Iranian Journal of Applied Language Studies
Vol 11, No 2, 2019, pp.30-56
http://ijals.usb.ac.ir

Non-native English Speaking Teachers' Pragmatic Criteria in the Holistic and Analytic Rating of the Agreement Speech Act Productions of Iranian EFL Learners

Minoo Alemi, Corresponding author, Associate Professor, West Tehran Branch, Islamic Azad University, Tehran, Iran, Email: minooalemi2000@yahoo.com

Mohammad Motamedi, Instructor, Sharif University of Technology, Tehran, Iran, Email: mohammed.motamedi@gmail.com

Abstract

Pragmatic rating is considered one of the novel and crucial aspects of second language education which has not been maneuvered upon in the literature. To address this gap, the current study aimed to inspect the matches and mismatches, to explore rating variations, and to assess the rater consistency between the holistic and analytic rating methods of the speech act of agreement in L2 by non-native English teachers. In this regard, 12 discourse completion tests (DCTs) for agreement accompanied by EFL learners' responses to each situation were rated by 50 non-native English teachers. Initially, they were asked to rate it holistically, and the content analysis of raters' comments revealed twelve agreement criteria. Grammatical structure was the prominent criterion which suggested that the raters were predominantly concerned with pragmalinguistics. In particular, the results of descriptive statistics demonstrated that there was a degree of divergence in the frequency of the criteria applied. Additionally, the teachers were asked to rate the pragmatic outputs analytically which showed that the raters were more consistent in the analytic phase. Finally, the findings indicated that there was a convergence between the two rating methods. The results of the present study implicated the necessity of rater training with regard to the rubric-based pragmatic rating. This study offers crucial pedagogical implications for syllabus designers, materials developers, language testers, and some suggestions for further research studies.

Keywords: Holistic and Analytic Rating, Pragmatic Rating Criteria, Speech Act, Agreement, Non-Native English Speaking Raters

Received: September 2018; Accepted: February 2019

1. Introduction

As one of the major components of language competence, pragmatics is deemed to be the study of language from the viewpoints of users, specifically, the alternatives they make, the limitations they face within using language in social interactions, and also the impacts their language use has on the interlocutors in an interaction (Crystal, 1997). As a matter of fact, understanding the appropriate use of language within different contexts and the attainment of the interactional skills has turned into the locus of attention for second language (L2) teachers and researchers (Barraja-Rohan, 2011). Basically, it has been remarked that pragmatics is teachable by some scholars (Bardovi-Harlig, 2001; Ishihara & Cohen, 2010; Rose & Kasper, 2001) and it has also been regarded a critical aspect of teaching (Eslami-Rasekh, 2005; Rose & Kasper, 2001), nevertheless the literature has been rather silent with respect to pragmatic assessment. It is worth mentioning that a few studies have bridged this gap (e.g., Alemi & Khanlarzadeh, 2016; Taguchi, 2011; Tajeddin & Alemi, 2014). To put it differently, it is an indispensable priority to pinpoint whether non-native raters apply the same standards and criteria as native raters do for rating speech acts produced by EFL learners. Indeed, the complex phase of rating pragmatic competence is to be elucidated by means of establishing a link between the raters, DCTs, and criteria for rating. Further, identifying the intercultural proficiency of language teachers is a variable which may affect the assessment and rating of pragmatic competence. Clearly, EFL learners' pragmatic knowledge is to be examined as their linguistic knowledge is to be. Therefore, both native and non-native raters should be provided with a yardstick and rigorous framework due to the fact that many factors such as background knowledge, gender, intercultural proficiency, and training may affect raters' judgment. As a result, the present study could illuminate and cover deficiencies

such as the absence of pragmatic strategies, criteria for raters, and rater variations by investigating the criteria which non-native raters adopt in holistic and analytic pragmatic rating methods regarding the agreement speech act.

2. Literature Review

2.1. ILP Rating

Pragmatics, as a branch of linguistics, is the study of language use in the context which is part of our knowledge of how to use language appropriately. Basically, pragmatic competence like other areas of linguistic competence is typically implicit and cannot be examined explicitly (Birner, 2012). One of the significant, though understudied, aspects of pragmatics is pragmatic assessment, which is still in its infancy (Rose & Kasper, 2001). This might be attributed to the lack of valid methods for testing interlanguage pragmatic knowledge or even to the fact that developing a measure of pragmatic competence in an EFL context is rather a complex task (Jianda, 2006).

Jianda (2006) assessed the pragmatic competence of both Chinese EFL learners, as well as English native speakers whose language proficiency was not the same. Overall, the results indicated that pragmatic competence is not directly related to language proficiency. Cohen (2008) also noted that there are more questions regarding testing pragmatics for instructional objectives than there are answers. He even indicates that pragmatic teaching is so important that failure to do so, can result in pragmatic failure and even cause misunderstanding in different situations.

