Scientific Research Publishing

Open Journal of Modern Linguistics, 2020, 10, 468-479
https://www.scirp.org/journal/ojml
ISSN Online: 2164-2834
ISSN Print: 2164-2818

Native vs. Nonnative Raters in Second Language Pronunciation Assessment of Guttural Sounds

Mahmoud S. Al Mahmoud

Department of English, Imam University, Riyadh, KSA
Email: msmahmoud@imamu.edu.sa

How to cite this paper: Al Mahmoud, M. S. (2020). Native vs. Nonnative Raters in Second Language Pronunciation Assessment of Guttural Sounds. *Open Journal of Modern Linguistics, 10,* 468-479.
https://doi.org/10.4236/ojml.2020.105028

Received: September 7, 2020
Accepted: September 27, 2020
Published: September 30, 2020

Abstract

This short paper examines inter-rater reliability of native vs. nonnative raters in their assessment of L2 Arabic speech by American learners. It is predicted that ratings provided by native speakers of Arabic would be more consistent and show less variance as opposed to ratings provided by nonnative speakers of Arabic. In a rating experiment, native and nonnative raters evaluated the "nativeness" of American learners' production of Arabic guttural consonants. A Pearson's correlation coefficient shows a significant strong inter-rater reliability in the judgments of native raters, and a poor one, although insignificant, in the judgments provided by the nonnative raters. Findings also indicate that overall native and nonnative rater groups produced comparable ratings, although no strong correlation could be established.

Keywords

Inter-Rater Reliability, L2 Pronunciation, Assessment, Arabic Gutturals, Nonnative Rating

1. Introduction

Over the years, considerable attention has been given to teaching Arabic as a foreign language in many schools and universities worldwide with examinations and interviews as common means of assessing L2 Arabic conversational abilities. However, these oral interviews and exams depend a great deal on the subjectivity of the raters. Even though a clearly defined scoring rubric is often provided to raters to assist in better assessing the test-takers' linguistic abilities with minimum bias, still studies have shown that inconsistencies in raters' judgments can vary substantially and are quite unsettling as they ultimately undermine the usefulness of such subjectively-scored tests.

This short paper endeavors to explore the effect of native vs. nonnative raters as a factor in the assessment of L2 pronunciation in Arabic learners. More specifically, it attempts to examine whether native raters' reliability and accuracy is different than that of nonnative ones when it comes to evaluating pronunciation of Arabic sounds by American learners. The paper is organized as follows. Section 2 reviews the literature on factors affecting raters' reliability, amongst which is the nativeness of the rater. Section 3 details the methodology of the experiment and shows a correlational relationship in the judgments of native and nonnative ratings. The results are discussed in Section 4, and a conclusion of the paper in Section 5.

2. Literature Review

Numerous studies have investigated various factors and variables that affect raters' reliability. Variables such as native vs. nonnative, trained vs. untrained and whether raters have different linguistic, EFL or occupational backgrounds have been considered. Barnwell (1989) states that students are usually assessed by "naïve" native speakers. His study of L2 American learners of Spanish yields that naïve nonnative speakers of Spanish who have not received any kind of training on how to assess (and what to assess precisely) are much harsher than ACTFL-trained raters. The reaction of native English speakers and native Spanish speakers to recorded speech of Puerto Rican learners of English is also examined. Fayer and Krasinski (1987) find out that native speakers of Spanish are less lenient than native English speakers. However, raters in the Fayer and Krasinski's study are neither trained teachers of English as a second language (ESL), nor are they trained in assessment. Shi (2001) examines native and nonnative EFL raters' judgments of Chinese students' English writings. Both groups of raters in her study are given the same scoring criteria and the same essays in order to find out if similar scores are obtained or not. The results of her study conclude that native and nonnative English teachers render similar scores; there were unmentionable differences in the evaluation of the Chinese EFL students' essays. Nonetheless, it is shown that nonnative teachers "attended more positively in their criteria to the content and language, whereas the native Chinese teachers attended more negatively to the organization and length of the essays" (Shi, 2001: p. 1).

