Pragmatic Rating of L2 Refusal: Criteria of Native and Nonnative English Teachers

Minoo Alemi and Zia Tajeddin

Many studies have shed light on rater criteria for assessing the performance of language skills (e.g., Eckes, 2005). However, the interface between rater assessment and interlanguage pragmatics (ILP) has remained largely unnoticed. To address this interface, this study explored the ratings native (NES) and nonnative English speaking (NNES) teachers assigned to second language (L2) refusal production and the criteria they applied in their ratings. To this end, 50 NES and 50 NNES teachers participated in rating L2 refusal production of EFL learners that included responses to a 6-item written discourse completion task. The data were analyzed qualitatively and quantitatively. Qualitative analysis showed that native teachers applied 11 criteria and nonnative teachers applied 6 criteria in their pragmatic ratings. Reasoning/explanation was the leading criterion in teacher assessment among native raters, whereas politeness was the main criterion for nonnative ratings. Quantitative analysis documented variation in the frequency of drawing on rating criteria and significant differences in ratings, with NNES teachers being more lenient and divergent in their ratings. The results suggest there is a gap between NES and NNES teachers in terms of rating criteria, strictness, and convergence in rating.

Plusieurs études ont fait la lumière sur les critères employés dans l'évaluation de compétences linguistiques (par ex. Eckes, 2005). Toutefois, peu de recherche a porté sur l'interface entre les évaluations et la pragmatique de l'interlangue. Pour aborder cette interface, l'étude a porté sur l'évaluation par des enseignants anglophones et des enseignants dont l'anglais n'était pas la langue maternelle de la production de refus en langue seconde, et de leurs critères d'évaluation. À cette fin, 50 enseignants anglophones et 50 enseignants dont la langue maternelle n'était pas l'anglais ont évalué la production de refus en L2 d'apprenants en anglais langue étrangère, y compris leurs réponses écrites à une tâche de complètement à six items. Les données ont été soumises à des analyses quantitatives et qualitatives. L'analyse qualitative a indiqué que, dans leurs évaluations, les enseignants locuteurs natifs d'anglais ont appliqué 11 critères et les enseignants dont la langue maternelle n'est pas l'anglais, 6 critères. Le critère d'évaluation principal des anglophones était le raisonnement/l'explication alors que pour les enseignants non natifs, c'était la politesse. L'analyse quantitative a révélé une variation dans la fréquence de recours aux critères d'évaluation et des différences significatives dans les évaluations, celles des enseignants non-natifs faisant preuve de moins de sévérité et de plus de divergence. Les résultats indiquent un écart entre les critères, la sévérité et l'homogénéité dans les évaluations des enseignants anglophones et enseignants non-anglophones.

Introduction

One component of pragmatic competence is to know how to perform a particular speech act. Among speech acts, *refusal* is highly complicated, primarily because it often involves lengthy negotiations and face-saving manoeuvres to accommodate the noncompliant nature of the speech act. Since refusal normally functions as a second pair part, it precludes extensive planning on the part of the refuser.

Against this backdrop, the study of rating speech acts such as refusal is salient on two grounds. First, refusal is a commonly used speech act in the process of communication and hence is a constituent of many pragmatic assessment tasks. Furthermore, as the speech act of refusal is realized differently across cultures and communicative situations, nonnative teachers should become familiar with native criteria for rating refusal production, particularly in the outer circle (Kachru, 1997), where there are no established local English norms for pragmatic appropriateness. Despite such saliency in the outer circle context, little research has been conducted to date on the criteria used by nonnative English speakers (NNESs) in rating refusal production as measured against the native English speaker (NES) baseline sociopragmatic and pragmalinguistic norms for rating the appropriateness of speech act production. This is particularly important in a foreign language context or in the expanding circle where there is no local variety of English and hence nonnative speakers are "norm dependent," that is, dependent on native speaker norms in their rating (Kachru, 1992). Accordingly, this study aimed to investigate the pragmatic rating of second language (L2) refusal production by nonnative teachers as measured against native English-speaking teachers' ratings.

Rating of Learner Productions

In performance assessment, our judgments are affected by our perceptual vantage points. The effects of rater perceptions introduce highly subjective factors that make ratings more or less inaccurate. Rater bias is a major problem when language raters judge learners' performance using criteria that are vague or highly subjective. Thus, if they use such rating criteria, it is likely that inconsistency and inaccuracy come into play. In fact, assessment of learners' performance is a complex process with many ramifications. Knoch, Read, and von Randow (2007) point out that raters' judgments are prone to various sources of bias and error that can ultimately undermine the quality of the ratings.

A number of studies using different psychometric methods have identified various rater effects (e.g., Myford & Wolfe, 2003, 2004) that need to be addressed if an acceptable level of reliability is to be maintained. Rater effects can be summarized as (a) the severity effect, (b) the halo effect, (c) the central tendency effect, (d) inconsistency, and (e) the bias effect (Myford & Wolfe, 2003). Studies focusing on language performance assessments, as reviewed by Eckes (2005), showed a significant range of rater effects. These studies, in particular, identified differences in raters' severity or leniency (e.g., Engelhard, 1994; Engelhard & Myford, 2003; Lumley & McNamara, 1995). These differences were found to be resistant to rater training (Barrett, 2001; Lumley & McNamara, 1995; Weigle, 1998) and to persist in raters for a long time (Fitzpatrick, Ercikan, Yen, & Ferrara, 1998). Furthermore, researchers identified significant effects for rater-ratee interaction (Kondo-Brown, 2002; Lynch & McNamara, 1998), rater-task type interaction (Lynch & McNamara, 1998; Wigglesworth, 1993), and rater-criteria interaction (Wigglesworth, 1993).

