Comparing foreign accent in L1 attrition and L2 acquisition: Range and rater effects

Monika S. Schmid[1] and Holger Hopp[2]

[1] Rijksuniversiteit Groningen, [2] Universität Mannheim

m.s.schmid@rug.nl, hhopp@rumms.uni-mannheim.de

Address for correspondence:

Dr. Holger Hopp

Universität Mannheim

Anglistische Linguistik

Schloss EW 266

68131 Mannheim

Phone: ++49-621-181-3160

Fax: ++49-621-181-2336

Running head: Comparing foreign accent in L1 attrition and L2 acquisition

Abstract

This study examines the methodology of global foreign accent ratings in studies on L2 speech production. In three experiments, we test how variation in raters, ranges of speech samples as well as instructions and procedure affects ratings of native and foreign accents in predominantly monolingual speakers of German, non-native speakers of German as well as long-term emigrants from Germany, i.e. L1 attriters. The findings show that rater differences do not result in systematic changes in rating patterns. In contrast, range effects and effects of familiarity with accented speech lead to shifts in absolute and relative ratings. Including more strongly foreign-accented samples leads to better judgments for the entire group of bilinguals compared to natives. Similarly, lower familiarity with foreign accents results in more variable and more strongly foreign-accented judgments. We discuss the implications for research on L2 pronunciation as well as for the interpretation of nativeness in L2 studies and language testing more generally.

Keywords: Bilingualism, L2 speech production, L1 attrition, L2 acquisition, methodology

Introduction

One of the most controversial issues in research on multilingualism is whether whether late, i.e. post-pubescent, second language learners (L2 learners) can ever become fully native-like in the L2. Many studies suggest that nativelike performance is achievable in some areas, for example lexical semantics, but more variable in areas like morphosyntax or articulatory phonetics. The present paper focusses on methodological issues of how it can be established whether two populations - e.g. monolingual natives and L2 learners - perform 'the same' or 'differently' in what is arguably the most challenging task in (late) L2 learning, namely attaining a native accent in L2 pronunciation.

In spite of the central importance of the debate about nativeness and its theoretical implications for second language research and language testing, the exact nature of the 'native target' as well as questions of how to establish, measure and assess it, have received relatively little attention (Davies, 2003; Davies, Hamp-Lyons & Kemp, 2003; Harding, 2013). With respect to L2 phonology, there are two possible approaches to determining to what extent an L2er has attained a 'native accent'. The first applies measurements either to the output (e.g. Flege, 1987) or to articulatory settings during speech production (see Mennen et al., 2011 for an overview). The second consists of eliciting native speaker ratings of segments of speech produced by L2 learners for their accentedness.

Both approaches have advantages and drawbacks. Phonetic characteristics of the sound stream or articulatory features such as lip rounding, jaw position etc. can be measured accurately and objectively. Such measurements are, however, extremely time consuming, can only be applied to individual phonemes and, in the case of physical measurements of the articulatory apparatus, can be quite intrusive and thus potentially impact on speech production (Mennen et al., 2011).

Global foreign accent ratings (FARs), on the other hand, can be elicited quickly, from a large range of raters at the same time, and provide a holistic assessment of all aspects of speech including suprasegmentals, intonation, fluency and so on. Explicitly or implicitly, global foreign accent forms part in assessing pronunciation skills in second-language oral production tests and is implicated in assessing fluency and accuracy in oral proficiency testing (e.g. Harding, 2013; Taylor & Geranpayeh, 2011). However, global foreign accent ratings rest on the implicit assumption that foreign accent can be captured on a unidimensional, linear scale of native to foreign accents, or assigned to one of a given number of discrete categories. This view, however, is an extreme oversimplification of a phenomenon that is comprised of a large variety of features. Accent is detectable not only in segmental articulation but on the subsegmental and suprasegmental level (stress, pitch, tone) as well as in intonation, speech rate and rhythm (Munro, 1995; Southwood & Flege, 1999). Furthermore, features such as disfluency markers and hesitation phenomena also play a role in the assessment of foreign versus native speech (Dewaele, 1996; Lennon, 1990).

A study by McDermott (1986) suggests that these components of foreign accent do not necessarily covary and that individual raters may rely on different features when judging foreign accent. Such differences between raters may impact on absolute and relative judgements of foreign accent. For instance, rater 1 may perceive speech sample X as more (or less) accented than rater 2, they may differ in how they assess sample X in relation to sample Y, and the distance between ratings on a Likert scale may not be the same for both raters (Kuiken & Vedder, this volume). In addition, differences in experimental design, for example concerning the range of accents represented in the samples to be rated, may influence FARs. For instance, a very slightly accented speaker may stand out among a majority of native samples, but may be perceived within the native range when a large number of more heavily accented speakers are to be rated at the same time. As Southwood and Flege (1999) point out, there are no physical units in which accent can be measured, and there is therefore the danger in a FAR experiment of the raters applying a contraction bias, i.e., overestimating small differences and underestimating large ones (Southwood & Flege, 1999, 344).

Before conclusions about (non-)nativeness can be drawn from global foreign accent ratings, it is therefore important to establish the extent to which rater and range effects may impact on the results. However, the literature that systematically discusses the methodological underpinnings of FAR studies and the consequences of methodological variation in assessments of foreign accent for L2 oral production testing is extremely limited to date. From an SLA perspective, we provide a review of previous studies that have used different methodologies in FAR experiments, and we discusses how these affect the overall outcomes. We then present the results of a series of experiments that address the issue of whether variation in raters, ranges of samples, instructions and procedures has significant effects on absolute and relative foreign accent ratings. In systematic comparisons based on comparatively large samples of speakers and raters, we show how methodological differences affect the validity of foreign accent ratings and, by consequence, the conclusions about (non-) nativeness in L2 speech production that have been drawn on their basis.

Methodological considerations

Any researcher wishing to conduct a foreign accent rating experiment will have to give consideration to a number of methodological issues concerning the material to be rated, the speakers to be rated, the raters and the rating scales. In many previous studies, these decisions appear to have been informed largely by convenience and are often not explicitly motivated or even reported, even though there is increasing evidence to show that seemingly minor variations in the setup can impact the outcome (for an overview, see Jesney, 2004).

Materials

There are three ways in which speech samples to be used in foreign accent ratings have been elicited: (delayed) repetition, reading and extemporaneous speech. In the first, speakers are presented auditorily with a sentence recorded by a native speaker and are asked to repeat it (e.g. Elliott, 1995; Jilka, 2000). This technique, which was common in the 1990s (Jesney, 2004), was later adapted in favour of a delayed strategy of presenting a short question recorded by one speaker, followed by an answer from another. The question is then repeated, cueing the participant to repeat the model answer sentence (Flege, Birdsong, Bialystok, Mack, Sung & Tsukada, 2006; Flege, Munro & McKay, 1995; Guion, Flege & Loftin, 2000; MacKay, Flege & Imai, 2006). The repetition condition can probably be seen as the 'easiest' type of speech for the L2 learners who are primed directly with a native model. It has never been tested to what extent foreign accented speech produced under these conditions correlates with accentedness in other speech situations. Unless it can be established that one is a good predictor of the other, experiments or assessments based on this type of material are open to criticisms such as those from Long (2005), who points out that some studies of ultimate attainment of L2 learners rely on 'language-like behavior' (p. 297) or the fact that some advanced L2 learners may be able to perform to native levels 'on certain well-defined tasks or within certain restricted domains' (p. 289) but fail to replicate this performance under the more taxing demands of real-life language use.