Other studies focus on the background of raters as an important facet, and whether training raters yields sustainable reliability. Jacobs, Zinkgraf, Wormuth, Hartfiel, & Hughey (1981) maintain that training, accompanied by the use of a well-defined scoring rubric, contributes to neutralizing the differences in raters' backgrounds, and to "ensure more consistent interpretation and application of the criteria and standards for determining the communicative effectiveness of writers." (Jacobs et al., 1981: p. 43). Shohamy, Gordon & Kraemer (1992) investigate inter-rater reliability of professional EFL teachers and nonprofessional ones, "lay raters". Their study examines raters' scores before and after training. It is concluded that while training has an ostensible effect on raters, since trained raters demonstrated higher inter-rater reliability, teaching background as a factor held little significance. However, the findings of Hadden (1991) contradict Shohamy et al. (1992) as well as Barnwell (1989). Hadden concludes that non-teachers produce higher ratings of students' second language English communicative skills than do ESL teachers (see also Lumley and McNamara, 1995).

To further explore whether training plays an important role in improving consistency in raters' assessment of ESL compositions, Weigle (1994) conducts an experiment on 16 experienced and inexperienced raters, with experience defined as participating in a previous rating process. The raters' assessment of the compositions in pre- and post-training sessions shows that "the training process was effective in two important positive ways: 1) it helped the raters understand and apply the intended rating criteria, and 2) it modified the raters' expectations in terms of the characteristics of the writers and the demands of the writing tasks" (Weigle, 1994: p. 214).

Brown (1995) explores the effect of various linguistic and occupational backgrounds on the assessment of oral language tests for a Japanese Language Test designed for tour guides. Thirty three native and near-native raters are employed with multiple different backgrounds either in teaching Japanese as a foreign language, or in tour guiding (the first being the linguistic and the latter being the occupational experience). Results confirm that "there is little evidence that native speakers are more suitable than nonnative speakers or that raters with teaching background are more suitable than those with an industry background" (Brown, 1995: p. 13). Similarly, in the speaking assessment of four L1 Japanese students, Caban (2003) observes linguistic and educational training factors that might have a direct effect on raters' assessment. She maintains that interviews are always rated by human observers, and as such, it becomes almost impossible to avoid subjectivity. Bias, she emphasizes, can be caused by factors like age, L1, gender and educational background. In her study, 83 raters are asked to rate four nonnative students. The students are interviewed and asked by the raters to perform certain role-plays. The scoring categories for each student included grammar, fluency, pronunciation, content, appropriateness and overall intelligibility. The findings indicate significant differences amongst the raters; it is suggested, however, that these differences are not directly related to the L1 background of the raters nor are they related to their academic training.

Still others such as Charney (1984) have argued that training may have negative effects on raters. "Raters can be trained to agree on ratings" or they could agree to a certain rating based on superficial aspects of the text such as the stylistics, handwriting and organization rather than the content. Similarly, it could be argued that training raters might lead them to become so restricted in their judgments to the provided scoring rubrics, thus ignoring any relevant past experience in the field that might be essential in the rating process. This, nonetheless, can be circumvented if raters were more directly involved in the preparation and construction of the rubrics. Raters could decide, drawing back on their experience in the field, how to develop a suitable rubric that provides better criteria for evaluation.

To sum up, all of these studies taken together imply that the use of appropriate scoring rubrics and proper training of raters will most likely lend more consistency, hence reliability to raters' judgments. While notable differences amongst raters with dissimilar linguistic, educational and occupational backgrounds exist, inconsistencies in assessment are still found among uniform raters who share common backgrounds and are equally trained in assessment and/or teaching, hence, the need to examine the effect of native language on raters' ability to produce reliable judgments of L2 pronunciation in this study.