Rater effects need more attention, as they are sources of systematic variance in observed ratings associated with raters rather than ratees (Cronbach, 1995; Hoyt, 2000; Myford & Wolfe, 2003). As a result, rater effects that are irrelevant to the construct being rated threaten the validity of the assessment procedure (Bachman, 2004; Messick, 1989, 1995; Weir, 2005). Two rater effects related to the main theme of this study are severity and inconsistency. The former occurs when raters are found to rate either too harshly or too leniently, as compared with other raters or established baseline ratings. The latter is exhibited when raters tend to rate in terms of different criteria or the inconsistent application of criteria. For example, they might favour a certain group of test takers or mainly apply one criterion at the expense of others. The variability of ratings as a result of these effects has been addressed in studies on speaking and writing (e.g., Schaefer, 2008; Shi, 2001). One source of rater variability is the status of the rater as a native or nonnative speaker. It is very important to determine whether native English- speaking and nonnative English-speaking raters use the same criteria for rating tasks. However, the results of studies comparing NES and NNES who rated oral and written language performance vary. Barnwell (1989) found that NESs were harsher in their evaluations than NNESs, whereas others, such as Fayer and Krasinski (1987), found that NNES raters were more severe. For instance, Fayer and Krasinski investigated Puerto Rican learners of English speech act production and gave their samples to two groups of raters: NES and Puerto Rican speakers. Their results revealed that NNES raters were harsher, especially with respect to pronunciation errors, than NES raters.

Although the literature is replete with references to native speaker assessment of speaking and writing performance, it seems that only two studies on rater variability are related to pragmatic rating (Taguchi, 2011; Youn, 2007). Taguchi studied native speakers' ratings of two types of speech acts produced by EFL learners. The data revealed similarities and differences in the raters' use of pragmatic norms and social rules in evaluating the appropriateness of speech acts. Focusing on Korean as a foreign language, Youn's study showed different degrees of severity in native Korean raters' ratings of speech act performance. However, there is no mention of native raters' criteria compared with nonnative raters' on pragmatic assessment. As a result, this issue is still underexplored.

Refusal: Nature and Strategies

Refusal functions as a response to an initiating act and is considered to be a speech act in which "a speaker fails to engage in an action proposed by the interlocutor" (Chen, Ye, & Zhang, 1995, p. 121). Refusal is a face-threatening act because it contradicts the listener's wants. The negotiation of refusal entails frequent attempts at directness or indirectness and also other degrees of politeness appropriate to the situation (Eslami, 2010). In addition, refusal behaviours vary across cultures, and pragmatic transfer occurs as learners rely on their "deeply held native values to carry out complicated and face-threatening speech acts like refusals" (Beebe, Takahashi, & Uliss-Weltz, 1990, p. 68). Hence, a proper understanding and production of refusal and, in turn, its rating require a certain amount of culture-specific knowledge.

As refusal is face-threatening, it usually involves a long negotiated sequence, and its form and content vary, depending on situational variables such as power, distance, and imposition. Saying "no" to requests, invitations, offers, and suggestions is a kind of dispreferred action that is typically complex, mitigated, indirect, and accompanied by prefaces, hesitations, repairs, apologies, and accounts (e.g., Levinson, 1983; Pomerantz, 1984).

Various strategies should be employed to avoid offending one's interlocutors. Takahashi and Beebe (1987) noted that an inability to say "no" politely will lead to an offense. Due to the different nature of this speech act, as well as some degree of risk-taking involved in refusing, pragmatic knowledge helps EFL learners realize appropriate strategies. However, a layer of complexity related to cultural issues exists and, in some cases, such as found in Ishihara and Tarone's (2009) study, L2 speakers intentionally resist what they perceive as native-speaker norms.

Beebe et al. (1990) categorized refusal into semantic formulas and adjuncts appropriate for refusal strategies. This taxonomy includes both direct and indirect strategies. In the direct category, two semantic formulas are included. They are performative (e.g., *I refuse it*) and nonperformative statements (e.g., *I can't*). In indirect strategies, there are 11 semantic formulas: statement of regret, wish, excuse/reason/explanation, statement of alternative, set condition for future or past acceptance, promise of future acceptance, statement of principle, statement of philosophy, attempt to dissuade interlocutor, acceptance that functions as a refusal, and avoidance.

The Current Study

This study was aimed at investigating native English speaking raters' and nonnative English speaking raters' criteria for rating the EFL learners' pragmatic production of refusals. To do so, the following research questions were addressed:

1. What criteria are used by native and nonnative English speaking raters in rating the speech act of refusal produced by EFL learners?
2. Is there any significant difference between native and nonnative English speaking raters in rating the speech act of refusal produced by EFL learners?

