CrossMark

Studies in Second Language Acquisition, 2016, **38**, 587–605.
doi:10.1017/S0272263115000418

METHODOLOGICAL CHOICES IN RATING SPEECH SAMPLES

Mary Grantham O'Brien
University of Calgary

Much pronunciation research critically relies upon listeners' judgments of speech samples, but researchers have rarely examined the impact of methodological choices. In the current study, 30 German native listeners and 42 German L2 learners (L1 English) rated speech samples produced by English-German L2 learners along three continua: accentedness, fluency, and comprehensibility. The goal was to determine whether rating condition, that is, (a) whether each speech sample is rated along all three continua after it is heard once or (b) whether all speech samples are rated along one continuum before being rated along the next continuum, and continuum order (e.g., whether participants rate speech samples for accentedness before comprehensibility or fluency) have an effect on listeners' ratings. Results indicate no significant overall effects of rating condition or continuum order, but there is evidence of rating condition effects by listener group. The results have implications for laboratory and classroom assessments of L2 speech.

INTRODUCTION

Listeners continuously assess the speech of their interlocutors, and pronunciation is among the most salient aspects to guide listeners in their judgments of both the speech that is produced and the speakers who produced it (e.g., Hayes-Harb & Hacking, 2015). Listeners have been shown to reliably and consistently rate the speech of a range of speakers, but research tends to rely more on native than nonnative listeners.

I'd like to thank Rahel Cramer and Paulina Maczuga for their assistance with data collection and Tak Fung for his statistical advice.

© Cambridge University Press 2015

https://doi.org/10.1017/S0272263115000418 Published online by Cambridge University Press

One reason that so few studies make use of nonnative speaker raters may be that their ability to perceive speech stream cues may be related to their L2 proficiency level. That is to say, less proficient learners may not be able to pick up on errors that are present in speech samples. Along the same lines, they may regard certain unfamiliar lexical items or grammatical features as incorrect.[1] It may therefore not be surprising that nonnative listeners tend to apply different evaluation criteria as they perform their ratings (e.g., Kim, 2009; Riney, Takagi, & Inutsuka, 2005). Numerous studies have examined the extent to which listener characteristics (e.g., Gui, 2012; Isaacs & Trofimovich, 2011; Kang, 2012; Winke & Gass, 2013) and acoustic or other linguistic measures (e.g., Gut, 2009; Isaacs, Trofimovich, Yu, & Chereau, 2015; O'Brien, 2014; Riney, Takada, & Ota, 2000; Saito, Trofimovich, & Isaacs, forthcoming; Saito, Webb, Trofimovich, & Isaacs, forthcoming; Trofimovich & Baker, 2006; Trofimovich & Isaacs, 2012) play a role in listeners' assessments of speech, and the results point to the importance of examining a range of variables that may affect a listener's judgment.

Less attention has been paid to methodological issues in the rating of L2 speech. One notable exception is Isaacs and Thomson's (2013) mixed-methods study, in which the researchers investigated the effects of rating scale length (i.e., five- vs. nine-point scales) and listener experience (i.e., experienced ESL teachers vs. novice raters) in the assignment of accentedness (i.e., the degree to which pronunciation deviates from an expected norm), fluency (i.e., fluidity of speech), and comprehensibility (i.e., the perceived ease with which an utterance is understood) ratings assigned to L2 English speech samples. The results indicated that the patterns of ratings did not differ across the five- and nine-point rating scale conditions, and there were no statistically significant differences in ratings provided by experienced and novice raters. Nonetheless, an analysis of raters' verbal reports indicated that the experienced raters were better able to isolate errors in pronunciation.

Studies performed to date have not investigated the effect of two potentially significant methodological issues: the effects of rating condition and continuum order, both of which are the focus of this study. As such, this study makes an important contribution to discussions surrounding methodology employed in pronunciation research.

Perceived Accentedness

Accentedness is the "degree to which the pronunciation of an utterance sounds different from an expected pronunciation pattern" (Munro, Derwing, & Morton, 2006, p. 112). This study examines the notion of foreign accentedness, that is, deviation from an expected pronunciation

https://doi.org/10.1017/S0272263115000418 Published online by Cambridge University Press

pattern that is rooted in a speaker's native language. Speaking with a foreign accent is a reality for most L2 learners who begin learning languages after early childhood (e.g., Long, 1990). Although researchers including Abrahamsson and Hyltenstam (2008) and Bongaerts, Mennen, and van der Slik (2000) point to exceptions to this general finding, those late learners in these studies who did achieve a nativelike accent differed from members of the general population in terms of their language aptitude, motivation, and intensity of input. A foreign accent acts as an immediate marker of nonnative speaker status, even when speech samples are very short (e.g., Flege, 1984) or when they are played backward (Munro, Derwing, & Burgess, 2010). Because of the salience of features of foreign accented speech, even nonspeakers of a target language are able to reliably rate speech for foreign accentedness (Major, 2007).

Researchers investigating the concept of accentedness have pointed to the ability of a variety of types of listeners to scale accents similarly to expert native speakers. These include teachers (e.g., Trofimovich, Gatbonton, & Segalowitz, 2007), naïve native speakers (e.g., Kang, Rubin, & Pickering, 2010), and nonnative speakers (e.g., Flege, 1988; Kang, 2012; Mackay, Flege, & Imai, 2006; Munro et al., 2006). Nonetheless, the majority of studies investigate the extent to which native English listeners are able to detect and scale foreign accents in L2 English (e.g., Anderson-Hsieh, Johnson, & Koehler, 1992; Chakraborty & Goffman, 2011; Chen, 2010; Moyer, 2011; Trofimovich & Isaacs, 2012), although a few accentedness studies have also been conducted on other target languages (e.g., Leikin, Eviatar, & Sapir, 2009; Pinget, Bosker, Quené, & de Jong, 2014; Reinisch & Weber, 2012). A few studies have investigated native listeners' (Molnar, 2010; Moyer, 1999; O'Brien, 2004) and nonnative listeners' (O'Brien, 2014) foreign accent ratings of L2 German speech. A common finding was that listeners differentiated native from nonnative speech, regardless of their status as a native or nonnative listener. Wilkerson (2010), who compared ratings assigned by native and nonnative listeners, found that L2 learners were better able than native speakers to determine whether speech samples were produced by native or nonnative speakers of German. She suggests that nonnative listeners hold "the speech samples to a different standard" than native listeners (p. 148), and although she did not examine speech stream characteristics, the author posits that nonnative listeners tune in to different speech stream cues than native speakers do.