3. Method

The purpose of this paper is to contribute to the afore-mentioned wealth of literature on raters' reliability. In particular, the native vs. nonnative factor is taken up here. The study explores the inter-rater reliability of both native and nonnative Arabic raters in their assessment of L2 learners of Arabic pronunciation. The following research question is of concern to this study:

1) RQ: Do native Arabic raters yield more consistent judgments than nonnative ones?

To address the research question properly, two main hypotheses will be tested:

2) H1: Inter-rater reliability will be significantly higher in native raters than in nonnative raters.

3) H2: There will be no positive significant correlation between overall native raters' and nonnative raters' judgments.

The particular effects of the native and nonnative raters are chosen mainly for three reasons. The fact that studies, which looked at these factors, show conflicting results necessitates further research in this regard. Second, very few studies, if any, have been conducted on raters' assessment of L2 Arabic students' pronunciation. Third, this study aspires to examine the claim that native raters fair better on reliability than their nonnative peers because of their knowledge and perception of the L1; since various studies have not so conclusively established this claim, it remains a speculation. Note that Hypothesis 2 here follows from H1; if inter-rater reliability is high amongst native raters compared to nonnative ones, then it is expected that judgments of the native and nonnative groups will be different, and not be correlative.

3.1. Participants

Four raters took part in this study. Two of them are native speakers of Arabic, one with a college degree and one with a master's. Both native raters (NR) are well-educated in Modern Standard Arabic (MSA) and only one of them has some teaching experience through private tutoring. The other two raters were native speakers of American English. One of them holds a master's degree in teaching Arabic as a second language and the other has just graduated from college with a degree in Arabic literature. All raters are males and their ages range from 24 - 30. The recruitment of the participants is done through the researcher's personal contacts with the Arabic Dept. at Georgetown University, Washington DC. None of the raters has ever participated in a rating process before and, given their little background in assessment, they can be fairly described as professionally untrained. The raters were invited to take part in this study and upon their consent, a simple interview was held with each one of them to collect some demographic and background information.

3.2. Materials

The material for this study is based on data drawn from another ongoing work on the production of Arabic guttural consonants by American learners of Arabic. The experiment examines the accuracy of ten L2 students' pronunciation of the Arabic guttural consonants.[1] American learners of Arabic in different levels (beginner and intermediate) were digitally recorded reciting a list of ten Arabic minimal pair words in a sound treated-lab using a clip-on PRO 7 Electret condenser microphone and the Audacity (Audacity Team, 2008) recording and editing software (version 1.2.4). The words are organized in minimal pairs and they contained the targeted sounds (gutturals) both word initially and word finally. The minimal pairs exhibit all three vowels commonly found in MSA, namely /i/, /o/ and /a/. The minimal pairs are constructed of nonsense words to eliminate any familiarity effects where test takers rely on previous knowledge of the stimuli in pronouncing them. The recording session lasted less than 20 minutes for all participants, after which the data were collected and presented to the raters to judge.

3.3. Data Analysis

All four raters are asked to take part in the rating process by listening intently to each utterance and providing judgments of the pronunciation stimuli produced by the ten American L2 learners of Arabic. The raters have been instructed that the rating process is limited to the pronunciation of guttural consonants only, thus excluding other aspects of pronunciation such as vowel quality, voicing or even other nonguttural consonants. A scoring rubric as criteria for raters to follow in their assessment task has been provided. The scoring rubric included a scale from 1 - 5 as in Table 1.

Raters listened to each of the 100-recorded utterances (10 words per subject); they were given five seconds per item to rate with no restriction on replays. The rating session lasted fifteen minutes approximately for each rater and their judgments were then recorded and documented to reveal any discrepancies in the assessment of scores, if any.

[1] Guttural sounds are produced at the back of the throat. The guttural consonants in Arabic include: the glottals /ʔ/, /h/, the uvulars /q/, /χ/, /ʁ/ and the pharyngeals /h̲/, /ʕ/. Gutturals are generally considered rare sounds in language, and are often more difficult to pronounce by American learners of Arabic.