Method

Participants

One group of participants included 50 educated native teachers of English from the United States, the United Kingdom, Canada, and Australia. The homepage data and the background information they provided clearly showed that they were NESs from these four countries. They were faculty members teaching ESL at different language centres in international universities. The other group consisted of 50 NNES teachers. Each had at least three years of teaching experience and held an MA degree in applied linguistics. The nonnative teachers were from different language centres in Iran, where English is taught as a foreign language. Both groups were asked to participate in this study via e-mail. Both groups included male and female teachers.

Instrument

A written discourse completion test (WDCT) was used to collect the data in this study, as it is a common measure to elicit learners' production of pragmatics. It was made up of six refusal situations reflecting different degrees of formality, power relation, and distance (see Appendix). The situations included educational contexts, workplace contexts, and daily-life contexts. In terms of power status and familiarity, the situations were marked by equal and unequal power relations, as well as familiar and unfamiliar interlocutors. Each situation was followed by a response given by an EFL learner. A number of EFL learners were asked to provide a response to each situation. Of the responses, one was selected by the researchers for each situation to ensure that the responses to the six situations varied in their degrees of pragmatic appropriateness. Thus the focus in the selection procedure was placed on pragmatic failure or appropriateness rather grammatical inaccuracy, as reflected in the choice of words *unsatisfactory* and *appropriate* in the rating scale. Every response was followed by a rating scale ranging from 1 (*very unsatisfactory*) to 5 (*most appropriate*). Below the rating scale for each response, there was a space entitled "criteria" so that the raters could write comments on the pragmatics criteria they applied to the rating of the response to each situation.

Data Collection Procedure

The refusal WDCT was administered in paper format to about 20 EFL students. They were studying for a BA program in English literature or translation in an Iranian university, and their L1 was Persian. The responses to each situation were reviewed by the researchers and one response selected for each situation. After this selection, the WDCT was sent electronically to NES teachers to rate the appropriateness of responses on a 5-point Likert scale and to write the criteria for their rating in comment format. The questionnaire was first uploaded to the *SurveyMonkey®* site, and native ESL teachers in different universities in the United States, the United Kingdom, Canada, and Australia were asked via e-mail to complete the questionnaire on that site electronically. Of 800 teachers contacted through e-mails, 50 filled out the questionnaire completely. Of the 106 nonnative teachers contacted, 50 completed the rating sheets and returned the WDCT with their rating comments.

Data Analysis

The current study investigated the rating of L2 refusal production by NES and NNES English teachers. In part, it used the content analysis technique to analyze the data. To derive the criteria that both native and nonnative raters considered in rating EFL learners' refusal production, the content of their comments about the pragmalinguistic and sociopragmatic appropriateness or infelicity of each response was analyzed. The analysis of criteria based on the comments consisted two steps. The first was a careful analysis of refusal strategy frameworks based on a modified version of Beebe et al.'s (1990) taxonomy. Although the strategies in that framework represented refusal production rather than functioning as a rating rubric, they helped to identify in the raters' comments criteria related to the (in)appropriateness of refusal in terms of the underrepresentation, overrepresentation, or nonrealization of certain strategies in response to a situation in the WDCT. The second source of insight was Brown and Levinson's (1987) politeness model, in which strategies of positive and negative politeness are depicted. The model contributed to the analysis of the criteria relevant to the violation of politeness in refusal production reflected in the raters' comments. In the quantitative part of the data analysis, frequency counts and *t*-tests were conducted to measure the difference between the refusal ratings of native raters and nonnative raters.

Results

Refusal Rating Criteria

Research Question 1 was concerned with the criteria used by NES and NNES teachers in rating the speech act of refusal produced by EFL learners. To derive the criteria that both NES and NNES raters used, the content of their comments stating the reasons for the pragmatic appropriateness of each response was analyzed. This analysis resulted in 11 criteria for rating refusal. The criteria, as described below, show that both NES and NNES teachers specified pragmatic, rather than grammatical, features as a source of their rating of refusal production.

(1) Brief apology. This refusal criterion is important as it prepares the interlocutor for an upcoming refusal. Two examples of this criterion derived from NES and NNES rating comments are given below.

NES comment: *I would add an apology before refusing the invitation.*
NNES comment: *An apology is needed before any refusal.*

(2) Statement of refusal. The second refusal criterion, a statement of refusal, is a head act expressing the refusal and giving a clear idea of rejection to an interlocutor. An example of the application of this criterion by NES raters is given below. NNES raters did not use this criterion.

NES comment: *A proper refusal should include a statement of refusal in terms that are both specific and in a tone appropriate to the social relationship between the one refusing and the requester or inviter on certain occasions.*

(3) Offer suitable consolation. This criterion, an offer of suitable consolation, follows the head act to mitigate the refusal. Like the previous criterion, NNES raters did not employ this criterion to rate the WDCT.

NES comment: *If I were her, I would offer a suitable consolation and say "Could we possibly have lunch some other day?"*

(4) Irrelevancy of refusal. This criterion focused on the irrelevancy of a refusal. In some cultures, refusal is so indirect that the addressee cannot understand whether it is a refusal or an acceptance of an offer or invitation.