The decision of whether to rely on material that is read or on spontaneous speech may be a more difficult one. On the one hand, presenting speakers with words, phrases, sentences or paragraphs to read out elicits samples that are directly comparable and can also be manipulated to contain specific sounds or constructions that the researcher is interested in, and which L2 speakers might choose to avoid in extemporaneous speech. On the other hand, reading aloud may again rely on different skills than speaking freely, for instance a higher degree of monitoring and literacy. In any case, potential differences in script or grapheme-phoneme correspondence between the L1 and the target language add variables to the task that are specific to reading, yet independent of L2 speech production as such. This may be taken to imply that the best basis for a comparison of accentedness between natives and L2 learners is free speech.

The second question is how long these samples should be. As Flege (1984) demonstrated, native speakers are more than two-thirds accurate at identifying non-natives in segments of a mere 30ms (containing only the initial voiceless alveolar plosive). Accuracy increases to 89%, however, when full phrases are presented. Most studies appear to follow this approach and choose full phrases or sentences, usually resulting in clips of 10-20 seconds long. To the best of our knowledge, there are no studies investigating whether judgments become more accurate when longer stretches of text are used. As sample length increases, there is more diversity in the aspects of the samples on which raters can focus, and longer samples may result in an overlong overall experiment causing rater fatigue and therefore a decrease in accuracy. Therefore, the present study uses samples ranging from 10 to 20 seconds in length.

Speakers

Southwood and Flege (1999) established that listeners are in principle capable of perceiving foreign accents on a linear scale and partitioning it into equal-seeming increments in a given experimental setting. However, human perception of gradient scales is amenable to context effects and so are judgements of linguistic stimuli. For instance, Sprouse (2008) showed that judgements of sentence acceptability are non-linearly modulated by the levels of acceptabilty of previous sentences. Similarly, it is to be expected that a speaker with a mild foreign accent will be rated as less accented if his speech sample is embedded among or follows more heavily accented speakers, but will stand out as a non-native among all native samples.

There is, so far, only one study investigating the impact of range effects on foreign accent ratings, namely Flege and Fletcher (1992) who manipulated the proportion of native controls (20% versus 40%) and concluded that a higher proportion of identifiably native speakers leads to the perception of stronger foreign accents for the nonnatives. While the proportion of native speakers has thus been shown to affect relative FARs for non-natives, there are currently no investigations of how the presence of more strongly accented samples among the non-native speakers themselves impacts on the perception of the less accented ones, or vice versa. Since the relative degrees of accentedness represented in the samples may influence the absolute judgements of foreign accent, we vary the range of foreign accented samples in the present study.

A second question concerns the nature of the baseline, i.e. the native controls. Most studies have chosen monolingual speakers as the reference norm. It was recently pointed out that this might constitute an unrealistic target (Hopp & Schmid, 2013), since even beginning bilinguals show some impact of L2 transfer to the L1, as Chang (2012) has demonstrated for native English speakers enrolled in a six-week intensive beginner’s course of Korean, who exhibited phonetic drift, i.e. a change away from the native norm, in their L1. Hopp and Schmid (2013) therefore argue that a control group of L1 speakers who are themselves experienced bilinguals may be better suited, in particular for studies with the aim of investigating questions of ultimate attainment. In the present study, we therefore use different groups and samples of non-native speakers, predominantly monolingual native speakers as well as long-term L1 attriters in order to test for range effects in the relative assessment of foreign accent.

Raters

Studies on foreign accent vary greatly in the number of raters used. Some studies use only one or a few raters (e.g. Snow & Hoefnagel-Hoehle, 1977) while others employ hundreds of judges rating the samples (e.g. Anderson-Hsieh & Koehler, 1988). The raters are typically native speakers of the language to be judged. However, some studies have used advanced or near-native L2 speakers and report that they make judgements similar to those from predominantly monolingual native speakers (e.g. Elliott, 1995). A recent series of studies by Major (2007, 2010) investigated how (non-) nativeness among raters and language context affects the perception of foreign accent by comparing native raters of Brazilian Portuguese living in Brazil and in the US, L2 speakers of Brazilian Portuguese in Brazil and the US as well as monolingual English native speakers in the US with no knowledge of Portuguese. All groups, including the monolingual English speakers, were found to reliably discriminate between native and non-native speech samples, and the major differences between groups depended on whether the raters currently lived in a Brazialian Portuguese-speaking community, i.e. whether they regularly listened to the language to be judged. Other studies on holistic oral production assessment find that rater familiarity with the particular language combinations in the speakers affects ratings among raters who are native (e.g. Carey, Mannell & Dunn, 2011; Winke, Gass & Myford, 2012) as well as non-native (Xi & Mollaun, 2011, Zhang & Elder, 2011) speakers of the language to be rated. In addition, several studies report that familiarity with regional accents and dialects that may occur in the speech samples affects ratings of foreign accent (e.g. Bongaerts et al., 1997; Flege, Frieda & Nozawa, 1997). Equally, raters who are familiar with foreign-accented speech in general make more nuanced judgements than judges without much experience of non-native speech (Thompson, 1991).

Whereas these findings indicate that judges need to be familiar with the types of samples they are asked to judge, general training on phonetics does not seem to affect foreign accent ratings. Groups of phonetically trained judges and untrained raters display great agreement in their overall ratings (e.g. Bongaerts et al., 1997; Hopp & Schmid, 2013), even though interrater reliability was found to be higher among trained raters in some studies (Thompson, 1991, though see Xi & Mollaun, 2011).

In most studies, speakers and raters were matched in age, region and level of education. There is little evidence, though, that differences in these factors engender variation in foreign accent ratings. In what is still the only systematic study of individual rater differences, McDermott (1986) reported that background variables accounted for differences in raters' emphasis on different aspects in rating foreign accent, e.g. rhythm, hesitation, loudness, etc. They did not, however, lead to systematic variation in the overall assessment of foreign accent (see also Munro & Derwing, 1995). Of the many predictors tested, only age, sex and ethnic diversity of the neighbourhood of the raters predicted accentedness ratings, with younger, male raters who live in ethnically homogeneous all-English-speaking neighbourhoods giving the strictest ratings. Since McDermott's findings were based on ratings of samples elicited from nine non-native speakers only and no native speaker controls were used in the study, research remains to be done on interrater differences in the assessment of foreign accent. By consequence, Experiment 1 in the present study aims to test whether there are individual differences between raters in their ratings.

Procedure

One of the basic problems in FAR studies is the fact that notions such as "foreign accent" and/or "native speaker", which are subjective, fuzzy and intuitive concepts, are rarely if ever defined in the instructions given to the raters. In addition, there is little information in published studies on whether raters were told about differences between foreign and regional accents, and whether they received information as to whether they should take regional accents into account when judging for nativeness.

As for scales, by far the majority of FAR studies use three- to ten-point Likert scales with labelled endpoints (see Jesney, 2004). A few studies have used other measurement scales, such as sliding scales or Magnitude Estimation. Southwood & Flege (1999) directly compared findings from Magnitude Estimation and a 7-point Likert scale and report high correlations.