Broadly speaking, the significance of rater-related issues has been deemed vital within assessment and rating, as it can challenge the validity of the test (Bachman, 2004). Additionally, McNamara (1996) stresses the point that the reliability of ratings is one of the most fundamental issues within the

performance-based assessment. Taguchi (2011) challenged the misconception formed about native rater consistency as far as raters are about to evaluate L2 pragmatic competence. In brief, he remarked that native speakers do not form a uniform category as some of the raters are predominantly focused on linguistic forms such as the use of politeness markers or level of directness. On the other hand, some are mostly concerned with non-linguistic aspects, for instance, the use of positive-negative politeness strategies as well as semantic moves or content of speech. Speaking of politeness, it is worth mentioning that three kinds of politeness system are integrated into different contexts, which are called " deference politeness system", "the solidarity politeness system", and "the hierarchical system" which are best known as "distance", "imposition" and " power" (Brown & Levinson, 1981; Scollon & Scollon, 2001).

Tajeddin and Alemi (2014) investigated the criteria of native teachers in the assessment of second language pragmatics in the case of apology and found that the expression of apology, situation explanation, repair offer, promise for the future, and politeness were among the primary features that native raters took into account while assessing L2 pragmatics. It was also implied that in spite of considering politeness, raters' biases toward assessing pragmatics have different degrees which are in line with the result of Taguchi's study (2011).

To further stress the criteria that raters take into account, Alemi, Eslami-Rasekh, and Rezanejad (2014), also explored rater criteria and consistency among Iranian EFL teachers' criteria concerning an EFL learner's production of compliments elicited through WDCTs. In fact, this study was conducted about the teacher's gender and teaching experience. Interestingly, seven major criteria were found for the teachers' ratings, namely "politeness", "interlocutors' characteristics and relationship", "variety and range", "socio-pragmatic appropriateness", "sincerity", "complexity", and "linguistic appropriateness".

In addition, Alemi and Khanlarzadeh (2016) conducted a study so as to explore the pragmatic rating criteria which non-native Iranian raters applied regarding the request speech act. By and large, nine criteria were noted by the raters. Additionally, in their study, raters' gender and prior teaching experience did not play a pivotal role in their pragmatic assessment. Also, Alemi and Khanlarzadeh (2017) did another study on the speech act of request, with the addition of both native and non-native raters. Overall, the results proved that there were major differences between the native and non-native teachers' rating patterns.

Last but not least, Alemi and Motamedi (2019) explored a different study concerning the speech act of disagreement. Having employed both holistic and analytic rating methods, they assessed the rating variations and consistency among non-native English teachers. The results of their study indicated that the raters were more consistent in the analytic rating. Further, it was concluded that non-native raters especially the in-service ones need assessment rubrics as a yardstick to help them provide a fair score to the pragmatic production of EFL learners.

2.2. Holistic and Analytic Scales

Rating scales are regarded as major instruments for assessing the performance of test-takers which yield valuable scoring results for test takers and test developers with respect to scoring which is both valid and reliable. More importantly, they are used to eliminate the impact of biases of the raters (Carr, 2000). According to Hunter, Jones, and Randhawa (1996), holistic scoring gained popularity when Godshalk, Seinford, and Coffman's (1966) study of College Entrance Board assessments revealed a high reliability of scores. In line with this argument, the holistic method of assessing speaking performance is

considered as a general and rapid impression. On the other hand, Weigle (2002) indicated that the analytic method of scoring is the counterpart of the holistic method which incorporates various features of a performance as elements for rating aims. Basically, the analytic scale considers numerous aspects of test-taker performance more accurately. As a result, it seems to be a more reliable and objective scoring method than the holistic scoring. Besides, in the analytic scale, raters approach the task of rating more consistently when they have an array of analytical criteria at their disposal (Hamp-Lyons, 1991). As a result, analytic scoring deals with various domains of a construct. Both holistic and analytic methods of scoring are basically utilized to evaluate a test taker's proficiency (Shohamy, 1995). Generally, in a holistic approach, teachers rate like the Likert scale and make comments to explain why they gave that score. However, in the analytic approach, a table of criteria is provided for raters which they choose and rate. As Carr (2000) put it, the holistic scoring is conducted in a unitary way following the overall assessment. However, in the analytic scoring, raters take into account several sub-scales to evaluate various traits of a performance. For example, some sub-scales can be taken into account to analytically assess a student's language performance such as choice and use of pragmatic tone, strategies for a speech act, choice and use of discourse markers, level of formality. On the other hand, in the holistic approach, a learner's performance is evaluated holistically and thus it avoids distinguishing between linguistic and cultural issues (Ishihara & Cohen, 2010).

A large number of studies have been conducted in order to address the differences between these two rating scales by many scholars (e.g., Bachman, Lynch, & Mason, 1995; Douglas & Smith, 1997). Furthermore, it is noteworthy that the holistic method has been favored over the analytic method due to its easier and applicable rating scale. It is believed that simplicity of this scoring

process is less time-consuming than the complicated ones. In line with this argument, the convenient and quick analysis of holistic methods encourage raters to apply it more frequently. However, teachers might fail to consider the details of learners' errors since the method is a one-shot rating scale (Carr, 2000). On the other hand, the analytic method of rating provides diagnostic information regarding the test takers' ability (Xi, 2007). As a result, the comprehensive and rich data of this method help raters to become cognizant of the learners' oral incompetence (Fulcher, 2003). To put it differently, teachers' understanding of the students' area of weaknesses help them to manage the course in a way to improve their performance. Thus, feedback plays a crucial role in terms of students' areas of deficiency

2.3. Agreement Studies

The speech act of agreement has been viewed as one of the most frequent and digestible acts within a language and more specifically within pragmatics. Despite the favor devoted to the speech of disagreement, the agreement has not been touched upon much in the related literature. As Pomerantz (1984) holds, the agreement comes to the fore when two or more users regard the intended referent as if they look upon it similarly. Generally speaking, agreeing and disagreeing are not merely linguistic actions but also social actions, since they are intimately related to the relationship between two speakers. Therefore, the agreement reflects an alliance between interlocutors (Johnson, 2006).