Table 1. Scoring rubric used by raters.

Score	*Interpretation*
4 - 5	the subject has pronounced the guttural consonant the way a native speaker would
2 - 3	the subject has identified the correct consonant but failed to pronounce it accurately (nonnative like pronunciation)
1	the subject did not pronounce the right consonant. Instead he or she pronounced another similar consonant (i.e. within the guttural family)
0	the subject has pronounced a totally different consonant that is not even remotely related to the targeted one (i.e. outside of the guttural family)

An analysis for each rater group as well as a comparison of the ratings given by the native and the nonnative speakers in the experiment were conducted. Native raters gave out higher ratings of the students' pronunciations (61.3%) than nonnative raters (55.4%). However, the mean difference between the two raters in each rater group varied significantly. The two native raters had a mean difference of 0.07, while the nonnative raters had a higher mean difference of 0.96. The divergence between the performance of the two rater groups measured to 0.89. Table 2 represents the mean difference between the pair raters in the native group for each subject.

Table 2. Mean difference between native raters.

Subjects / *Raters*	1	2	3	4	5	6	7	8	9	10	*Average Total*
NR1	3.4	1.7	1.7	2.9	3.6	3.2	3.7	3.4	3.6	3.8	**3.1**
NR2	3.5	1.4	2.3	2.7	3.1	2.9	3.5	3.2	3.5	4.2	**3.03**
Mean Dif.	0.1	0.3	0.6	0.2	0.5	0.3	0.2	0.2	0.1	0.4	**0.07**

The inter-rater reliability between the first NR and the second NR can be shown by computing the correlation of scores provided by each rater. This is represented in Figure 1.

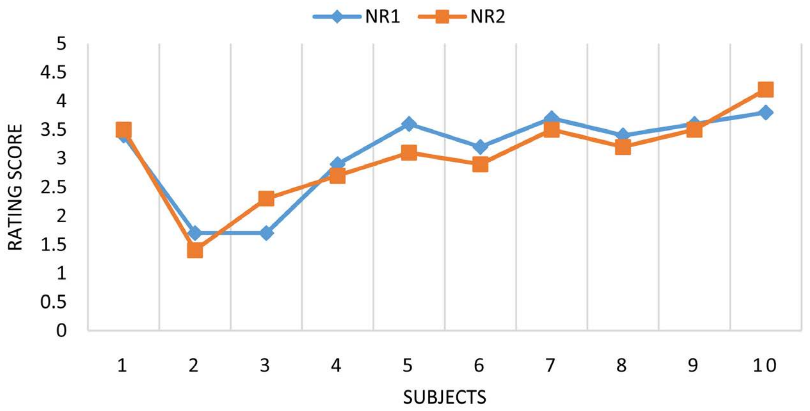

Figure 1. Correlation between judgments of the two native raters.

The scores in Table 2 were submitted to a Pearson's correlation coefficient measure, a parametric test of the strength of the relationship between two variables. Results are shown in Table 3.

Table 3. Coefficient values of the native ratings.

Variables	*Coefficient* (*r*)	*N*	*T-statistic*	*DF*	*Sig.* ($p < 0.05$)
*NR*1/*NR*2	0.904	10	5.99	8	0.0003

The test produced a highly significant measure of covariance, $r = 0.90$, $p < 0.001$. This indicates a high level of correlation and reliability between the two native raters' assessment scores. The two ratings overlap to the extent of r^2 (0.817), which is a strong relationship.

The mean difference between the two raters in the nonnative group is shown in Table 4.

Table 4. Mean difference between nonnative raters.

