English Language Teaching; Vol. 9, No. 5; 2016
ISSN 1916-4742 E-ISSN 1916-4750
Published by Canadian Center of Science and Education

Comparison of Native and Non-native English Language Teachers' Evaluation of EFL Learners' Speaking Skills: Conflicting or Identical Rating Behaviour?

Emrah Ekmekçi[1]

[1] Faculty of Education, Ondokuz Mayıs University, Samsun, Turkey

Correspondence: Emrah Ekmekçi, Ondokuz Mayıs University, Faculty of Education, Kurupelit Campus, Samsun, Turkey. Tel: 90-362-312-1919. E-mail: emrah.ekmekci@omu.edu.tr

Received: March 3, 2016 Accepted: April 4, 2016 Online Published: April 6, 2016

doi: 10.5539/elt.v9n5p98 URL: http://dx.doi.org/10.5539/elt.v9n5p98

Abstract

Assessing speaking skills is regarded as a complex and hard process compared with the other language skills. Considering the idiosyncratic characteristics of EFL learners, oral proficiency assessment issue becomes even more important. Keeping this situation in mind, judgements and reliability of raters need to be consistent with each other. This study aims to compare native and non-native English language teachers' evaluation of EFL learners' speaking skills. Based on the oral proficiency scores in the final exam conducted at a state university in Turkey, the study analysed the scores given by native and non-native English language teachers to 80 EFL students attending preparatory classes in the 2014-2015 academic year. 3 native and 3 non-native English language teachers participated in the study. Data were collected through an analytic rating scale and analysed with the help of *independent samples t-test* and *Pearson product-moment correlation test*. Pearson product-moment correlation test (calculated as 0,763) indicated that the raters had high inter-rater reliability coefficients. T-test results revealed that there is no statistically significant difference in the total scores given by both groups of teachers. The study also investigated the different components of speaking skills such as fluency, pronunciation, accuracy, vocabulary, and communication strategies with regard to the existence of significant difference between the scores. The only component which created a statistically significant difference was found to be pronunciation, which was expected prior to the research. In line with the overall findings of the study, it can be concluded that native and non-native English language teachers display almost identical rating behaviour in assessing EFL students' oral proficiency.

Keywords: Rating behaviour, oral proficiency, speaking skill, assessment

1. Introduction

Assessing language skills constitutes a substantial place in the language learning process since learners somehow need to authenticate what they have gained so far. Speaking, being one of the productive skills, had been neglected in terms of assessment until the advent of communication-based methods in which communicative competence is desired for success in general language proficiency. However, assessment of speaking skill has always been a thorny component of language testing as it is not easy to specify the traits of oral proficiency and determine reliable, valid, and practical methods to assess the related traits (Brown, 2001). For this reason, it needs a special attention in order to provide a reliable and valid speaking assessment together with the raters' internal consistency. As it is known, speaking tests are conducted by experienced native English speaking language teachers (henceforth 'NESTs') throughout the world. *Test of English as a Foreign Language Internet-Based Test* (TOEFL IBT), for example, which is one of the most widely respected English-language tests in the world includes speaking section conducted by NESTs having British, North American, New Zealander, and Australian accents. However, English is spoken almost all over the world by non-native speakers as well, and as Crystal (2003) states non-native speakers now outnumber native speakers by a ratio of 3 to 1. This fact brings about the inclusion of non-native English speaking language teachers (henceforth 'NNESTs' as oral proficiency assessors in various institutions in the world.

In Turkey where English is spoken and taught as a foreign language (EFL), NESTs are hired widely in both private and state institutions. This means that NESTs and NNESTs are to work collaboratively and assess

language skills together in some cases. Private institutions advertise NESTs to induce language learners to enrol in their courses. State institutions, mostly universities, employ NESTs in preparatory schools within the framework of Fullbright agreement between Turkish Fullbright Commission and the Turkish Higher Education Council. A study investigating the contributions of NESTs to universities reports that English language teaching assistants contribute to students' language development by providing authentic language and increasing intercultural competence of the students (Ekmekçi, 2015). In addition, they constitute a good source and opportunity for a real communication in the class. As the case stated here makes it clear, it is inevitable for institutions to prepare proficiency exams for learners conducted by both NESTs and NNESTs working cooperatively. This means that both NESTs and NNESTs have to work together and prepare reliable and valid language tests.