Relying on Pomerantz's (1984) stance on the agreement, Schegloff (1996) remarks that the practice of agreeing with interlocutors via repeating what they have uttered proved to constitute the action of confirming an allusion which is in a way acknowledging both its content as well as its prior inexplicit conveyance. In other words, even though agreeing with another's explanation of an inexplicit

message might not always propose a prior proclivity to conveying it, agreeing employing repeating might be a practice that takes care of that. In fact, repetition is conducted to a point of prior orientation to convey, although it might have an uncertain relation to the proclivity that is addressed in the prior talk.

To investigate the productive level of the learners' use of the speech act of agreement, Al-Khanaifsawy (2016) administered a questionnaire to twenty Iraqi EFL university learners and found that the instructors' teaching methods do not equip the learners with pragmatic competence with respect to the speech of agreement. However, this study was limited to the context of Iraq where its generalizability is somehow under question.

In spite of the fact that a number of studies have been conducted with regard to rater issues, few studies have managed to explore the interface between raters' criteria and interlanguage pragmatic assessment. In addition, since most interlanguage pragmatic studies have been primarily concerned with the use of production tasks (Bardovi-Harlig, 2001), it accentuates the fact that more research needs to be carried out with regard to the judgment and perception of speech acts (Olshtain & Blum-Kulka, 1985). The review of related studies on ILP rating and agreement discloses the significance of these variables in any educational setting. The crux of the matter lies within the fact that almost no study has managed to consider the rater's criteria as well as their variations guided by the two rating methods of holistic and analytic, regarding the speech act of agreement. In fact, equipping raters with reliable criteria helps them to have a more consistent and productive evaluation of the learners' interlanguage pragmatic knowledge. Therefore, the following research questions were posed so as to fulfill the gap in the literature:

1. What are the criteria used by non-native English speaking raters for rating the speech act of agreement produced by EFL learners?

2. What are the variations in ratings of non-native English speaking raters in relation to the speech act of agreement produced by EFL learners?
3. Is there any significant relationship between the analytic and holistic ratings of non-native English speaking raters in relation to the speech act of agreement produced by EFL learners?

3. Method

3.1. Participants

The participants of this study consisted of 50 non-native English speaking teachers and 12 Iranian upper-intermediate EFL learners (both males and females ranging between 20 to 30 years old) studying at a private language institute in Tehran. The students had already passed the elementary and intermediate levels. The raters were EFL teachers (19 males and 31 females) chosen based on convenience sampling. The teachers were categorized into three levels of 1-5 years, 6-11 years, and more than 11 years of teaching experience so as to have a rich variety. Additionally, 22 of the participants held Ph.D. and the rest were MA students of Applied Linguistics studying at different universities. The underlying assumption for choosing teachers majoring in Applied Linguistics was that pragmatics or discourse analysis is one of the key and major courses integrated within the curriculum; however, some of the teachers are not entirely acquainted with pragmatics.

3.2. Instruments

To elicit EFL learners' perceptions of the agreement speech act, a WDCT was developed. To be exact, items 1 and 2 were adopted from Tajeddin and Alemi (2015), and the rest of the items were developed by the researchers. Thereafter,

the WDCT was judged by two language experts so as to add to the validity of the DCTs. The DCTs were predominantly made up of six situations with different degrees of power, distance, and imposition (Brown & Levinson, 1987). In particular, two types of WDCTs were designed for this research, the holistic one in which the raters were supposed to read the learners' answers in each situation and rate its appropriateness according to rating scale (1. Poor, 2. Fair, 3. Proficient, and 4. Native-like).

Moreover, the raters were asked to write their applied rating criteria for assessing the pragmatic performance of EFL learners. In the second (analytic) phase, the raters were asked to rate the appropriateness of the same responses in situations via a different method. Thus a table with assessment rubrics adopted from Ishihara and Cohen (2010) was provided for the raters.

3.3. Data Collection Procedure

The present study drew on the principles of mixed methods design. In particular, the design employed for the current study was sequential exploratory. First, both hard and soft versions of the DCTs were developed and distributed among the EFL learners. Then, the soft version of answered WDCTs was created with the help of Microsoft Word and sent to non-native raters via email and other social networks (e.g., LinkedIn, Telegram). Initially, the non-native raters were supposed to rate pragmatic outputs holistically. Then, they were asked to rate them analytically based on the established scale and the assessment rubrics adopted from Ishihara and Cohen (2010).