There is furthermore diversity in the type of assessment raters are asked to make. Many studies explicitly ask to rate the degree of (foreign) accent on the scale provided (e.g. Flege, MacKay & Piske, 2002), while others label the endpoints of the scale as “native “ and “non-native” (e.g. Moyer, 1999), “very good pronunciation” and “very poor pronunciation” (e.g. Yeni-Komshian, Flege & Liu, 2000) or more vaguely in relative terms as “close to native English” and “less close to native English” (e.g. Magen, 1998), respectively. In some studies, the scale conflates accentedness and intelligibility (e.g. Anderson-Hsieh & Koehler, 1988) or treats these as separate dimensions to be judged individually (e.g. Munro & Derwing, 1995). Finally, several studies have raters categorize speakers as “native speakers” or “non-native speakers” (e.g. Asher & Garcia 1969, Flege, 1984) and add confidence judgements to these categorizations (e.g. De Leeuw et al., 2010; Hopp & Schmid, 2013).

Studies also differ in the amount of information about the type of samples the raters listen to. Some studies explicitly mention the L1 of the non-native speakers in the samples (e.g. Flege et al., 1995), but most do not provide information about the linguistic backgrounds of the speakers. Sometimes, information about the relative amount of native and non-native samples is provided (e.g. Flege et al., 1995), whereas most studies do not tell the raters whether there will be native samples among the stimuli.

In sum, then, there is considerable variation between studies in the procedure, although their results are interpreted in similar terms. Systematic comparisons of effects of differences in instructions, scales and background information of the raters are lacking to date. In consequence, the aim of Experiment 3 in the present study is to assess whether the type of scale and the type of instruction given to raters has systematic effects on rating behaviour.

The study

The present study consists of a series of experiments in which the same (sub)sets of speech samples from predominantly monolingual (n = 20) and bilingual (n = 80) speakers of German were assessed in a variety of settings and conditions. The speech samples were taken from Hopp & Schmid (2013) who investigated the effects of individual differences in *speakers* on foreign accent. In this study, we explore the effects of differences in *raters and methods* on foreign accent. Specifically, the aim of the individual experiments is to assess how variation among raters (Experiment 1), variation in the degree of accent among the speakers to be rated, i.e. range effects (Experiment 2, where two subsets of the original sample were rated) and variation in the type of instructions and scales used in foreign accent ratings (Experiment 3) affect the accuracy and reliability of foreign accent ratings.

Materials

All experiments in the present paper used the speech samples described in Hopp & Schmid (2013) that were elicited from three groups: native speakers, late L1 attriters, and late L2 learners. Specifically, the groups comprised 20 predominantly monolingual speakers of German (controls), 20 L1 English and 20 L1 Dutch late L2 learners of German (L2 learners) as well as 20 native Germans who had emigrated to English-speaking Canada and 20 German natives who had emigrated to the Netherlands as adults (L1 attriters,). All speakers performed a narrative-descriptive task designed to elicit free speech (for details, see Hopp & Schmid, 2013). From these narratives, speech samples ranging between 10 and 20 seconds in length were extracted.

All samples comprised full sentences and did not contain lexical borrowings or any non-target expressions or structures that would allow easy identification of non-native speech on the basis of other than phonetic or phonological factors. Foreign accent ratings which had been elicited in two precursor studies (De Leeuw et al., 2010; Hopp, 2007) on different samples from the same recordings correlated very highly with the FARs elicited by Hopp & Schmid, indicating that the samples used here were representative of the speakers' overall production.

Experiment 1

Procedure

The listeners made two judgments for each speech sample. The first binary judgment determined native versus nonnative speaker status (in answer to the question "Is this person a native speaker of German?"). The second judgment expressed the level of confidence on a 3-point scale. This resulted in an operative 6-point Likert scale: 6 = certain of nonnative speaker status, 5 = semicertain of nonnative speaker status, 4 = uncertain of nonnative speaker status, 3 = uncertain of native speaker status, 2 = semicertain of native speaker status, and 1 = certain of native speaker status. Hence, a low FAR reflects a speaker who was perceived as native or near-native, whereas a high FAR indicates that the speaker was rated as having a noticeable foreign accent in his or her German speech.

Experiment 1 was conducted in two sessions in which the samples were presented in different pseudorandomized orders.

Listeners

Two groups of listeners took part in the foreign accent assessment in two separate sessions: 76 listeners took part in the first session, and 73 listeners took part in the second. All 149 listeners were first-year students at the Department of English at the University of Mannheim, Germany. They had received no specific phonetic training. Only those listeners who reported not to have been exposed to languages other than German in childhood were retained for analysis. In all, 130 German listeners were analyzed: 68 in the first group and 62 in the second.

Results

The control group of predominantly monolingual native speakers received a mean FAR of 2.36 (sd = .95), the L1 attriters received a mean FAR of 2.79 (sd = 1.25), and the L2 learners scored a mean of 3.94 (sd = 1.46). In a one-way analysis of variance, the group differences were highly significant with a strong effect size (F(2, 98) = 14.033, p < .001, η2 = 0.47). Post-hoc comparisons (Tukey HSD) showed that the L2 learners were significantly different from both the control speakers and the L1 attriters (p < .001) but that there was no difference between L1 attriters and controls at the group level (p = .258). Despite these differences, there was considerable variation in the individual FARs across and within group.

In order to test to what extent this variation was due to rater effects in the perception of foreign accent, we assessed interrater reliability. This measure is difficult to assess in large samples, as Cronbach's α invariably increases with the number of scales tested (Field, 2005, 668). We therefore opted for the compromise of randomly dividing the sample into 10 groups of 13 raters each, and assessing the reliability of the 10 average FARs, which led to a Cronbach's α of .989. Pearson correlations of the individual groups ranged from .865 to .961.

In order to assess whether the relative judgements of speakers and groups varied according to rater characteristics, we divided the raters into three near equal-sized subgroups depending on the ratings they had given to the native control group of speakers. The first group comprised the most lenient raters, i.e. the 43 judges who gave native speakers an average FAR between 1.05 and 1.94. The second group (n = 44) were the intermediate raters, who had judged the native speakers between 1.95 and 2.35, and the third group were the strictest raters (n = 44), i.e. the judges who rated the native speakers as predominantly non-native in accent (FAR range: 2.4 – 4.15). Fig. 1a depicts the Likert-Scale FAR for each population and rater group and shows that all three rater groups differentiated reliably between the populations, but that the strictest raters located all groups higher on the rating scale. For all rater groups, the L1 attriters were statistically indistinguishable from the natives (all p's>.05), and the L2 group was significantly different from both other groups. Cronbach's α for the three group scales was excellent (.984), which suggests that the different rater groups perceived the differently accented samples similar in relation to each other, although the range of abolute FAR values was located differently on the Likert scale (Figure 1a).

To check this, we converted the Likert-scale FARs into ranks for each of the three rater groups by giving rank 1 to the speaker who had received the lowest FAR and rank 100 to the speaker with the highest FAR in each group. Figure 1b shows that, when ranked FARs are used, differences between rater groups level out.

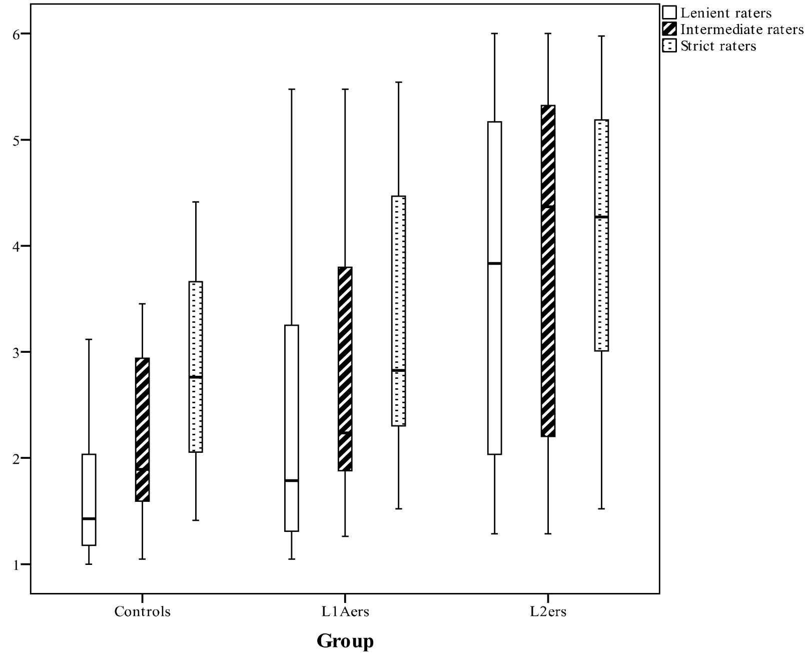

Fig. 1a: Likert-scale foreign accent ratings of the three populations by lenient, intermediate and strict raters

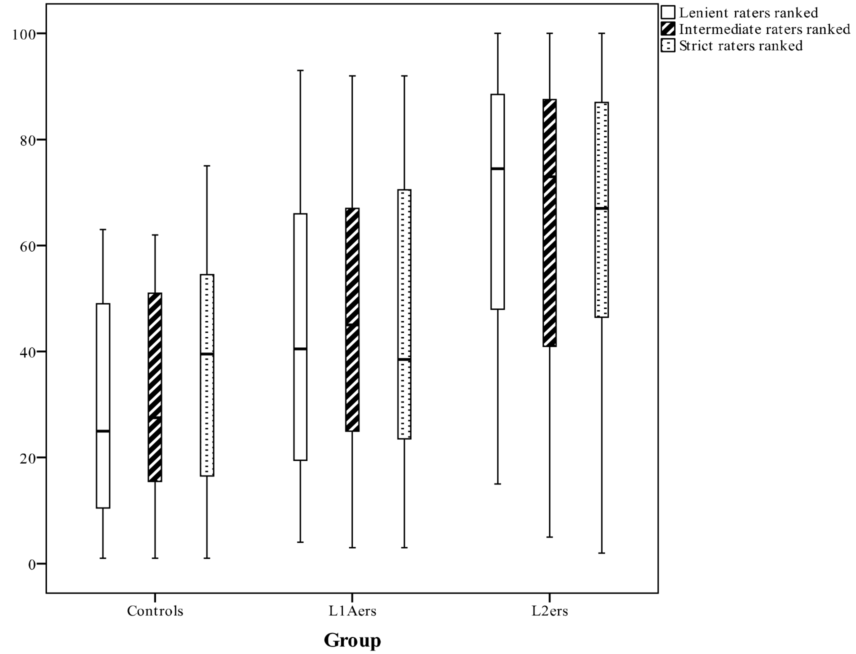

Fig. 1b: Ranked foreign accent ratings of the three poulations by lenient, intermediate and strict raters

Next, we explored whether familiarity of the raters with the languages or phonetic training affected FARs. Recall that the speakers had German as their L1/L2 and either Dutch or English as their other language(s). The raters were more familiar with English as well as English accents in German than Dutch or Dutch-accented speech because all raters were first-year students of English and had daily exposure to English. In addition, contact with English (speakers) is abundant in modern-day Germany, whereas contact with Dutch (speakers) is comparatively rare in regions like the South-West of Germany that are far from the Dutch border and where the experiment took place. In order to test whether familiarity with the language combination affected the stability of ratings, we recomputed the interrater reliability for the 10 randomly selected groups separately for the monolingual controls, the L1/2 English and the L1/2 Dutch speakers. Interrater reliability was lowest for the controls (α = .969), probably due to the fact that this sample (n = 20) was smaller than the other two (n = 40). For these two groups, reliability was virtually identical, at α = .988 for the L1/2 English and .990 for the L1/2 Dutch speakers. In paired-samples t-tests of the Pearson's correlations across individual rater groups between the L1/2 Dutch and L1/2 English speakers, only one of the 10 comparisons reached significance at the Bonferroni adjusted alpha level of 0.005, with the L1/2 English group showing higher correlations in judgements across rater groups than the L1/2 Dutch group.

To further check whether experiential differences with the languages involved influence FARs, we correlated the FARs of the 130 native German raters and the FARs of the 18 raters who stated having other or additional native languages than German. Non-native raters showed excellent within-group interrater reliability with a Cronbach's α of .896. The correlation in ratings between the native raters and the non-native raters was high at r = .931. However, in a Repeated Measures ANOVA with the factors Speaker Group and Rater Group, we find a significant interaction of Speaker and Rater Group (F(1,98) = 7.254, p = .008). Post-hoc comparisons reveal that this interaction reflects significantly higher, i.e. more strongly foreign-accented, ratings by the non-native raters for the native control speaker group (mean FAR = 2.82, sd = 0.84) than by the natives (F(1,39) = 5.487, p = .024). Non-native and native raters did not differ in their ratings of the bilingual speaker groups (all p's > .4). These group differences suggest that a lower degree of experience with the target language German affects the ratings for native speakers, yet it does not influence FARs for the bilingual speakers.

Finally, we investigated whether general familiarity with phonetics and phonology affected ratings by comparing the FARs of the 130 native raters in Experiment 1, who had not received any phonetic training, with the FARs given to the L1 attriter group by the raters in De Leeuw et al. (2010), who had had phonetic training and who rated samples from the narratives of the same speakers as were tested here. Again, the FARs correlated strongly (r = .839), which suggests that general training in phonetics does not lead to higher accuracy in the perception of a foreign accent.

In sum, the high correlations of (ranked) FARs across rater differences in Experiment 1 demonstrate that individual differences in rater strictness, raters' backgrounds or training do not have a systematic influence on FARs. However, familiarity of the raters with the target as well as the contact language affect the relative ratings in the sense that lower degrees of familiarity engender lower levels of nativeness judgements for monolingual speakers.

Experiment 2

Experiment 2 addresses the issue of range effects in FARs by testing the extent to which variation in the relative degrees of foreign accentedness in the speech samples affects listeners' judgements. Experiment 2a used those samples from L1 attriters and L2 learners that had been rated the most native-like in Experiment 1, and Experiment 2b presented samples from the L1 attriters and L2 learners that had been rated at the low end of the foreign accent scale and thus represented a homogenously foreign-accented sample.

Experiment 2a

Materials

For Experiment 2a, the ten samples that had received the most native-like FARs in Hopp & Schmid (2013) were selected from each group. In total, this means that 30 samples with an average FAR of 1.64 (sd = .28) for the natives, 1.59 (sd = .18) for the L1 attriters and 1.90 (sd = .33) for the L2 learners were chosen. There was a significant difference between the original FARs of the groups in a one-way ANOVA (F(2,27) = 3.992, p = .030).

Listeners and procedure

For Experiment 2a, 54 students from the same population as in Experiment 1 listened to the 30 samples. None of the raters had participated in Experiment 1. The FARs of fifty listeners who had had exposure only to German in childhood were analysed. The procedure was identical to Experiment 1.

Results

In Experiment 2a, the native group received a mean FAR of 1.84 (sd = 0.35), the L1 attriters a mean FAR of 1.85 (sd = 0.5), and the L2 group obtained a FAR of 2.06 (sd = 0.45). Importantly, the differences between the groups found in Experiment 1 were no longer significant in Experiment 2a (F(2,27) = 0.803, p = .458). Repeated measures ANOVAs were conducted on the FARs from Experiment 1 and the FARs awarded in Experiment 2a. In the analysis, Group was included as a covariate. Mauchly’s Test of Sphericity was not significant, indicating that there was no violation of the assumption of sphericity. For the most native-like speakers in Experiment 2a, the analysis was significant with a weak effect size (F(1, 29) = 4.464, p = .044, η^2 = .138). No interaction between the factor (Experiment 1 versus Experiment 2a) and Group was found (F(1, 29) = 0.064, p = .802, η^2 = .002).

The visual representation of the FARs allocated in Experiment 1 and Experiment 2 in Fig. 2a shows, first, that restricting the sample to the most native-like speakers in Experiment 2a resulted in a somewhat wider distribution of scores, in particular for the attriters. The L1 attriters had originally been distributed across the smallest range in Experiment 1, but conversely spread across the largest range in Experiment 2a.

Experiment 2a bears out that range effects influence (a) the absolute ratings of speakers in the three groups and (b) relative ratings between groups. In Experiment 2a, the best L2 speakers received more native-like ratings than in Experiment 1, and the formerly observed group differences between L2 learners, on the one hand, and native controls and L1 attriters, on the other, disappeared. Once the most strongly accented samples from the L2 group were removed, then, L2 speakers were perceived as native-like at the top end of the accentedness scale. We next test whether removing the most strongly accented samples also leads to range effects in absolute and relative group ratings at a lower end of the accentedness spectrum.

Experiment 2b

Materials

For Experiment 2b, thirty samples from Hopp & Schmid (2013) were selected at the low end of FARs for each group. The ten native speakers rated as least native like (mean FAR: 2.83, sd = 0.58) were included. The two bilingual populations were matched on their original FAR ratings, in other words, the ten L1 attriters who were rated the least native-like (mean FAR: 4.72, sd = 0.56) were selected. From the range delimited by these 10 attriters, 10 L2 learners were selected (mean FAR: 4.89, sd = 0.57). A one-way ANOVA showed a significant difference between groups (F(2,28) = 38.291, p < .001) where both bilingual populations differed from the controls (p < .001) but not from each other (p = .792, Tukey HSD).

Listeners and procedure

In Experiment 2b, 48 students from the same population as in Experiment 1 took part. None of the raters had participated in any of the previous experiments. The FARs of 38 listeners who had only been exposed to German in childhood were analysed. The procedure was identical to Experiments 1 and 2a.

Results

In Experiment 2b, the native speakers who had scored lowest on the native scale in Experiment 1 scored a mean FAR of 1.84 (sd = 0.43), the L1 attriter group a mean FAR of 4.28 (sd = 1.1), and the L2 group a mean FAR of 3.30 (sd = 0.77). A one-way ANOVA shows significant differences between groups (F(2,27) = 22.808, p < .001), and post-hoc Tukey comparisons show that all groups differ significantly from each other. Importantly, now the L2 group is rated as significantly less accented than the L1 attriters in Experiment 2b (p = .031), even though both groups were rated indistinguishably in Experiment 1. A repeated measures ANOVA on FARs from Experiment 1 and 2b was then conducted. Mauchly's Test of Sphericity was not significant, indicating that there was no violation of the assumption of sphericity. The result was significant with a medium effect size (F(1, 29) = 10.823, p < .01, η^2 = .279). The interaction between the factor (Experiment 1 versus Experiment 2b) and group was marginally significant ((F(1, 29) = 3.338, p = .078, η^2 = .107). As Figure 2b illustrates, the range of scores increased for the L1 attriters, while both controls and L2 learners were rated across a more narrow range. In addition, these two latter populations were both rated distinctly more towards the native end of the spectrum, while for the attriters the upper end of the scale remained the same.

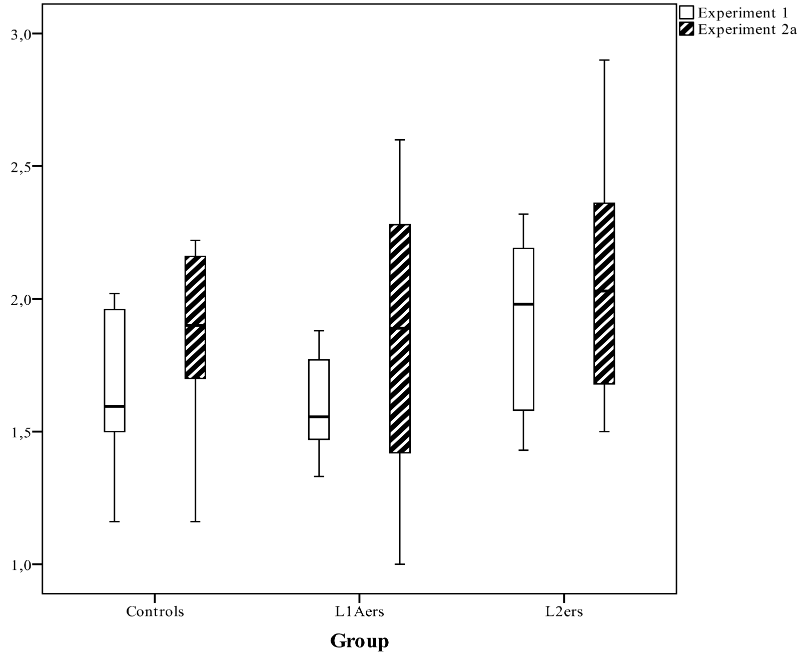

Fig. 2a: FARs received in Experiment 1 and Experiment 2a by the 10 speakers rated most native-like in their cohorts in Experiment 1

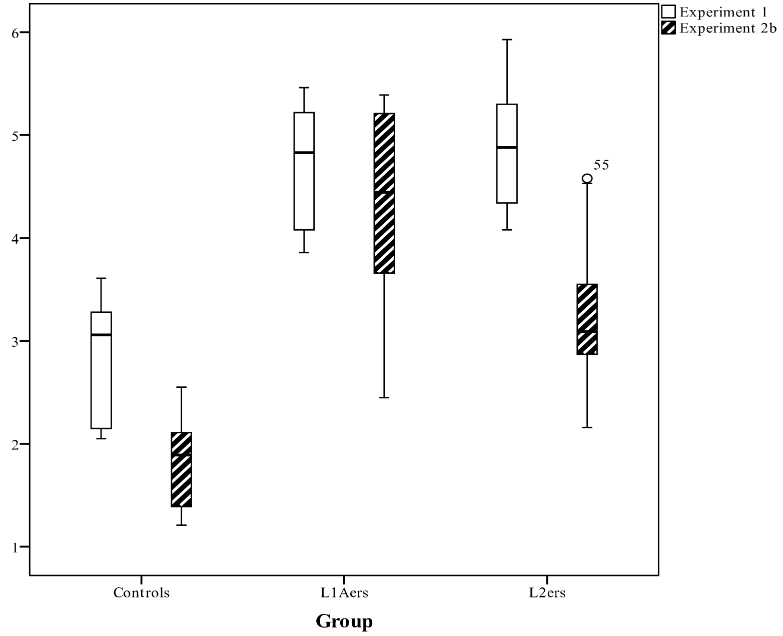

Fig. 2b: FARs received in Experiment 1 and Experiment 2b by the 10 speakers rated least native-like among the controls and attriters in Experiment 1 and 10 L2 learners matched for their FAR in Experiment 1 with the attriters

Similar to Experiment 2a, removing the most strongly accented L2 samples led to an improvement of the L2 group ratings *vis à vis* the other groups in lower ranges on the accentedness spectrum in Experiment 2b. Both experiments thus furnished evidence that differences in the selection of samples lead to range effects in the absolute and relative judgements of bilingual speakers.