As compared with the concrete aspects of assessing skills and micro-skills such as grammar, pronunciation, reading, writing, and listening, both groups of teachers are expected to administer oral proficiency tests in harmony with each other in terms of being fair and objective towards the students. If NESTs expect EFL students to reach native-speaker-like competence and if they evaluate them considering the native speaker norms, it is quite possible that each group of teachers can assess students' oral proficiency from different points of views and this may result in inconsistency between teachers as language assessors. For this reason, it is of utmost importance to administer oral proficiency tests in which internal consistency between the raters is high enough not to aggrieve learners due to misevaluation.

1.1 Assessing Oral Proficiency

Constructing reliable, valid, and practical tests of oral production ability has been one of the challenges of large-scale communicative assessment (Brown, 2001). Preparation and assessment of productive skills take time and effort. As Brown (2001) points out:

"The best test of oral proficiency involves a one-on-one tester/test-taker relationship, "live" performance (as opposed to taped), a careful specification of tasks to be accomplished during the test, and a scoring rubric that is truly descriptive of ability." (p. 395).

The above-mentioned 'one-on-one tester' and 'live performance' nature of assessing oral proficiency constitute the complexity and hardship of speaking tests. For this reason, teachers who intend to assess oral proficiency of EFL learners in particular should be competent enough. However, as Madsen (1983) states, most of the teachers do not even try to evaluate the speaking skill or they are not aware of what criteria to choose to assess oral communication. As ingredients of speaking tests, grammar, vocabulary, pronunciation, fluency, and appropriateness of expression are taken into consideration. Some other factors such as listening comprehension, reasoning ability, correct tone, etc. are to be included in speaking test as well (Madsen, 1983). A lot of disagreement arises from how to assess these components separately. Limiting the range of speaking activities to be assessed is suggested as a possible solution for this problem. Through guided and controlled techniques, elicitation can be provided.

Another crucial issue in assessing oral proficiency is the choice and construct of scales. Scales to be employed in assessing oral proficiency play great roles since which speech functions will be assessed is of great importance. Cohen (1994) suggests some scales in order to rate oral language ability. These scales focus on sociocultural, sociolinguistic, and grammatical ability of language learners. The sociocultural scale assesses speakers' language ability considering their culture, age, sex, occupations, social class, roles, and status in the interaction. The sociolinguistic scale involves assessment on speakers' choice of appropriate linguistic forms to express speech acts such as complaint, request, getting permission, apology, refusal, and etc. Speakers' utterances, words, phrases, and sentences are evaluated on the basis of acceptability as for scale assessing grammatical ability. As it is clear from what has been emphasized so far, testing oral proficiency is a controversial issue in terms of the criteria to choose, employing the appropriate scale and which components to include during assessment process. In addition to criteria and scale-related controversial issues, qualifications and identity of the oral proficiency raters should also be taken into consideration.

1.2 Related Literature

Research on NEST and NNESTs focuses on their contribution to language learning process, students' perceptions, language teaching experience, judgements and behaviours of raters, and differences in teaching styles. A great bulk of research is related to the comparison of NESTs and NNESTs in terms of superiority of one another in teaching and contributing to different language skills (Merino, 1997; Walkinshaw & Oanh, 2014; Moh'd Albakrawi, 2014; Gurkan & Yuksel, 2012). However, the comparative studies concerning oral proficiency assessment of EFL students are quite limited. Huang (2013), for example, explores the effects of accent