NES comment: *This sounds like an acceptance of the apology, not a refusal.*
NNES comment: *It is an apology acceptance, not refusal!*

(5) Explanation/Reasoning. The fifth refusal criterion was an explanation that follows the head act to justify the refusal. After refusing an offer, an invitation, a suggestion, or a piece of advice, some explanation is needed to soften the face-threatening effect.

NES comment: *A bit more effort to explain the reason would be required here.*
NNES comment: *In my opinion, frankly speaking and elaborating on the main issue and reason is better than evading the issue.*

(6) Cultural problem. Because pragmatic competence is highly dependent on culture, cultural misinterpretation occurs in EFL contexts. NNES raters did not apply this criterion in their ratings.

NES comment: *This might be something cultural but I could never say so.*

(7) Dishonesty. This criterion is sometimes misinterpreted as indirectness in refusal and may result in offering false excuses rather than giving reasons. Only NES raters referred to this criterion in their comments.

NES comment: *Being more honest with your reasons for not wanting to ride together would have been easier for the old friend to take.*

(8) Thanking. The eighth criterion was thanking, which is a mitigation device to soothe the face-threatening effect and to console the hearer. NNES raters gave no comment representing this criterion.

NES comment: *First, you should thank her for the invitation and then explain the reason for refusing it.*

(9) Postponing to another time. This criterion reflected the need for mitigation so that the face-threatening effect could be softened by postponing the offer or request to another time. Both NES and NNES raters used this criteria to rate the appropriateness of WDCT responses.

NES comment: *The speaker should say, "Can I take you up on your offer some other time?"*

NNES comment: *The speaker could postpone the invitation to [an]other time politely.*

(10) Statement of alternative. This criterion was used to evaluate learners' success/failure in giving other choices after a refusal, in order to ease the situation for the hearer. The raters' comments below document the significance of this criterion.

NES comment: *You should say "We can arrange something else for some other time."*

NNES comment*:* *In order to not hurt your friend, it's better to ask her to have copies of your notes instead of lending them.*

(11) Politeness. The last criterion was politeness, the interpretation of which varies in different cultures. In fact, its interpretation depends on the values of social distance, dominance, and degree of imposition in a given context. The severity of its violation varies cross-culturally. Both NES and NNES raters took this criterion into account in rating WDCT responses.

NES comment: *I would not criticize [the] other person without knowing more about the circumstances they are in, so I find this response a bit rude.*

NNES comment*:* *He should politely reject the suggestion to show the respect.*

Table 1 shows the frequency of NES and NNES raters' criteria for the total WDCT and across all six situations. As some NNES teachers failed to provide criteria for their pragmatic rating in certain WDCT contexts, the total number of their criteria was far less than that of NES teachers. This shows that the former group had comparatively lower pragmatic awareness of the rationale behind the (in)appropriateness of the refusal produced in a WDCT situation.

In general, the raters' comments on the refusals across the situations manifested many sources of inappropriateness, such as lack of *explanation*, *politeness*, *cultural problem*, *postponing to another time*, using *brief apology expression* appropriately, *offering repair*, and *thanking*. NES and NNES raters did not agree in their ratings of most of the refusal cases. Moreover, their criteria were different, due to lack of awareness in terms of appropriate refusal on the one hand and English sociocultural norms on the other. Results of the study indicate that, to make an accurate assessment of students' performance, NES and NNES teachers frequently applied a variety of relatively stable criteria that remained applicable from situation to situation. The criteria common across situations were *explanations*, *politeness*, *cultural problems*, *speech act appropriateness*, and *offer compensation*.

NES-NNES Refusal Ratings

Research Question 2 was raised to investigate the difference between NES and NNES teachers in rating the speech act of refusal produced by EFL learners. To address the research question, descriptive statistics were calculated and *t*-test procedures conducted. Table 2 presents descriptive statistics for refusal rating by NES and NNES raters. As shown in the table, the overall mean refusal rating was 2.59 for NES raters and 3.29 for NNES raters. The highest mean for native ratings across situations was 3.32 and the lowest was 2.06, while the highest mean for non-native raters was 4.02 and the lowest was 2.56. Table 2 shows that NNES raters' ratings for all situations were higher than those of NES raters. Furthermore, standard deviations of NNES ratings for the total WDCT and all six situations therein were found to be greater, showing less convergence in their ratings compared with the NES ratings.

Next, an independent-samples *t*-test was conducted to compare the difference in refusal rating between NES raters and NNES raters (Table 3). As displayed in the table, there was a significant between-group difference in total refusal ratings ($t = 7.21$, df = 98, $p = .000$). NES and NNES manifested variation in their ratings across all situations except for Situation 4 ($t = 0.30$, df = 98, $p = .76$). As multiple *t*-tests were applied for the analysis of Research Question 2, to avoid Type I error the Bonferroni method was used to arrive at adjusted alpha-level. The results showed that the differences in five situations (all except Situation 6) remained significant after the Bonferroni correction.

Table 1
Frequency of Refusal Criteria among NES and NNES Raters.