3.4. Data Analysis

The qualitative part of this study was carried out through thematic content analysis so as to find the criteria that teachers deemed appropriate. In particular, throughout this process, the comments given by the teachers were analyzed and the dominant patterns were identified. Furthermore, the quantitative part drew on the descriptive analysis (e.g., frequency of criteria), inferential analysis including inter-rater reliability of raters' rating.

4. Results

4.1. Rating Criteria

To answer the first research question, several criteria were elicited utilizing content analysis. The dominant criteria are illustrated here.

(1) *Grammatical Structure*: The first and foremost criterion is intimately tied with the raters' perceptions regarding issues such as grammatical points, syntactic elements, as well structure of the sentences. Strikingly, the raters dedicated the largest portion of their attention to this parameter. The following example indicates this criterion.

> *The response does not have any grammatical mistakes, and the words which have been used are at the level of high proficiency, the implemented structure and wording accord to the context.*

(2) *Formality*: Obviously, the stylistic variations that we adopt in different contexts vary according to the kind of people we are talking with. For instance, the way we make a request from our professor highly varies from the way we do so from our friend. In the same vein, various styles are adopted when it comes to

agreeing with someone. The instance below stated by a rater is worth paying attention to:

> *As a rule of thumb, we human beings tend to be more relaxed and intimate with our family members.*

(3) *Justification and Reasoning*: The underlying assumption behind this criterion is that the responses produced by the learners should be comprehensive enough in terms of content and justification. Moreover, sufficient information and clues to prove or justify the agreement are crucially important as it equips the text more understandable. The following example displays one of the rater's comments regarding this criterion:

> *Too general response and not well-reasoned*

(4) *Vocabulary*: This criterion refers to the lexis. Basically, the scope and breadth of lexis and the kind of wording a learner uses while expressing agreement are regarded as a prevailing point. The following comment by one of the raters justifies the aforementioned points.

> *Using a wide range of vocabulary convinced me to rate this utterance as proficient.*

(5) *Appropriate Use of the Speech Act*: Through this criterion, raters asserted that some of the responses made by the learners do not indicate any relevance. In other words, an agreement was not truly expressed by the learners which might be due to the fact they have not yet been acquainted with the proper way to agree with someone. One of the examples of this criterion is given below:

> *I would rate it as poor since it has no air of agreement! To me, it is an ironic way of telling some "you might like it, but not me, because I do not like rainy weather."*

(6) *Vagueness or Clarity*: This criterion refers to the fact that the responses which are made by the speakers need to be clear and congruent with what has been asked. In other words, to provide a communicatively rich discourse, the meaning, and intention of the speaker is to be transferred smoothly and without confusion. The following comment by a rater sheds light on this issue.

Lacks sufficient information and does not provide enough pragmatic clues for better understanding.

(7) *Appropriate Length of Production*: Generally, the responses are supposed to be appropriate when we consider its length. As a result, responses that are unnecessarily long or short are likely to convey an unclear message. The mentioned comment by one the raters clearly addresses this point:

Laconic! It is a fairly short answer but very precise.

(8) *Social Distance*: This criterion refers to the degree of social distance that exists between the interlocutors. There are several factors that influence the realization of social distance with the hearer, namely the power of interlocutor and the relationship that exists between them. An Example of this criterion is presented below:

It shows that the student knows the power distance and the context of language use.

(9) *Politeness:* This criterion is seen as one of the predominant and widely acknowledged factors in our discourse. In this regard, one may easily be attached the label 'rude' as soon as the face of the interlocutor is threatened. The following is an example of this criterion remarked by one of the raters.

The response is not rude or impolite based on their relationship.

(10) *L1 Transfer:* This criterion refers to the cross-linguistic influence of the learners' first language while expressing agreements. The following example refers to this point.

> *First language thinking fashion comes to be interfering the second language performance.*

(11) *Social Status*: When rating agreement responses of the learners, the raters deemed the social status, age, and gender of the interlocutors as highly important factors to be aware of. It is noteworthy that this criterion was rarely noted by the raters. The instance provided below clarifies the point.

> *Fair because of the respect he has for the teacher but the student could give a better response.*

(12) *Organization*: This criterion received the lowest degree of attention compared to other criteria. For a response to seem appropriate, it is supposed to have an organized body with a beginning (introduction), body, and a conclusion. Thus, the topic is developed well employing an organized agreement. The comment below addresses this issue.

> *Pragmatically, the answer is well organized. First, the learner has shown his/her agreement with his/her instructor's comments. Then, he /she has provided the instructor with his/her justification. Finally, he/she stated a concluding remark to mark the end of his/her speech.*

Having extracted the mentioned criteria, the quantitative analysis of the data was conducted to measure the frequency as well as the percentage of the criteria in each situation. Table 1 illustrates the non-native raters' frequency of each of the six situations.