Previous studies have found that a range of speech stream characteristics is associated with accentedness ratings. These may include subsegmental phenomena (e.g., González-Bueno, 1997), segmental errors (e.g., Riney et al., 2005), and a number of prosodic parameters (e.g., Anderson-Hsieh et al., 1992). Prosodic errors that have been shown to influence accent ratings include word stress (e.g., Chakraborty & Goffman, 2011) and sentence stress errors (e.g., Hahn, 2004), articulation

https://doi.org/10.1017/S0272263115000418 Published online by Cambridge University Press

and/or speech rate (e.g., Kang et al., 2010; Munro, 1995; Munro & Derwing, 2001), pitch range (e.g., Kang, 2010), and pause frequency and duration (e.g., Chen, 2010; Trofimovich & Baker, 2006).

Perceived Fluency

Fluency is a term that is often used to describe the speech of both native and nonnative speakers, but the use of the term—both its definition and its operationalization—differs between researchers and members of the population at large (see, e.g., Koponen & Riggenbach, 2000). For the purposes of the current research, it is important to highlight two of the main ways in which the term is used. The first, which is commonly used by researchers, corresponds to Lennon's (1990) narrow definition of fluency. That is, it is "an impression on the listener's part that the psycholinguistic process of speech planning and speech production are functioning easily and efficiently," and disfluency markers like pause time, repetitions, filled pauses, and speech rate "make the listener aware of the production process under strain" (p. 391). When speech samples are rated according to this definition of fluency, listeners are directed to listen for speech stream cues that disrupt the fluidity of speech. The second definition, common in lay usage, is Lennon's (1990) broad definition of fluency, which corresponds to oral proficiency (p. 389). As such, when listeners make use of this definition of fluency, they may tune into morphological, syntactic, or lexical errors. The current study operationalizes perceived fluency according to Lennon's narrow definition.

Much of the research on the topic has shown that fluency is related to planning and automaticity of processing (e.g., Bortfeld, Leon, Bloom, Schober, & Brennan, 2001; Segalowitz, 2007). As such, even native speakers of a given language do not always produce fluent utterances (e.g., Riggenbach, 1991). Speakers—whether native or nonnative—make use of pauses and other disfluencies in the planning and formulation of utterances (Clark & Fox Tree, 2002). Nonetheless, it has been shown that L2 learners tend to pause more in their L2s than in their L1s (Riantseva, 2001), and L2 speech is usually slower than that of native speakers (e.g., Gut, 2009). Speaking fluently has practical implications for both speakers and listeners. Derwing, Munro, and Thomson (2008) note that "[f]luency problems can … be exasperating for L2 speakers" (p. 360). Along the same lines, Rossiter (2009) notes that fluent speakers may keep their listeners more engaged than those who are disfluent.

Much of the research investigating perceived fluency relies on the judgments of native listeners, although Rossiter (2009) employed three types of listeners in her investigation of L2 English fluency: ESL teachers,

https://doi.org/10.1017/S0272263115000418 Published online by Cambridge University Press

novice native listeners, and nonnative English listeners from a variety of L1 backgrounds. She found that the nonnative listeners rated speech samples more harshly (i.e., they rated the samples to be less fluent) than did the novice native listeners. Kormos and Dénes (2004) also examined listener effects. They investigated the difference in fluency ratings assigned to English L2 speech by native and nonnative (teacher) raters. Unlike Rossiter (2009), they found no significant difference in ratings assigned by listeners in the two groups. To date, most of the research into the perceived fluency of L2 speech has been performed on L2 English, although researchers have investigated fluency in other L2s including Dutch (Cucchiarini, Strik, & Boves, 2002; de Jong, Steinel, Florijn, Schoonen, & Hulstijn 2013) and French (Préfontaine, 2013).

Fluency researchers often look to the speech stream characteristics that may be associated with fluency ratings assigned by listeners, and most of them include cues that negatively affect the fluidity of speech. These include frequency of filled and/or unfilled pauses (e.g., Brand & Götz, 2011), pause duration (e.g., Kormos & Dénes, 2004; Lennon, 1990), speech rate (Derwing, Rossiter, Munro, & Thomson, 2004; Riggenbach, 1991), mean length run (e.g., Cucchiarini et al., 2002; Préfontaine, 2013; Towell, Hawkins, & Bazergui, 1996), and self-repetition (e.g., Lennon, 1990; Rossiter, 2009). Some studies have found that listeners may rely on other cues when they rate speech along the perceived fluency continuum. For example, the ratings assigned by the listeners in Freed (1995), Kormos and Dénes (2004), and Rossiter (2009) were correlated with errors in pronunciation, grammar, and vocabulary.

Comprehensibility

While learners regularly report a desire to speak their L2 without an accent or more fluently, the focus of many researchers and classroom practitioners has turned to comprehensibility, which has been determined to be a more realistic and important goal inside and outside the L2 classroom (e.g., Derwing & Munro, 2009a). In the current study, comprehensibility is operationalized according to definitions provided by Derwing and colleagues. That is, it is the "ease or difficulty with which a listener understands … accented speech" (Derwing et al., 2008, p. 360), and it corresponds to "judgments on a rating scale of how difficult or easy an utterance is to understand" (Derwing & Munro, 1997, p. 2).