Situation	Group	BA	SOR	OSC	IOR	E/R	CP	D	T	PAT	SOA	P	TOTAL
1	NNES	0	0	0	10	0	0	0	0	0	0	3	13
	NES	0	0	0	32	0	0	0	0	0	0	0	32
2	NNES	0	0	0	0	10	0	0	0	0	0	10	20
	NES	3	1	1	0	28	5	5	1	0	0	5	49
3	NNES	2	0	0	0	6	0	0	0	5	1	14	28
	NES	1	2	3	0	8	24	1	0	7	5	1	52
4	NNES	3	0	0	0	2	0	0	0	0	6	15	26
	NES	2	1	2	0	9	0	1	0	0	5	30	50
5	NNES	1	0	0	0	10	0	0	0	2	0	7	20
	NES	3	2	2	0	18	4	1	2	2	0	10	44
6	NNES	3	0	0	0	8	0	0	0	2	0	8	21
	NES	9	1	1	0	20	0	0	6	2	0	12	51
Total NNES		9	0	0	10	36	0	0	0	9	7	57	128
Total NES		18	7	9	32	83	33	8	9	11	10	58	278
Percentage	NNES	7.03%	0%	0%	7.81%	28.13%	0%	0%	0%	7.03%	5.54%	44.53%	
	NES	6.47%	2.52%	3.24%	11.51%	29.86%	11.87%	2.88%	3.24%	3.96%	3.60%	20.86%	

Note. BA = brief apology; SOR = statement of refusal; OSC = offer suitable consolation; IOR = irrelevancy of refusal; E/R = explanation/reasoning; CP = cultural problem; D = dishonesty; T = thanking; PAT = postponing to another time; SOA = statement of alternative; P = politeness.

Table 2
Descriptive Statistics of Ratings by NES Raters and NNES Raters for Refusal

Situation	Group	*N*	Mean	Std. Deviation	Situation	Group	*N*	Mean	Std. Deviation
Situation 1	NNES	50	3.18	1.47	Situation 4	NNES	50	3.00	1.05
	NES	50	2.08	1.10		NES	50	2.94	.93
Situation 2	NNES	50	3.50	1.15	Situation 5	NNES	50	3.50	1.01
	NES	50	2.58	.83		NES	50	2.56	.79
Situation 3	NNES	50	4.02	1.13	Situation 6	NNES	50	2.56	1.16
	NES	50	3.32	1.02		NES	50	2.06	.77
Total	NNES	50	3.29	1.29					
	NES	50	2.59	.83					

Table 3
T-tests of Mean Differences in Refusal Ratings by NES Raters and NNES Raters

	Levene's Test for Equality of Variances		*t*-test for Equality of Means						
								95% Confidence Interval of the Difference	
	F	Sig.	*t*	*df*	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Situation 1	12.49	.001	4.23	98	.000	1.10	.26	.58	1.61
Situation 2	8.11	.005	4.58	98	.000	.92	.20	.52	1.31
Situation 3	.28	.597	3.24	98	.002	.70	.21	.27	1.12
Situation 4	1.31	.254	.302	98	.763	.06	.19	-.33	.45
Situation 5	3.40	.068	5.17	98	.000	.94	.18	.58	1.30
Situation 6	15.55	.000	2.53	98	.013	.50	.19	.10	.89
Total	.01	.910	7.21	98	.000	.70	.09	.50	.89

These results, in conjunction with those related to the frequency of criteria, indicate that NES and NNES teachers differed from each other not only in their application of rating criteria to evaluation of refusals made in different WDCT situations, but also in assigning scores to rate the appropriateness of refusals.

Discussion

Similar to language assessment in general, pragmatic assessment can be affected by three main variables: test task, rater characteristics, and rating criteria. Many studies have shed light on the third variable (rating criteria) for assessing the performance of language skills (e.g., Eckes, 2005; Gamaroff, 2000). However, the interface between rating criteria and pragmatics has remained largely unexamined. Hence, this study was conducted in a multiple-raters setting with NES and NNES raters to explore the impact of raters on pragmatic assessment of the speech act of refusal in terms of rating criteria and rating scores.

The first objective of this study was to discover how NES and NNES teachers rated L2 refusal production and what criteria they applied to the evaluation of its appropriateness. With regard to raters' rating criteria, the results of this study showed that NES and NNES teachers applied certain criteria to evaluate the appropriateness of L2 refusal production. Many of these criteria are pragmatically general or universal, in that they can be applied to the assessment of other speech acts. Salient instances of such criteria were *explanation* and *politeness*. The largely homogeneous rating criteria, particularly among NES teachers from different nationalities, lend further support to the universality of many pragmatic criteria.

Besides general pragmatics rating criteria, such as politeness and explanation, this study shed light on criteria specific to refusal, including *brief apology*, *state of refusal*, and *offer suitable consolation*. The findings indicate that speech act rating required an awareness of specific criteria involved in the appropriateness of a particular speech act. Unlike rating language skills, which largely depends on a set of general criteria, pragmatics rating is, to some extent, shaped by the nature of a particular speech act and the criteria specifically related to it. It follows that both groups of raters drew on two types of criteria in their rating: pragmatically general criteria and speech-act-specific criteria. A very revealing aspect of this study comes from the finding that most of the refusal-rating criteria corresponded to the strategies needed to produce refusal. This is strong evidence in favor of rating validity. Raters need to use the components of a construct and the strategies underlying performance to maximize their rating validity. In the case of refusal, such correspondence strengthens the validity of pragmatic rating.