Subjects *Raters*	1	2	3	4	5	6	7	8	9	10	*Average Total*
*NNR*1	2.2	1.4	2.7	2.2	1.6	2.4	2.1	2.8	2.9	2.6	**2.29**
*NNR*2	4.2	3.3	1.6	2.9	3.0	2.2	3.6	3.5	3.8	4.4	**3.25**
Mean Dif.	2.0	1.9	1.1	0.7	1.4	0.2	1.5	0.7	0.9	1.8	**0.96**

Interestingly, the extent to which the two raters agreed in their judgments becomes clear when the correlation between the NNRs pair is examined as illustrated in Figure 2.

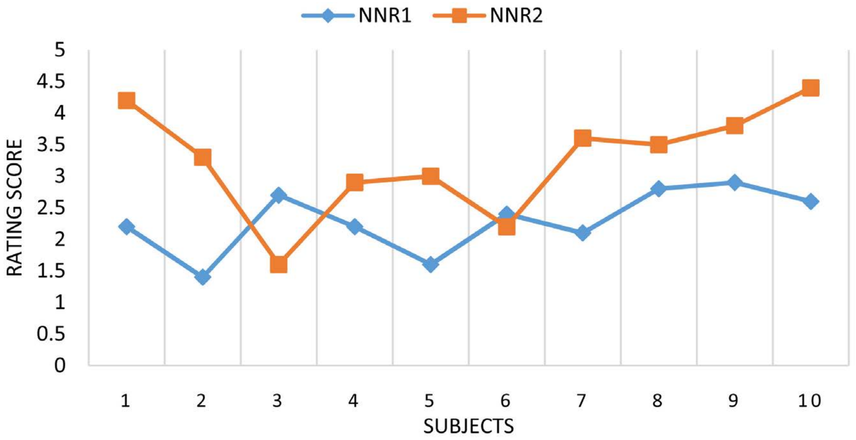

Figure 2. Correlation between judgments of the two nonnative raters.

The Pearson's correlation test shows that the pronunciation ratings provided by the two nonnative subjects have a weak correlation.

The coefficient value in Table 5 is low and indicates a poor correlation between the two ratings of the nonnative subjects, $r = 0.009$, with extremely low overlap ($r^2 = 0.00008$). However, it should be noted that this lack of covariance in the nonnative data is not significant as it failed to reach the significance level, $p > 0.05$.

Table 5. Coefficient values of the nonnative ratings.

Variables	*Coefficient* (*r*)	*N*	*T-statistic*	*DF*	*Sig.* ($p < 0.05$)
NNR1/NNR2	0.009	10	0.025	8	0.98

To calculate the inter-rater reliability between the native and nonnative groups, the scores of the native raters as well as the nonnative raters were collapsed and averaged. Mean differences were computed to show the degree of divergence between the two rater groups. Table 6 summarizes the average scores of each rater group and the mean differences.

Table 6. Mean difference between the native and nonnative rater groups.

Subjects *Raters*	1	2	3	4	5	6	7	8	9	10	*Average Total*
NR Group	3.45	1.55	2	2.8	3.35	3.05	3.6	3.3	3.55	4	**3.065**
NNR Group	3.2	2.35	2.15	2.55	2.3	2.3	2.85	3.15	3.35	3.5	**2.77**
Mean Dif.	0.25	0.8	0.15	0.25	1.05	0.75	0.75	0.15	0.2	0.5	**0.295**

It is clear from Table 6 that the greatest discrepancy between the two rater groups exists in their averaged scores for subject 5 as indicated by the mean difference of 1.05. The smallest difference of 0.15 between the two rater groups is found in their ratings of subjects 3 and 8. The inter-reliability or agreement between the two rater groups is illustrated in the correlational graph in Figure 3.

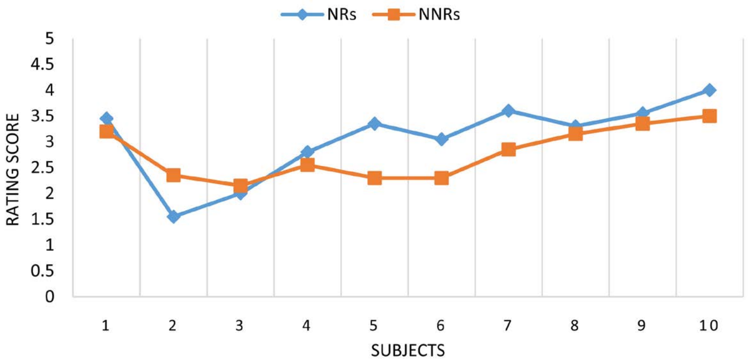

Figure 3. Correlation between judgments of the native and nonnative rater groups.

A two-tailed Pearson's covariance test reveals a moderate level of correlation between the native and nonnative rater groups' scores (Table 7).

Table 7. Coefficient values of the native and nonnative groups.

Variables	*Coefficient* (r)	*N*	*T-statistic*	*DF*	*Sig.* ($p < 0.05$)
NRs/NNRs	0.73	10	3.00	8	0.017

The Pearson's coefficient value here suggests a significant level of correlation between the two rater groups, $r = 0.73$, $p < 0.05$, albeit moderate. A value of 0.60 and above is generally accepted as a reliable measure of correlation in the field of second language research. In other words, the two group's ratings overlapped 53% of the time ($r^2 = 0.5329$).

4. Discussion

Recall that the main research question of this paper asked whether native Arabic raters yield more consistent judgments than nonnative ones. Hypothesis 1 predicted that native inter-raters' judgments of American L2 learners of Arabic would be more reliable than would judgments of nonnative raters. It is expected, therefore, that the rating scores provided by the native raters would be closely matched, or highly similar. This is borne out in the results. The numbers in Table 2 show that the ratings scores of native rater 1 (NR1) averaged 3.1 out of 5.0 (62%), while native rater 2 (NR2) averaged 3.03 out of 5.0 (60.6%); in other words, NR1 yielded slightly higher ratings than NR2. The mean difference between these two averaged assessment ratings is quite minimal 0.07, indicating comparable performance between the two raters. The reliability between the two native raters is significant $r = 0.90$, $p < 0.001$, with a high degree of overlap, r^2 (0.817), i.e. the two ratings overlapped 81.7% of the time. Hypothesis 1 also gives rise to the prediction that nonnative raters' assessment scores of American L2 learners' pronunciation would be variant. Looking at Table 4, nonnative rater 1 (NNR1) averaged 2.29/5.0 (52%), and nonnative rater 2 (NNR2) averaged higher, 3.25/5.0 (65%). The mean difference between the two nonnative raters amounted to 0.96, which is quite large and suggests that their assessment scores lacked covariance, hence reliability. The degree of overlap is a meager $r^2 = 0.00008$, meaning only 0.0008% of similarity. However, it is important to note that this lack of reliability in the nonnative ratings has not reached the level of significance, $r = 0.009$, $p > 0.05$, and can only therefore be regarded as expressive of a tendency.

The NR group achieved much higher level of inter-rater reliability than did the NNR group. In other words, as the figures indicate, consistency in the judgments of the two raters in the NR group is almost 14 times higher than it is in the NNR group. The performance of NNR2 was very close to that of the NRs; however, it appears that because NNR1 gave such low ratings, the reliability within the NNR group suffered much. These results are partially supportive of Hypothesis 1, which posits that native raters yield more inter-rater reliability in their judgments of L2 pronunciation than do nonnative ones.

Hypothesis 2 states that no positive significant correlation between native and nonnative rater groups should exist in their judgments of L2 pronunciation. Thus, it is predicted that the ratings of the two rater groups would be different. A cursory look at the results in **Table 6** reveals that the native group produced an average of 3.06/5.0 (61.3%), and the nonnative rater group produced an average of 2.77/5.0 (55.4%). The mean difference between the two rater groups amounted to 0.295, which is decent and suggests some correlation. However, this level of correlation although significant is moderate, $r = 0.73$, $p < 0.05$, since it amounted to only 53% of the time, $r^2 = 0.5329$. Thus, it is concluded that the results of this experiment disconfirm Hypothesis 2. The native rater group provided higher overall ratings of American L2 learners' pronunciation data than did the nonnative one. The ratings of the two groups in fact significantly showed small but positive correlation. This shows that the native and nonnative subjects preformed similarly on the judgment task, which is contrary to what Hypothesis 2 assumes. It is important to note that the level of similarity between the two rater groups is slightly above chance level (53%); in addition, recall that the performance of NNR2 is exceptionally higher than NNR1, and is almost akin to that of the native raters'. This could have arguably led to the ratings of the NNR group being similar in some degree to the ratings of the NR group. Hence, due to the small number of subjects and the large difference in performance between the nonnative raters, it is best to interpret such correlation as suggestive.

Although speculative, the reason why native raters performed better than their nonnative peers could be attributed to the fact that native raters have acquired a fully developed sense of the language. Given that almost all native speakers transition through very similar developmental stages in the acquisition of their first language, their intuition as well as their acute ability to perceive and categorize the sounds of their language are quite heightened and might have contributed positively to their more unified judgments. It can be argued, on the contrary, that nonnative raters lack this perceptual discriminability of second language sounds and may, therefore, resort to guesstimating in some cases, which definitely leads to arbitrary ratings, as seen presumably in the ratings of NNR1 in this study.

There are subtle differences between the productions of different sounds and sometimes these variations are hard to perceive. It is commonly assumed that performance amongst nonnative speakers of a certain L2 background is highly variable from one person to another. In rare cases do we find some nonnative speakers whose command of a foreign language, especially its phonology, is considered exceptional (cf. Ioup, Boustagui, El Tigi, & Moselle, 1994; Long, 1990; Moyer, 1999; Patkowski, 1994, for studies that report on cases of ultimate attainment in phonology). The variation in the performance of L2 speakers is gradable and could be reflective of the discrepancies in their ratings. On the other hand, almost all native speakers achieve one uniform level of nativeness. It is impossible to say that person A is more native-like than person B when both A and B are native speakers of the same language. True they may differ in their eloquence or oration skills but their ability to perceive and produce their L1 speech sounds should be comparable.

Alternatively, it could be the case that nonnative raters may have developed a high sense of caution towards perceiving and producing L2 sounds as expressed in the harshness of NNR1 who notably produced the lowest rating judgment amongst all native and nonnative raters. It is possible that being over corrective, NNR1 dismissed most of the pronunciation stimuli as being less native-like thus assigning them the lowest of scores. Whatever the reason might be, the results of this study seem to be in contradiction of the findings of Fayer and Krasinski (1987) who reported that native speakers of Spanish were much harsher than nonnative English ones. Native (Spanish) raters produced lower ratings in their evaluation of Puerto Rican L2 speech than did nonnative (English) raters. Nonetheless, the current results appear more in line with those of Brown (1995) who found little evidence that native speakers are more suitable than nonnative speakers in the assessment of oral language tests for the tour guide Japanese Language Test. That is, both native and nonnative rater groups in Brown's study performed quite similarly, and no significant difference, just as in this study, between them was observed. The findings obtained here are also supported by Shi (2001) who concludes that native English teachers and nonnative ones rendered similar rating scores, and that marginal differences between the two rater groups in their evaluation of the Chinese EFL students exist. Kobayashi (1992), however, provides conflicting results of how English native speakers were more accurate than Japanese native speakers in their corrections of ESL compositions written by Japanese students.

5. Conclusion

This short paper sets out to examine the effect of native vs. nonnative as a factor on the assessment of L2 pronunciation. It explores whether the assessment of American L2 learners of Arabic speech by native Arabic raters yields more inter-rater reliability than by nonnative Arabic raters. A rating experiment in which native and nonnative rater groups provided judgments of L2 Arabic students' utterances is carried out. Results show that while native raters exhibit significantly higher inter-rater reliability with a large degree of correlation, nonnative raters' poor reliability and lack of correlation are insignificant. Findings also suggest that overall native and nonnative groups behaved similarly in their judgments of L2 pronunciation task, although no strong correlation is obtained. Although the results of this study reaffirmed former studies, more conclusive evidence is still needed. The small number of raters in this experiment coupled with the nature of the stimuli and the raters' diverse linguistic background may have contributed to inter-reliability of the raters.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

Barnwell, D. (1989). Naïve Native Speakers and Judgments of Oral Proficiency in Spanish. *Language Testing, 6,* 152-163. https://doi.org/10.1177/026553228900600203

Brown, A. (1995). The Effect of Rater Variables in the Development of an Occupation-Specific Language Performance Test. *Language Testing, 12,* 1-15. https://doi.org/10.1177/026553229501200101

Caban, H. (2003). Rater Group Bias in the Speaking Assessment of L1 Japanese ESL Students. *Second Language Studies, 21,* 1-44.

Charney, D. (1984). The Validity of Using Holistic Scoring to Evaluate Writing: A Critical Review. *Research in the Teaching of English, 18,* 65-81.

Fayer, J. M., & Krasinski, E. (1987). Native and Nonnative Judgments of Intelligibility and Irritation. *Language Learning, 37,* 313-326. https://doi.org/10.1111/j.1467-1770.1987.tb00573.x

Hadden, L. (1991). Teacher and Nonteacher Perceptions of Second Language Communication. *Language Learning, 41,* 1-24. https://doi.org/10.1111/j.1467-1770.1991.tb00674.x

Ioup, G., Boustagui, E., El Tigi, M., & Moselle, M. (1994). Re-Examining the Critical Period Hypothesis: A Case Study of Successful Adult SLA in a Naturalistic Environment. *Studies in Second Language Acquisition, 16,* 73-98. https://doi.org/10.1017/S0272263100012596

Jacobs, H. L., Zinkgraf, S.A., Wormuth, D.R., Hartfiel, V F., & Hughey, J. B. (1981). *Testing ESL Composition: A Practical Approach.* London: Newbury Publishers House.

Kobayashi, T. (1992). Native and Nonnative Reactions to ESL Compositions. *TESOL Quarterly, 28,* 81-121. https://doi.org/10.2307/3587370

Long, M. (1990). Maturational Constraints on Language Development. *Studies in Second Language Acquisition, 12,* 251-285. https://doi.org/10.1017/S0272263100009165

Lumley, T., & McNamara, T. F. (1995). Rater Characteristics and Rater Bias: Implications for Training. *Language Testing, 12,* 54-71. https://doi.org/10.1177/026553229501200104

Moyer, A. (1999). Ultimate Attainment in L2 Phonology, the Critical Factors of Age, Motivation, and Instruction. *Studies in Second Language Acquisition, 21,* 81-108. https://doi.org/10.1017/S0272263199001035

Patkowski, M. (1994). The Critical Age Hypothesis and Inter-Language Phonology. In M. Yavas (Ed.), *First and Second Language Phonology* (pp. 209-221). San Diego: Singular Publishing Group.

Shi, L. (2001). Native and Nonnative-Speaking EFL Teachers' Evaluation of Chinese Students' English Writing. *Language Testing, 18,* 303-325. https://doi.org/10.1177/026553220101800303

Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The Effect of Raters' Background and Training on the Reliability of Direct Writing Tests. *Modern Language Journal, 76,* 27-33. https://doi.org/10.1111/j.1540-4781.1992.tb02574.x

Weigle, S. (1994). Effects of Training on Raters of ESL Compositions. *Language Testing, 11,* 172-197. https://doi.org/10.1177/026553229401100206