Experiment 3

Experiment 3 addresses the question of whether variation in the instructions given to the raters incurs differences in accent ratings. In Experiment 3, the full set of 100 samples from Experiment was used, and raters were given three different scales for judgements.

Materials

The materials were identical to Experiment 1 but pseudo-randomized in a different order in each of the three sub-experiments described here.

Procedure

Experiment 3a used the same rating criterion as Experiments 1 and 2, i.e. listeners were asked whether the speakers had German as their L1 (“Hat diese Person Deutsch als Muttersprache?” – *Does this person have German as his/her native language?).* However, instead of making a categorical rating and giving a separate confidence rating, raters made FAR judgements on a six-point Likert scale whose endpoints were labelled “yes, definitely” and “no, definitely not”, respectively. In Experiment 3b, listeners were explicitly asked to assess foreign accent in German (“Hat diese Person einen fremdsprachlichen Akzent?” – *Does this person have a* *foreign accent?*). Listeners gave responses on a six-point Likert scale with endpoints labelled "yes, strongly" and "no, not at all". Finally, Experiment 3c elicited categorical judgements and asked listeners to categorize the speakers as native speakers, L1 attriters and L2 learners of German, respectively. As in Experiment 1, the raters indicated their level of confidence in a second step on a three-point scale. In this experiment, the raters were told before the experiment that they would encounter 100 samples in total, of which 20 were native speakers, 40 attriters and 40 L2 speakers of German. In Experiments 3a-c, the presentation of the stimuli was pseudorandomized differently for each group and otherwise identical in procedure to Experiment 1.

Listeners

All experiments used listeners from the same population as the previous experiments. No raters had participated in any of the previous experiments, nor did any raters participate in more than one of Experiments 3a-c. For Experiment 3a, FARs from 82 listeners were obtained, and 67 raters with German as the only L1 were retained for analysis. For Experiment 3b, 77 listeners were used, 54 of whom were German monolinguals whose data were used for analysis. Finally, in Experiment 3c, 70 raters made judgements, and we analysed the data from 61 raters who had only had exposure to German in childhood.

Scale effects

All three experiments in Experiment 3 tested the full set of 100 samples used in Experiment 1. The FARs from Experiment 1, 3a and 3b are summarized in Table 1.

Table 1: Foreign accent ratings in Experiments 1, 3a and 3b

	Controls		L1 attriters		L2 learners	
	mean	sd	mean	sd	mean	sd
Experiment 1	2.23	0.75	2.84	1.28	3.94	1.46
Experiment 3a	2.36	0.75	2.85	1.18	3.74	1.37
Experiment 3b	1.73	0.43	2.44	1.22	2.88	1.24

The results from Experiment 3a and 3b correlated strongly with those from Experiment 1 (r = .959 and r = .922, respectively) and with each other (r = .905). In order to determine whether there were nevertheless any differences between the three settings, the results from Experiment 3a and 3b were compared to the results from Experiment 1 by means of a Repeated Measures ANOVA with Group as a covariate. Mauchly's Test of Sphericity was not significant for either test, indicating that the data did not violate the assumption of sphericity. For the comparison of Experiment 1 and 3a, the repeated measures ANOVA was significant with an extremely weak effect size (F (1, 99) = 4.085, p = .046, η^2 = .04). The interaction of the factor Experiment and Group was also significant with a weak effect (F (1, 99) = 10.427, p < .01, η^2 = .096). This interaction stems from the native control group being judged as slightly more foreign accented and the L2 learners as slightly less accented in Experiment 3a compared to Experiment 1.

For Experiments 1 and 3b, the repeated measures ANOVA was significant with a weak effect size (F (1, 99) = 8.973, p < .041, η^2 = .087). The interaction of the factor (Experiment 1 versus Experiment 3b) and Group was significant with a medium effect (F (1, 99) = 21.933, p < .01, η^2 = .184). This interaction reflects that all groups are judged as less foreign-accented in Experiment 3b than in Experiment 1.

A comparison of the FARs from Experiment 1, Experiment 3a and Experiment 3b showed that all three were able to differentiate the populations (Experiment 1: F(2, 97) = 14.056, p < .001; Experiment 3a: F(2, 97) = 10.584, p < .001; Experiment 3b: (F(2, 97) = 6.633, p < .01). The effect size, however, was somewhat stronger in Experiment 1 than in

Experiments 3a/b (η^2 = .23 versus .18/.12). Furthermore, post-hoc tests (Tukey) revealed that while the L2 learners were rated as different from the controls in all three experiments (p < .001) and there was no significant difference between controls and L1 attriters in any experiment, L1 attriters and L2 learners differed in Experiment 1 and 3a (p < .01) but not in 3b (p = .068). Hence, when raters were asked to assess the strength of foreign accent rather than making the categorization 'native speaker or not', the group differences between L1 attriters and L2 learners disappeared.

Finally, Experiment 3c asked for categorical allocations of the speakers to groups of native speakers in Germany, L1 attriters and L2 learners. Table 2 presents the overall number and percentage of classifications in the respective categorizations. The group differences are significant (χ^2 = 401.571, p <.001) .

Table 2: Foreign accent ratings in Experiment 3c (percentages of ratings)

	Identified as native		Identified as L1 attriter		Identified as L2 learner	
	n	%	n	%	n	%
Native controls	580	47.7	436	35.8	201	16.5
L1attriters	825	33.9	984	40.4	624	25.7
L2 learners	571	23.5	797	32.7	1066	43.8

Overall, identification accuracy is slightly higher for natives and L2 learners, yet, it did not reach 50% for any of the groups. While relatively few native speakers were identified as L2 learners, roughly the same proportion of L1 attriters, i.e. native speakers from birth, were categorized as L2 learners (M = 25.7%) as there were L2 learners who were categorized as native speakers (M = 23.5%). When looking at the confidence ratings (1 = uncertain, 3 = certain) of the accurate categorizations in Experiment 3c, raters differed in having strong confidence in identifying natives (mean = 2.51, sd = 0.3) and L2 learners (mean = 2.46, sd = 0.25), on the one hand, but less confidence in identifying L1 attriters (mean = 1.95, sd = 0.34), on the other hand. This difference was significant (F(2, 180) = 67,114, p< .001; Tukey HSD

CG versus L1 attriters = .708, CG versus L2 learners < .001, L1 attriters versus L2 learners < .001). Experiment 3c thus bears out that, when raters are told about the nature and relative size of the speakers they judge, they can more easily and more confidently identify native speakers and non-native speakers than L1 attriters.

In sum, whereas variation in scales does not have systematic effects on absolute or relative FARs, differences in instructions yield changes of FARs, in particular of the L2 learners relative to the L1 attriters. In all previous experiments, the speakers in the L1 attriter group received the most variable ratings across raters and across experiments which leads to L1 attriters being perceived as non-native speakers in a large number of cases.

General discussion

The present study found that FARs are remarkably robust against individual variation in raters, differences in phonetic training as well as experience with the target language and the potentially interfering other language (Experiment 1). Although variation in raters can lead to shifts in the absolute assessments of the strength of foreign accents on a given scale, with some raters apparently being more strict in the threshold of who they judge to be native or nativelike, we found that raters come to make similar relative judgements across samples, such that group differences among speakers replicate across differences in raters. These findings stand in seeming contrast to earlier reports by McDermott (1986) that raters focus on different aspects for the same speakers in a multidimensional phenomenon like foreign accent. If they do, the present findings suggest that such individual emphases wash out in a sufficiently large sample of speakers and raters and collapse in an integrative foreign accent scale.

In contrast, Experiment 2 revealed range effects on FARs depending on the range of foreign accentedness present in the samples. Experiment 2a found that excluding the relatively more foreign-accented samples led to a compression of the FARs across groups at the high end of the native-like spectrum, with the L2 group being indistinguishable from natives and L1 attriters. In turn, Experiment 2b demonstrated that testing a homogenous selective set of bilinguals in the more strongly accented range revealed differences between groups that did not surface in a more diverse sample. In other words, range effects can have different consequences and either lead to the levelling of group differences or the emergence of differences in speaker ratings that are not observed when a larger range of samples is presented.

In both experiments, range effects affected the relative rankings of L1 attriters and L2 learners in that the absence of the most strongly accented L2 learners resulted in an increase of FARs of the L2 group *vis à vis* the other groups. These findings complement the results reported in Flege and Fletcher (1992) who showed that the removal or the reduction in size of a native speaker group led to more native-like ratings for L2 groups. Similarly, Experiment 2 illustrated that the removal or reduction of the most strongly-accented L2 samples brought about more native-like ratings for the L2 group in absolute and relative terms. Hence, Long's (1990, 2005) observation about range effects in foreign accent experiments holds true both ways, in that variation at either end of the accentedness spectrum results in shifts of the relative differences of ratings of bilingual speakers. Interestingly, this phenomenon appears to have differentially affected the ratings for L2 learners but not the ratings for the attriters in the present study. It is possible that the effects of crosslinguistic influence in language attrition affects in part different phonetic or phonological features than in advanced L2 acquisition. These features may stand out more prominently in a more local comparison of samples that are perceived as more similar in a larger number of samples. However, it would be necessary to conduct closer phonetic analyses of these samples to identify potential qualitative differences between accentedness in L1 attrition and late L2 acquisition.

Experiment 3 investigated effects of the variation in instructions and scales. Cross-experiment comparisons of Experiments 1 and 3a showed that effects of different scales were minimal as long as the same reference point for accent ratings was preserved, namely the native standard. However, Experiment 3b asked raters to assess whether the speakers had a foreign accent, and it yielded more native-like absolute FARs across all groups and led to a levelling of the relative differences between the L1 attriter and the L2 learner groups. We suggest that these differences between experiments asking for an assessment of nativeness and those asking for degree of foreign accent indicate that raters have a relatively homogeneous implicit standard of nativeness against which they can make proportional judgments of non-native accentedness. However, they differ in their understanding of foreign accentedness. Such variation is likely due to differences in the familiarity with and exposure to foreign accents among the raters (see also McDermott, 1986; Munro, Derwing & Morton, 2006). Against varying implicit standards of foreign accentedness, judgements of less or non-foreign accentedness will not be proportional and reliable across raters. As a consequence, the group differences between the bilingual groups in Experiments 1 and 3a did not replicate in Experiment 3b, since both L1 attriters and L2 learners were rated the same in terms of foreign accent. In this respect, differences in the instructions can result in consequences similar to the range effects observed in Experiment 2. We suggest that both effects likely reflect the same cause, namely, differences in the foreign-accentedness standard. Range effects occur when the native or the foreign accent standards are varied in the samples, and raters will be familiarized to a different range of accents in the course of the experiment. In similar terms, effects of instructions occur when the frame of reference changes from a native to a non-defined foreign accent standard, and raters need to rely on their implicit notions of these terms based on their familiarity with native and foreign accented speech.

It appears that familiarity may also account for the overall pattern of greater variability and lower confidence found for the ratings of the L1 attriter group. Seeing that the raters were German students in their early twenties in Germany, they were unlikely to have had much or consistent contact with long-term German emigrants, i.e. speakers like those in in the L1 attriter group. As a result, the raters may have interpreted the degrees of cross-linguistic influence perceptible in the speech of the L1 attriters differently on a FAR scale (see also Thompson, 1991; Winke, Gass & Myford, 2012 for familiarity effects). As a consequence, the wider range of variation of judgements for the L1 attriter group across experiments may, in part, not only reflect varying degrees of cross-linguistic influence occasioning foreign accent in L1 attrition, yet also varying degrees of familiarity of the raters with attrited speech. However, future research with other rater groups is necessary to disentangle these factors.

The findings of the current study have practical consequences for future research using global FARs or similar assessments of oral production in second language acquisition research and language testing. First, the robustness of FARs across raters suggests that it is likely sufficient to use a relatively small sample of ten to twenty raters. Second, the range effects we find among the bilingual speakers indicate that it is necessary to pay careful attention in the selection of speakers, since the inclusion or exclusion of relatively strongly accented speakers can affect the FARs across the board and, in turn shift, level or reveal relative differences between groups of speakers. Any study that draws conclusions from group differences should consider the degree to which this finding may be affected or even caused by range effects in the selection of samples. Third, our study shows that raters should be familiar with the reference points of the scales as well as the type of speakers, since lower degrees of familiarity will lead to greater and potentially spurious variability in FARs. In this respect, the present results add to and complement the findings about the impact of non-native accents for assessing proficiency and fluency in oral production tests (e.g. Carey, Mannell & Dunn, 2010; Harding, 2012; Major et al., 2003; Winke, Gass & Myford, 2012). Familiarity of the raters with the scales and the (non-native) accents to be judged needs be taken into account. Raters should be trained on the relevant types of scales and accents to reduce rater biases (Harding, 2012; Winke, Gass & Myford, 2012) and thus increase the validity of pronunciation ratings (Taylor & Geranpayeh, 2011). Future studies on global foreign accent ratings or holistic oral production and pronunciation assessments should take these findings into consideration or at least report the relevant points and justify the methodological choices they make.

Finally, the methodological findings presented in this study corroborate the results obtained in Hopp & Schmid (2013): Although bilingual native speakers, i.e. late L1 attriters, and late L2 learners of German differ in their perceived degrees of foreign accent at the group level, there is considerable overlap in accent between late L1 attriters and late L2 learners, with many late L1 attriters being rated as non-native and many late L2 learners judged to be native-like. The present study bears out that this finding is reliable across differences in raters, ranges, scales and instructions. These results then highlight that the phenomenon of nativeness in late L2 speech production ultimately requires an explanation in terms of the characteristics of late bilingual speakers and does not reduce to methodological artefacts or variation in raters.

Acknowledgements

The order of the authors is arbitrary, and both authors contributed equally to this study. Part of the research presented here was supported by NWO grant 275-70-00. We would like to thank the audience at EUROSLA 22 in Stockholm as well as the guest editors of this special issue and two anonymous reviewers for helpful comments.

References

Anderson-Hsieh, J., & Koehler, K. (1988). The effect of foreign accent and speaking rate on native speaker comprehension. *Language Learning*, *38*, 561-613.

Asher, J.J., & García, R. (1969). The optimal age to learn a foreign language. *The Modern Language Journal*, *53*, 334-341.

Bongaerts, T., van Summeren, C., Planken, B., & Schils, E. (1997). Age and ultimate attainment in the pronunciation of a foreign language. *Studies in Second Language Acquisition*, *19*, 447-465.

Carey, M.C., Mannell, R.H. & Dunn, P.K. (2010). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing, 28(2),* 201-219.

Chang, C.B. (2012). Rapid and multifaceted effects of second-language learning on first-language speech production. *Journal of Phonetics, 40,* 249-268.

Davies, A. (2003). *The native speaker: Myth and reality.* Multilingual Matters: Bristol.

Davies, A., Hamp-Lyons, L., & Kemp, C. (2003). Whose norms? International proficiency tests in English. *World Englishes, 22,* 571–584.

Dewaele, J.-M. (1996). Les phénomènes d'hésitation dans l'interlangue française: analyse de la variation interstylistique et interindividuelle. *Rassegna Italiana da Linguistica Applicata, 28*, 87-103.

Elliott, A.R. (1995), Field independence/dependence, hemispheric specialization, and attitude in relation to pronunciation accuracy in Spanish as a foreign language. *The Modern Language Journal, 79*, 351-371.

Field, A. (2005). *Discovering statistics ssing SPSS*. London. Sage.

Flege, J.E. & Fletcher, K.L. (1992). Talker and listener effects on degree of perceived foreign accent. *Journal of the Acoustical Society of America*, *91,* 370-389.

Flege, J.E. (1984). The detection of French accent by American listeners. *The Journal of Acoustical Society of America, 76*, 692-707.

Flege, J.E. (1987). The production of ‘new’ and ‘similar’ phones in a foreign language: evidence for the effect of equivalence classification. *Journal of Phonetics, 15,* 47-65.

Flege, J.E., Birdsong, D., Bialystok, E., Mack, M., Sung, H., & Tsukada, K. (2006). Degree of foreign accent in English sentences produced by Korean children and adults. *Journal of Phonetics, 34,* 153-75.

Flege, J.E., Frieda, E.M. & Nozawa, T. (1997). Amount of native-language (L1) use affects the pronunciation of an L2. *Journal of Phonetics, 25,* 169-86.

Flege, J.E., MacKay, I.R.A., & Piske, T. (2002). Assessing bilingual dominance. *Applied Psycholinguistics, 23,* 567-598.

Flege, J.E., Munro, M.J., & MacKay, I.R.A. (1995). Factors affecting strength of perceived foreign accent in a second language. *Journal of the Acoustical Society of America, 97,* 3125-3134.

Guion, S.G., Flege, J.E., & Loftin, J.D. (2000). The effect of L1 use on pronunciation in Quichua-Spanish bilinguals. *Journal of Phonetics*, *28*, 27-42.Harding, L. (2012). Accent, listening assessment and the potential for a shared-L1 advantage: A DIF perspective. *Language Testing, 29(2),* 163-180.

Harding, L. (2013). Pronunciation assessment. *The Encyclopedia of Applied Linguistics.* Oxford: Wiley-Blackwell.

Hopp, H. (2007). *Ultimate attainment at the interfaces in second language acquisition: Grammar and processing.* Unpublished doctoral dissertation, University of Groningen.

Hopp, H. & Schmid, M.S. (2013). Perceived foreign accent in first language attrition and second language acquisition: The impact of age of acquisition and bilingualism. *Applied Psycholinguistics, 34, 361-394*

Jesney, K. (2004). *The use of global foreign accent rating in studies of L2 acquisition*. MS, Language Research Centre University of Calgary.

Jilka, M. (2000). Testing the contribution of prosody to the perception of foreign accent. *Proceedings of New Sounds (4th International Symposium on the Aquisition of Second Language Speech).* Retrieved from http://ifla.uni-stuttgart.de/index.php?article_id=59, 14.11.2012.

Leeuw, E. de, Schmid, M.S. & Mennen, I. (2010). Perception of foreign accent in native speech. *Bilingualism: Language and Cognition, 13*, 33-40.

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

Long, M.H. (1990). Maturational constraints on language development. *Studies in Second Language Acquisition*, *12*, 251-285.

Long, M.H. (2005). Problems with supposed counter-evidence to the Critical Period Hypothesis. *International Review of Applied Linguistics in Language Teaching, 43*, 287-317.

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.

Magen, H.S. (1998). The perception of foreign-accented speech. *Journal of Phonetics, 26,* 381-400.

Major, R.C. (2007). Identifying a foreign accent in an unfamiliar language. *Studies in Second Language Acquisition, 29,* 539–556.

Major, R.C. (2010). First language attrition in foreign language perception. *International Journal of Bilingualism, 14,* 163-183.

Major, R. C., Fitzmaurice, S. F., Bunta, F., & Balasubramanian, C. (2002). The effects of nonnative accents on listening comprehension: Implications for ESL assessment. *TESOL Quarterly, 36(2),* 173–190

McDermott, W.L.C. (1986). *The scalability of degrees of foreign accent*. Unpublished doctoral dissertation, Cornell University, Ithaca, NY.

Mennen, I., Scobbie, J.M., de Leeuw, E., Schaeffler, S., & Schaeffler, F. (2010). Measuring language-specific phonetic settings. *Second Language Research, 26,* 13-41.

Moyer, A. (1999). Ultimate attainment in L2 phonology: The critical factors of age, motivation, and instruction. *Studies in Second Language Acquisition, 21,* 81-108.

Munro, M.J. (1995). Nonsegmental factors in foreign accent. *Studies in Second Language Acquisition, 17,* 17-34.

Munro, M.J., & Derwing, T.M. (1995). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and Speech, 38,* 289-306.

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

Snow, C.E. & Hoefnagel-Höhle, M. (1977). Age differences in the pronunciation of foreign sounds. *Language and Speech, 20*, 357-365.

Southwood, M.H. & J.E. Flege. (1999). Scaling foreign accent: direct magnitude estimation versus interval scaling. *Clinical Linguistics & Phonetics, 13,* 335-349.

Sprouse, J. (2008). Magnitude estimation and the non-linearity of acceptability judgments. In N. Abner & J. Bishop (eds), *Proceedings of the West Coast Conference on Formal Linguistics (WCCFL).* (pp. 397-403). Somerville. MA: Cascadilla Proceedings Project.

Taylor, L., & Geranpayeh, A. (2011). Assessing listening for academic purposes: Defining and operationalising the test construct. *Journal of English for Academic Purposes, 10(2),* 89–101.

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

Winke, P., S. Gass & C. Myford (2012). Raters‘ L2 background as a potential source of bias in rating oral performance. *Language Testing, 30(2)*, 231-252.

Xi, X. & Mollaun, P. (2011). Using raters from India to score a large-scale speaking test. *Language Learning, 61(4),* 1222-1255.

Yeni-Komshian, G.H., Flege, J.E., & Liu, S. (2000). Pronunciation proficiency in the first and second languages of Korean-English bilinguals. *Bilingualism: Language and Cognition, 3,* 131-149.

Zhang, B. & Elder, C. (2011). Judgements of oral proficiency by non-native and native English speaking teacher ratings: Competing or complementary constructs? *Language Testing, 28(1),* 31-50.