The findings from this study also revealed variability in different situations among teachers as evidenced by the frequency of criteria reported. The frequency-based variability may be a determining factor affecting the rating of pragmatic performance. The most frequent criterion mentioned by NES raters was *explanation,* which can be attributed to the nature of the refusal speech act. However, NNES raters applied *politeness* as the main criterion. It seems that NNES raters mostly regarded *politeness* as a general criterion and hence overused it to justify any inappropriate production of refusal; as a result, they lost sight of the fact that appropriate apology required the provision of specific reasoning to refuse or reject a suggestion, an invitation, an offer, or a piece of advice. Variation in the frequency of rating criteria reported across situations is also a manifestation of divergence existing in evaluating the appropriateness of L2 refusal production in each single situation. For instance, in situation 1, all NES raters applied *irrelevancy of refusal* because the L2 learner had not produced a refusal; however, NNES raters mostly applied *politeness* and felt sympathy with the interlocutor, a domestic servant, in that WDCT situation. The finding of this study is in line with that of Taguchi (2011), which revealed divergent focus among raters of different nationalities in their use of pragmatic norms when evaluating appropriateness of speech acts.

The second objective of the study was to explore the ratings that NES and NNES English-speaking teachers assigned to refusal production. Results showed that NNES raters manifested different rating behaviour by consistently overrating refusals across situations and thus being inclined towards leniency in rating. This NES-NNES difference in refusal rating can be explained in terms of variation in their perceptions of such variables as power, social status, and preferred refusal strategies by native and nonnative speakers (e.g., Félix-Brasdefer, 2003; Takahashi, 1996; Takahashi & Beebe, 1987). Nonnative speakers' perception of social status, for example, is among the factors that influence their estimation of appropriateness of L2 learners' refusal production. This factor was considered in rating refusals in the current study. For instance, in Situation 3, native raters commented on "bother" as a cultural problem; however, nonnatives commented on the lack of politeness. This is in line with Sadler and Eröz's (2001) results that showed that Turkish speakers refused less frequently than speakers of other languages, but if they did, refusals were definitely followed by an excuse or explanation. Moreover, the findings of Al-Issa (2003) indicate that indirect strategies were favoured more by the Jordanians than the Americans. In Honglin's (2007) study of American and Chinese participants, the results revealed that the Americans were more direct than the Chinese in their refusals, but that the Chinese considered refusals as face-threatening acts and used politeness strategies in their refusals. In essence, the Americans tried to solve the problem, while the Chinese tried to restore the relationship between interlocutors. This point is true of Situation 4, in which native raters evaluated a refusal in terms of its directness, whereas non-native raters were more concerned with politeness and preserving the relationship with the interlocutor. However, despite variation in the types of criteria NES and NNES raters employed to measure the appropriateness of refusal in Situation 4, the ratings were largely similar. This indicates that similarity in ratings does not entail the application of the same criteria. Whereas rating scores are a product-oriented measure of difference between NES and NNES raters, the analysis of the criteria leading to rating scores, that is, a process-oriented approach, is necessary for an in-depth understanding of raters' rating behaviour. Similarly, in Youn's (2007) study, the results revealed that each rater showed unique bias patterns, depending on the test type and speech act.

Conclusions and Implications

The study revealed the criteria employed for refusal production ratings by NES and NNES raters. The findings showed that NES raters applied 11 criteria while assessing L2 refusal production. The criteria common across situations in refusal for both NES and NNES raters were *brief apology, irrelevancy of speech act, explanations, postponing to another time, statement of alternative, and politeness.* Although NES and NNES raters gave different weights to politeness, it was among the most frequently employed criteria in both groups. The frequent reference of raters to this criterion is compatible with the general perception of politeness as the main measure of pragmatic appropriateness. Emphasized in pragmatic literature, politeness seems to be the principle overriding the other criteria for pragmatic appropriateness.

The premise that politeness is considered a pragmatic universal, and hence has cross-linguistic and cross-cultural realizations, can contribute to convergence on pragmatic rating. However, in view of the fact that there are variations in the perception of both sociocultural norms and pragmalinguistic realizations of politeness, the application of the politeness criterion to the rating of speech acts showed variability among NES and NNES raters. Mostly, NNES raters mentioned *politeness* as a leading criterion while NES raters highlighted *explanation* for the speech act of refusal. Moreover, mere mention of the criterion of politeness may be misleading because variation arises when it comes to the evaluation of the degree of politeness observed in the production of a speech act.