Table 1. *Frequency of Agreement Holistic Criteria in Different Situations among Non-native Raters*

Situation	GS	FOR	JR	VOC	AUS	VC	ALP	SD	POL	L1T	SS	ORG	Total
1	30	6	9	11	5	6	2	3	4	2	1	2	81
2	32	8	9	4	8	11	1	4	3	1	0	1	82
3	22	10	13	5	5	3	11	5	3	0	0	0	77
4	12	12	6	12	5	3	6	4	6	0	3	0	69
5	6	8	6	6	11	6	6	4	4	0	0	0	57
6	28	9	6	7	6	6	3	6	4	2	0	0	77
Total	130	53	49	45	40	35	29	26	24	5	4	3	443
Percentage	29.34	11.96	11.06	10.15	9.02	7.90	6.54	5.86	5.41	1.12	0.90	0.67	100

Note: NNES: Non-native English Speaker; GS: Grammatical Structures; FOR: Formality; JR: Justification and Reasoning; VOC: Vocabulary; AUS: Appropriate Use of the Speech Act; VC: Vagueness and Clarity; ALP: Appropriate Length of Production; SD: Social Distance; POL: Politeness; L1T: L1 Transfer; SS: Social Status; ORG: Organization

As Table 1 indicates, the criterion "*Grammatical Structure*" was the most frequent criterion deemed by the raters. Based on Table 1, the frequency of occurrence of the criteria across the six situations as well as the length of the comments was almost dissimilar among the raters. Overall, 29.34% of the raters considered grammatical structure as the dominant factor for their ratings. Besides, (n=11.96%) of the participants regarded "*Formality*" as the next important criterion for rating. The criterion of "*Justification and Reasoning*" was deemed by (n=11.06%) of the respondents. Further, the criterion of "*Vocabulary*" was noted by (n=10.15%) of the teachers. The criteria "*Appropriate Use of the Speech act*", "*Vagueness or Clarity*", and "*Appropriate Length of Production*" were respectively mentioned by 9.02%, 7.90%, and 7.90 of the raters. Additionally, the following rated criteria were "*Social Distance*" and "*Politeness*" with each having mentioned 5.86% and 5.41% respectively. Last but not least, the least frequent criteria stated as reasons for ratings by the

raters were "*L1 Transfer*", "*Social Status*", and "*Organization*" which were stated by 1.12%, 0.90%, and 0.67% of the raters correspondingly.

In addition, to answer the second research question and to explore the variations in ratings of NNESRs, descriptive analysis was employed which included the calculation of mean and standard deviation of the rating scores for the total DCTs and each situation.

As Table 2 presents, the descriptive statistics for the total agreement-holistic DCTs and the six situations thereof for non-native raters indicate that a score on each situation ranged from 1 (poor) to 4 (native-like). Besides, the mean (M) rating of the 50 non-native raters for total DCTs was 2.23. It shows that their overall evaluation of agreements in the six situations fell within the " fair" point on the scale. Although standard deviation (SD) for the total situation was comparatively low (SD=.46), the distance between minimum score (1) and maximum score (4) provides a rough account of divergence in the holistic rating of agreements.

Table 2. ***Descriptive Statistics for Agreement-Holistic(AH)***

Speech Act	N	Minimum	Maximum	Mean	Std. Deviation
A1	50	1.00	4.00	1.94	.62
A2	50	1.00	4.00	1.82	.72
A3	50	1.00	4.00	2.20	.86
A4	50	1.00	4.00	2.20	.86
A5	50	1.00	4.00	2.68	.98
A6	50	1.00	4.00	2.52	1.15
Total AH				2.23	.46

By the same token, the descriptive statistics for the total agreement-analytic DCTs and the six situations are presented in Table 3. As Table 3 indicates the mean (M) rating of the 50 non-native raters for total DCTs was

2.40. It shows that their overall evaluation of agreements in the six situations fell within the "fair" point on the scale. Although standard deviation (SD) for the total situation was comparatively low(SD=.35), the distance between minimum score (1) and maximum score (4) provides a rough account of divergence in the analytic rating of agreements.

Table 3. *Descriptive Statistics for Agreement-Analytic(AA)*

Speech Act	N	Minimum	Maximum	Mean	Std. Deviation
AA1	50	1.33	3.25	2.42	.44
AA2	50	1.33	3.25	2.33	.44
AA3	50	1.42	4.00	2.40	.61
AA4	50	1.33	4.00	2.66	.67
AA5	50	1.33	3.75	2.26	.68
AA6	50	1.00	3.50	2.32	.60
Total AA				2.40	.35

Finally, to answer the third research question and to inspect whether non-native raters were different in their assessment of the EFL learners' agreements, several intraclass correlation coefficients were calculated. As can be seen in Table 4, there was a significant intraclass correlation in the agreement-holistic group, ICC=.46, F=1.36, df= (49,245), p=.01 (see the descriptive statistics related to this group in Table 2).

Table 4. *Intraclass Correlation Coefficient: Agreement-Holistic*

	Intraclass Correlation	95% Confidence Interval		F Test			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Average Measures	.46	-.10	.54	1.36	49	245	.01

Similarly, the intraclass correlation coefficient was also explored for the analytic phase of the speech act of agreement. As shown in Table 5, the third group, agreement-analytic, was consistent in the rating of the speech act of agreement considering the significant intraclass correlation in Table 5, ICC= .65, F=2.85 df=(49,245), p=.00 (see the descriptive statistics related to this group in Table 3).

