech characteristics, rater experience, and speaker first language background. *Journal of Second Language Pronunciation*.

Abstract

This study investigates how Mandarin and Slavic language speakers' comprehensibility, accentedness, and fluency ratings, as assigned by experienced teacher-raters and novice raters, align with discrete linguistic measures, and raters' accounts of influences on their scoring. In addition to examining mean ratings in relation to rater experience and speaker first language background, we correlated ratings with segmental, prosodic, and temporal measures. Introspective reports were segmented, coded, enumerated, and submitted to loglinear analysis to elucidate influences on ratings. Results showed that ratings were strongly correlated with prosodic goodness and moderately correlated with segmental errors, implying the importance of both segmentals and prosody in L2 speech ratings. Experienced teacher-raters provided lengthier reports than novice raters, producing more comments for all coded categories where an error was identified except for pausing (a dysfluency marker). This may be because novice raters observed little else about the speech or struggled to pinpoint or articulate other features.

Keywords: Accent, Comprehensibility, English as a Second Language, Fluency, Pronunciation Assessment, Raters, Rating Scales, Speech Perception

1. Introduction

A growing body of second language (L2) pronunciation research examining global perceptual constructs (e.g., comprehensibility, accentedness, fluency) in relation to discrete linguistic measures (e.g., segmental accuracy, temporal measures) has exerted a sustained influence on L2 speaking assessment research over the past decade (Isaacs & Harding, 2017). If we accept the view that both speakers and listeners play a role in successfully exchanging oral messages (Schiavetti, 1992) and share communicative responsibility (Rajadurai, 2007), a few points logically follow. This includes needing to better understand what features of L2 speech are salient to different types of listeners. We also need to examine whether listeners' beliefs about which linguistic features inform their assessments match what is actually present in learner speech.

In traditional L2 pronunciation research, ratings of global perceptual constructs are often measured using 9-point numerical scales, with brief, relativistic descriptors anchoring the scales on each end (e.g., no accent/extremely strong accent; Derwing & Munro, 1997). These scales have the advantage of being user-friendly, jargon-free, and accessible to raters who may lack specialist knowledge of pronunciation. Further, ratings obtained using these Likert-type scales consistently yield high interrater reliability across studies, even without listener training (Munro, 2018). However, such scales provide raters with little guidance on how to interpret score levels. Even if there is exact rater agreement on a score assigned to an L2 speaking performance, it does not *necessarily* follow that raters arrived at the same score for the same reasons or interpreted the constructs in the same way (Douglas, 1994). Indeed, a fundamental principle in psychometrics is that reliability is a prerequisite for construct validity but is an insufficient condition for it (Bannigan & Watson, 2009). Therefore, it is important to establish

what lies beneath listeners' impressionistic judgments and scoring decisions.

Variability is integral to the rating process, with ratings of speech involving both L2 learners and raters who vary on many characteristics (e.g., cognitive, attitudinal). Raters interact with the speech elicitation task and scoring system in different ways to generate a score (Upshur & Turner, 1999). If numerous deviations from native patterns were to co-occur in a speech sample, raters may tune into different constellations of deviations (Munro, 2018). They then need to filter their impressions through the artifact of a scoring system, with descriptors necessarily underrepresenting the complexity of performances (Lumley, 2005). Variability in L2 learner performance on the trait being measured is desirable, so that learners' ability levels can be differentiated and reflected in the scoring. The criteria that raters use to assign meaning to scale levels are important to investigate in research contexts, where, in contrast to many high-stakes assessment settings that use extended scale descriptors, raters receive scant guidance from rating scales and little rater training. Hence, they need to arrive at their own understanding of what the scale levels mean in terms of performance features during real-time scoring. To date, few L2 pronunciation studies have used introspective methods to probe listeners' accounts of influences on their scoring decisions. Derwing and Munro (2009) elicited listeners' written reports about preferences for L2 recorded voices, which had been pre-rated at different L2 comprehensibility and accentedness levels. Other researchers have used introspective reports to extend quantitative findings about the relationship between discrete linguistic measures and global L2 speech ratings (e.g., Foote & Trofimovich, 2018; Isaacs & Trofimovich, 2012).

The current study contributes to this emerging body of research, combining raters' verbalizations with other sources of evidence to illuminate their responses to L2 speech. More specifically, we analyze experienced teacher-raters' accounts compared to those of novice raters

(undergraduate students) and how their ratings align with linguistic measures derived from the L2 speech samples. Eliciting ratings from experienced teacher-raters and novice listeners in settings where English is used as a lingua franca is ecologically valid due to likely interactions involving L2 speakers inside and/or outside of the classroom (Rose & Galloway, 2019), although only teachers would likely formally assess their speech.

The variability associated with rater experience is not viewed as a threat to validity in this study (see Isaacs & Thomson, 2013, for a discussion of the rater experience construct in L2 pronunciation research). Rather, it is regarded as a rich source of information that allows reflection on our understanding of global constructs often examined in pronunciation research (Chalhoub-Deville, 1995). Listeners are by far the best resource for better understanding such constructs, which, by definition, relate to listener perceptions of L2 speech. Thus, examining listeners' interpretations of the focal constructs, listening and rating processes and strategies, and how their perceptions align with linguistic characteristics of spoken productions (e.g., word choice, grammar) is essential for better understanding the L2 abilities we are attempting to measure.

2. The Current Study

This study brings together insights from two disciplines: language testing research on systematic sources of variance in human scoring, and L2 pronunciation research on the linguistic properties underlying global perceptual constructs. The goal is to examine the linguistic variables that underlie comprehensibility, accentedness, and fluency ratings. We examine how listeners' ratings align with both discrete L2 speech measures (e.g., segmental error counts, speaking rate), and listener reports of linguistic features that they attend to, grouped by listener experience and speaker first language (L1) background variables. These

aims are distilled into the following research questions:

1. Which discrete L2 pronunciation and fluency measures are most related to listeners' global ratings of comprehensibility, accentedness, and fluency?
    - Does listener experience play a role?
    - Do learners' L1 backgrounds influence the listener?
2. How do listeners' perceptions of the linguistic influences on their judgments relate to these global L2 speech ratings?
    - Does listener experience play a role?
    - Do learners' L1 backgrounds influence the listener?

3. Method

3.1 Research design

In holistic rating, raters condense their impressions of a complex L2 performance into a single rating. Previous research has established that even highly trained raters may draw on different criteria to make scoring decisions, which may or may not be reflected in the scale descriptors (Lumley, 2005). Multiple sources of evidence were needed to elucidate this research problem. Therefore, a concurrent mixed methods design was used (Creswell & Plano Clark, 2017). To address the first research question, experienced teacher-raters' versus novice raters' global pronunciation and fluency ratings of L2 Mandarin and Slavic language speakers' utterances were statistically examined in relation to segmental, prosodic, and temporal measures. For research question two, an inductive coding scheme was generated from raters' introspective reports. The coded comments were then quantified and counts of coded categories for experienced teacher-raters versus novice listeners and Mandarin versus Slavic language speakers were obtained. The highest frequency codes were then subjected to quantitative

analysis to test for between-group differences.

3.2 L2 speakers

Speech samples were elicited from 38 adult newcomers to Canada (27 females, 11 males, M_{age} = 39.4 years; 29–52). Half were L1 Mandarin speakers, who reported first exposure to English at a mean age of 14.3 years (7.0) and had resided in Canada for 16.7 months on average (11.9). The other half were L1 Slavic speakers (13 Russian, 3 Serbo-Croatian, 2 Ukrainian, 1 Polish), whose first reported English exposure was at a mean age of 16.2 years (11.8), with 15.6 months' Canadian residency on average (10.7). All were assessed at beginner English levels on the Canadian Language Benchmarks (CLB levels 1-4 of the instrument; Pawlikowska-Smith, 2000) and were enrolled in the government-funded Language Instruction for Newcomers to Canada (LINC) program at the time of the study. Mandarin and Slavic language speakers were matched for proficiency level based on the English as a Second Language (ESL) class in which they were registered. Placement decisions had been based on both CLB level and results from an in-house English proficiency test, which assessed L2 grammatical and lexical knowledge, literacy skills, and aural/oral performance.

Table 1 shows Mandarin and Slavic language speakers' self-reported L2 English exposure and estimated proficiency levels, obtained from questionnaire items administered at the beginning of data collection. Mandarin learners estimated speaking and listening to English outside class a greater proportion of the time than did Slavic language speakers but perceived having extended conversations with L1 English speakers less often and assessed their overall proficiency at a lower level. However, none of these self-report measures were statistically significant, $t(36) = |.19–1.78|$ $p > .05$, suggesting that the L1 groups were matched on language-related variables.

Table 1. Mandarin and Slavic language speakers' reported English language exposure and proficiency

	L1 Mandarin			L1 Slavic	
Self-report measures	*M*	*SD*		*M*	*SD*
Percent of time speaking English outside of class	35.8	25.5		34.2	26.5
Percent of time listening to English language media	80.5	24.6		71.6	33.5
Number of extended conversations with native English speakers per week[a]	1.8	2.4		2.5	3.0
English listening/speaking proficiency[b]	3.9	1.5		4.7	1.3
English reading/writing proficiency[b]	4.9	1.8		5.2	1.3

Note: [a]An extended conversation was defined as ≥10 min; [b]Measured on a 9-point Likert-type scale (*1 = extremely poor, 9 = extremely proficient*).

3.3 Speech elicitation and data preparation

Speech samples were audio recorded on several speaking tasks in a quiet room using a Marantz PMD661 SD recorder (duration: ≤ 40 mins). This article will report on performance on one task, an eight-frame picture narrative often used to elicit adult ESL learners' extemporaneous speech samples in L2 pronunciation and fluency research (Derwing & Munro, 2013). The essential plot elements were the collision of a man and a woman carrying similar suitcases on the street, their retrieval of the wrong suitcase and eventual discovery that they had accidently exchanged suitcases. The speakers were given a minute to look over the visual prompt before describing the picture sequence. After normalizing the speech samples for peak amplitude and removing any dysfluencies that had preceded the storytelling (e.g., false starts,

hesitations), the first 20 seconds of each narrative were excised from the recordings and randomized in preparation for rating ($M_{duration}$ = 27.1 s; *SD* = 2.3). The speech sample of a male native English speaker was included about two thirds of the way through the set of recordings for all randomizations to verify that listeners' ratings corresponded to the correct speech sample in the printed response sheet. Once this was established, the native speaker's ratings were excluded from subsequent analyses.

3.4 Raters

Forty native English speakers, who reported having normal hearing, participated as raters. Half were experienced ESL teachers (14 females, 6 males; $M_{experience}$ = 9.7 years; *SD* = 5.1), who either held or were pursing graduate degrees in applied linguistics from a Canadian English-medium university. These experienced teacher-raters reported teaching ESL for 13.9 hrs/week on average before commencing their studies (*SD* = 8.47). However, they varied in their teacher training, with 13 having taken a pronunciation course for teachers, 16 an L2 assessment course, and two with no training in these areas. The remaining 20 raters (15 females, 5 males), henceforth referred to as novice raters, were pursuing graduate degrees in nonlinguistic disciplines (e.g., political science, law, epidemiology) and uniformly had no assessment training.

The raters indicated their age range from a list in a background questionnaire due to some raters' sensitivity about age reporting during piloting. The experienced teacher-raters were the older demographic, with two raters in their 20s, 10 in their 30s, five in their 40s, and three in the 50 years or over age category. In contrast, 15 novice raters were in their 20s and only five were over 30. As a precondition for participating, only raters who reported never having learned Chinese or Russian (the most common Slavic L1 in the study) and who did not have notable

exposure to members from either language community (e.g., through family relations, extended travel) could take part.

At the beginning of data collection, recruited raters were asked about their L1 accent familiarity in a background questionnaire (*1 = extremely unfamiliar*, *9 = extremely familiar*). They reported significantly greater familiarity with Mandarin speakers' English (M = 4.38, SD = 2.52) than that of Russian speakers (M = 3.28, SD = 2.21), $t(39) = 3.65$; $p = .001$, with significant effects retained when raters were broken down into experienced teacher, $t(19) = 2.44$; $p = .025$, and novice groups, $t(19) = 2.89$; $p = .009$. Table 2 shows that experienced teacher-raters reported interacting significantly more with L2 speakers as a proportion of their total time than did novice raters, $t(38) = 3.02$, $p < .005$. This is unsurprising, since teaching time was subsumed in experienced teacher-raters' estimates but was, by definition, absent from novice raters' estimates. Experienced teacher-raters also reported significantly greater exposure than novice raters to the English speech of both Mandarin learners, $t(38) = 3.15$, $p < .001$, and Slavic language speakers, $t(38) = 2.20$, $p < .002$.

Table 2. Experienced teacher-raters' and novice raters' self-reported mean interactions with L2 speakers and exposure to the L2 English of Mandarin and Slavic speakers

	Experienced		Novice	
Self-report measures	*M*	*SD*	*M*	*SD*
Percentage of time interacting with L2 speakers	39.0	16.83	22.5	17.73
Exposure to Mandarin-accented speech[a]	5.60	2.39	3.15	2.03
Exposure to Slavic-accented speech[a]	4.35	2.43	2.20	1.28

[a]Measured on a 9-point Likert-type scale (*1 =extremely familiar*, *9 = extremely unfamiliar*).

3.5 Rating sessions

The rating sessions were conducted individually in a quiet office, with a short break to mitigate rater fatigue (duration: ≤ 2 hrs). After hearing each speech sample, raters recorded scores on separate numerical scales for comprehensibility (very hard/very easy to understand), accentedness (heavily accented/not accented at all), and fluency (very dysfluent/very fluent), with descriptors at scale anchors. As part of a larger study examining rating scale length (Isaacs & Thomson, 2013), half of each rater group was arbitrarily assigned to either a 5-point or 9-point rating scale length condition. In order to establish a baseline understanding about the constructs they were rating, we provided raters with explicit definitional guidance. Comprehensibility was defined as how easy the L2 speech is to understand (Derwing & Munro, 1997); accentedness denoted how different the speech sounds from that of a native speaker of North American English (Isaacs & Thomson, 2013); and fluency referred to the smoothness and rapidness of the oral delivery, corresponding to Lennon's (1990) narrow sense of the term and reflecting temporal phenomena (e.g., speech rate, hesitations). After familiarizing raters with the speaking prompt and rating procedures, they received general feedback on their ratings of four practice items (2 native English speakers, 1 Mandarin speaker, 1 Slavic language speaker) based on comparisons with mean scores that had previously been assigned by an independent group of raters in Derwing, Thomson, and Munro (2006). Specifically, they were told whether their ratings were considerably harsher, considerably more lenient, or roughly the same compared to mean scores assigned by the previous group of raters. In all cases, the researcher highlighted that there were no right or wrong answers and raters were not directed to adjust their scoring as a result.

Introspective reports were elicited for the linguistic factors that experienced teacher-

raters and novice raters reportedly attended to when listening to and rating the speech (Gass & Mackey, 2000). Half of the raters in each rater group completed verbal protocols during their first listening. Procedurally, this involved the researcher pausing immediately after each recording so raters could articulate their thoughts while completing their ratings or reflecting on their scoring. If a halting silence occurred, the researcher prompted raters to continue verbalizing their thoughts with the probe, "what are you thinking?" However, raters were the ultimate arbiters of the amount of commentary they delivered, indicating when they were ready to proceed to the next recording using verbal or nonverbal signals (M_{duration} of listening, rating, and verbal protocols = 39 min and 34 min for experienced teacher-raters and novice raters respectively, range: 25–57 min). Because the additional cognitive demand of having raters verbalize their thoughts while scoring is not representative of rating procedures (Lumley, 2005), the other half of the raters provided scores without verbalizing their thoughts during their first listening. This was a timed condition, with a 7-second interval between speech samples (duration: 18 min).

Raters performed a second listening immediately after finishing their first set of ratings, consulting their scores. When the recording was paused, raters articulated what they remembered thinking about the rating process or their impressions of the speech. For half of the raters not in the verbal protocol condition described above, these delayed recalls were their only opportunity to comment on factors that had fed into their listening and scoring. However, the time lapse meant that the introspective reports were removed from their initial thought processes when rating (Ericsson & Simon, 1993). Finally, at the end of the session, all raters were interviewed about their scoring behavior, think-aloud experience, interpretations of the

constructs, and perceived influences on their judgments. The interview data are not discussed in this article.

3.6 Rating scale normalization

Table 3 shows the equivalencies that we used to scale the 9-point scale down to a 5-point scale in preparation for data analysis. Isaacs & Thomson (2013) found that rater consistency was similar across scale length condition, the distributions of rating outcomes for each rated measure were virtually identical, and rater preference for using 9- versus 5-point scales was mixed, with no rater consensus achieved. Therefore, we pooled ratings across scale length condition using the normalized scales.

Table 3. Original and normalized scales for comprehensibility, accentedness, and fluency ratings.

	Scale levels								
Original 5-point scales	1		2		3		4		5
Original 9-point scales	1	2	3	4	5	6	7	8	9
Normalized 9-point scales	1	1.5	2	2.5	3	3.5	4	4.5	5

3.7 Deriving discrete linguistic measures from the L2 speech samples

In order to examine the discrete pronunciation and fluency measures that most strongly relate to experienced teacher-raters' and novice raters' global ratings for the two learner groups, we obtained segmental, prosodic, and temporal measures from the speech. For speech segments, a phonetically-trained research assistant annotated orthographically transcribed recordings to indicate error locations and type, specifically vowel and consonant substitutions, deletions, and additions. When marking substitutions, the research assistant was told to ignore instances where

a non-English sound was substituted for English in a way that did not impact intelligibility (e.g., a trilled ‘r’ in place of an English ‘r’ was acceptable, as were palatalized fricatives in place of English ‘h’). The second author, a phonetician, then verified the annotated transcripts, noting any differences of opinion. He agreed with the assistant’s assessment in 93% of cases. There was greatest agreement on consonantal errors (97%), with less agreement on vowels (88%), which are notoriously ambiguous (McAndrews & Thomson, 2007). After considering each discrepancy, when consensus was not possible, the second author’s judgment stood. This only affected a few decisions related to vowels and one related to a consonant error. In most cases where there was disagreement, vowel productions were determined to be ambiguous and were subsequently accepted as correct. Previous studies have used blind randomized assessment of discrete speech tokens produced from a word list using a forced-choice decision task (e.g., Thomson & Isaacs, 2009). We did not feel that this approach was suitable for the current study, since unpredictable speech tokens arising in extemporaneous narratives were the focus rather than discrete items targeting specific sounds. In the final analysis, there were an average of 4.2 vowel errors (range: 0-10) and 4.6 consonant errors (range: 0 -10), per 20 second L2 speech sample.

We computed ratios of correctly pronounced segments over segmental incidence, tabulated separately for vowels and consonants in content versus function words. We distinguished between these word types because Zielinski’s (2008) in-depth analysis revealed little role for function words in intelligibility breakdowns. However, her study analyzed only three L2 learners’ speech samples. Further, Munro and Derwing (2006) provided evidence supporting the functional load hypothesis in relation to comprehensibility, albeit with the potential confound that in their stimuli, high functional load errors solely occurred in content words, which, by definition, are more consequential for meaning than function words.

Therefore, we examined error prevalence for vowels and consonants in content versus function words and related this to the mean L2 speech ratings. We also computed the percent of correctly pronounced segments in pruned content versus function words (i.e., with all dysfluencies removed), with vowels and consonants counted separately.

Prosody was captured by eliciting three pronunciation experts' prosodic goodness ratings using 9-point scales (1 = extremely non-native prosody; 9 = native-like prosody) following Derwing, Rossiter, Munro, and Thomson, 2004. The experts were L2 pronunciation researchers and teachers with phonetic training and at least 15 years' residence in the Canadian province where the speech samples had been collected. Cronbach's alpha was used confirm high internal consistency (.90) for the resulting prosodic goodness ratings. Drawing on Derwing et al. (2006), we examined two temporal measures using Sound Studio 3: (1) speaking rate, operationalized as the total number of uttered syllables over speech sample duration, and (2) pruned syllables per second, operationalized as the proportion of uttered syllables per second with all dysfluencies removed (e.g., self-repetitions, self-corrections). We used 400 milliseconds as the minimum threshold for counting silent pauses or fillers (see Derwing et al., 2004; Riggenbach, 1991).

3.8 Analysis of introspective reports

The verbal protocol and delayed recall data were orthographically transcribed and verified by a second researcher. Words with irregular pronunciation that raters had recalled or imitated from the speech samples were written with phonemic symbols or underlined for stress. To examine the linguistic aspects that experienced teacher-raters and novice raters reportedly attended to in Mandarin and Slavic language speakers' utterances, the first author inductively generated a coding scheme in an iterative process based on raters' verbatim comments. Twenty

verbal protocols and 20 delayed recalls were subjected to coding and enumeration so that only one set of comments per rater was included for each speech sample (i.e., their first think-aloud opportunity). Coded categories and subcategories included: (1) segmental errors, identifying, where possible, error type (epenthesis, substitution, deletion) and whether vowels or consonants were implicated; (2) word pronunciation difficulty, in which raters expressed difficulty with or a pronunciation irregularity of a lexical item, but the error source could not be identified from the rater's comment; (3) word stress; (4) pitch, intonation, or voice quality (including pleasant/strange voice); (5) rhythm and linking (e.g., smooth/choppy speech); (6) pausing and other hesitation markers, specifying whether the comment pertained to filled or unfilled pauses where possible; and (7) speech rate or pacing (fast/reasonable vs. slow/halting delivery). Positive and negative comments about categories 3 through 7 were tallied separately. The coding scheme also captured general comments about the global rated measures (comprehensibility, accentedness, fluency), speakers' presumed personality attributes extrapolated from the speech (e.g., confidence), and rater processes or strategies. Comments about storytelling ability, grammatical use, syntactic complexity, and lexical appropriateness were not included in the coding scheme, although Derwing and Munro (1997) and Isaacs and Trofimovich (2012) have shown that comprehensibility pertains to more than simply pronunciation and fluency phenomena. Because raters were not constrained in the length of their introspective reports and we used a balanced design, the recording time or number of words uttered was not controlled for in subsequent analyses.

After obtaining the research team's feedback on the coding scheme, the following refinements were made. Pronoun errors were interpreted as grammatical rather than lexical errors, self-repetition was classified under the pausing/hesitations category, and stuttering fell

under rhythm/linking. A second coder then applied the coding scheme to the data, recording frequencies separately for rater experience and speaker L1. Exact intercoder agreement was obtained 93% of the time for the main categories, with differences of opinion resolved through discussion. Discrepant codes were assigned, for example, when one coder interpreted “stops” to mean plosives, whereas a closer reading revealed that the rater was, in fact, referring to stops and starts. Comments about lexical retrieval difficulties resulting in dysfluency or inadequate information produced, which were a source of coding inconsistency, were ultimately assigned the pausing/hesitation code, except for instances when the rater directly referred to slow speech or processing as being an issue, in which case speech rate/pacing was selected. In ambiguous cases when an error type could not be classified based on the rater’s account, the audio recordings of the introspective reports were consulted to check the fidelity of the transcription and coding interpretation.

After finalizing the frequency counts, the five main coded categories that, together with subcategories, were most frequent in the data were submitted to loglinear analysis using SAS 9.4 GENMOD and CATMOD procedures. This yielded a crosstabulation of categorical variables using chi-square tests for statistical significance and maximum likelihood estimation (Stevens, 2009). All other statistical analyses were computed using SPSS 24.

4. Results

4.1 Preliminary analyses

Before addressing the research questions, we conducted three preliminary analyses. First, intraclass correlations for ratings of comprehensibility (.964), accentedness (.965), and fluency (.972) revealed high internal consistency. Next, an independent samples *t*-test, conducted to examine whether there were scoring differences for raters assigned to the verbal

protocol versus delayed recall conditions, which was an artifact of the research design, revealed no significant differences, $t(38) = |.01–1.38|$, $p > .05$. Therefore, we pooled ratings across introspective report conditions and ran Pearson correlations between the three global rated measures. The moderate to strong associations in Table 4 suggest that these constructs are related yet distinct.

Table 4. Correlations between L2 comprehensibility, accentedness, and fluency ratings

Rated measures	1	2	3
1 Comprehensibility			
2 Accentedness	.71**		
3 Fluency	.65**	.61**	

*$p \leq .05$, **$p \leq .01$, two-tailed

4.2 Rater experience and speaker L1 in relation to global ratings and discrete measures

A series of partially repeated measures ANOVAs were conducted, with speaker L1 a within-subjects' factor and rater experience a between-subjects factor. For comprehensibility ratings, we found a significant main effect for speakers' L1, $F(1,38) = 248.026$, $p < .001$, partial $\eta^2 = .867$, but not for rater experience. For accentedness ratings, we found significant main effects for speakers' L1, $F(1,38) = 233.156$, $p < .001$, partial $\eta^2 = .860$, but not for rater experience. For fluency ratings, we found a significant main effect for speakers' L1, $F(1,38) = 230.681$, $p = .<001$, partial $\eta^2 = .859$, but not for rater experience. There were no significant interaction effects.

In sum, there were no significant group differences in how experienced teacher-raters and novice raters scored all speakers, but pooled across raters, the speakers' L1 did affect comprehensibility, accentedness and fluency ratings. Slavic language speakers were rated as

significantly more comprehensible, significantly less accented and significantly more fluent compared to Mandarin speakers. These findings are not surprising given the extremely strong Pearson correlations between mean ratings provided by the experienced teacher-raters and novice raters (see Figure 1). Figure 2 shows mean ratings by L1 background for the experienced teacher and novice groups combined.

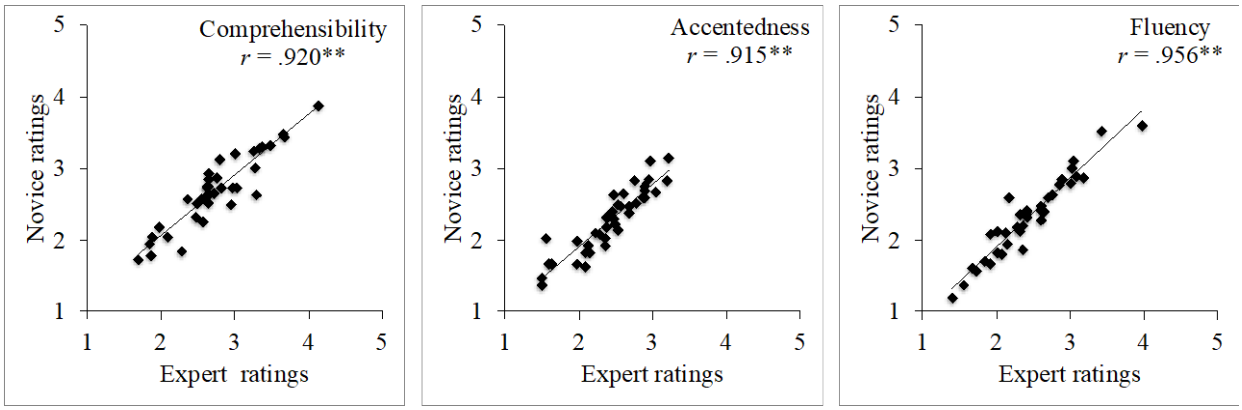

Figure 1. Scatterplots of mean expert x novice raters' scores for each L2 speaker using normalized comprehensibility, accentedness and fluency scales.

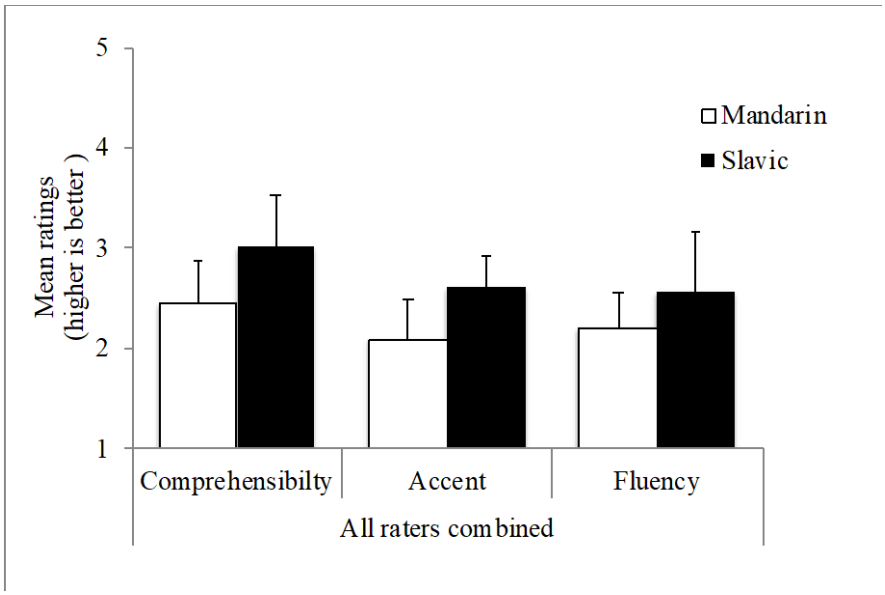

Figure 2. Mean comprehensibility (Comp.), accentedness (Acc) and fluency (Flu) ratings on the normalized scales by speakers' L1 background. Bars enclose ±1 *SD*.

Next, we computed correlations between the three global rated measures pooled across all raters and segmental accuracy (in content and function words), prosodic goodness, and the temporal measures (pruned syllables/s and speaking rate). Results revealed a nearly perfect correlation between prosodic goodness and L2 comprehensibility ratings, r = .98. Strong correlations were also revealed between prosodic goodness and ratings of both fluency, r = .91, and accentedness, r = .83. The proportion of correctly pronounced segments in content words was moderately associated with ratings for accentedness, r = .58, comprehensibility, r = .55, and fluency, r = .36. However, in function words, there was a very weak to no relationship with any of the three global rated constructs. The ratio of segmental errors over segmental incidence for vowels, consonants, and both are presented in Figures 3 and 4 for comprehensibility and accentedness ratings, respectively. The correlation is slightly higher for vowel than consonant accuracy measures, particularly for accentedness. Finally, both temporal measures strongly correlated with fluency, with a moderate relationship with comprehensibility and a moderate to weak association with accentedness.

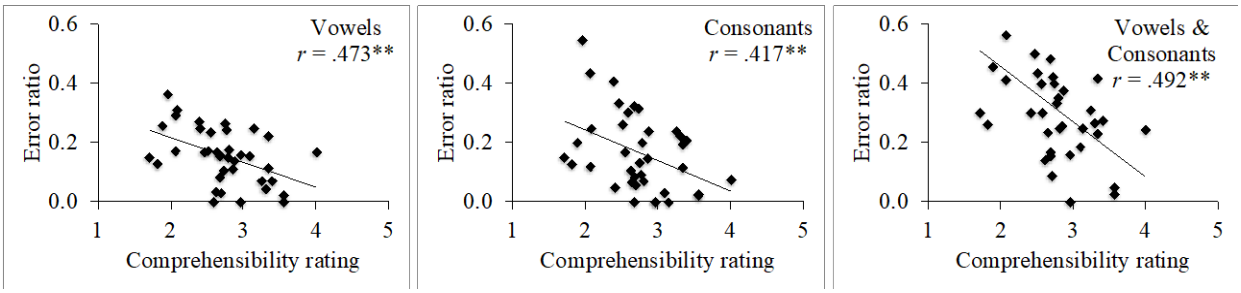

Figure 3. Ratio of segmental errors (vowels, consonants, or combined) to total errors in relation to comprehensibility ratings.

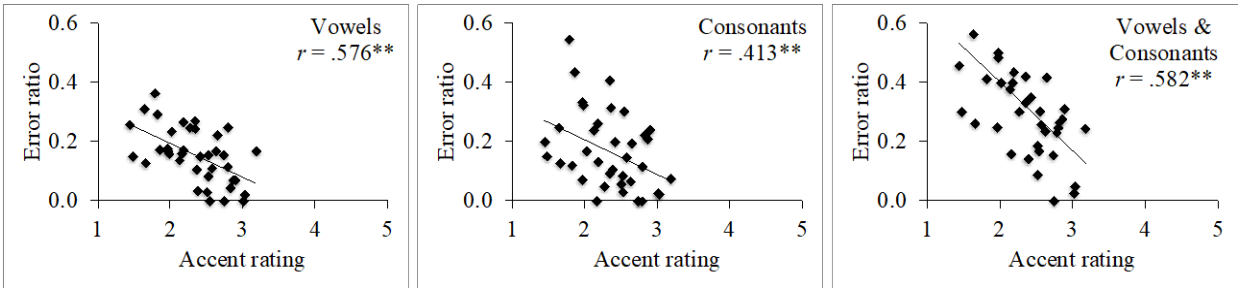

Figure 4. Ratio of segmental errors (vowels, consonants, or combined) to total errors in relation to accentedness ratings.

We then broke these findings down by the two independent variables of interest. For experienced-teacher raters versus novice raters, the overall patterns of association were similar (see Table 5). However, the temporal measures were more strongly associated with novice than experienced raters' overall perceptual judgments, whereas prosodic goodness was more strongly related to experienced teachers' than novice raters' fluency judgments. Table 6 shows a much stronger relationship between the two temporal measures and both comprehensibility and accentedness ratings for the L1 Slavic compared to Mandarin speakers. This implies that the overall ratings of the Mandarins' speech productions are not captured as well by these measures.

Table 5. Correlations between mean L2 comprehensibility, accent, and fluency ratings, and discrete speech measures grouped by rater experience

	Comprehensibility		Accentedness		Fluency	
	Experienced	Novice	Experienced	Novice	Experienced	Novice
Pruned content word segmental accuracy	.55**	.52**	.56**	.57**	.38*	.33*
Pruned function word segmental accuracy	.21	.16	.28	.27	-.01	-.02
Prosodic goodness	.96**	.96**	.82**	.80*	.93**	.87**
Speaking rate	.54**	.55**	.32	.36*	.79**	.81*
Pruned syllables/s	.58**	.62**	.37*	.46*	.78**	.83**

*$p \leq .05$, **$p \leq .01$, two-tailed

Table 6. Correlations between mean comprehensibility, accent, and fluency ratings, and discrete speech measures grouped by L1 background

	Comprehensibility		Accentedness		Fluency	
	Mandarin	Slavic	Mandarin	Slavic	Mandarin	Slavic
Content word segmental accuracy	.425	.379	.362	.551*	.214	.277
Function word segmental accuracy	.189	-.112	.145	.115	.041	-.240
Prosodic goodness	.976**	.975**	.760**	.807**	.839**	.953**
Speaking rate	.450	.756**	.254	.518**	.742**	.876**
Pruned syllables/s	.419	.797**	.174	.637**	.713 **	.869**

*$p \leq .05$, **$p \leq .01$, two-tailed

4.3 Analysis of the factors that raters reportedly take notice of when rating L2 speech

Having clarified the relationship between global L2 speech ratings and discrete measures in relation to rater experience and speaker L1, we sought to examine the factors to which experienced teacher-raters versus novice raters reportedly attend to when rating Mandarin and Slavic language speakers' utterances (research question 2). Table 7 shows frequency counts of the coded comments and loglinear analysis results for the five main categories that were most frequent. Figures 5 and 6 show counts of coded categories or subcategories by experience and L1, respectively.

Table 7. Frequencies of coded comments and loglinear analysis[a] by rater experience and speaker L1

	Mandarin Experienced	Slavic Experienced	Mandarin Novice	Slavic Novice
Total segmental errors comments Experience: χ^2 (1,39) = 20.95, $p < .0001$ L1: χ^2 = (1,39) = 11.53, $p = .0007$	**109**	**63**	**53**	**43**
Total vowel errors Experience: $\chi^2 = 8.88$, $p = .003$ L1: $\chi^2 = 6.85$, $p = .009$	48	24	22	18
Epenthesis	15	4	4	3
Substitution	26	15	12	8
Deletion	2	-	1	-
Error source unclear	5	5	5	7
Total consonant errors Experience: $\chi^2 = 14.61$, $p = .0001$ L1: $\chi^2 = 6.22$, $p = .0126$	55	34	26	18
Epenthesis	8	5	5	2
Substitution	27	20	14	16
Deletion	12	1	4	-
Error source unclear	8	8	3	-
Segmental error unclassifiable	6	5	5	7
Word pronunciation difficulty (unclassifiable pronunciation errors) Experience: $\chi^2 = -4.78$, $p = .037$ L1: $\chi^2 = 29.88$, $p < .0001$	**26**	**5**	**42**	**8**
Total rhythm/linking comments	**26**	**25**	**20**	**18**
Good rhythm/linking Experience: $\chi^2 = 4.80$, $p = .0284$	12	18	4	11
Poor rhythm/ linking L1: $\chi^2 = 5.54$, $p = .0185$	14	7	16	7
Total pausing-related comments Experience: $\chi^2 = -4.89$, $p < .0001$	**60**	**84**	**117**	**118**
Silent pauses	7	12	22	19
Filled pauses	12	20	30	35
Dysfluency source unclear	41	52	65	64
Total speech rate comments Experience: $\chi^2 = 19.41$, $p < .0001$	**50**	**55**	**21**	**28**
Fast/reasonable pace Experience: $\chi^2 = 5.4$, $p = .020$ L1: $\chi^2 = 4.05$, $p = .044$	9	15	2	8
Slow pace Experience: $\chi^2 = 14.06$, $p < .001$	41	40	19	20
Total comments about confidence Experience: $\chi^2 = 14.42$, $p < .0001$ L1: $\chi^2 = 3.99$, $p = .0457$	**23**	**40**	**12**	**14**
Speaker confident Experience: $\chi^2 = 12.62$, $p = .0004$	16	29	7	9
Speaker unconfident	7	11	5	5

[a]Only statistically significant main effects are shown for the chi-square results ($p \leq .05$). No significant interaction effects were detected.

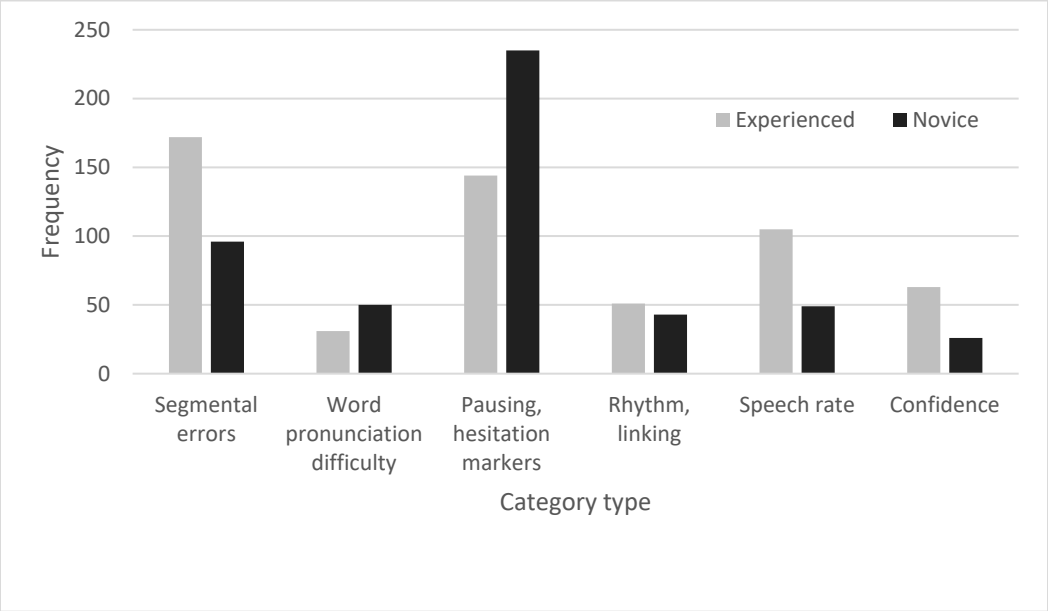

Figure 5. Frequency of coded comments by category type grouped by rater experience.

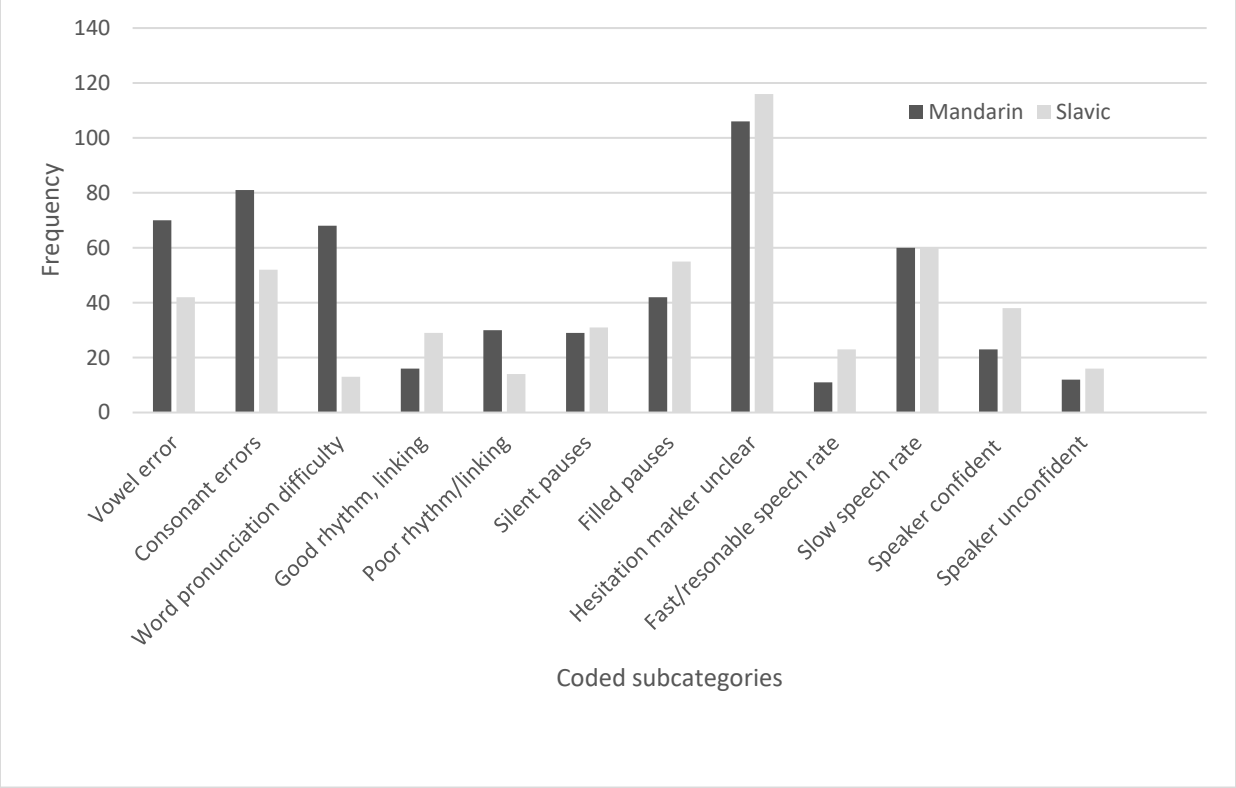

Figure 6. Frequency of coded comments for subcategories grouped by speakers' L1.

Experienced teacher-raters' introspective reports were longer than those of novice raters, producing significantly more comments for all coded categories and subcategories. The exceptions to this were comments about pausing and "word pronunciation difficulty," in which a pronunciation irregularity was signaled in the comments but the specific error type could not be identified in the coding based on the rater's account (e.g., "mispronounced a couple of words that made the words incomprehensible"). This may be because novice raters observed little else about the speech or lacked the vocabulary with which to pinpoint other features. Conversely, experienced teacher-raters were more precisely able to articulate the error source or more frequently imitated a lexical item such that the error type could be identified. Experienced teacher-raters may also have been more invested in the task than novice raters, which could partially account for their lengthier verbalizations. Overall comment frequencies about rhythm and linking revealed no rater group differences. However, experienced teacher-raters made significantly more positive comments about these elements than novice raters. They also commented more about how confident the speaker sounded. Frequency counts for word stress and pitch/intonation/voice were too low to be included in the loglinear analysis.

Mandarin speakers received more comments about segmental errors than Slavic language speakers, with higher frequency counts for consonants than vowels in the contingency table. There was a main effect for L1 for both vowels and consonants with a larger effect size for vowels. This could suggest that the vowel errors that raters pinpointed for Mandarin speakers may have been more salient or consequential compared to the more numerous consonant errors identified. Raters also appeared to struggle with word pronunciation when listening to Mandarin compared to Slavic language speakers and provided more negative comments on rhythm/linking for Mandarins. However, pausing was commented on significantly

more frequently for Slavic language speakers. Raters also noted a fast/reasonable speech rate more often for Slavic language speakers, although comments about slow paced speech and pausing were nonsignificant across groups. Finally, more comments extrapolating speakers' confidence levels from the speech samples were made for L1 Slavic than Mandarin speakers.

5. Discussion

5.1 Rater experience

This mixed methods study examined one rater characteristic (experience) and one speaker variable (L1) in relation to L2 comprehensibility, accentedness, fluency ratings, how segmental, temporal, and prosodic measures relate to these constructs, and raters' reported influences when scoring the speech. Our first main finding that experienced teacher-raters' and novice raters' scores were not significantly different echoes Bongaerts, van Summeren, Planken, and Schils' (1997) nonsignificant result for accentedness. However, it contradicts both Thompson (1991), who found that experienced teacher-raters were harsher judges than novice raters for accentedness, and Rossiter (2009), who found that experienced teacher-raters were more lenient than novice raters for fluency. None of these studies examined comprehensibility, accentedness, and fluency together. A methodological explanation for these inconsistent findings across studies includes differences in how experienced and novice raters were operationalized, L2 speaker characteristics (e.g., L1 background , L2 proficiency), the speaking task(s) used, rater characteristics (e.g., accent familiarity), the rating scales used, the way that rater severity was computed, and statistical power. A systematic review or meta-analysis synthesizing the rater experience variable could help clarify the strength of the evidence and provide further methodological considerations.

Experienced teacher-raters' and novice raters' mean comprehensibility, accentendness, and fluency ratings were strongly correlated with the experts' pooled goodness-of-prosody ratings, with a near perfect correlation for comprehensibility. This finding is consistent with research emphasizing the importance of prosodic features for comprehensibility (Isaacs & Trofimovich, 2012; Saito, Trofimovich, & Isaacs, 2016) and, for some L2 learners, intelligibility (Derwing & Munro, 1997; Hahn, 2004). However, two limitations need to be acknowledged. First, we did not apply a low pass filter for prosodic goodness ratings, which would have isolated prosodic phenomena and removed the distraction of segmental and morphosyntactic errors for the expert raters (Derwing & Munro, 1997). Therefore, the strength of association between prosodic goodness, comprehensibility, and other measures in this study should be treated with caution. Another limitation is that the more objective measure of intelligibility, which, by definition, captures actual rather than perceived listener understanding, was not examined here.

Next, we found that researcher-coded segmental accuracy ratios were moderately related to raters' mean L2 accentedness and comprehensibility ratings, with a larger role for vowels than consonants, particularly for accentedness. This result, especially for comprehensibility, which applied linguists widely consider an appropriate goal for L2 pronunciation teaching and assessment (Isaacs & Harding, 2017), implies that segments should not be ipso facto discounted in favor only of prosodic instruction. This view is consistent with previous research demonstrating a role for high functional load segmental errors in impeding comprehensibility (Munro & Derwing, 2006), distinguishing between different L2 speaking levels (Kang & Moran, 2014), and detracting from some L1 groups' comprehensibility (Suzukida & Saito, 2019).

Whereas accurately pronounced pruned segments in content words were moderately correlated for both experienced teacher-raters' and novice raters' L2 accentedness and comprehensibility ratings, in function words, this measure had a nonsignificant relationship with the global rated measures. This suggests that Zielinski's (2008) finding that function words are rarely implicated in intelligibility breakdowns extends to comprehensibility. Put simply, segmental errors in content words are a more robust measure (and more consistent with the meaning-laden nature of comprehensibility) than segmental error measures that also include function words. Consequently, we suggest that function words be removed from segmental accuracy measures or, alternatively, that functional load or some other way of gauging error locus or gravity be taken into account.

Correlations between the global rated measures and two temporal measures (pruned syllables per/s and speaking rate) were marginally higher for novice than experienced teacher-raters. This finding roughly aligns with results from the introspective reports. Although experienced teacher-raters verbalized their thoughts more fully than novice raters, the sole category where the frequency of novice raters' comments exceeded that of experienced teacher-raters was for pausing. This may be because pausing was particularly salient and disruptive for novice raters. Alternatively, pausing may have been easier for them to discuss than other linguistic phenomena, for which they lacked the vocabulary, or may have served as the default option when they had little else to say. As for experienced teacher-raters, previous research has shown that that even teachers who have served as accredited examiners or textbook authors can have difficulty with pronunciation-related terminology (Foote, Isaacs, & Trofimovich, 2013; Isaacs, Trofimovich, Yu, & Chereau, 2015). This finding did not apply uniformly to the experienced teacher-raters in our study, with nearly a third reporting pronunciation training.

Whereas some used technical terms in their introspective reports to refer to pronunciation and fluency phenomena (e.g., “sibilants,” “semivowel,” “primary stress”), others used more colloquial language (e.g., “mangles vowel sounds,” “r’s… swallowed,” “putting noise in between what he’s saying” for filled pauses). Such variability within the experienced teacher group is noteworthy. However, there were still overall differences with the novice group in terms of talk quantity, linguistic features emphasized, and likely pronunciation literacy levels.

The only other coded category where the frequency of comments for novice raters was higher than for experienced teacher-raters was for word pronunciation difficulty, designating an unclassifiable error type. This suggests that novice raters may have struggled to recall or articulate the source of a pronunciation difficulty that they had noticed. Such explanations are speculative, and it would be useful to examine raters’ accounts of their observations and processes using the follow-up interviews. Similarly, as most existing L2 pronunciation and fluency research on rater experience has been primarily quantitative (e.g., Rossiter, 2009; Saito, Trofimovich, Isaacs, & Webb, 2017), future studies could triangulate statistical findings with qualitative data to better understand rater orientations.

Although we have emphasized differences between experienced teacher-raters and novice raters above, the correlations patterns between discrete linguistic features and global speech measures was similar, with correlations coefficients at most only .06 different between groups. These values were less divergent than in Rossiter’s (2009) L2 fluency development study, suggesting the need for further investigation. Future research could also compare ESL teachers’ scoring behaviour and perspectives with those of people who do not spend their working days with L2 speakers but, nonetheless, interact with them regularly (e.g., as work colleagues).

5.2 Speaker L1 background

The Slavic language speakers were rated significantly higher than their Mandarin peers for comprehensibility, accentedness, and fluency ratings, despite both rater groups reporting significantly more exposure to Mandarin- than Russian-accented English. This familiarity effect would likely have advantaged the Mandarin speakers (Browne & Fulcher, 2017), but they were still judged more harshly. Bongaerts, Mennen, & van der Slik (2000) suggest that such results may be partially explained by the phonological distance between learners' L1 and L2. Despite being potentially more familiar to listeners, Mandarin accented English may contain more divergences from English than Slavic accented English. For example, with a few exceptions, Mandarin disallows coda consonants. Transferred to English, dropping coda consonants and/or vowel insertion could have a strong effect on Mandarin learners' comprehensibility relative to Slavic language speakers' utterances, which would not contain the same error types (McAndrews & Thomson, 2017). Ultimately, familiarity with a particular accent cannot, on its own, predict how accented or comprehensible speech in that accent is to listeners. Phonological distance is also known to play a role (Bradlow, Clopper, Smiljanic, & Walter, 2010). While Bradlow et al (2010) did not explicitly measure the phonological distance between Russian/Ukrainian and English and Mandarin an English, they did examine phonological distances between other Slavic languages (Slovene and Croatian) and English and between Cantonese and English. Their evaluation concluded that the Slavic languages are phonologically much more similar to English than Cantonese is to English.

The relationship between the temporal measures and listeners' L2 comprehensibility and accentedness and fluency ratings was moderate for Slavic language speakers, whereas for Mandarin speakers there was a significant correlation between temporal measures and fluency

ratings, but not with comprehensibility and accentedness ratings. For prosodic goodness, all correlations were strong, but the association was stronger for the Slavic language than Mandarin speaking group. Finally, for content word segmental accuracy, the sole significant relationship was for Slavic language speakers' accentedness ratings. This suggests that raters may have been preoccupied by extraneous features of Mandarins' speech not accounted for by the segmental, prosodic, and temporal measures examined. For example, none of the measures captured morphosyntax or task execution, which could have been subject to L1 differences. It could also be that raters were overwhelmed by the amount of divergence of Mandarin learners' speech due to its typological dissimilarity with English, such that the linguistic measures were less related to the global rated constructs than for Slavic language speakers. Further research could incorporate a wider range of linguistic measures and gauge their sensitivity in capturing the variance in L2 speaking performances for different L1 groups. Saito, Webb, Trofimovich, & Isaacs (2016), for example, focused on a set of lexical measures in relation to L2 comprehensibility and accentedness ratings. More research investigating macro-level discourse measures using longer speech samples would also be useful.

Although not statistically significant, the association between comprehensibility and content word segmental accuracy was higher for Mandarin than for Slavic language speakers. The loglinear analysis revealed significant main effects for word pronunciation difficulty and segmental errors, with frequencies of coded comments higher for Mandarin than Slavic language speakers. Although consonant-related comments were more numerous for both L1 groups, the effect size was higher for vowels, in line with the correlation analysis in Figures 3 and 4. This finding supports previous pronunciation research on L1 effects emphasizing the contribution of segmental errors to Mandarin speakers' comprehensibility (Crowther,

Trofimovich, Saito, & Isaacs, 2015).

There were no significant L1 group differences for the frequency of rater comments about dysfluency markers by L1. However, pure frequency counts of coded comments suggest that filled pauses may have been more perceptually salient for Slavic than Mandarins language speakers. It may be that L1 influence in the articulation of fillers was more noticeable for Slavic language speakers (de Boer & Heeren, 2019), although formant frequencies were not obtained and filled pause duration only indirectly factored into the pruned syllables measure. Whereas significantly more comments were generated about Mandarin speakers' poor rhythm or linking in the introspective reports, Slavic language speakers received significantly more comments about having fast or reasonably paced speech. Raters also commented more about Slavic language speakers' confidence, although the number of positively or negatively coded comments did not translate into significant L1 differences.

5.3 Concluding remarks

This study moves beyond most existing L2 pronunciation and fluency research by examining not only linguistic measures drawn from L2 speech samples, but also raters' accounts of the linguistic features they reportedly pay attention to when scoring L2 speech. Ensuring that raters interpret the focal constructs in the same way while taking into account construct-relevant features is important for construct validity, with implications for rater screening and training in research and assessment settings. We acknowledge that examining the frequency of raters' comments, which they are conscious of and willing/able to articulate, is an imperfect proxy of what they are actually attending to (Ericsson & Simon, 1993). Further, listeners may not understand their own analytic processes (Munro, 2018), and post-hoc reporting is prone to rationalization and face-saving strategies. Because methods for examining

what goes on in raters' minds in light of their interaction with L2 speaking performances, tasks, and scoring systems are indirect, research evidence needs to be triangulated using multiple data sources to paint a more complete picture. In addition, moving beyond observational studies to examine causal relationships between linguistic deviations and ratings using experimental or quasi-experimental designs would be desirable.

We suggest that L2 pronunciation research would benefit from greater exploration of rater processes. Most existing studies focus on which linguistic measures/dimensions account for the variance in global L2 speech ratings without examining how raters arrive at their scoring decisions (e.g., Saito et al., 2016). Future research could incorporate an eye-tracking component to examine rater fixations on different scale bands, be they numerical scales or more elaborated descriptors. The resulting evidence could then be triangulated with other data sources (e.g., stimulated recalls, interviews, ratings). In sum, we highlight here the importance of investigating individual and group differences in listeners' approaches to rating. Such research could elucidate key methodological issues in running experiments or operational L2 assessments with a pronunciation or fluency component (e.g., O'Brien, 2016 found no scale sequencing effects).

Finally, this study has focused on linguistic measures derived from L2 speech and raters' introspective reports. However, variables extraneous to the properties of L2 speaking performances may also be reflected in ratings, posing problems for score interpretation. For example, so-called rater effects, such as listeners' exposure to or attitudes toward L2 accented speech, could influence their scoring decisions (e.g., Winke, Gass, & Myford, 2013). However, negative rater judgments should not automatically be dismissed as prejudicial (Munro, 2018). Future research should ideally examine rater characteristics or orientations that could threaten

the validity of the L2 abilities being measured within the same research program as construct-relevant factors.

Acknowledgments

This research was supported by fellowships awarded from the Social Sciences and Humanities Research Council of Canada to both the first and second authors. Thanks are due to Garrett Byrne, Daniel Clement, Marc Garellek, Joseph Hartfeil, Donna Pearce, and Monika Spak, who assisted with data transcription or coding, Carolyn Turner and Pavel Trofimovich for their input on the content of the paper, Tracey Derwing and Murray Munro for sharing their speaking prompt, the teachers and administrators who facilitated participant recruitment, and our participants, who spoke candidly about their impressions. We are also grateful to the *JSLP* editor and special issue editors for their insightful comments on previous versions of this article.

References

Bannigan, K., & Watson, R. (2009). Reliability and validity in a nutshell. *Journal of Clinical Nursing*, *18*(23), 3237-3243.

Bongaerts, T., Mennen, S., & van der Slik, F. (2000). Authenticity of pronunciation in naturalistic second language acquisition: The case of very advanced late learners of Dutch as a second language. *Studia Linguistica*, *54*(2), 298-308.

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*(4), 447-465.

Bradlow, A., Clopper, C., Smiljanic, R., & Walter, M. A. (2010). A perceptual phonetic similarity space for languages: Evidence from five native language listener groups. *Speech Communication*, *52*(11-12), 930-942.

Browne, K., & Fulcher, G. (2017). Pronunciation and intelligibility in assessing spoken fluency. In T. Isaacs & P. Trofimovich (Eds.), *Second language pronunciation: Interdisciplinary perspectives* (pp. 37-53). Bristol, UK: Multilingual Matters.

Chalhoub-Deville, M. (1995). Deriving oral assessment scales across different tests and rater groups. *Language Testing*, *12*(1), 62–70.

Creswell, J. W., & Plano Clark, V. L. (2017). *Designing and conducting mixed methods research* (3rd ed.). Thousand Oaks, CA: SAGE.

Crowther, D., Trofimovich, P., Saito, K., & Isaacs, T. (2015). Second language comprehensibility revisited: Investigating the effects of learner background. *TESOL Quarterly*, *49*(4), 814–837.

de Boer, M., & Heeren, W. (2019). The speaker-specificity of filled pauses: A cross-linguistic study. *Proceedings of the International Congress of Phonetic Sciences (ICPhS) 2019* (pp. 607-611). Melbourne, Australia: Australasian Speech Science and Technology Association.

Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition, 19*(1), 1–16.

Derwing, T. M., & Munro, M. J. (2009). Comprehensibility as a factor in listener interaction preferences: Implications for the workplace. *Canadian Modern Language Review*, *66*(2), 181–202.

Derwing, T. M., & Munro, M. J. (2013). The development of L2 oral language skills in two L1 groups: A 7-year study. *Language Learning*, *63*(2), 163-185.

Derwing, T. M., Rossiter, M. J., Munro, M. J., & Thomson, R. I. (2004). Second language fluency: Judgments on different tasks. *Language Learning*, *54*(4), 665–679.

Derwing, T. M., Thomson, R. I., & Munro, M. J. (2006). English pronunciation and fluency development in Mandarin and Slavic speakers. *System*, *34*(2), 183-193.

Douglas, D. (1994). Quantity and quality in speaking test performance. *Language Testing*, *11*(1), 125–144.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (Rev. ed.). Cambridge, MA: MIT Press.

Foote, J. A., Isaacs, T., & Trofimovich, P. (2013, June 3-5). *Developing a teacher-friendly assessment tool for L2 comprehensibility*. Canadian Association of Applied Linguistics (ACLA/CAAL) conference, Calgary, AB.

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*(2), 253-278.

Gass, S. M., & Mackey, A. (2000). *Stimulated recall methodology in second language research*. Mahwah, NJ: Lawrence Erlbaum.

Hahn, L. D. (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly*, *38*(2), 201–233.

Isaacs, T., & Harding, L. (2017). Research timeline: Pronunciation assessment. *Language Teaching*, *50*(3), 347–366.

Isaacs, T., & Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly*, *10*(2), 135-159.

Isaacs, T., & Trofimovich, P. (2012). Deconstructing comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings. *Studies in Second Language Acquisition*, *34*(3), 475–505.

Isaacs, T., Trofimovich, P., Yu, G., & Chereau, B. M. (2015). Examining the linguistic aspects of speech that most efficiently discriminate between upper levels of the revised IELTS pronunciation scale. *IELTS research reports online series*, *4*.

Kang, O., & Moran, M. (2014). Functional loads of pronunciation features in nonnative speakers' oral assessment. *TESOL Quarterly*, *48*(1), 176-187.

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning*, *40*(3), 387–417.

Lumley, T. (2005). *Assessing second language writing: The rater's perspective*. Frankfurt: Peter Lang.

McAndrews, M. M., & Thomson, R. I. (2017). Establishing an empirical basis for priorities in pronunciation teaching. *Journal of Second Language Pronunciation, 3*(2), 267-287.

Munro, M. J. (2018). Dimensions of pronunciation. In O. Kang, R. Thomson, & J. Murphy. The *Routledge handbook of contemporary English pronunciation* (pp. 413-431). New York: Routledge.

Munro, M. J., & Derwing, T. M. (2006). The functional load principle in ESL pronunciation instruction: An exploratory study. *System, 34*(4), 520–531.

O'Brien, M. G. (2016). Methodological choices in rating speech samples. *Studies in Second Language Acquisition, 38*(3), 587-605.

Pawlikowska-Smith, G. (2000). *Canadian Language Benchmarks 2000: Theoretical framework*. Ottawa, ON: Centre for Canadian Language Benchmarks.

Rajadurai, J. (2007). Intelligibility studies: A consideration of empirical and ideological issues. *World Englishes, 26*(1), 87–98.

Riggenbach, H. (1991). Toward an understanding of fluency: A microanalysis of non-native speaker conversations. *Discourse Processes, 14*(4), 423–441.

Rose, H., & Galloway, N. (2019). *Global Englishes for language teaching*. Cambridge: Cambridge University Press.

Rossiter, M. J. (2009). Perceptions of L2 fluency by native and non-native speakers of English. *Canadian Modern Language Review, 65*(3), 395–412.

Saito, K., Trofimovich, P., & Isaacs, T. (2016). Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels. *Applied Psycholinguistics*, *37*(2), 217-240.

Saito, K., Trofimovich, P., Isaacs, T., & Webb, S. (2017). Re-examining phonological and lexical correlates of second language comprehensibility: The role of rater experience. In T. Isaacs & P. Trofimovich (Eds.), *Second language pronunciation assessment: Interdisciplinary perspectives* (pp. 131–146). Bristol, UK: Multilingual Matters.

Saito, K., Webb, S., Trofimovich, P., & Isaacs, T. (2016). Lexical correlates of comprehensibility versus accentedness in second language speech. *Bilingualism: Language and Cognition*, *19*(3), 597–609.

Schiavetti, N. (1992). Scaling procedures for the measurement of speech intelligibility. In R. D. Kent (Ed.), *Intelligibility in speech disorders* (pp. 11-34). Amsterdam: John Benjamins.

Stevens, J. P. (2009). *Applied multivariate statistics for the social sciences* (5th ed.). New York: Taylor & Francis.

Suzukida, Y., & Saito, K. (2019). Which segmental features matter for successful L2 comprehensibility? Revisiting and generalizing the pedagogical value of the Functional Load principle. *Language Teaching Research*. Advance online publication. https://doi.org/10.1177/1362168819858246

Thomson, R. I., & Isaacs, T. (2009). Within-category variation in L2 English vowel learning. *Canadian Acoustics*, *37*, 138=139.

Thompson, I. (1991). Foreign accents revisited: The English pronunciation of Russian immigrants. *Language Learning*, *41*(2), 177–204.

Upshur, J. A., & Turner, C. E. (1999). Systematic effects in the rating of second-language speaking ability: Test method and learner discourse. *Language Testing*, *16*(1), 82–111.

Winke, P., Gass, S., & Myford, C. (2013). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing*, *30*(2), 231-252.

Zielinski, B. W. (2008). The listener: No longer the silent partner in reduced intelligibility. *System*, *36*(1), 69–84.

Author information

Talia Isaacs (corresponding author)

UCL Centre for Applied Linguistics

UCL Institute of Education

University College London

20 Bedford Way

London, United Kingdom

WC1H 0AL

Tel.: +44 (0) 207 612 6348

talia.isaacs@ucl.ac.uk

Ron I. Thomson

Department of Applied Linguistics

Brock University

500 Glenridge Avenue

St. Catharines, ON

Canada

L2S 3A1

Tel: 905-688-5550, ext. 5842

rthomson@brocku.ca