The results of a host of comprehensibility studies indicate that listeners are indeed able to understand some accented speech (e.g., Derwing & Munro, 1997). In Derwing and Munro's (2009b) workplace study, employees were asked to choose between two foreign-accented

https://doi.org/10.1017/S0272263115000418 Published online by Cambridge University Press

speech samples, one of which was more comprehensible than the other. The listeners preferred the more comprehensible speech samples. The authors thus concluded that the listeners placed a higher value on comprehensible than on less accented speech. With the exception of Munro et al. (2006) and Derwing and Munro (2013), few comprehensibility or intelligibility studies have made use of L2 listeners. The results of these studies have shown that L2 listeners perform similarly to native speakers. As is the case with studies investigating foreign accentedness and fluency, most comprehensibility and intelligibility studies investigate the production of L2 English speech (but see Caspers, 2010; Neuendorf, 2010; O'Brien, 2014 for exceptions).

A variety of speech stream characteristics including segmental accuracy (e.g., Braun, Lemhöfer, & Mani, 2011), a combination of word stress and segmental errors (Caspers, 2010), sentence accents (e.g., Hahn, 2004), and speech rate (Kang, 2010) have been shown to correlate with comprehensibility ratings. Trofimovich and Isaacs (2012), in their comprehensive study into the speech stream cues that might affect comprehensibility ratings, found that comprehensibility ratings assigned to L2 English speech samples were primarily predicted by grammatical accuracy, lexical richness (i.e., type frequency), and word stress. Similarly, Munro and Derwing (1995) found that grammatical errors were significantly correlated with ratings indicating less comprehensible speech.

RESEARCH QUESTIONS

Previous research has found that the concepts of perceived accentedness, fluency and comprehensibility not only share common features but that listeners are also able to distinguish among the three when they rate speech samples. This previous research, however, relies primarily on native speaker raters (but see, e.g., Kim, 2009; Munro et al., 2006; Rossiter, 2009; Wilkerson, 2010) who evaluate English L2 speech (but see, e.g., Caspers, 2010; de Jong et al., 2013; Freed, 1995; Gut, 2009; Moyer, 1999). A common finding is not only that accented speech may be comprehensible but also that incomprehensible speech is often heavily accented (e.g., Munro & Derwing, 1995).

Research making use of human raters who rate speech samples along more than one continuum (accentedness, fluency, comprehensibility) employs one of two methodologies: multiple ratings are completed together (e.g., Isaacs & Trofimovich, 2011) or separately (e.g., Derwing & Munro, 1997). The current study builds on research into methodological issues in the assessment of L2 speech, and the methodological focus of this study is the rating procedure. The current study thus seeks answers to two key questions:

https://doi.org/10.1017/S0272263115000418 Published online by Cambridge University Press

1. To what extent do perceived accentedness, fluency, and comprehensibility ratings assigned by German native listeners and English-German L2 listeners differ according to rating condition? That is to say, does it make a difference whether listeners who assess speech samples (a) listen to a sample only once before they assess its accentedness, fluency, and comprehensibility, or (b) rate a large number of samples along one continuum before moving onto the next continuum and then the final continuum?

The answer to this question is methodologically important. Because few previous studies report on the specific methodology employed, and no study has compared the results of one rating condition (i.e., *together*: three ratings in tandem after one exposure to a speech sample vs. *separate*: ratings along the three continua separately from one another) compared to the other, we do not know whether ratings will differ if listeners rate the speech samples along all three continua at once compared to if listeners rate the samples separately along each continuum. If the results indicate no differences by rating condition, researchers will be well justified in carrying out the more efficient rating task that involves only one exposure to each speech sample. If, however, there are differences, researchers will be encouraged to carefully consider the role of rating condition in future studies.

2. What is the effect of continuum order on the results obtained?

Order effects are methodologically important, as they enable us to determine whether ratings along each continuum differ according to whether they are performed first, second, or third. The study builds on the results of Munro and Derwing (1994), who report that listeners rated speech samples more harshly in later tasks (i.e., when they rated speech samples for the third and fourth times). Although previous studies control for continuum order by balancing out the number of times listeners rate along a specific continuum first, second, or third, we know of no study to date that has investigated whether listeners who rate speech samples along one continuum (e.g., accentedness) before another (e.g., comprehensibility) provide different ratings from those who rate the samples in the opposite order.

METHOD

Speech Samples

Speech samples rated in the current study are the same 48 twenty-second samples as those used in O'Brien (2014). These were gathered as part of a larger study investigating the use of prosodic cues among native speakers

https://doi.org/10.1017/S0272263115000418 Published online by Cambridge University Press

and L2 learners of German. They were produced by 48 German speakers, 24 of whom were German native speakers living in Germany. The remaining 24 speakers were English-German L2 learners, six of whom were living in Germany at the time of the study. The German native speakers (6 males, 18 females) were university students who were an average of 23.9 years of age (range 21–30 years). The English-German L2 learners (12 males, 12 females) were university students with a mean age of 22.8 years (range 21–30 years). Half of the L2 learners' speech samples were produced by speakers with intermediate (CEFR level B1) proficiency in German, and half were produce by speakers with intermediate to advanced (CEFR level C1 or C2) proficiency in German, as determined by the results of an online multiple-choice proficiency test that focuses on grammar and vocabulary knowledge (Goethe Institut, 2004).

All of the speakers were recorded reading the suitcase narrative,[2] a picture story illustrating a collision between a man and a woman, both of whom are carrying identical green suitcases. This accident causes them to drop their bags. It is only when they reach their final destinations that each realizes that he or she has picked up the bag of the other. This task has been widely used in previous studies investigating L2 speech production (e.g., Derwing & Munro, 2009b; Isaacs & Trofimovich, 2011).

Participants

Listeners in the current study included 30 German native speakers and 42 English-German L2 learners. In order to test for the effect of rating condition (i.e., ratings performed *together* or *separately*), listeners from each L1 group were further divided into two groups (*together* vs. *separate*). An overview of the participants by group is provided in Table 1.

The only significant difference between the L2 listener groups is in length of immersion experience, with the *together* group having spent significantly more time in a German immersion setting ($t(40) = -1.858$, $p = .001$).

Procedure

The first 20 seconds of speech—excluding hesitations like filled or unfilled pauses at the onset of the recording—from each speech sample were prepared for use in the rating task, along the lines of previous studies (e.g., Derwing et al., 2008). As in other research, listeners rated the speech samples along nine-point Likert scales (e.g., Chakraborty & Goffman, 2011;

https://doi.org/10.1017/S0272263115000418 Published online by Cambridge University Press

Table 1. Means and standard deviations (in parentheses) of listener demographic information and L2 proficiency by rating condition group

Demographic variable	German L1 listeners separate (n = 15)	German L1 listeners together (n = 15)	English-German L2 listeners separate (n = 25)	English-German L2 listeners together (n = 17)
Gender	13 f,2 m	8 f, 7 m	15 f, 10 male	10 f, 7 m
Age	25.60 (2.03)	25.93 (4.46)	21.08 (3.12)	20.41 (1.94)
L2 proficiency	NA	NA	13.36 (5.57)	14.29 (3.55)
Age of learning	NA	NA	19.36 (2.6)	17.12 (3.62)
Immersion experience	NA	NA	.06 (1.49)	.60 (1.45)

Note. f = female; m = male; age = current age in years; L2 proficiency = score out of 30 points on the German proficiency test (Goethe Institut, 2004); NA = not applicable; age of learning = age of learning German as an L2; immersion experience = time living in a German-speaking region in years.

Derwing et al., 2004; Isaacs & Trofimovich, 2011; Kennedy & Trofimovich, 2008; Munro & Derwing, 1995). The three scales in the study include perceived accentedness, where 1 = no accent and 9 = extremely strong accent; perceived fluency, where 1 = extremely fluent and 9 = extremely disfluent; and comprehensibility, where 1 = easy to understand and 9 = extremely difficult to understand. Given the recommendations of researchers to familiarize listeners with rating scales and procedure (e.g., Derwing et al., 2004; Rossiter, 2009), participants were asked to rate three spontaneous speech samples along each of the continua before they began the experimental task. The samples used in the training, which were gathered as part of another production study, were produced by speakers with a proficiency level equivalent to those speakers whose samples were used in the current study. In each instance, one native speaker sample was included among the training samples. Feedback that focused on speech stream characteristics associated with each of the continua was given to listeners after they rated each of the training stimuli. For example, listeners were told to focus primarily on segmental and prosodic (especially word stress) pronunciation errors as they rated speech samples for accentedness. They were told to listen for pauses (filled and unfilled), repetitions, and overall speed of delivery when they rated the speech samples for fluency. Listeners were directed to only focus on how well they understood utterances when they rated for comprehensibility. As they rated the speech samples during the training session, they were asked to point out specific speech stream characteristics that caused them to rate the samples as they did. The researcher also pointed out specific instances of each type of speech stream characteristic associated with each continuum as they worked through the training.

https://doi.org/10.1017/S0272263115000418 Published online by Cambridge University Press

As shown in Table 1, a total of 40 listeners (15 native speakers, 25 L2 learners) rated each of the speech samples in three separate tasks (once for accentedness, once for fluency, and once for comprehensibility), and 32 listeners (15 native speakers, 17 L2 learners) rated each of the speech samples along all three continua together after having heard each stimulus only once. The presentation of stimuli was randomized, and the continuum order was balanced across participants. All participants took part in the study in their home countries, and they were paid a small honorarium of $15 or €12 for their participation.

Data Analysis

Once it was determined that listeners in both groups distinguished among the speech samples produced by native German speakers and English-German L2 learners (i.e., that ratings assigned by both listener groups to speech samples produced by native speaker and L2 learners differed significantly along each continuum), it was decided that the focus of the study should be on the ratings assigned to the English-German L2 learners, along the lines of Derwing et al. (2004).

Because each rater provided three ratings for each speech sample, there are potential correlations among ratings. It is essential to make use of a statistical procedure that can deal with dependence among the outcomes. For this reason, data were analyzed through an unbalanced repeated measure design with interval outcome variables using a Generalized Estimating Equation (i.e., GEE under Genlin procedures in SPSS version 21) to determine whether and the extent to which English-German L2 listeners and German native listeners differed in their ratings of English-German L2 learners' speech along each of the three continua. The dependent variable in all analyses was rating, and raters were treated as repeated measures. We applied a Bonferroni adjustment in the order analysis to account for the fact that multiple comparisons were performed simultaneously on the data.

RESULTS

Given that the research questions focus on the differences in ratings assigned by native and nonnative listeners by rating condition and order, it is helpful to begin by examining how listeners in the two groups rated the speech samples overall. The mean ratings assigned by native listeners and English-German L2 listeners are provided Table 2. An analysis of the results indicates that German native listeners and English-German L2 listeners did not differ significantly in their accentedness ratings.

https://doi.org/10.1017/S0272263115000418 Published online by Cambridge University Press

Table 2. Mean ratings and standard deviations (in parentheses) by L1 group

Rating continuum	German L1 listeners	English-German L2 listeners
accentedness	5.52 (2.54)	5.85 (2.47)
fluency	4.8 (2.12)	5.58 (2.32)
comprehensibility	3.76 (2.18)	4.57 (2.35)

Nonetheless, they did differ in their ratings of fluency (B = -.794, SE = .191, $\chi^2(1)$ = 17.232, p < .001) and comprehensibility (B = -.835, SE = .3256, $\chi^2(1)$ = 6.65, p = .010), with English-German L2 listeners assigning significantly harsher ratings on both continua.

The Effect of Rating Condition

The results of analyses investigating the effects of rating condition indicate that there is no effect of rating condition along any of the continua when the data are collapsed across listeners. That is to say, accentedness, fluency, and comprehensibility ratings assigned to L2 speech samples did not differ according to whether listeners performed the ratings along all three continua together after having heard the speech sample once or whether they rated speech samples along each continuum separately from the other two. The story becomes a bit more nuanced, however, if we analyze rating condition effects by listener groups. Ratings by listener groups and rating condition are provided in Table 3.

An analysis of the results indicates no effects of rating condition by language group for ratings along the accentedness continuum. This finding is in line with those presented for listeners as a whole: accentedness ratings are not significantly different whether they are performed together or separately for both German native listeners and English-German L2 listeners. The same can be said for comprehensibility ratings

Table 3. Mean ratings and standard deviations (in parentheses) by rating condition and listener group

Rating continuum	German L1 listeners separate	German L1 listeners together	English-German L2 listeners separate	English-German L2 listeners together
accentedness	5.74 (2.72)	5.29 (2.33)	5.69 (2.39)	6.09 (2.58)
fluency	5.13 (2.23)	4.47 (1.95)	5.56 (2.32)	5.6 (2.32)
comprehensibility	3.35 (2.11)	4.17 (2.18)	4.92 (2.28)	4.06 (2.36)

https://doi.org/10.1017/S0272263115000418 Published online by Cambridge University Press

performed by German native listeners and fluency ratings performed by English-German L2 listeners. There is no significant effect of rating condition on ratings assigned along these continua. Nonetheless, we see one important rating condition effect for German native listeners: their ratings along the fluency continuum differ significantly by rating condition (B = -.679, SE = .2873, $\chi^2(1)$ = 5.582, p = .018). That is to say, German native listeners who rated fluency together with accentedness and comprehensibility assigned ratings indicating that speech samples were more fluent than did those who performed the tasks separately along each of the three continua. A slightly significant rating condition effect was found for English-German L2 listeners along the comprehensibility continuum (B = -.863, SE = .4475, $\chi^2(1)$ = 3.721, p = .054). That is to say, English-German L2 listeners who rated the speech samples along all three continua together showed a tendency to rate the speech samples as being more comprehensible than did those who rated comprehensibility separately from accentedness and fluency.

The Effect of Order

The results examining the effect of rating order (i.e., whether participants performed ratings along a given continuum first, second, or third) on ratings indicate no effects of rating order on the ratings assigned. That is to say, listeners rated speech samples similarly along the accent continuum when they rated for accentedness before the other two ratings, between the two ratings, and as the last of the three rating types. The same is true for fluency and comprehensibility ratings.

DISCUSSION

Native and L2 Listener Ratings

German native listeners and English-German L2 listeners showed similar patterns in their ratings: accentedness ratings were harshest, and these were followed by fluency and then comprehensibility ratings. Thus, the results support previous research that has shown that accented speech may indeed be comprehensible (e.g., Munro & Derwing, 1995) as well as that which posits that L2 learners are competent raters (e.g., Derwing & Munro, 2013; Rossiter, 2009; Wilkerson, 2010). As in previous studies investigating the role of L1 listener status in accent ratings, the English-German L2 listeners and German native listeners in the current study did not differ from one another in their accentedness ratings (e.g., Flege, 1988; MacKay et al., 2006; Munro et al., 2006).

https://doi.org/10.1017/S0272263115000418 Published online by Cambridge University Press

German native listeners and English-German L2 listeners did differ in the ratings they assigned to the speech samples along the fluency and comprehensibility continua. The English-German L2 listeners assigned significantly harsher ratings to the speech samples along these two continua than did the German native listeners. The finding for fluency ratings is in line with the results of Rossiter (2009), whose advanced L2 English listeners assigned the harshest ratings of all three participant groups. This is in spite of the fact that Rossiter's listeners were advanced EFL learners. To date only Derwing and Munro (2013) and Munro et al. (2006) have compared the perceived comprehensibility ratings assigned by native listeners to those of L2 listeners. The findings from these studies indicate that L2 listeners' responses are in line with those of native listeners. Thus, although the comprehensibility results in the current study may seem to stand counter to those of previous studies, they do support the results of research highlighted by Rossiter (2009), whereby L2 learners have been shown to evaluate the errors made by other L2 learners more severely than native speakers do (e.g., Derwing, Rossiter, & Ehrensberger-Dow, 2002; Hughes & Lascaratou, 1982). One important reason why the comprehensibility results are not in line with those of previous studies may have to do with a factor raised at the outset of the paper: L2 learners' relatively low proficiency in German. The mean proficiency level of participants in the current study corresponds to the CEFR B1 level, indicating that participants exhibited on average low-intermediate proficiency. As such, they may have applied different criteria when performing their ratings, due, among other things, to more limited lexical and grammatical knowledge. The fact that the fluency ratings assigned by L2 learners in the current study were in line with those of previous research may have to do with the fact that knowledge of a language is not required when participants rate speech samples according to the narrow definition of fluency.

Methodological Issues

At first glance, it may appear as though the methodological issues taken up in the current paper—rating condition and order effects—play little to no role in the ratings assigned by listeners. While this is the case for the group of listeners as a whole, significant rating condition effects were found for German native listeners in their fluency ratings, and the rating condition effects for English-German L2 listeners were slightly significant for comprehensibility. In both instances, listeners assigned lower ratings—indicative of more fluency and greater comprehensibility—when they rated the speech samples along all three continua at once as opposed to when they performed the ratings separately. To date no study has

https://doi.org/10.1017/S0272263115000418 Published online by Cambridge University Press

specifically investigated the effects of rating condition on rating outcomes, but these results may support the findings of Munro and Derwing (1994), who found that listeners assigned harsher ratings to utterances that they listened to more often. Similarly, participants in the current study who performed the ratings together heard each speech sample only one time and then performed all of the ratings. Those who performed the ratings separately heard each speech sample three times.

No effects of order were found for either listener group. The results of the current study therefore support those of Derwing and Munro (1997), who found no order effects when listeners rated each speech sample twice. Order effects—like rating condition effects—may have to less to do with order as such and may instead be due to the number of times a listener hears a given utterance.

IMPLICATIONS

Nonnative speakers teach and assess L2 learners on a regular basis, and this study adds to the body of literature that indicates that L2 learners—in this case L2 learners of German who speak English as an L1—rate the speech of fellow L2 learners similarly to native speakers along the accentedness scale. Although German native listeners were significantly more lenient in their ratings than the English-German L2 listeners along the fluency and comprehensibility continua, participants in the two listener groups showed similar trends in all of their ratings: accentedness ratings were harsher than fluency ratings, which were harsher than comprehensibility ratings.

As noted by Isaacs and Thomson (2013), studies that examine methodological issues in pronunciation research have implications for both L2 pronunciation researchers and for the language testing community (p. 157). Although future research is required on the topic, the results of the current study imply that performing ratings along three continua after hearing a speech sample once is at least as effective as rating a series of stimuli along each continuum separately. In fact, it may be that listeners may become more irritated by or sensitive to errors in the speech stream the more often they hear them.

The results indicate that both native and nonnative raters, all of whom were trained on the operationalization of the three concepts—accentedness, fluency, and comprehensibility—distinguished among the terms. This study therefore supports a recommendation made by Munro and Derwing (1995): developers of assessment schemes should consider which of the concepts matter for the type of assessment being carried out. Such schemes should provide independent guidelines for assessing each of the concepts as opposed to including a single category like "pronunciation." Whereas many formal assessment schemes

https://doi.org/10.1017/S0272263115000418 Published online by Cambridge University Press

like the Common European Framework of Reference (Council of Europe, 2001) and the *ACTFL Proficiency Guidelines* for speaking (American Council on the Teaching of Foreign Languages, 2012) clearly distinguish among the concepts—especially accent and comprehensibility, research has shown that difficulties may arise when the concepts are not clearly operationalized in assessments of nonnative speakers (e.g., Hanna & Allen, 2012).

Future research on the topic would benefit from making use of L2s that have not been investigated in previous studies as well as from L2 listener raters from a wider variety of L1 backgrounds. An additional factor that should also be considered in future studies includes L2 listeners' proficiency, as this may play an especially important role in comprehensibility ratings.

Received 4 July 2015
Accepted 2 October 2015
Final Version Received 5 October 2015

NOTES

1. I'd like to thank an anonymous reviewer for bringing up this important point.
2. The "suitcase narrative" was used with permission from M. J. Munro and T. M. Derwing.

REFERENCES

Abrahamsson, N., & Hyltenstam, K. (2008). The Robustness of aptitude effects in near-native second language acquisition. *Studies in Second Language Acquisition*, *30*, 481–509. http://dx.doi.org/10.1017/S027226310808073X.

American Council on the Teaching of Foreign Languages. (2012). *ACTFL Proficiency Guidelines*. http://actflproficiencyguidelines2012.org/speaking.

Anderson-Hsieh, J., Johnson, J., & 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. http://dx.doi.org/10.1111/j.1467-1770.1992.tb01043.x.

Bongaerts, T., Mennen, S., & van der Slik, F. (2000). Authenticity of pronunciation in naturalistic second language acquisition: The case of very advanced learners of Dutch as a second language. *Studia Linguistica*, *54*, 298–308. http://dx.doi.org/10.1111/1467-9582.00069.

Bortfeld, H., Leon, S. D., Bloom, J. E., Schober, M. F., & Brennan, S. E. (2001). Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. *Language and Speech*, *44*, 123–147.

Brand, C., & Götz, S. (2011). Fluency versus accuracy in advanced spoken learner language: A multi-method approach. *International Journal of Corpus Linguistics*, *16*, 255–275. http://dx.doi.org/10.1075/ijcl.16.2.05bra.

Braun, B., Lemhöfer, K., & Mani, N. (2011). Perceiving unstressed vowels in foreign-accented English. *The Journal of the Acoustical Society of America*, *129*, 376–387. http://dx.doi.org/10.1121/1.3500688.

Caspers, J. (2010). The influence of erroneous stress position and segmental errors on intelligibility, comprehensibility and foreign accent in Dutch as a second language. *Linguistics in the Netherlands*, *27*, 17–29. http://dx.doi.org/10.1075/avt.27.03cas.

Chakraborty, R., & Goffman, L. (2011). Production of lexical stress in non-native speakers of American English: Kinematic correlates of stress and transfer. *Journal of Speech,*

https://doi.org/10.1017/S0272263115000418 Published online by Cambridge University Press