Table 5. *Intraclass Correlation Coefficient: Agreement-Analytic*

	Intraclass Correlation	95% Confidence Interval		F Test			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Average Measures	.65	.47	.78	2.85	49	245	.00

Further, the Pearson Product Moment Correlation Coefficient was computed as the estimate of inter-rater reliability of non-native English speaking teachers' holistic and analytic rating. This was computed between the mean of analytic rating components and the mean of holistic ratings. As it is shown in Table 6, the Correlation Coefficient is .50 which is statistically significant, suggesting that there was a convergence between the two ratings.

Table 6. *Interrater Reliability of Raters' Ratings for Agreement*

		AGH	AGA
AGH	Pearson Correlation	1	509**
	Sig. (2-tailed)		.000
	N	50	50
AGA	Pearson Correlation	509**	1
	Sig. (2-tailed)	.000	
	N	50	50

** Correlation is significant at the 0.01 level (2-tailed)

5. Discussion

Since pragmatic assessment has been widely contingent upon the overall effectiveness of speech acts, the use of rating scales is not adequately taken into account in speech act research. Seemingly, we deeply care and evaluate the linguistic knowledge of EFL learners, although the pragmatic knowledge of them is not greatly assessed and known to raters.

The primary aim of this study was to investigate the criteria in the holistic and analytic rating of EFL learners' production of the agreement by NNERs. The findings revealed that the raters employed twelve criteria while rating L2 agreement productions, although the weight given to the criteria was not consistent. The findings also showed that the criterion "grammatical structure" was the dominant one among the raters. The frequent use of this criterion is in light of the fact that Iranian raters are deeply concerned with the pragmalinguistic aspects of language (Alcon-Soler & Martinez-Flor, 2008; Alemi & Motamedi, 2019). Some mismatches were also found between the holistic and analytic ratings, namely pragmatic tone, use of discourse markers, epistemic stance markers, and the level of imposition. In addition, raters favored some new criteria such as justification and reasoning, vagueness or clarity, and L1 Transfer. Also, among the new favored criteria, "justification and reasoning" was correspondingly mentioned in other speech acts through different labels such as "explanation" by Tajeddin and Alemi (2014) in the case of apology by native raters, Tajeddin, Alemi, and Razzaghi (2014) in the case of apology and perceptions of impoliteness by both native English speakers and EFL learners, Fraser (1981), Olshtain and Cohen (1983), Holmes (1990), Sydorenko, Maynard, and Guntly (2015), and Alemi and Khanlarzadeh (2016). Besides, many of the discussed criteria have been addressed in the previous speech act studies. For example, Blum-Kulka and Olshtain's (1984) research on request and

apology addressed factors such as directness. Furthermore, criteria such as politeness, linguistic appropriacy have been remarked in several studies (e.g., Eslami-rasekh, Jafari, & Mehregan, 2012; Jalilifar, 2009; Murphy & Neu, 1996).

Employing criteria from pragmalinguistics, sociopragmatics, and metapragmatic aspects of language imply that non-native raters consider all of them. However, the raters considered pragmalinguistics features such as "grammatical structure" more often than sociopragmatics and metapragmatics ones which is in line with what the remarks of Alcon-Soler and Martinez-Flor (2008), who indicated that EFL teachers are predominantly concerned with pragmalinguistics knowledge. On the other hand, it stands in contrast with the results of other ILP studies (e.g., Alemi & Tajeddin, 2013; Alemi, Eslami-Rasekh, & Rezanejad, 2015) in which the focus of native raters was mostly on sociopragmatics features.

The findings of the current study also revealed that there was a great dispersion among non-native raters with regard to the rating methods. In particular, the degree and variations that they adopt when the rating is rather different, manifesting itself in different degrees of divergence and convergence. Divergence of non-native raters within a single situation and across situations reveals the subjective nature of scoring in general and pragmatic divergence particularly. This might be due to the fact that raters require training so as to be qualified to rate appropriately. Clearly, as Weigle (1994) pointed out, non-native raters who attend the training program are more capable of adjusting their evaluations concerning the native-speakers' benchmark.

Regarding the last research question, the raters managed to show less consistency when they were asked to rate holistically; however, their consistency maximized when they were asked to rate analytically. This was evident through the higher mean and lower standard deviation of raters within the analytic

method. Besides, the results suggested that the respondents were more consistent in the analytic assessment as it helped to minimize the differences in terms of rater variability and maximize the consistency among the raters comparatively speaking. The results are in line with Alemi and Motamedi (2019) about the speech act of disagreement.

Finally, non-native raters employed a relative convergence between the two rating methods. However, it is obvious that non-native raters applied a wide range of criteria with large variations and frequencies which were not constant across all situations. The inter-rater reliability among the two rating methods suggested that the convergence was not quite well (.50). To be more precise, a higher degree of convergence is demanding to call this a significant convergence. Therefore, the non-native raters' preferred criteria were not entirely the same in every single situation. Similarly, Youn (2007) also claimed that each rater tends to propose a unique pattern that might be contingent upon the test type and speech act.

6. Conclusion

This study suggested twelve different criteria adopted by nonnative raters while rating (holistically and analytically) agreement outputs. The set of criteria consisted of: "grammatical structure", "formality", "justification and reasoning" , "vocabulary", "appropriate use of the speech act", "vagueness or clarity", "appropriate length of production", "social distance", "politeness", "L1 transfer", "social status", and "organization". The tendency of non-native raters to the three dominant aspects of pragmatics, namely pragmalinguistics, sociopragmatics, and metapragmatics, which were already embedded in the analytic scale, was seen. However, the inclination of raters to pragmalinguistics

and sociopragmatics was more observable. In particular, pragmalinguistics attracted the highest attention among NNERs.