Generally, NNES raters in this study were more lenient than NES raters, which highlights the need for more pragmatically informed ratings by NNES teachers. This can be achieved through rater training programs in which the focus would be on helping NNES teachers recognize effective criteria for rating pragmatic production and paving the way for increasing accuracy in their ratings. Because rating criteria play a significant role in pragmatic assessment, NNES teachers in EFL contexts such as Iran, where there is insufficient pragmatic awareness of such criteria for speech act production in English, should be encouraged to participate in training programs that aim to raise their pragmatics rating consciousness so that their rating criteria more closely approximate those of NES raters. Such a program may include video clips demonstrating native speakers' production of refusals, as well as less appropriate refusals performed by nonnative speakers, along with the rating scores and rating criteria assigned by native teacher raters. NNES teachers, particularly in an EFL context where there are neither any established local English norms nor any variety of world English to function as a frame of reference, usually apply different rating criteria to assess the same pragmatic production. In fact, raters may have a different understanding of the construct being measured, and such differences may have a direct influence on the ratings they assign to test takers' performance in the testing context. As evident from the results of this study, the scores that NES and NNES raters assigned to students' performance were different, with NNES teachers being more lenient than NES raters. This suggests that the two groups applied different rating criteria to rate the same construct; this, in turn, signifies the need for rating training.

NNES teachers should become conscious of rating criteria through training programs to increase their accuracy in interlanguage pragmatic (ILP) rating as measured against the benchmark. Therefore, rater training should be implemented in teacher education programs to alter the assessment practice of teachers, and decision makers need to take training programs into consideration for EFL raters. Furthermore, the significance of the politeness criterion has implications for the rating of pragmatics. Provided that this criterion features highly in pragmatics rating and consequently affects raters' judgment of pragmatic appropriateness, NNES teachers should have sufficient pragmalinguistic and sociopragmatic competence underpinning their perception of politeness. Although NES raters are comparatively more homogeneous in this regard, pragmatics rating by NNES raters, which is most common in an EFL context, requires a good understanding not only of the pragmalinguistic realization of politeness but also of L2 social norms and conventions, particularly those diverging from L1 politeness norms. As for NES raters, it should be noted that, had they been from the same national background, they would likely have manifested more homogeneous rating behaviour. This should be taken into account in the interpretation of the NES data in this study.

The Authors

Minoo Alemi holds a PhD in Applied Linguistics and is a faculty member of Sharif University of Technology, Iran. She is currently doing her postdoctoral research on Robot-Assisted Language Learning (RALL). Her areas of interest include discourse analysis, interlanguage pragmatics, and materials development.

Zia Tajeddin is associate professor of applied linguistics at Allameh Tabataba'i University, Iran, and the Director of Iranian Interlanguage Pragmatics SIG. His areas of interest include (critical) discourse analysis, interlanguage pragmatics, L2 learner/teacher identity, and sociocultural theory.

References

Al-Issa, A. (2003). Sociocultural transfer in L2 speech behaviors: Evidence and motivating factors. *International Journal of Intercultural Relations, 27*(5), 581–601.

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

Barnwell, D. (1989). “Naive” native speakers and judgments of oral proficiency in Spanish. *Language Testing*, *6*(2), 152–163.

Barrett, S. (2001). The impact of training on rater variability. *International Education Journal, 2*(1), 49–58.

Beebe, L. M., Takahashi, T., & Uliss-Weltz, R. (1990). Pragmatic transfer in ESL refusals. In R. C. Scarcella, E. S. Andersen, & S. D. Krashen (Eds.), *Developing communicative competence in a second language* (pp. 55–73). Cambridge, MA: Newbury House.

Brown, P., & Levinson, S. (1987). *Politeness*. Cambridge, UK: Cambridge University Press.

Chen, X., Ye, L., & Zhang, Y. (1995). Refusing in Chinese. In G. Kasper (Ed.), *Pragmatics of Chinese as native and target language* (pp. 119–163). Honolulu, HI: University of Hawai’i Press.

Cronbach, L. J. (1995). Giving method variance its due. In P. E. Shrout & S. T. Fiske (Eds.), *Personality research, methods, and theory: A Festschrift honoring Donald W. Fiske* (pp. 145–157). Hillsdale, NJ: Lawrence Erlbaum.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly, 2*(3), 197–221.

Engelhard, G., Jr. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement, 31*(2), 93–112.

Engelhard, G., Jr., & Myford, C. M. (2003). *Monitoring faculty consultant performance in the Advanced Placement English Literature and Composition Program with a many-faceted Rasch model* (College Board Research Report No. 2003-1). New York, NY: College Entrance Examination Board.

Eslami, Z. R. (2010). Refusals: How to develop appropriate refusal strategies. In A. Martínez-Flor & E. Usó-Juan (Eds.), *Speech act performance: Theoretical, empirical and methodological issues* (pp. 217–236). Amsterdam: John Benjamins.

Fayer, J. M., & Krasinski, E. (1987). Native and nonnative judgments of intelligibility and irritation. *Language Learning*, *37*(3), 313–326.

Félix-Brasdefer, J. C. (2003). Declining an invitation: A cross-cultural study of pragmatic strategies in American English and Latin American Spanish. *Multilingua: Journal of Cross-Cultural and Interlanguage Communication, 22*(3), 225–255.

Fitzpatrick, A. R., Ercikan, K., Yen, W. M., & Ferrara, S. (1998). The consistency between raters scoring in different test years. *Applied Measurement in Education, 11*(2), 195–208.

Gamaroff, R. (2000). Comment: ESL and linguistic apartheid. *ELT Journal, 54*(3), 297–298.

