THE EFFECT OF EARLY BILINGUALISM ON PERCEIVED FOREIGN ACCENT

Leona Polyanskaya*, Mikhail Ordin

Fakultät für Linguistik und Literaturwissenschaft, Universität Bielefeld
leona.polyanskaya@uni-bielefeld.de, mikhail.ordin@gmail.com

ABSTRACT

High degree of between-rater variability in pronunciation assessment is often reported in literature. However, human assessments of pronunciation skills of second language (L2) learners are used in standardized language-proficiency tests. Besides, these scores are used as a reference point in evaluating computer-based systems for pronunciation teaching and testing. Therefore it is important to be aware of rater-related factors that might affect the degree of perceived foreign accent in L2 speech. We used Cronbach's alpha and inter-class coefficient to estimate the between-rater agreement of 10 native English speakers who assessed accentedness in L2 utterances. We found that early immersion into bilingual environment might affect the degree of perceived foreign accent. This finding can be explained by interaction of two linguistic systems in early language acquisition, when phoneme prototypes are formed based on language-specific fine phonetic details.

Keywords: accentedness, foreign accent, inter-rater agreement, bilingualism, pronunciation.

1. INTRODUCTION

There is a large body of work dedicated to perception of foreign-accented speech and to the factors affecting the degree of the foreign accent (FA) in L2. Unfortunately, these studies almost exclusively deal with the speaker-related factors. The issue of the listener-related factors is very rarely discussed. However, many studies report substantial variability in between-raters agreement in pronunciation assessment. At the same time, human scores, often from a single rater, are often used to evaluate pronunciation of L2 learners in standardized tests on language proficiency or to evaluate the performance of computer-based pronunciation teaching programmes that are supposed to reproduce the human scores of pronunciation assessment [5]. Therefore, the issue of listener-related factors in perception of the FA should not be ignored. Below we describe measures of inter-rater agreement and provide a literature overview on inter-rater agreement in assessment of L2 pronunciation.

1.1. Measures of inter-rater agreement

Only when agreement between raters is high can we consider the obtained assessment to be reliable, i.e., generalizable and reproducible. That means, similar pronunciation scores can be expected when similar utterances are given for evaluation. Two most frequently used measures of reliability are Interclass correlation coefficient (ICC) and Cronbach's alpha (α). Other measures of reliability are not considered because the researchers of the later reviewed papers did not use them, and for comparability of the results we used those measures in our study that had been used for similar purposes before.

ICC is a ratio of between-rater variance to the total variance, and in case of measuring inter-rater agreement, it is equal to standardized α, if the variances in scores from each rater are equal [4, 10]. This measure does not show the exact agreement between raters, but it indicates the degree of consistency. The values for ICC and alphas vary between 0 and 1.0, with 1 meaning highest degree of consistency (i.e., no variance in scores), and 0 meaning no agreement at all. High values do not mean absolute agreement in score between raters, but rather indicate similar pattern in assessment, i.e., consistency of scores. That is, if one listener gives higher rating to pronunciation of one learner than another, then other listeners will also give higher rating to the first learner than the second learner, although the exact scores may differ. Possibly, two listeners will give different scores to the same learner, if one is more strict than the other. However, when we measure consistency, we are more focussed on whether the learner who receives higher scores compared to other learners from one rater will also receive higher scores from other raters.

1.2. Reported inter-rater agreement in assessment of pronunciation in L2

The reports on inter-rater consistency in assessment of L2 pronunciation reveal a high degree of variability. Morrow [13] and Tayler and Falvey [17] mentioned frequent disagreements between human raters in pronunciation assessment, and also pointed

to frequent cases low within-rater consistency, i.e., to differences in L2 pronunciation scores when the same utterance has been presented for assessment to the same rater several times. The authors say that inconsistencies arise due to lack of precise criteria regarding how to assess pronunciation and even what to assess.

In many studies qualitative claims regarding the degree inter-rater consistency have been substantiated with empirical analysis. Franco et al. [9, 14], for example, asked 10 professional French teachers and certified language testers to evaluate 5089 L2 sentences on a 5-point scale, ranging from "strongly non-native" to "almost native". Ratings given by five out of ten listeners were selected for further analysis. Ratings by five other teachers were not considered due to very high within-rater variability. The teachers were not consistent when they had to evaluate the same sentence several times, 130 sentences were presented twice for assessment to each rater, and these responses were used to calculate the intra-rater consistency. To estimate the inter-rater reliability between the other five teachers, ICC was calculated separately for the assessments given to individual L2 sentences, r = .65, and to L2 learners, r = .8. The average correlation between the scores given by each individual listeners and by the mean scores averaged across all listeners was r=.76 for sentences and r=.87 for speakers. Correlations between pairs of listeners varied between r=.86 and r=.55. Neumeyer et al. [14] say that "this level of correlation is acceptable within the limitations of the experimental design", although admitting high inter-rater variability in L2 pronunciation scoring patterns. The authors also report on the project "Autograder" by Greenberg, Baker and Lowe, in which the averaged pair-wise correlation between 10 raters assessing L2 pronunciation on a 7-point scale was even lower, r = .55.

De Wet et al. [3] report very low inter-rater agreement in pronunciation assessment r=.3 (6 raters, 45 L2 speakers, 5-point scale). They explain this huge disagreement between the raters by a homogeneous level of proficiency of their L2 leaners, saying that lower reliability in pronunciation assessment is expected when the students have a similar level of language mastery.

On the contrary, Cucchiarini et al. [2] reported very high inter-rater agreement (measured as Cronbach's alpha), ranging between .87 and .97 for three groups of raters, each group consisting of three listeners. The listeners were native speakers of the target language, phoneticians and speech and language therapists specializing in pronunciation problems, who had to evaluate 80 speakers on an 11-point scale, each speaker producing 10 sentences.

Koren [11] reports high agreement between two raters (TEFL teachers) who assessed pronunciation on a 5-point scale, α = .94. Magen [12] reported inter-rater reliability r = .727 (10 monolingual American English listeners, 72 sentences, each sentence presented for evaluation four times) and r = .685 (10 monolingual American English listeners, 128 sentences, each sentence presented for evaluation three times).

Piske et al. [15] recruited nine raters from four different provinces in Canada to evaluate English produced by native Italians. They found very high agreement between raters, lowest pairwise correlation r = .88 and interclass rho = .99. Yet in the earlier study Flege and Fletcher [7] the authors carefully suggest that the observations they make regarding nativeness assessment of L2 speech "probably will generalize to other native English speaking listeners". Flege, Munro and MacKay (1995) used ratings given by five listeners from Canadian province of Ontario and 5 other listeners from elsewhere in Canada. All listeners were naïve native speakers of Canadian English. They had to assess 264 L2 productions by moving a slider between two extreme positions marked as "near native" and "strongest accent". The position of the slider between these two extremes was further converted into a value on a 255-point scale. This study reported substantial and influential idiosyncrasies in assessment scores in different raters, but unfortunately, inter-rater reliability was not estimated directly. The authors suggested that those individuals who have encountered many varieties of English are willing to tolerate a higher degree of FA. Calculating inter-rater reliability would be very helpful to measure consistency of the scoring patterns. It can well be the case that those individual who had had more exposure to more accents of English are less strict compared to those raters who had only been familiar with standard varieties. However, this does not mean that the scoring patterns of tolerant and strict raters differ. It may be the case, that all listeners rated pronunciation of some learners higher than others, with significant differences in mean scores. In other words, the raters might agree that speaker A has better pronunciation than speaker B, although the absolute pronunciation scores might have differed depending on the raters' background. To estimate consistency, we need different measures and methods than those used to find significant differences in mean scores given by tolerant and strict raters. Thompson [18] also found that inexperienced native raters perceive higher degree of FA in L2 speech, or willing to tolerate lower degree of FA than experienced raters. This might probably be explained by the fact that

experienced raters have more familiarity with and exposure to different non-native accents than inexperienced raters. However, as the measures of inter-rater agreement are missing in the study, we cannot be certain if the difference is only in scores, or also in scoring patterns. On the contrary, Bongaerts et al. [1] did not find significant difference between experienced and non-experienced raters.

Southwood and Flege [16] used two different methods to evaluate the degree of FA in L2 English spoken by Italians. They asked 10 naïve native English listeners to assess L2 pronunciation on a 7-point scale and using direct magnitude estimation (DMA) technique[1]. Inter-rater consistency was substantially higher when L2 pronunciation was assessed on the scale r=.85) compared to using DMA technique, r=.58). This shows that the method of assessment might also affect the consistency in scoring patterns.

The reviewed results are not always comparable across studies and reveal differences in inter-rater consistency of L2 pronunciation. We placed this issue in the focus of our study.

2. METHOD

We recorded 30 German learners of English who had had no exposure to other languages in childhood and adolescence. The participants grew up in Nord-Rhein Westfalia in or near the city of Bielefeld. This area is considered not to have a marked regional accent. The speakers were equally spread in L2 mastery from lower-intermediate to advanced levels, as judged by three TEFL teachers, native speakers of English.

We used sentence elicitation task with picture prompts to avoid reading mode and yet obtain comparable utterances. 20 pictures, each accompanied with a descriptive sentence, were embedded into PowerPoint slides. We asked the participants to flap through the slides at their own pace and to memorize the sentences (e.g. "The dog is eating a bone"). Afterwards we showed the participants the same pictures and asked them to retrieve the sentence from memory. These utterances were recorded in PCM format at 44 kHz, 16 bit in the sound booth of the studio at Bielefeld University.

We recruited 10 native English speakers to act as expert listeners (table 1). 6 listeners were from Belfast area in Northern Ireland (one being a fluent speaker of German, learnt German late after puberty), 2 speakers were from the US residing in Italy one was a Canadian residing in Slovenia, and 1 southern British speaker residing in Croatia. Residents of Slovenia and Croatia were fluent speakers of the ambient languages, but they acquired these languages late in adulthood. Two raters had grown up in bilingual environment in early childhood, both were recorded in Belfast area. One of these raters grew up in English-Irish Gaelic environment, and the other lived in Indonesia in early childhood. Both claim not to be able to speak or understand any other language but English at the time testing, regardless of having lived in bilingual environment till the age of five.

Table 1: Listeners and their background

Rater ID	**1**	**2**	**3**	**4**	**5**
Native dialect of English	Irish	Irish	Irish	Irish	Irish
Residence	Belfast	Belfast	Belfast	Belfast	Belfast
Mon/bilin	mon	mon	mon	bil (Gaelic)	mon
Fluent in a foreign language	no	no	no	no	no
Rater ID	**6**	**7**	**8**	**9**	**10**
Native dialect of English	Irish	South British	Can	Am	Am
Residence	Belfast	Croatia	Sloven	Italy	Italy
Mon/bilin	Bil (Indon)	mon	mon	mon	mon
Fluent in a foreign language	no	yes	yes	no	no

The raters were instructed to evaluate how native each sentence sounded on the scale from 6 (native or native-like) to 1 (the strongest FA). The sentences were presented in random order. Before doing the test, the listeners had a voice line-up of fifteen sentences to get familiar with the range of accents which will be presented. The scores were used to calculate α as a measure of consistency between scoring patterns of different listeners.

3. RESULTS

α = .424 showed very low level of agreement between the raters. Inter-item correlation matrix revealed lowest correlations between bilingual raters and monolingual raters. All other correlations in the matrix were higher (table 2). Removing ratings of any listener will increase α. However, the best inter-rater agreements will be achieved if the scores of bilingual raters are removed (from .324 to .711 if the scores of rater 4 are removed and to .788 if the scores of rater 6 are removed).

[1] In DMA approach, the listener hears a standard accented sentence (chosen by the experimenter) and is told that this sentence has a value of 100. The listener hears the next sentence and has to give it a numerical value relative to the standard. If a listener perceives the next sentence twice as accented, he has to give it a value of 200. Then the next sentence is presented, which also has to be compared with the standard. To keep the listener calibrated, the standard stimulus is presented after every 10 evaluations.

Table 2: Inter-item correlation matrix with α values when removing the scores of a particular listener. Bilingual listeners are included.

Rater ID	**1**	**2**	**3**	**4**	**5**
Inter-item correlation	.482	.45	.537	.24	.614
α if rater is removed	.519	.509	.488	.711	.464
Rater ID	**6**	**7**	**8**	**9**	**10**
Inter-item correlation	.21	.844	.34	.783	.677
α if rater is removed	.788	.442	.55	.468	.482

Excluding the scores of both bilinguals from the dataset improved the consistency measure to α=.882, which shows high intra-rater agreement [4, 10]. Moreover, removing scores of any other rater does not result in substantial increase of α, i.e., in improvement of inter-rater consistency (table 3). That said, all other listeners had very similar scoring patterns, regardless of their native dialect (Canadian, Irish or American) or fluency in L2 acquired in adulthood. Thus, we can only conclude that the scoring patterns of bilinguals differ from those of monolinguals, which indicates that the early immersion into bilingual environment might affect the degree of perceived FA in L1 of the listener in adulthood.

Table 3: Inter-item correlation matrix with α values when removing the scores of a particular listener. Bilingual listeners are excluded.

Rater ID	**1**	**2**	**3**	**5**
Inter-item correlation	.582	.536	.637	.695
α if rater is removed	.874	.886	.87	.863
Rater ID	**7**	**8**	**9**	**10**
Inter-item correlation	.927	.406	.873	.739
α if rater is removed	.844	.89	.852	.861

4. DISCUSSION

Based on the results, we conclude that early immersion in bilingual environment affects the perception of accentedness in L2 speech. It should be noted that the ability to speak a second language in adulthood in not that important as the immersion into multilingual environment from birth till 5 years of age. Listeners 4 and 6 were not bilingual in a common sense of this word, i.e., they claimed to have completely forgotten the language of one of their parents and the ambient language of their social environment in childhood. However, their patterns of scoring a degree of FA in L2 English differed from those of the listeners who had grown up in monolingual environment. A probabilistic explanation that the individuals with early bilingual exposure are simply exposed to more FAs in their day-to-day lives and therefore perceive accentedness differently does not reflect the complexity of the situation. Among the raters who participated in the assessment experiment, we had Canadian and Southern British English speakers who had resided in Slovenia and Croatia for many years, Americans who had resided in Italy for one to two years prior to the experiment. They had resided in multilingual international communities where English is used as lingua franca, and their familiarity with accented L2 English was substantial. However, their scoring patterns did not differ from those with less familiarity with FAs. A plausible explanation for their different scoring patterns is bilingual input in infancy and in early childhood, when babies have to pay attention to the minute differences within a language (in order to segment the continuous speech into discrete units, to extract the linguistic properties of their ambient language, e.g. word order, stress placement, morphological type, etc.); and minute differences between languages (for the purposes of languages discrimination). These fine phonetic details are the basis for the phoneme prototypes formed during first years of life. Exposure to different and a wider range of phonetic cues probably results in different prototypes, and thus bilinguals apply different criteria to what they think is accented.

Our finding might be relevant in the fields related to assessing pronunciation in L2 speech. If bilinguals assess the degree of FA differently from monolinguals, this evaluation will vary between geographical areas depending on whether the area is characterized by multilingualism. As most societies in the modern world are bilingual or multilingual, the scoring patterns in pronunciation assessment may also vary. It might be more useful to assess the intelligibility of L2 speech than approximation of L2 pronunciation to the target represented by a certain variety of English. It might also be problematic to use human raters as the reference point with which performance of computer-based pronunciation testing modules is supposed to be compared. Many studies in speech technologies are aimed at comparing automatically obtained pronunciation assessments with the scores given by humans [5]. The usefulness of these comparisons and the reliability of software evaluation are based on the assumption that human scores are consistent. We have noticed that consistency can only be claimed – to a certain extent – for monolingual listeners, while individuals with early immersion in bilingual environment might have different internal constructs of FA and thus will vary in the degree of the perceived FA.

7. REFERENCES

[1] Bongaerts, T., van Summeren, C., Planken, B., Schils, E. 1997. Age and ultimate attainmentin the pronunciation of a foreign language. *Studies in second language acquisition* 19, 447-465.

[2] Cucchiarini,C., Strik, H., Boves, L. 2000. Different aspects of expert pronunciation quality ratings and their relation to scores produced by speech recognition algorithms. *Speech Communication* 30, 109-119.

[3] de Wet, F., Van der Walt, C., Niesler, T. 2009. Automatic assessment of oral language proficiency and listening comprehension. *Speech Communication* 51, 864-874.

[4] DeVellis, R. F. 2003. *Scale Development: Theory and Applications*. Thousand Oaks, CA:

[5] Eskenazi, M. 2009. An overview of spoken language technology for education. *Speech Communication* 51, 832-844.

[6] Flege, J. 1988. Factors affecting degree of perceived foreign accent in English sentences. *J. Acoust. Soc. Am*. 84, 70-79.

[7] Flege, J., Fletcher, K. 1992. Talker and listener effects on degree of perceived foreign accent. *J. Acoust. Soc. Am*. 91, 370-389.

[8] Flege, J., Munro, M., MacKay, I. 1995. Factors affecting the strength of perceived foreign accent in a second language. *J. Acoust. Soc. Am*. 97, 3125-3134.

[9] Franco, H., Neumeyer, L., Digalakis, V., Ronen, O. 2000. Combination of machine scores for automatic grading of pronunciation quality. *Speech Communication* 30, 121-130.

[10] Kline, P. 1999. *The handbook of psychological testing*. London: Routledge.

[11] Koren, S. 1995. Foreign language pronunciation testing: A new approach. *System* 23, 387-400.

[12] Magen, H. 1998. The perception of foreign-accented speech. *Journal of Phonetics* 26, 381-400.

[13] Morrow, K. 2004. Insights from the Common European Framework. Oxford: OUP.

[14] Neumeyer, L., Franco, H., Digalakis, V., Weintraub, M. 2000. Automatic scoring of pronunciation quality. *Speech Communication* 30, 83-93.

[15] Piske, T., MacKay, I., Flege, J. 2001. Factors affecting degree of foreign accent in an L2: a review. *Journal of Phonetics* 29, 191-215.

[16] Southwood, H., Flege, J. 1999. Scaling foreign accent: direct magnitude estimation versus interval scaling. *Clinical Linguistic and Phonetics* 13, 335-349.

[17] Taylor, L., Falvey, P. (2007). *IELTS Collected Papers: Research in speaking and writing assessment*. Cambridge: CUP

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