Basically, raters demonstrated a higher level of consistency within the analytic rating method compared to holistic. This higher consistency may be due to the fact the participants of our research especially the in-service ones were not comprehensively aware of pragmatic assessment and appropriacy.

Even though there was a relative convergence among the two rating methods, there were significant differences between them too. To be clear, the divergence obtained from the two rating methods indicated that non-native raters are not fully familiar with the pragmatic rating. As a result, it indicates that EFL raters significantly require training programs so as to enhance their accuracy in ILP rating.

The findings of this study have significant implications. First and foremost, it can assist EFL teachers to broaden their horizons in testing and open their eyes to the fact that assessment is not just restricted to linguistic aspects, but other issues should be noticed as well. EFL learners' pragmatic knowledge needs to be examined as their linguistic knowledge is to be. Consequently, a comprehensive set of criteria is required to rate this ability. More precisely, these sets of criteria for each specific speech act help raters to measure the learners' interlanguage pragmatic knowledge more consistently and efficiently. Exploring non-native raters' criteria in assessing pragmatic knowledge can reveal the differences in ratings; and by running workshops regarding standard pragmatic rubrics, we can help non-native raters to reset their standards and make them native-like in order to prevent raters' inconsistencies and to add test fairness.

This study was limited to the speech act of agreement only, and the level of the designed tasks of this study was for upper-intermediate EFL learners since learners in lower levels could not understand the pragmatic stimuli and produce

the intended responses. The sample size could also be larger if the researcher had an access to a wider range of participants. On the whole, pragmatic assessment and more importantly ILP rating are areas which are relatively unexplored in the EFL context. Conclusively, further studies need to be conducted to investigate the effect of teacher variables such as age, gender, teaching experience, background, and field of study. Also, other speech acts such as condolence, cursing, etc. need close attention.

References

Alcon Soler, E., & Martinez Flor, A. (Eds.). (2008). *Investigating pragmatics in foreign language learning, teaching and testing*. Clevedon: Multilingual Matters.

Alemi, M., Eslami Rasekh, Z., R., & Rezanejad, A. (2014). Rating EFL learners' interlanguage pragmatic competence by non-native English speaking teachers. *Procedia-Social and Behavioral Sciences, 98,* 171-174.

Alemi, M., Eslami Rasekh, Z., & Rezanejad, A. (2015). Iranian non-native English speaking teachers' rating criteria regarding the speech act of compliment: An investigation of teachers' variables. *Journal of Teaching Language Skills, 6*(3), 21-49.

Alemi, M., & Khanlarzadeh, N. (2016). Pragmatic assessment of request speech act of Iranian EFL learners by non-native English speaking teachers. *Iranian Journal of Language Teaching Research*, *4*(2), 19-34.

Alemi, M., & Khanlarzadeh, N. (2017). Native and non-native teachers' pragmatic criteria for rating request speech act: The case of American and Iranian EFL teachers. *Applied Research on English Language, 6*(1), 67-84.

Alemi, M., & Motamedi, M. (2019). Pragmatic criteria in the holistic and analytic rating of the disagreement speech act of Iranian EFL learners by non-native English speaking teachers. *Journal of Teaching Language Skills*, *38*(1), 1-36.

Alemi, M., & Tajeddin, Z. (2013). Pragmatic rating of L2 refusal: Criteria of native and non-native English teachers. *TESL Canada Journal*, *30*(7), 63-81.

Al-Khanaifsawy, A. N. (2016). Investigating Iraqi EFL learners' use of the speech act of agreement. *Adab-alkufa. 27*(1), 11-30.

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing, 12*(2), 238-257.

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge, UK: Cambridge University Press.

Bardovi-Harlig, K. (2001). Evaluating the empirical evidence: Grounds for instruction in Pragmatics?. In K. Rose & G. Kasper (Eds.), *Pragmatics in language teaching* (pp. 13-32). Cambridge, UK: Cambridge University Press.

Barraja-Rohan, A. M. (2011). Using conversation analysis in the second language classroom to teach interactional competence. *Language Teaching Research*, *15*(4), 479-507.

Birner, B. J. (2012). *Introduction to pragmatics*. Malden, MA: Wiley-Blackwell.

Blum-Kulka, S., & Olshtain, E. (1984). Requests and apologies: A cross-cultural study of speech act realization patterns (CCSARP). *Applied Linguistics*, *5*(3), 196-213.

Brown, P., & Levinson, S. (1987). *Politeness: Some universals in language usage.* Cambridge: Cambridge University Press.

Carr, N. T. (2000). A comparison of the effects of analytic and holistic rating scale types in the context of composition tests. *Issues in Applied Linguistics, 11*(2), 207-229.

Cohen, A. D. (2008). Teaching and assessing L2 pragmatics: What can we expect from learners? *Language Teaching*, *41*(2), 213-235.

Crystal, D. (1997). *The Cambridge encyclopedia of language* (2nd ed.). Oxford: MIT Press.

Douglas, D., & Smith, J. (1997). *Theoretical underpinnings of the Test of Spoken English Revision Project* (TOEFL Monograph Series Report No. 9). Princeton, New Jersey: Educational Testing Service.

Eslami-Rasekh, Z. (2005). Raising the pragmatic awareness of language learners. *ELT Journal, 59*(3), 199-208.

Eslami, A., Jafari, D. S., & Mehregan, M. (2012). How do you react to the breakdown after it happens? Do you complain about it? : A contrastive study on the complaint behavior in American English and Persian. *Procedia-Social and Behavioral Sciences*, *47*, 34-40.

Fraser, B. (1981). On apologizing. In F. Coulmas (Ed.), *Conversational Routine: Explorations in standardized communication situations and pre-patterned speech* (pp. 259-271). New York: Mouton.

Fulcher, G. (2003). *Testing second language speaking.* London: Pearson Education Limited.

Godshalk, F. I., Swineford, F., & Coffman, W.E. (1966). *The measurement of writing ability*. New York: College Entrance Examination Board.

Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 241-276). Norwood, NJ: Ablex.

Holmes, J. (1990). Apologies in New Zealand English. *Language in Society, 19*(2), 155-199.

Hunter, D. M., Jones, R. M., & Randhawa, B. S. (1996). The use of holistic versus analytic scoring for large-scale assessment of writing. *The Canadian Journal of Program Evaluation*, *11*(2), 61-85.

Jalilifar, A. (2009). Request strategies: Cross-sectional study of Iranian EFL learners and Australian native speakers. *English Language Teaching, 2*(1), 46-61.

Jianda, L. (2006). Assessing EFL learners' interlanguage pragmatic knowledge: Implications for testers and learners. *Reflections on English Language Teaching 5* (1), 1-22.

Johnson, F. (2006). Agreement and disagreement: A cross-cultural comparison. *BISAL, 1*, 41-67.

Ishihara, N., & Cohen, A. D. (2010). *Teaching and learning pragmatics: Where language and culture meet*. Harlow, UK: Pearson Education.

McNamara, T. F. (1996*). Measuring second language performance*. London: Longman.

Murphy, B., & Neu, J. (1996). My grade's too low: The speech act set of complaining. In S. M. Gass, & J. Neu (Eds.), *Speech acts across cultures: Challenges to communication in a second language* (pp. 191-216). Berlin: Mouton de Gruyter.

Olshtain, S., & Cohen, A. D. (1983). Apology: A speech act set. In N. Wolfson & E. Judd (Eds.), *Sociolinguistics and language acquisition* (pp. 18-35). Rowley, MA: Newbury House.

Olshtain, E., Blum Kulka, S. (1985). Degree if approximation: Nonnative reactions to native speech act behavior. In S. M. Gass, & C. G. Madden (Eds.), *Input in second language acquisition* (pp. 303-325). Rowley, MA: Newbury House.

Pomerantz, A. (1984). Agreeing and disagreeing with assessments: Some features of preferred / dispreferred turn shapes. In J. Atkinson & J. M. Heritage (Eds.), *Structures in Social Action* (pp. 57-101). Cambridge: Cambridge University Press.

Rose, K. R., & Kasper, G. (Eds.). (2001). *Pragmatics in language teaching*. New York: Cambridge University Press.

Schegloff, E. A. (1996). Confirming allusions: Toward an empirical account of action. *American journal of sociology*, *102*(1), 161-216.

Scollon, R., & Scollon, S.W. (2001). 27 discourse and intercultural communication. In D. Schiffrin, D. Tannen, & H. Hamilton (Eds.), *The handbook of discourse analysis*, (pp. 538-547). Oxford: Blackwell.

Shohamy, E. (1995). Performance assessment in language testing. *Annual Review of Applied Linguistics, 15,* 188-211.

Sydorenko, T., Maynard, C., & Guntly, E. (2015). Rater behaviour when judging language learners' pragmatic appropriateness in extended discourse. *TESL Canada Journal*, *32*(1), 19-41.

Taguchi, N. (2011). Rater variation in the assessment of speech acts. *Pragmatics, 21*(3), 453-471.

Tajeddin, Z., & Alemi, M. (2014). Criteria and bias in native English teachers' assessment of L2 pragmatic appropriacy: Content and FACETS analyses. *The Asia-Pacific Education Researcher*, *23*(3), 425-434.

Tajeddin, Z., Alemi, M., & Razzaghi, S. (2014). Cross-cultural perceptions of impoliteness by native English speakers and EFL learners: The case of apology speech act. *Journal of Intercultural Communication Research*, *43*(4), 304-326.

Tajeddin, Z., & Alemi, M. (2015). *Functional communication in English*. Tehran: Jungle Press.

Weigle, S. C. (1994b). Effects of training on raters of ESL compositions. *Language Testing, 11*(2), 197-223.

Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.

Xi, X. (2007). Evaluating analytic scoring for the TOEFL Academic Speaking Test (TAST) for operational use. *Language Testing, 24*(2), 251-286.

Youn, S. J. (2007). Rater bias in assessing the pragmatics of KFL learners using facets analysis. *Second Language Studies, 26*(1), 85-163.