Language, and Hearing Research, *54*, 821–835. http://dx.doi.org/10.1044/1092-4388(2010/09-0018.
Chen, H. (2010). Second language timing patterns and their effects on native listeners' perceptions. *Concentric: Studies in Linguistics*, *36*, 183–212.
Clark, H. H., & Fox Tree, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition*, *84*, 73–111.
Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press. http://www.coe.int/t/dg4/linguistic/source/framework_en.pdf.
Cucchiarini, C., Strik, H., & Boves, L. (2002). Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. *Journal of the Acoustical Society of America*, *111*, 2862–2873. http://dx.doi.org/10.1121/1.1471894.
de Jong, N. H., Steinel, M. P., Florijn, A., Schoonen, R., & Hulstijn, J. H. (2013). Linguistic skills and speaking fluency in a second language. *Applied Psycholinguistics*, *34*, 893–916. http://dx.doi.org/10.1017/S0142716412000069.
Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition*, *19*, 1–16.
Derwing, T. M., & Munro, M. J. (2009a). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching*, *42*, 476–490. http://dx.doi.org/10.1017/S026144480800551X.
Derwing, T. M., & Munro, M. J. (2009b). Comprehensibility as a factor in listener interaction preferences: Implications for the workplace. *The Canadian Modern Language Review/La Revue Canadienne Des Langues Vivantes*, *66*, 181–202. http://dx.doi.org/10.3138/cmlr.66.2.181.
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*, 163–185. http://dx.doi.org/10.1111/lang.12000.
Derwing, T. M., Munro, M. J., & Thomson, R. I. (2008). A longitudinal study of ESL learners' fluency and comprehensibility development. *Applied Linguistics*, *29*, 359–380. http://dx.doi.org/10.1093/applin/amm041.
Derwing, T. M., Rossiter, M. J., & Ehrensberger-Dow, M. (2002). "They speaked and wrote real good": Judgements of non-native and native grammar. *Language Awareness*, *11*, 84–99. http://dx.doi.org/10.1080/09658410208667048.
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. http://dx.doi.org/10.1111/j.1467-9922.2004.00282.x.
Flege, J. E. (1984). The detection of French accent by American listeners. *Journal of the Acoustical Society of America*, *76*, 692–707. http://dx.doi.org/10.1121/1.391256.
Flege, J. E. (1988). Factors affecting degree of perceived foreign accent in English sentences. *Journal of the Acoustical Society of America*, *84*, 70–79. http://dx.doi.org/10.1121/1.396876.
Freed, B. F. (1995). What makes us think that students who study abroad become fluent? In B. F. Freed (Ed.), *Second language acquisition in a study-abroad context* (pp. 123–148). Amsterdam: John Benjamins.
Goethe Institut. (2004). *Einstufungstest* [Placement test]. http://www.goethe.de/cgi-bin/einstufungstest/einstufungstest.pl.
González-Bueno, M. (1997). Voice-onset-time in the perception of foreign accent by native listeners of Spanish. *International Review of Applied Linguistics*, *35*, 251–267. http://dx.doi.org/10.1515/iral.1997.35.4.251.
Gui, M. (2012). Exploring differences between Chinese and American EFL teachers' evaluations of speech performance. *Language Assessment Quarterly*, *9*, 186–203. http://dx.doi.org/10.1080/15434303.2011.614030.
Gut, U. (2009). *Non-native speech: A corpus-based analysis of phonological and phonetic properties of L2 English and German*. Frankfurt: Peter Lang.
Hahn, L. D. (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly*, *38*, 201–223. http://dx.doi.org/10.2307/3588378.
Hanna, P. L., & Allen, A. (2012). Educator assessment: Accent as a measure of fluency in Arizona. *Educational Policy*, *27*, 711–738. http://dx.doi.org/10.1177/0895904811429293.
Hayes-Harb, R., & Hacking, J. F. (2015). Beyond rating data: What do listeners believe underlies their accentedness judgments? *Journal of Second Language Pronunciation*, *1*, 43–64. http://dx.doi.org/10.1075/jslp.1.1.02hay.

https://doi.org/10.1017/S0272263115000418 Published online by Cambridge University Press

Hughes, A., & Lascaratou, C. (1982). Competing criteria for error gravity. *ELT Journal*, *36*, 175–182. http://dx.doi.org/10.1093/elt/36.3.175.

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. http://dx.doi.org/10.1080/15434303.2013.769545.

Isaacs, T., & Trofimovich, P. (2011). Phonological memory, attention control, and musical ability: Effects of individual differences on rater judgments of second language speech. *Applied Psycholinguistics*, *32*, 113–140. http://dx.doi.org/10.1017/S0142716410000317.

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 Report Series, 4. https://www.ielts.org/pdf/Isaacs%20et%20al%20final.pdf.

Kang, O. (2010). Relative salience of suprasegmental features on judgments of L2 comprehensibility and accentedness. *System*, *38*, 301–315. http://dx.doi.org/10.1016/j.system.2010.01.005.

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. http://dx.doi.org/10.1080/15434303.2011.642631.

Kang, O., Rubin, D., & Pickering, L. (2010). Suprasegmental measures of accentedness and judgments of language learner proficiency in oral English. *The Modern Language Journal*, *94*, 554–566. http://dx.doi.org/10.1111/j.1540-4781.2010.01091.x.

Kennedy, S., & Trofimovich, P. (2008). Intelligibility, comprehensibility and accentedness of L2 speech: The role of listener experience and semantic context. *The Canadian Modern Language Review*, *64*, 459–489. http://dx.doi.org/10.3138/cmlr.64.3.459.

Kim, Y.-H. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing*, *26*, 187–217. http://dx.doi.org/10.1177/0265532208101010.

Koponen, M., & Riggenbach, H. (2000). Overview: Varying perspectives on fluency. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 5–24). Ann Arbor: University of Michigan Press.

Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, *32*, 145–164. http://dx.doi.org/10.1016/j.system.2004.01.001.

Leikin, M., Ibrahim, R., Eviatar, Z., & Sapir, S. (2009). Listening with an accent: Speech perception in a second language by late bilinguals. *Journal of Psycholinguistic Research*, *38*, 447–457. http://dx.doi.org/10.1007/s10936-009-9099-1.

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning*, *40*, 387–417. http://dx.doi.org/10.1111/j.1467-1770.1990.tb00669.x.

Long, M. H. (1990). Maturational constraints on language development. *Studies in Second Language Acquisition*, *12*, 251–285. http://dx.doi.org/10.1017/S0272263100009165.

Mackay, I. R. A., Flege, J. E., & Imai, S. (2006). Evaluating the effects of chronological age and sentence duration on degree of perceived foreign accent. *Applied Psycholinguistics*, *27*, 157–183. http://dx.doi.org/10.1017/S0142716406060231.

Major, R. C. (2007). Identifying a foreign accent in an unfamiliar language. *Studies in Second Language Acquisition*, *29*, 539–556. http://dx.doi.org/10.1017/S0272263107070428.

Molnar, H. (2010). The influence of the age factor on pronunciation competence in a second language: Results of a pilot study with learners of German as a second language [Der Einfluss des Faktors Alter auf die Aussprachekompetenz in der L2: Ergebnisseeiner Pilotstudie mit DaZ-Lernern]. *Zeitschrift für Interkulturellen Fremdsprachenunterricht*, *15*, 42–60. http://zif.spz.tu-darmstadt.de/jg-15-1/beitrag/Molnar1.htm.

Moyer, A. (1999). Ultimate attainment in L2 phonology. *Studies in Second Language Acquisition*, *21*, 81–108.

Moyer, A. (2011). An investigation of experience in L2 phonology: Does quality matter more than quantity? *The Canadian Modern Language Review/La Revue Canadienne Des Langues Vivantes*, *67*, 191–216. http://dx.doi.org/10.3138/cmlr.67.2.191.

Munro, M. (1995). Nonsegmental factors in foreign accent: Rating of foreign speech. *Studies in Second Language Acquisition*, *17*, 17–33. http://dx.doi.org/10.1017/S0272263100013735.

https://doi.org/10.1017/S0272263115000418 Published online by Cambridge University Press

Munro, M. J., & Derwing, T. M. (1994). Evaluations of foreign accent in extemporaneous and read material. *Language Testing*, *11*, 253–266. http://dx.doi.org/10.1177/026553229401100302.
Munro, M., & Derwing, T. M. (1995). Foreign accent, comprehensibility and intelligibility in the speech of second language learners. *Language Learning*, *45*, 73–97.
Munro, M. J., & Derwing, T. M. (2001). Modeling perceptions and the accentedness and comprehensibility of L2 speech. *Studies in Second Language Acquisition*, *23*, 451–468.
Munro, M. J., Derwing, T. M., & Burgess, C. S. (2010). Detection of nonnative speaker status from content-masked speech. *Speech Communication*, *52*, 626–637. http://dx.doi.org/10.1016/j.specom.2010.02.013.
Munro, M. J., Derwing, T. M., & Morton, S. L. (2006). The mutual intelligibility of L2 speech. *Studies in Second Language Acquisition*, *28*, 111–131. http://dx.doi.org/10.1017/S0272263106060049.
Neuendorf, J. A. (2010). *The intelligibility and comprehensibility of learner speech in Russian: A study in the salience of pronunciation, lexicon, grammar and syntax* (Unpublished doctoral dissertation). Bryn Mawr College.
O'Brien, M. G. (2004). Pronunciation matters. *Die Unterrichtspraxis: Teaching German*, *37*, 1–9. http://dx.doi.org/10.1111/j.1756-1221.2004.tb00068.x.
O'Brien, M. G. (2014). L2 learners' assessments of accentedness, fluency and comprehensibility of native and nonnative German speech. *Language Learning*, *64*, 715–748. http://dx.doi.org/10.1111/lang.12082.
Pinget, A-F., Bosker, H. R., Quené, H., & de Jong, N. H. (2014). Native speakers' perceptions of fluency and accent in L2 speech. *Language Testing*, *31*, 349–365. http://dx.doi.org/10.1177/0265532214526177.
Préfontaine, Y. (2013). Perceptions of French fluency in second language speech production. *Canadian Modern Language Review*, *69*, 324–348. http://dx.doi.org/10.3138/cmlr.1748.
Reinisch, E., & Weber, A. (2012). Adapting to suprasegmental lexical stress errors in foreign-accented speech. *Journal of the Acoustical Society of America*, *132*, 1165–1176. http://dx.doi.org/10.1121/1.4730884.
Riazantseva, A. (2001). Second language proficiency and pausing: A study of Russian speakers of English. *Studies in Second Language Acquisition*, *23*, 497–526.
Riggenbach, H. (1991). Towards an understanding of fluency: A microanalysis of nonnative speaker conversations. *Discourse Processes*, *14*, 423–441.
Riney, T. J., Takada, M., & Ota, M. (2000). Segmentals and global foreign accent: The Japanese flap in EFL. *TESOL Quarterly*, *34*, 711–737. http://dx.doi.org/10.2307/3587782.
Riney, T. J., Takagi, N., & Inutsuka, K. (2005). Phonetic parameters and perceptual judgments of accent in English by American and Japanese listeners. *TESOL Quarterly*, *39*, 441–466. http://dx.doi.org/10.2307/3588489.
Rossiter, M. J. (2009). Perceptions of L2 fluency by native and non-native speakers of English. *The Canadian Modern Language Review*, *65*, 395–412. http://dx.doi.org/10.3138/cmlr.65.3.395.
Saito, K., Trofimovich, P., & Isaacs, T. (forthcoming). Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels. *Applied Psycholinguistics.* http://dx.doi.org/10.1017/S0142716414000502.
Saito, K., Webb, S., Trofimovich, P., & Isaacs, T. (forthcoming). Lexical correlates of comprehensibility versus accentedness in second language speech. *Bilingualism: Language and Cognition*. http://dx.doi.org/10.1017/S1366728915000255.
Segalowitz, N. (2007). Access fluidity, attention control, and the acquisition of fluency in a second language. *TESOL Quarterly*, *41*, 181–186. http://dx.doi.org/10.1002/j.1545-7249.2007.tb00047.x.
Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics*, *17*, 84–119. http://dx.doi.org/10.1093/applin/17.1.84.
Trofimovich, P., & Baker, W. (2006). Learning second language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech. *Studies in Second Language Acquisition*, *28*, 1–30. http://dx.doi.org/10.1017/S0272263106060013.
Trofimovich, P., & Isaacs, T. (2012). Disentangling accent from comprehensibility. *Bilingualism: Language and Cognition*, *15*, 905–916. http://dx.doi.org/10.1017/S1366728912000168.

https://doi.org/10.1017/S0272263115000418 Published online by Cambridge University Press

Trofimovich, R., Gatbonton, E., & Segalowitz, N. (2007). A dynamic look at L2 phonological learning: Seeking processing explanations for implicational phenomena. *Studies in Second Language Acquisition*, *29*, 407–448. http://dx.doi.org/10.1017/S027226310707026X.

Wilkerson, M. (2010). Identifying accent in German: A comparison of native and nonnative listeners. *Die Unterrichtspraxis*, *43*, 144–153. http://dx.doi.org/10.1111/j.1756-1221.2010.00076.x.

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. http://dx.doi.org/10.1002/tesq.73.

https://doi.org/10.1017/S0272263115000418 Published online by Cambridge University Press