familiarity and language teaching experience on raters' judgements of non-native speech. The study includes three groups of raters who have different characteristics in terms of accent familiarity and teaching experience. They provided holistic (assessing overall proficiency) and analytic ratings for vocabulary, grammar, foreign accent, and content for 26 non-native speech samples. The results of the study do not reveal any significant effects of accent familiarity and teaching experience on raters' evaluation of non-native speech. However, raters who have accent familiarity and experience report that they are affected by these characteristics while making rating decisions. In another comparative study, Shi (2001) dwells on differences between native and non-native EFL teachers' criteria while evaluating Chinese students' writing tasks and the scores they gave. In the study, 46 teachers (23 Chinese and 23 English) evaluate 10 expository essays employing a 10-point scale. The researcher codes the qualitative data concerning the raters' explanations as general, content, organization, language and length. The findings of the study reveal that there is no statistically significant difference between the scores given by native and non-native EFL teachers. The statistical analysis of the collected data in the study indicates that native English language teachers tolerate mistakes concerning the content and language whereas non-native teachers attend negatively to the organization and length of the essay. The study also raises some questions as to the validity of holistic scales.

Zhang and Elder's (2014) study explores native and non-native English-speaking teachers judgements of oral proficiency based on the College English Test-Spoken English Test. In the study, two rater groups' scores are evaluated through Many-facet Rasch measurement and content analysis. The results of the study reveal that even if NEST and NNESTs differ in their approaches to rating, overall rating behaviours show similarity. The intra- and inter-group consistency analyses indicate that there exists evidence of similarity rather than difference both within and between groups. Zhang and Elder (2010) conduct another research investigating judgements of oral proficiency by native and non-native raters. The research attempts to uncover the competing or complementary aspects of raters' behaviours. The results reveal no significant difference in raters' holistic judgements of the speech samples and each group of raters agree on the construct components of oral English proficiency to a great extent. Likewise, Kim (2009) conducts a study examining how native and non-native teachers assess students' oral English performance. The participants of the study are 12 Canadian native teachers and 12 Korean non-native teachers. Their rating behaviours are compared with regard to severity, internal consistency, and evaluation criteria. The results of the study indicate that most of the teachers in each group have acceptable levels of internal consistency except for one or two raters in each. Similar severity patterns are also exhibited across different tasks. In contrast to internal consistency and severity issues, the raters displayed some dissimilarities in the evaluation criteria they employed. In addition, native teachers' judgements are found to be more detailed compared to the non-native counterparts.

The above-mentioned studies represent comparative studies conducted on NEST and NNESTs. In consideration of the theoretical information and conducted research presented above, this current study aims at finding answers to the following research questions:

1) Do NESTs and NNESTs have inter-rater reliability in assessing EFL students' oral proficiency?
2) Is there a statistically significant difference between the overall scores assigned by NESTs and NNESTs?
3) Is there a statistically significant difference between the scores given for each component?

2. Method

The study is a quantitative one based on EFL students' final exam oral proficiency scores given by NESTs and NNESTs in a preparatory class at a state university in Turkey in the spring term of 2014-2015 academic year. Data were gathered from six scorers in the oral proficiency exam which was a part of English Proficiency Exam administered in School of Foreign Languages. Related information about the participants, data collection instrument, and data analysis process is presented in detail in the following sub-sections.

2.1 Participants

80 EFL students participated in the study. 45 of whom were prep-class students from Faculty of Education, English Language Teaching Department and 35 of whom from Faculty of Medicine. Both groups of students were attending compulsory preparatory English classes when the study was conducted. They were offered 8 class hours speaking course out of 26 a week by both NESTs and NNESTs. Subsequent to 31 weeks of English language education, students were taken to oral proficiency exam. 3 NESTs and 3 NNESTs evaluated students' oral proficiency in the final exam. Three groups of jury, each group consisting of one NEST and one NNEST, gave scores on the basis of an analytic scale containing five components; fluency, pronunciation, vocabulary, accuracy, and communication strategies. Each group of jury was expected to grade each student one by one and

independent from each other out of 20 points. For each component the raters are expected to grade ranging from 0 to 4. The participants' demographic distribution is presented in the following table:

Table 1. Demographic distribution of NESTs and NNESTs

NEST/NNEST	***Age***	***Gender***	***Course of study at university***
1. NEST	35	female	Secondary Teaching/Arts
2. NEST	26	female	English Literature / Photography
3. NEST	22	female	Languages/Political Science
4. NNEST	22	female	English Language Teaching
5. NNEST	36	female	English Language and Literature
6. NNEST	28	female	English Language Teaching

2.2 Data Collection Instrument

As data collection instrument, a 20 point-analytic scale was employed. Analytic scoring provides students more detailed information about students' performance. For this reason, analytic scales are preferred by most teachers and institutions. It is reported that analytic rubric is attached importance as it is based on explicit and detailed scoring. It provides more diagnostic information about students' abilities (Weigle, 2002). The scale used in this study consisted of five components; fluency, pronunciation, vocabulary, accuracy, and communication strategies. 4 points was allocated for each component.

2.3 Data Analysis

Data were analysed via SPSS package program. Initially, each rater's scores for 80 students were entered in the program separately. *Pearson-Product Moment Correlation* coefficients were statistically calculated via SPSS so as to find out the inter-rater reliability of the scores. After that, *Independent samples t-test* was applied to determine whether there is a statistically significant difference between the scores of two groups of raters, NESTs and NNESTs. T-test was also utilized for each component in the scale separately.

3. Findings and Discussion

Findings with regard to the first research question reveal that NESTs and NNESTs have acceptable inter-rater reliability since total correlation coefficient was calculated as 0,763 which means that there is a positive correlation between the raters. The following table indicates correlation coefficients calculated on the basis of scores given by NESTs and NNESTs for each component of the oral proficiency exam.

Table 2. Pearson product-moment correlation coefficients of NESTs and NNESTs

	NNESTS (Fluency)	NNESTS (Pronunciation)	NNESTS (Vocabulary)	NNESTS (Accuracy)	NNESTS Com. Str.)	NNESTS (Total)
NESTS (Fluency)	0,777					
NESTS (Pronunciation)		0,893				
NESTS (Vocabulary)			0,797			
NESTS (Accuracy)				0,799		
NESTS (Com. Str.)					0,633	
NESTS (Total)						0,763

As it is indicated in Table 2, correlation coefficients of each category were high enough to comment that both groups of raters displayed similar rating behaviours in assessing EFL students' speaking skills. However, it does not mean that there is no significant difference between the scores assigned by the groups. In order to highlight the issue of whether there is a statistically significant different between the scores or not and find answers to the second and third research questions, *independent samples t-test* was administered. Table 3 below shows the comparison of total scores assigned by NESTs and NNESts:

Table 3. Comparison of *total* scores given by NESTs and NNESTs

	N	Mean	SD	t	P
NESTs	80	8,35	1,685	-11,78	,309
NNESTs	80	11,70	1,905		

*p> ,05.

Independent samples t-test results reveal that there is not statistically significant difference between the total scores given by NESTs and NNESTs. Mean scores also indicate that NNESTs' scores are higher than those assigned by NESTs. Mean difference between the groups was found to be -3,35, which proves that both groups of raters statistically displayed identical rating behaviour. In order to scrutinize the issue, oral proficiency components were analysed as well based on the scores given separately. Table 4 indicates that there is not statistically significant difference between the scores with regard to the assessment of EFL students' fluency in English.

Table 4. Comparison of *fluency* scores given by NESTs and NNESTs

	N	Mean	SD	t	P
NESTs	80	1,88	,736	-6,06	,467
NNESTs	80	2,58	,725		

*p> ,05.

As it is obvious in Table 4, mean difference is -,700, which verifies that there is no statistically significant difference in the scores related to the component of fluency between the two groups of teachers. These findings indicate that NESTs and NNESTs have given similar scores in assessing the students' fluency.

Table 5. Comparison of *pronunciation* scores given by NESTs and NNESTs

	N	Mean	SD	t	P
NESTs	80	1,59	,650	-7,34	,030
NNESTs	80	2,45	,825		

*p< ,05.

Analysis of pronunciation scores given by NESTs and NNESTs indicates that there is a statistically significant difference between the groups. Mean of NNESTs was found to be 2,45 which is quite higher than those of NESTs. The mean difference -,863 was the highest of all components. This statistically significant difference in assessing pronunciation scores can be attributed to NESTs' unfamiliarity with the speech of EFL students. In fact, it is not surprising that they might have expected them to produce native-like utterances in terms of articulation of sounds. In this sense, pronunciation as a sub-skill can be put into a different category in the scale while assessing oral proficiency of EFL students.

The other component the analysis of which does not reveal a statistically significant difference is vocabulary section. As it is shown in Table 6 below, NESTs' mean score is 1,35 while NNESTs' is 2,10. The mean difference was found to be -,750. Although statistical analysis did not reveal any significant difference, this component was found to be the second in order in terms of differences between the scores of NESTs and NNESTs. This can be attributed to NESTs' expectations about lexical variety from EFL learners.

Table 6. Comparison of *vocabulary* scores given by NESTs and NNESTs

	N	Mean	SD	t	P
NESTs	80	1,35	,566	-7,26	,851
NNESTs	80	2,10	,722		

*p>,05.

Table 7 presented below indicates mean scores of both groups of teachers. NNESTs' mean score was calculated as 1,96 while NESTs' was 1,43. This mean difference (-, 537) is not big enough considering the other components. The statistical analysis showed that there is not significant difference between the scores assigned by the two groups. Prior to the research, it was actually expected that NNESTs and NESTs' scores could vary and the difference would be obvious since NNESTs usually attach too much importance to accuracy rather than other components. Contrary to our expectations, the mean difference was found the fourth among five components.

Table 7. Comparison of *accuracy* scores given by NESTs and NNESTs

	N	Mean	SD	t	P
NESTs	80	1,43	,652	-4,54	,205
NNESTs	80	1,96	,834		

*p> ,05.

The last component the details of which are presented in Table 8 below is communication strategies. T-test analysis reveals that there is a statistically significant difference between the scores given by NESTs and NNESTs. The mean difference was calculated as -,500, which was the least of all components. This result can suggest that both groups of teachers almost displayed similar rating behaviours as to EFL learners' use of communication strategies.

Table 8. Comparison of *communication strategies* scores given by NESTs and NNESTs

	N	Mean	SD	t	P
NESTs	80	2,11	,827	-3,80	,475
NNESTs	80	2,61	,834		

*p>,05.

Independent samples t-test results of five components indicate that NESTs and NNESTs' evaluation of EFL students' speaking skills is similar and reveals no statistically significant difference except for one component, *pronunciation*. As it is known, EFL students' exposure to target language is usually limited to the borders of the classroom. In other words, they do not have much chance to practise the language outside the classroom. For this reason, it seems to be really hard for them to improve oral proficiency in general and pronunciation in particular. This means that NNESTs can tolerate some mistakes as to the pronunciation since they may have difficulties in enhancing pronunciation skills. It is possible that they have also experienced pronunciation-related difficulties in the course of their language learning process. On the other hand, NESTs are not familiar with unusual or strange pronunciation of words. The reason for the significant difference in the scores they give may be related to this unfamiliarity and this may have been reflected to their rating behaviour in the assessment of speaking skill.

The critical point concerning the research was NESTs' majors and their abilities of assessing language skills. Except for one NEST participating in the research, the other two NESTs were graduate of language-related departments. They also gained experiences in terms of assessing oral proficiency thanks to speaking quizzes and mid-term exams administered within the academic year. They were also in coordination with NNESTs about all the courses they were assigned to teach and scales employed in the exams. For this reason, NESTs' majors are not regarded to have had a great influence on their rating behaviour. Huang's (2013) study verifies this in that the

study found no significant effects of accent familiarity and teaching English as a foreign language (TEFL) experience. In addition, Huang (2013) states that:

"… raters with accent familiarity or TESL/TEFL experience self-reported that those experiences affected their rating decisions. Many raters believed that those experiences enhanced their listening comprehension and error detection. In particular, virtually all ESL/EFL teachers reported to be influenced by their teaching experience, and approximately one third of them felt that they were resultantly more lenient in assessing nonnative speech." (p.783).

Taken together, NESTs and NNESTs' overall assessment scores for EFL students are similar to a great extent. There appeared some minor differences between the scores of both groups of teachers, but that did not make a significant effect on the different dimensions of oral proficiency exam.

4. Conclusion

The current study compares NESTs and NNESTs' assessment of EFL students' oral proficiency. To this end, we collected data based on the final exam conducted in the School of Foreign Languages at a state university in Turkey. 80 EFL students were evaluated on the basis of an analytic scale and scores given by both groups of teachers were analysed and compared via independent samples t-test. T-test results revealed that there was not a statistically significant difference in the total scores assigned by NESTs and NNESTs. Five components of the analytical scales were also analysed separately. The scores given for only one component, pronunciation, were found to have statistically significant difference.

The results indicated that NESTs and NNESTs displayed similar rating behaviours substantially. The only contradicting behaviour was in the assessment of pronunciation component of the scale. Considering the vast majority of NNESTs all over the world, the study constitutes an important example for the reliability of NNESTs' rating behaviour compared to that of NESTs'. This means that both groups of teachers have their own characteristics with regard to assessment of language skills and they can complement each other in many assessment-related issues. In fact, our study cannot be said to represent the whole native and non-native speaking teachers' population, but our findings suggest some hints regarding the cooperation of both groups of teachers in assessing language skills.

Our research also revealed similar findings with the previously studies conducted by Huang (2013), Shi (2001), and Zhang and Elder (2010; 2014). Each of the above-mentioned studies has limitations peculiar to themselves, but the common point of all is complementary characteristics and similar assessment tendencies of both groups of teachers.

Our study is limited to restricted numbers of NESTs and NNESTs. Further studies can be conducted with larger groups of teachers. Moreover, the current study particularly examines oral proficiency scores of EFL teachers. Other language skills with different variables such as rater training, sex, age, majors, and etc. can be compared in prospective research.

References

Brown, H. D. (2001). *Teaching by Principles: An Interactive Approach to Language Pedagogy*. White Plains, NY: Longman.

Cohen, A. D. (1994). *Assessing Language Ability in the Classroom* (2nd ed.). Boston: Heinle & Heinle.

Crystal, D. (2003). *English as a Global Language* (2nd ed.). London: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511486999

Ekmekçi, E. (2015). Contributions of English Language Teaching Assistants to Universities in Turkey: A Case Study. *The Online Journal of Quality in Higher Education, 2*(3), 31-38.

Gurkan, S., & Yuksel, D. (2012). Evaluating the contributions of native and non-native teachers to an English Language Teaching program. *Procedia-Social and Behavioral Sciences, 46*, 2951-2958. http://dx.doi.org/10.1016/j.sbspro.2012.05.596

Huang, B. H. (2013). The effects of accent familiarity and language teaching experience on raters' judgments of non-native speech. *Sysyem, 41*, 710-785. http://dx.doi.org/10.1016/j.system.2013.07.009

Kim, Y. H. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing, 26*(2), 187-217. http://dx.doi.org/10.1177/0265532208101010

Madsen, H. S. (1983). *Techniques in Testing*. Oxford: OUP.

Merino, I. G. (1997). Native English-speaking teachers versus non-native English-speaking teachers. *Revista alicantina de estudios ingleses, 10*, 67-79.

Moh'd Albakrawi, H. T. (2014). Is there a difference between native and non-native English teachers in teaching English? *Journal of Scientific Research and Studies, 16*, 87-94.

Shi, L. (2001). Native-and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing. *Language Testing, 18*(3), 303-325. http://dx.doi.org/10.1177/026553220101800303

Walkinshaw, I., & Oanh, D. H. (2014). Native and Non-Native English Language Teachers: Student Perceptions in Vietnam and Japan. *SAGE Open, 4*(2), 1-9. http://dx.doi.org/10.1177/2158244014534451

Weigle, S. C. (2002). *Assessing writing.* Cambridge: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511732997

Zhang, Y., & Elder, C. (2014). Investigating native and non-native English-speaking teacher raters' judgements of oral proficiency in the College English Test-Spoken English Test (CET-SET). *Assessment in Education: Principles, Policy & Practice, 21*(3), 306-325. http://dx.doi.org/10.1080/0969594X.2013.845547

Zhang, Y., & Elder, C. (2010). Judgments of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary constructs? *Language Testing, 28*(1), 31-50. http://dx.doi.org10.1177/0265532209360671