Honglin, L. (2007). A comparative study of refusal speech acts in Chinese and American English. *Canadian Social Science, 3*(4), 64–67.

Hoyt, W. T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological Methods, 5*(1), 64–86.

Ishihara, N., & Tarone, E. (2009). Subjectivity and pragmatic choice in L2 Japanese: Emulating and resisting pragmatic norms. In N. Taguchi (Ed.), *Pragmatic competence* (pp. 101–128). Berlin, Germany: Mouton de Gruyter.

Kachru, B. B. (1992). World Englishes: Approaches, issues and resources. *Language Teaching, 25*(1), 1–14.

Kachru, B. B. (1997). World Englishes and English-using communities. *Annual Review of Applied Linguistics*, *17*, 66–87.

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.

Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing, 12*(1), 26–43.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing, 19*(1), 3–31.

Levinson, S. C. (1983). *Pragmatics.* Cambridge, UK: Cambridge University Press.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12*(1), 54–71.

Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing, 15*(2), 158–180.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*(4), 386–422.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement, 5*(2), 189–227.

Plough, I. C., Briggs, S. L., & Van Bonn, S. (2010). A multi-method analysis of evaluation criteria used to assess the speaking proficiency of graduate student instructors. *Language Testing, 27*(2), 235–260.

Pomerantz, A. (1984). Agreeing and disagreeing with assessments: Some features of preferred/dispreferred turn shapes. In J. M. Atkinson & J. Heritage (Eds.), *Structures of social action: Studies in conversation analysis* (pp. 57–101). Cambridge, UK: Cambridge University Press.

Sadler, R. W., & Eröz, B. (2001). "I refuse you!" An examination of English refusals by native speakers of English, Lao, and Turkish. *Arizona Working Papers in SLAT, 9*, 53–80.

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing, 25*(4), 465–493.

Shi, L. (2001). Native- and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing. *Language Testing*, *18*(3), 303–325.

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

Takahashi, S. (1996). Pragmatic transferability. *Studies in Second Language Acquisition, 18*(2), 189–223.

Takahashi, T., & Beebe, L. M. (1987). The development of pragmatic competence by Japanese learners of English. *JALT Journal, 8*, 131–155.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263–287.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach.* New York, NY: Palgrave Macmillan.

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing, 10*(3), 305–319.

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

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

Appendix
Refusal Rating

In the following situations, an English language learner was supposed to make refusals. Please read the EFL learner's answer in each situation and rate its appropriateness according to the following rating scale. Then provide your criteria and reasons for the selection of a particular point (1, 2, 3, 4, or 5) on the scale.

1. very unsatisfactory 2. unsatisfactory 3. somehow appropriate
4. appropriate 5. most appropriate

1. You arrive at your office and see that your cleaner is upset. You notice that he has bumped into an antique vase while cleaning the table and has broken the vase. He apologizes to you and wants to pay for it, but you don't accept his apology. What would you say?
Answer: *Hey, accidents happen. Don't worry about that. OK?*
1. very unsatisfactory 2. unsatisfactory 3. somehow appropriate 4. appropriate 5. most appropriate
Criteria:

2. You have just started working in a new company. The first day, you are walking in the hall and see one of your old friends from university. You come to know that he lives next to you, in the same neighborhood. He suggests that everyday you get to work together in your car, but you like to get to work alone, so you refuse his suggestion. What would you say?
Answer: *I'd love to but I can't. You are dear to me. I wish I could.*
1. very unsatisfactory 2. unsatisfactory 3. somehow appropriate 4. appropriate 5. most appropriate
Criteria:

3. You meet one of your professors in the hall at university. You like him very much and you think he is the best professor at that university. You go and greet him. He is very happy to see you and invites you to lunch at the university cafeteria. Unfortunately, you have promised your friends you would visit them for lunch, so you can't accept his invitation. What would you say?
Answer: *I'd love to sir, but I've promised some of my friends to meet them for lunch. I hope you don't mind. Can I bother you some other time? I really don't want to pass such a great offer.*
1. very unsatisfactory 2. unsatisfactory 3. somehow appropriate 4. appropriate 5. most appropriate
Criteria:

4. You are a junior in college. You attend classes regularly and take good notes. Your classmate often misses class, and asks you for the lecture notes. However, you need the notes yourself and can't lend them to her, so you refuse her request. What would you say?
Answer: *Actually I need the notes myself. Why don't you try to attend the classes regularly?*
1. very unsatisfactory 2. unsatisfactory 3. somehow appropriate 4. appropriate 5. most appropriate
Criteria:

5. You are going to refuse an invitation offered to you by your colleague to an art gallery. What would you say?
   Answer: *Sorry, but I can't come. I've some things I should take care of, you know.*
   1. very unsatisfactory 2. unsatisfactory 3. somehow appropriate 4. appropriate 5. most appropriate
   Criteria:

6. You are trying to reject an invitation offered to you by your older sister to her house for a dinner party. You are so busy and can't go there. How would you decline her invitation?
   Answer: *I'm so busy. Excuse me.*
   1. very unsatisfactory 2. unsatisfactory 3. somehow appropriate 4. appropriate 5. most appropriate
   Criteria: