Tabaran Institute of Higher Education
Iranian Journal of Language Testing
Vol. 4, No. 1, March 2014

ISSN 2251-7324

Pragmatic Rater Training: Does It Affect Non-native L2 Teachers' Rating Accuracy and Bias?

Zia Tajeddin[1], *Minoo Alemi*[2]

Abstract

Pragmatics assessment literature provides little evidence of research on rater consistency and bias. To address this underexplored topic, this study aimed to investigate whether a training program focused on pragmatic rating would have a beneficial effect on the accuracy of non-native English speaker (NNES) ratings of refusal production as measured against native English speaker (NES) ratings and whether NNES rating bias diminishes after training. To this end, 50 NNES teachers rated EFL learners' responses to a 6-item written discourse completion task (WDCT) for the speech act of refusal before and after attending a rating workshop. The same WDCT was rated by 50 NES teachers who functioned as a benchmark. Pre-workshop non-native ratings as measured against the native benchmark in terms of mean, SD, mean difference, and native/non-native correlation revealed that non-native raters tended to be more lenient and greatly divergent in rating total DCT and across items. Subsequent to training, however, non-native rating produced more accurate and consistent scores, indicating its approximation toward the native benchmark. To measure rater bias, a FACETS analysis was run. FACETS results showed that both before and after training, many of the raters were outliers. Besides, after training, a few raters became biased in rating certain items. From these findings, it can be concluded that pragmatic rater training can positively influence non-native ratings by getting them closer to those of natives and making them more consistent, but not necessarily less biased.

***Keywords**: pragmatic rater training, refusal, native English speaker teacher, non-native English speaker teacher, bias, FACETS*

1. Introduction

In rating ILP assessment tasks, a certain number of variables have an impact on assessment, among which rater variables stand out. However, studies focused on the rater assessment of interlanguage pragmatic competence have been rare. Therefore, there is a need to investigate not only raters' overall judgments and the possible difference between the rating of native and non-native raters but also rating consistency across raters. Moreover, the literature documents no

[1] Allameh Tabataba'i University, Iran. E-mail: zia_tajeddin@yahoo.com
[2] Sharif University of Technology, Iran. E-mail: minooalemi2000@yahoo.com

evidence if a rater training workshop has any impact on improving non-native L2 teachers' accuracy and bias in pragmatic rating.

Out of the speech acts to be rated by non-native raters, refusal was selected in this study due to three main reasons. First, the appropriate use of this speech act is truly vital in the process of communication because of its face-threatening nature and its frequent occurrence in the daily life. Furthermore, since the speech act of refusal differs in different cultures and under different communicative situations, non-native teachers should get familiar with native criteria in rating non-native refusal production. Finally, among speech acts, refusal is a complicated one primarily because it often involves lengthy negotiations and face-saving strategies to accommodate the noncompliant nature of the speech act. As refusal normally functions as a response to an interlocutor’s request, suggestion, or invitation, it precludes extensive planning on the part of the refuser (Gass & Houck, 1999).

Against the above backdrop, the researchers in this study aimed to explore non-native English speaker (NNES) teachers' pragmatic rating accuracy and bias before and after a workshop in their rating of L2 refusal production as measured against native English speaker (NES) teachers' rating baseline. To regard native speaker norms as the baseline is not totally incompatible with the notions developed within English as an international language (EIL). Learners of English as a foreign language are expected to conform to Inner Circle norms. As Seidlhofer and Jenkins (2003) rightly put it, for Expanding Circle consumption, the main effort remains to describe English as it is used among the British and American native speakers. In line with this argument, Kirkpatrick (2006) argues that a lingua franca model is the most sensible model in those common and varied contexts where the learners' major reason for studying English is to communicate with other non-native speakers; however, until we are able to provide teachers and learners with adequate descriptions of lingua franca models, teachers and learners will have to continue to rely on either native-speaker or nativized norms. The third argument comes from Kachru (1992a), who believes that while the Outer Circle is “norm-developing,” the Expanding Circle (which includes most countries like Iran, Egypt, Korea, and Denmark), is “norm-dependent” because it relies on the standards set by native speakers in the Inner Circle.

While accepting Kachru’s (1992b) three-part categorization of English use in Inner, Outer, and Expanding circles, the researchers in this study, drawing on the three arguments mentioned above, believe that in an EFL context like Iran, no local variety of English with established or emerging norms exist. This is unlike the situation in Outer-Circle countries, such as India, where local norms for accuracy and appropriateness have emerged. In essence, as LoCastro (2012) points out, there is no transparent answer to the question of whose norms need to be taught and learned by non-native speakers of English. Drawing on this controversy, the researchers think native speaker norms should be the point of departure in an EFL context like Iran where a lack of pragmatic training has resulted in teachers’ impoverished pragmatic understanding, poor perception of appropriateness in terms of pragmalinguistic and sociopragmatic features of speech act production, and the overuse of L1-based criteria for appropriateness which disregard formulaic routines, speech act strategies, and social norms largely specific to the native or even EIL-driven model of English. In consequence, the underlying assumption in this study is that the native perceptions of appropriateness in English production can be the most reliable frame of reference in the Iranian context where non-native EFL teachers’ knowledge bases in the pragmatics of L2 English have not been developed

(Tajeddin & Mohammad Bagheri, 2012) and where no locally shaped norms for English use are at work.

2. Review of Literature

2.1. Raters and Rating Variability

As raters, our judgments about language performance are affected by our own perceptual presuppositions which may vary in terms of rater background characteristics such as being trained (e.g. Hsieh, 2011) or being a native/non-native speaker (e.g. Wen, Liu, & Jin, 2005). The effects of rater perceptions introduce highly subjective factors that make many ratings more or less inaccurate. Rater bias is a major problem when language raters rate learners using scales that are vague or highly subjective; hence, if they use such rating scales, 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) argue that raters' judgments are prone to various sources of bias and error which can ultimately change the quality of the ratings. A number of studies using a range of psychometric methods have identified various rater effects (Myford & Wolfe, 2003, 2004) which need to be addressed if an acceptable level of reliability is to be maintained. The different rater effects can be summarized as (1) the severity effect, (2) the halo effect, (3) the central tendency effect, (4) inconsistency, and (5) the bias effect. Two of the rater effects which are related to the main themes of this study are the severity effect and the bias effect. The former occurs where raters are found to follow "a systematic pattern of rater behavior that manifests itself in unusually severe (or lenient) ratings," (Eckes, 2012, p. 273), as compared with other raters or established benchmark ratings. The latter, bias effect, is exhibited when raters tend to rate unusually harshly or leniently with regard to one aspect of the rating situation. For example, they might favor a certain group of test takers or they might always rate one category of the rating scale too harshly or leniently. The variability of ratings as a result of these two effects has been addressed in many studies on speaking and writing (e.g. Caban, 2003; Eckes, 2005; Johnson & Lim, 2009; Kim, 2009; Schaefer, 2008; Wigglesworth, 1993). Nevertheless, a small number of findings on rater variability are related to ILP rating (Liu & Xie, 2014, in this issue; Taguchi, 2011; Youn, 2007).

One source of rater variability is the status of the rater as being a native or non-native speaker. It is very important to determine whether native speaker (NS) and non-native speaker (NNS) raters use the same or different criteria for rating tasks. Studies comparing NS and NNS ratings of oral and written language performance vary in their results. As Barnwell (1989) points out, NES are harsher in their evaluations than NNES, whereas others found that the opposite is true and that NNES raters are more severe. For instance, Fayer and Krasinski (1987) investigated Puerto Rican learners of English speech act production and gave their samples to two groups of raters: NES and Puerto Rican speakers. Results revealed that NNES were stricter especially on pronunciation errors than native speakers of English. While the literature is replete with references to native and non-native speakers' ratings of listening and speaking performance, there is hardly any mention or comparison of native and non-native ratings of pragmatic performance.

2.2. Rater Training

Rater variability in language performance assessment is a serious problem, so rater training courses can be run to increase within-rater consistency. Rater training can reduce the variability of raters' behavior. In dealing with how to improve accurate rating among teachers, workshops have been suggested. Raters participating in workshops are introduced to assessment criteria and asked to rate a series of selected performances. In fact, training is intended to minimize the differences as a result of rater variability and to maximize the consistency among raters who are expected to focus on the appropriate criteria and to adjust their expectation in accordance with task requirements and learners' abilities (Weigle, 1994a). According to Nation and Macalister (2010), the goals of a training session are experiencing and evaluating exercises, producing materials or exercises, planning units of work, and above all solving problems. According to McIntyre (1993), training can attenuate extreme differences between raters in terms of severity, enhance the consistency of raters by decreasing random error, and counteract individual biases in relation to the various aspects of the rating situation such as the rating scale and candidates.

There is evidence that rater training can be effective by eliminating extreme differences in rater severity, increasing the self-consistency of raters, and reducing individual biases displayed by raters toward the aspects of the rating situation (McIntyre, 1993; Sugita, 2011; Weigle, 1994a, 1994b, 1998). However, the value of the effect of rater training has been questioned by a few researchers (e.g. Lumley, 2002, 2005). Barritt, Stock, and Clark (1986) and Huot (1990) believe that rater training induces raters to reach agreement but may cause them to ignore their experiences in rating which threaten the validity of their judgment. By contrast, Weigle (1994b) argues that reaching agreement does not jeopardize the validity of raters' ratings and is not necessarily raters' overriding concern. Although there are contradictory findings, most studies on rater behavior indicate that differences in raters' harshness persist after the training session (Cason & Cason, 1984). To remedy the problem, researchers such as Lumley and McNamara (1995) suggest that regular training sessions be held before administrating a large-scale test like ILETS to permit the raters to re-organize a set of criteria for their ratings. Lumley and McNamara studied the effect of rater characteristics and rater bias in terms of rater training. The implication of this study has been defined in terms of the multi-faceted Rasch measurement in understanding rater behavior and variability in the performance assessment context. Despite these inconclusive findings, rater training seems to have a crucial role in increasing systematicity of rater behavior, and the main goal of training is not to force raters into agreement with one another but to make them more self-consistent.

As regards language performance, a number of studies have been conducted on training effect and its persistence over time. Several effects of training on raters to assess the writing ability of learners have been explored (e.g. Weigle, 1994a; Wigglesworth, 1993). Findings show that those attending the training program become able to adjust their evaluations in accordance with the task requirements and learners' abilities (Weigle, 1994a). Congdon and McQueen (2000) report fluctuations in rater severity between pre- and post-rater training sessions. They recommend the need for dynamic training during the rating period in large-scale and high-stakes tests. Some studies have investigated how long rater training effects last. For instance, Lunz and Stahl (1990) found inconsistencies among participants even a day after the training period, which shows that the training effect did not last long. Lumley and McNamara (1995) also claimed lack of stability in rater behavior on the writing test one month after the training session. Barnwell (1989) compared untrained raters vis-à-vis trained ACTFL raters, and found that native speakers were harsher in their evaluation than were non-native speakers. In Fayer and Krasinski’s (1987) study, non-native raters were, however, harsher than native ones. Non-native raters in Shi’s (2001) study gave more negative comments on learners’ writing while native raters made significantly more positive comments.

Raters might indeed focus on different aspects of language performance, but which raters give ratings closer to the true score? Brown (1995) provided NES and high-proficient NNES language teachers with one day of training in the rating of an oral language test. He used the multiple-facet extension of the Rasch model and found that NNES raters were harsher than they should be with regard to politeness and pronunciation and that NNES raters’ scoring was more likely to overfit; that is, there was insufficient variability in the ratings they assigned. By contrast, NESs were more diverse in their use of rating scales and in their relative severity. These findings show that the result of NNES rating is closer to the rating scale whereas NESs take a more intuitive approach to rating. Elder, Barkhuizen, Knoch, and von Randow (2007) investigated rater responses to an online training program for L2 writing assessment. Their findings revealed limited overall gains in reliability and considerable individual variation in receptiveness to the training input. Finally, the findings in Knoch’s (2007) study indicate that, in terms of severity, training was successful in bringing the raters closer together in their ratings.

In sum, in the context of EFL, which is marked by a shortage of native English-speaking teachers, it is very common for non-native English-speaking teachers to participate in teaching and assessing pragmatic production. Moreover, there is great concern about the compatibility of the two groups' rating and their bias. Therefore, it is highly important to study non-native English teachers’ consistency and accuracy in pragmatic rating.

3. Purpose of the Study

The main aims of this study were to investigate whether a training program focused on pragmatic rating would have a beneficial effect on the accuracy of non-native English speaker (NNES) ratings of refusal production as measured against native English speaker (NES) baseline and whether NNES rating divergence and bias, i.e. their severity and leniency in scoring, would diminish after training. To achieve these aims, the following questions were raised:

a. Does pragmatic rater training have a positive effect on the non-native English teachers’ rating of refusal production as measured against the native baseline?

b. Does pragmatic rater training have a positive effect on non-native English teachers’ bias in rating refusal production?

4. Method

4.1. Participants

Participants were composed of 100 native and non-native teachers. Fifty of them were educated native teachers of English from the U.S, the U.K, Canada, and Australia. Their ratings acted as a baseline against which the accuracy of non-native raters’ rating was measured. The native English teachers were faculty members of different language centers (ESL teachers) in international universities. The other group consisted of 50 non-native English-speaking teachers from different language centers in the EFL context and with different teaching experiences. To have a homogenous group for the treatment part, non-native raters who held an M.A. degree in

Teaching English as a Foreign Language (TEFL) were contacted. They accepted to participate in the pragmatic rater training workshop to get familiar with pragmatic rating, including the rating of L2 English refusals. As many as 15 of them were male and 35 were female. Their teaching experience ranged from 5 years to 15 years, with a mean of 7.

4.2. Instrumentation

A written discourse completion test (WDCT) was used to collect the data in this study. A WDCT is a common measure used to assess L2 learners' pragmatic production (LoCastro, 2000; Taguchi, 2011). The WDCT in this study was made up of six refusal situations characterized by different degrees of formality, power, and distance between interlocutors in each situation. 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. Out of the responses, one was selected for each situation to ensure that the responses to the six situations would vary in their degrees of appropriateness. Every response was followed by a rating scale ranging from 1 to 5: 1=very unsatisfactory, 2=unsatisfactory, 3=somehow appropriate, 4=appropriate, and 5=most appropriate.

4.3. Pragmatic Rater Training Workshop

A six-hour workshop was designed for 50 non-native raters to teach them the appropriate criteria in rating speech acts and to make them more familiar with common patterns natives use for rating the appropriateness of WDCTs. The workshop was held one month after collecting pre-training NNES raters' data. The researchers themselves designed and presented this one-day workshop. In the morning session, the construct of pragmatic competence, pragmatic rating scales, and speech act production strategies, particularly those related to refusal, were taught. In the afternoon session, samples of NES and NNES rating criteria and the application of pragmatic rating were discussed. In this way, the inconsistency between natives' and non-natives' ratings was aimed to be reduced. The workshop was interactive, with the participants discussing the criteria they applied and the scores they assigned in rating refusals.

4.4. Data Collection and Analysis

The refusal WDCT was administered to 20 EFL students studying for a B.A. degree in English literature or translation. The responses to each situation were reviewed by the researchers to select one response to each situation in a way that the responses to the six situations would ultimately vary in the degree of appropriateness and that the raters would be given a chance of selecting different points on the Likert scale in rating responses across situations.

After choosing one response to each situation, the WDCT accompanied by respective responses was sent electronically to both native and non-native English teachers to rate the appropriacy of responses on a five-point Likert scale. For non-native teachers, the WDCT was emailed to teachers from various language centers. Out of over 100 ratings received, the 50 related to the teachers attending the training workshop were included in the final data analysis. As to native speakers, the WDCT was uploaded in the SurveyMonkey site, and native ESL teachers and university professors from different universities in the US, the UK, Canada, and Australia were asked via email to rate the WDCT on that site. Out of many teachers contacted, 50 native teachers completed rating sheets.

Data analysis was both descriptive and inferential by nature. The descriptive phase included the calculation of mean and standard deviation of the rating scores for the total WDCT and each situation thereof by native and non-native teachers. In addition, the inter-rater reliability between native and non-native ratings was calculated through intraclass correlation. This type of correlation is appropriate when there are multiple raters. In the inferential phase of the analysis, rater bias was measured through FACETS (Linacre & Wright, 1996). FACETS is a computer application which uses the many-facet Rasch model to measure consistency or bias in rating patterns. It follows from this that FACETS accepts differences in rating severity rather than expecting identical ratings by raters (Linacre, 1989). T-test was used to measure the differences between native and non-native ratings of refusal production.

5. Results

5.1. Refusal Rating Accuracy and Consistency

Non-native teachers' pragmatic rating accuracy was measured against native ratings. Descriptive statistics on refusal for non-native raters before and after the workshop and for native speakers are displayed in Table 1. As the table shows, the mean (M) rating of the 50 native raters acting as the benchmark for the total WDCT was 2.59. It means that their overall evaluation of refusal in the six situations fell at the "somehow appropriate" point on the scale. The mean value of 3.29 for non-native raters before the workshop means "appropriate," and that of 2.67 after the workshop can be regarded as "unsatisfactory." As the mean scores clearly show, non-native raters improved in rating the speech act of refusal after the workshop in that their rating became largely native-like, indicating that the training program affected their rating accuracy. The means across situations manifested the same trend. In all situations, except No. 4, non-native ratings got closer to the native benchmark. Besides, the ratings they assigned to WDCT responses declined which, along with the approximation of their ratings to the native benchmark, indicates they became less lenient and more accurate in their ratings.

Table 1. Descriptive statistics for refusal rating by native raters (N) and non-native raters before (NN pre) and after (NN post) the workshop

Situation	Group	No	Mean	Std. Deviation
Refusal 1	**N**	50	2.08	1.10
	NN pre	50	3.18	1.47
	NN post	50	2.06	.98
Refusal 2	**N**	50	2.58	.83

	NN pre	50	3.50	1.15
	NN post	50	3.00	.73
Refusal 3	**N**	50	3.32	1.02
	NN pre	50	4.02	1.13
	NN post	50	3.02	.94
Refusal 4	**N**	50	2.94	.93
	NN pre	50	3.00	1.05
	NN post	50	2.80	.95
Refusal 5	**N**	50	2.56	.79
	NN pre	50	3.50	1.01
	NN post	50	2.92	.66
Refusal 6	**N**	50	2.06	.77
	NN pre	50	2.56	1.16
	NN post	50	2.20	.83
Total Refusal	**N**	50	2.59	.83
	NN pre	50	3.29	1.29
	NN post	50	2.67	.84

The SD values presented in Table 1 are also revealing. The total SD for non-native raters was 1.29 before the workshop but dropped to .84 after the workshop. This shows more consistent rating in terms of SD. The observation of non-natives raters' SDs across situations substantiates the same improvement in consistency and native-likeness of ratings. The least variation based on SD occurred in situation 5 (SD=.66). More importantly, in situations 1, 2, 3, and 5, non-native ratings became more consistent and less variant than the native benchmark. This further shows the great impact of pragmatic rating training on SD-based rating variation. Generally, the variation was very high in all refusals among non-native raters before the workshop. For instance, the mean was 2.08 for natives and 3.18 for non-natives before the workshop and 2.06 after the workshop. As to SD, it was as high as 1.47 for non-natives before the workshop; however, it decreased to .98 after the workshop to get close to natives' SD of 1.10.

The second measure of non-native raters' rating accuracy was intraclass correlation. It was used to measure the inter-rater reliability between non-native and native ratings. The inter-rater reliability of refusal ratings between non-native raters before the workshop and native raters is shown in Table 2. From the table, it is evident that the ratings were not highly correlated with each other before the workshop (r=.30).

Table 2. Intraclass correlation coefficient between non-native raters before the workshop and native raters for refusal

	Intraclass Correlation	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.036	-.001	.096	1.442	49	539	.080
Average Measures	.306	-.015	.560	1.442	49	539	.080

However, there was improvement in inter-rater reliability after the workshop (see Table 3). While the inter-rater reliability index were .30 and non-significant at *p*>.05 before the workshop, it increased to .49 and became significant at *p*<.01 after the workshop. Therefore, there was a significant, albeit not high, correlation between native and non-native ratings of refusal after the workshop.

Table 3. Intraclass correlation coefficient between non-native raters after the workshop and native raters for refusal

	Intraclass Correlation	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.121	.061	.202	1.87	99	495	.000
Average Measures	.491	.491	.613	1.87	99	495	.000

Another measure of non-native rating accuracy was *t*-test. An independent-samples *t*-test was conducted to compare means of NES and NNES ratings before and after training to reveal the accuracy of NNES rating against the native benchmark. The results of the *t*-test, as reported in Table 4, display that there was a significant difference in total refusal WDCT between NES and NNES ratings before training (*t*(98) = 7.21, *p*<.01). Similar differences were found across situations. Except for situation 4, the non-native ratings diverged significantly from the native benchmark across five situations.

Table 4. T-test for refusal ratings by non-native raters before the workshop and native raters

	Levene's Test for Equality of Variances		t-test for Equality of Means					
	F	Sig.	t	df	Sig. (2-tailed)	Mean Diff.	Std. Error Diff.	95% Confidence Interval of the Diff.

								Lower	**Upper**
Refusal 1	12.49	.001	4.23	98	.000**	1.10	.26	.58	1.61
Refusal 2	8.11	.005	4.58	98	.000**	.92	.20	.52	1.31
Refusal 3	.28	.597	3.24	98	.002**	.70	.21	.27	1.12
Refusal 4	1.31	.254	.302	98	.763	.06	.19	-.33	.45
Refusal 5	3.40	.068	5.17	98	.000**	.94	.18	.58	1.30
Refusal 6	15.55	.000	2.53	98	.013*	.50	.19	.10	.89
Total Refusal	.01	.910	7.21	98	.000**	.70	.09	.50	.89

Note: * significant at *p*<.05; ** significant at *p*<.01

To delve into the differences between the ratings of native raters and non-native raters after the workshop, another independent-samples *t*-test was conducted. The results from Table 5 reveal that there was not any significant difference between the two groups in total ratings after training (*t*(98)=.61, df=98, *p*>.05). There was also consistency in terms of situations. While non-native raters' ratings before the workshop significantly differed from native ratings in five situations, rating accuracy improved to include only two situations documenting divergence between native and non-native ratings (situations 2 and 5). The comparison of pre-training and post-training results documents the positive impact of the rating workshop on non-native teachers' accuracy in total rating and ratings across refusal situations.

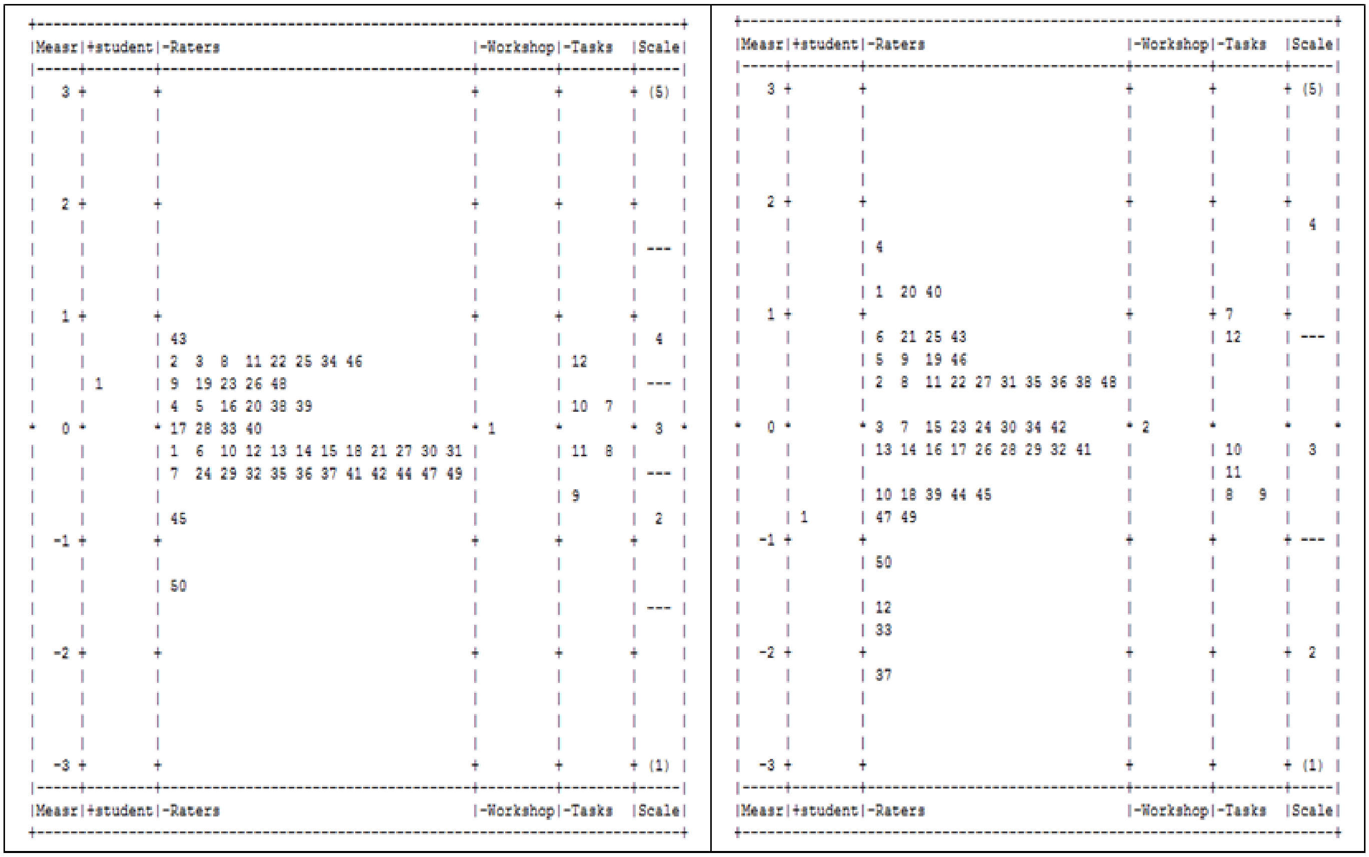
Table 5. T-test of refusal ratings by non-native raters after the workshop and native raters

	Levene's Test for Equality of Variances		**t-test for Equality of Means**						
	F	**Sig.**	**t**	**df**	**Sig. (2-tailed)**	**Mean Diff.**	**Std. Error Diff.**	**95% Confidence Interval of the Diff.**	
								Lower	**Upper**
Refusal 1	.06	.803	-.09	98	.924	-.02	.20	-.43	.39
Refusal 2	8.60	.004	2.68	98	.009**	.42	.15	.10	.73
Refusal 3	1.15	.285	-1.53	98	.129	-.30	.19	-.68	.08
Refusal 4	1.14	.286	-.74	98	.459	-.14	.18	-.51	.23
Refusal 5	6.95	.010	2.47	98	.015*	.36	.14	.07	.64
Refusal 6	3.35	.070	.87	98	.384	.14	.16	-.17	.45
Total Refusal	.75	.070	.61	98	.824	.08			

Note: * significant at *p*<.05; ** significant at *p*<.01

5.2. Rater Bias in Pragmatic Rating of Refusal

The second aim of this study was to employ FACETS to measure the effect of pragmatic rating training on non-native raters' rating bias. The results of FACETS analysis show that the raters, except for rater # 50, were bunched together, rating the refusal production in a rather similar, non-biased way (Figure 1). However, the training program resulted in the raters’ falling within a wider scope in terms of leniency and severity. Whereas before-training ratings predominantly ranged between -1 and +1, post-training ratings manifested more dispersion. In the latter, four raters (# 1, 20, 40, 4) were found to be rather biased toward strict rating, and four raters (# 50, 12, 33, 37) tended to be lenient.



```
+------------------------------------------------------------------------------+
|Measr|+student|-Raters                              |-Workshop|-Tasks  |Scale|
|-----+---------+-------------------------------------+---------+--------+-----|
|   3 +         +                                     +         +        + (5) |
|     |         |                                     |         |        |     |
|     |         |                                     |         |        |     |
|     |         |                                     |         |        |     |
|     |         |                                     |         |        |     |
|   2 +         +                                     +         +        +     |
|     |         |                                     |         |        |     |
|     |         |                                     |         |        | --- |
|     |         |                                     |         |        |     |
|     |         |                                     |         |        |     |
|   1 +         +                                     +         +        +     |
|     |         | 43                                  |         |        |  4  |
|     |         | 2  3  8  11 22 25 34 46             |         | 12     |     |
|     | 1       | 9  19 23 26 48                      |         |        | --- |
|     |         | 4  5  16 20 38 39                   |         | 10   7 |     |
*   0 *         * 17 28 33 40                         * 1       *        *  3  *
|     |         | 1  6  10 12 13 14 15 18 21 27 30 31 |         | 11   8 |     |
|     |         | 7  24 29 32 35 36 37 41 42 44 47 49 |         |        | --- |
|     |         |                                     |         | 9      |     |
|     |         | 45                                  |         |        |  2  |
|  -1 +         +                                     +         +        +     |
|     |         |                                     |         |        |     |
|     |         | 50                                  |         |        |     |
|     |         |                                     |         |        | --- |
|     |         |                                     |         |        |     |
|  -2 +         +                                     +         +        +     |
|     |         |                                     |         |        |     |
|     |         |                                     |         |        |     |
|     |         |                                     |         |        |     |
|     |         |                                     |         |        |     |
|  -3 +         +                                     +         +        + (1) |
|-----+---------+-------------------------------------+---------+--------+-----|
|Measr|+student|-Raters                              |-Workshop|-Tasks  |Scale|
+------------------------------------------------------------------------------+
```

```
+------------------------------------------------------------------------+
|Measr|+student|-Raters                        |-Workshop|-Tasks  |Scale|
|-----+---------+-------------------------------+---------+--------+-----|
|   3 +         +                               +         +        + (5) |
|     |         |                               |         |        |     |
|     |         |                               |         |        |     |
|     |         |                               |         |        |     |
|     |         |                               |         |        |     |
|   2 +         +                               +         +        +     |
|     |         |                               |         |        |  4  |
|     |         | 4                             |         |        |     |
|     |         |                               |         |        |     |
|     |         | 1  20 40                      |         |        |     |
|   1 +         +                               +         + 7      +     |
|     |         | 6  21 25 43                   |         | 12     | --- |
|     |         | 5  9  19 46                   |         |        |     |
|     |         | 2  8  11 22 27 31 35 36 38 48 |         |        |     |
|     |         |                               |         |        |     |
*   0 *         * 3  7  15 23 24 30 34 42       * 2       *        *     *
|     |         | 13 14 16 17 26 28 29 32 41    |         | 10     |  3  |
|     |         |                               |         | 11     |     |
|     |         | 10 18 39 44 45                |         | 8    9 |     |
|     | 1       | 47 49                         |         |        |     |
|  -1 +         +                               +         +        + --- |
|     |         | 50                            |         |        |     |
|     |         |                               |         |        |     |
|     |         | 12                            |         |        |     |
|     |         | 33                            |         |        |     |
|  -2 +         +                               +         +        +  2  |
|     |         | 37                            |         |        |     |
|     |         |                               |         |        |     |
|     |         |                               |         |        |     |
|     |         |                               |         |        |     |
|  -3 +         +                               +         +        + (1) |
|-----+---------+-------------------------------+---------+--------+-----|
|Measr|+student|-Raters                        |-Workshop|-Tasks  |Scale|
+------------------------------------------------------------------------+
```

Figure 1. Variable map for refusal before and after workshop

The greater bias in ratings after the rater training is also evident from the separation index presented in Table 6. While the index was .82 in pre-training rating, it increased to 1.60 after training. As a result, the workshop resulted in a great variability in rating.

Table 6. Rater measurement report for refusal before and after the workshop: Severity statistics

	Pre-workshop	Post-workshop

Separation index	.82	1.60
Reliability	.12	.47
Fixed chi-square	48.9; df= 49; *p*=.48	86.6; df= 49; *p*=.00

Note: chi-square test of homogeneity was not statistically significant before the workshop, but it was statistically significant after the workshop.

Bias analysis for the interaction between raters and WDCT situations/tasks did not show any bias before the workshop. However, as Figure 2 reveals, seven raters manifested different degrees of severity and leniency after the workshop because the measurement index went beyond the normal range -2 to +2. One rater was found to be too strict whereas six raters showed great leniency.

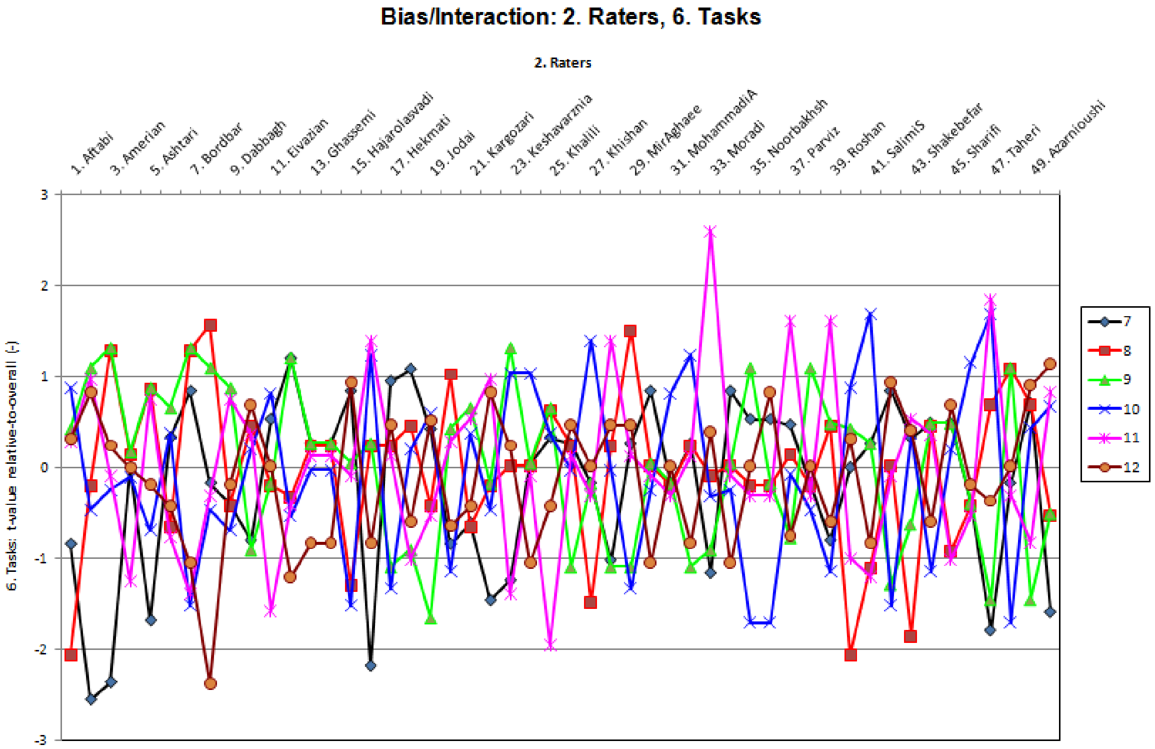


Figure 2. Bias analysis for rater-task interaction after the workshop

To locate the refusal tasks toward which these seven raters were biased, bias sizes were calculated for the six refusal tasks. The results (Table 7) revealed that there was bias primarily toward four tasks (# 1, 2, 5, 6), with more raters biased in rating task 2. Among the four tasks, tasks 1, 2, and 6 showed leniency-based interaction with raters 1, 2, 3, 8, 16, and 40. The greatest leniency-based bias interaction was observed in the case of rater 2 and task 1 (bias size=-4.01). By contrast, there was one strictness-based bias interaction, which occurred between rater 33 and task 5 (bias size=3.50).

Table 7. Bias analysis of rater-task interaction after the workshop

Rater	Tasks	Obs-exp	Bias	Model	t. score	Infit	Outfit

		average	Size	error		MnSq	MnSq
1	2	1.69	-3.25	1.58	-2.06	.0	.0
2	1	2.12	-4.01	1.58	-2.55	.0	.0
3	1	1.97	-3.72	1.58	-2.36	.0	.0
8	6	1.98	-3.75	1.58	-2.38	.0	.0
16	1	1.80	-3.43	1.58	-2.18	.0	.0
33	5	-1.88	3.50	1.35	2.60	.0	.0
40	2	1.69	-3.25	1.58	-2.06	.0	.0

Fixed (all = 0) chi-square: 228.2; df= 300; *p*= 1.00

6. Discussion

One of the main purposes of this study was to explore the differences between NSs and NNSs in pragmatic rating and the pragmatic accuracy of non-native raters as measured against that of the native raters' benchmark. The results of the present study revealed that there was a great dispersion among non-native raters before the workshop; however, this variation decreased after the workshop. In fact, the training session brought about more consistency among non-native raters as it can be confirmed by the amount of standard deviation and non-native teachers' rating means which approached those of native teachers. This decrease in rater variability subsequent to training is in line with findings reported by researchers like Wigglesworth (1993). However, it runs counter to a number of studies which showed the persistence of rater variability despite rater training (Engelhard, 1992, 1994; Lumley, 2002, 2005).

With regard to rating means, the disagreement between native and non-native ratings was significant before training. The dispersion of ratings was also high among non-native ratings before training. Like Knoch's (2007) findings which indicated that, in terms of severity, training was successful in bringing the raters closer together in their ratings, this study reveals more consensus among non-natives after training. Training was successful in the reduction of dispersion and the enhancement of inter-rater agreement with natives. The notion of compliance with the benchmark finally suggests that non-native raters can become aware of the criteria for their ratings through pragmatic rating workshops and accordingly modify their scoring behavior. Nevertheless, the persistence of divergence between NS/NNS ratings in two DCT situations, i.e. #2 and #5, after training shows that some L1 cultural norms and perceptions of appropriateness related to certain refusal contexts cannot be easily reshaped. Perceptions of appropriateness for certain refusal situations may vary greatly from one language to another. It follows that some L1 norms need more pragmatic awareness on the part of NNSs to undergo significant changes. Although NNS ratings for these two situations got closer to those of NSs, the change fell short of statistical significance.

As to rater training in this study, which refers to the workshop in which raters were exposed to assessment criteria and then had to rate a number of DCT samples based on the criteria they became conscious of, it served to reduce the amount of rater variability. In terms of variations in the pragmatic rating of non-native raters before and after the workshop, the results of *t*-tests demonstrated that non-native ratings approached the benchmark except in the case of those items which were in formal situations. For example, concerning situations 2 and 5, which were formal situations, it can be argued in view of the findings that non-native raters were not comfortable with the direct refusal strategy in formal situations and hence considered the responses as rather inappropriate despite attending the rating workshop. It follows that the effect of pragmatic training is limited in some respects.

The results of the current study also indicate that non-native raters displayed the patterns of overrating/leniency before attending the rater training program as measured against the benchmark (NNES refusal Mean=3.29 vs. NES refusal Mean=2.59). However, after attending the program they tended to do ratings that were more similar in accuracy to native ratings (NNES refusal Mean=2.67 vs. NES refusal Mean=2.59). With regard to rating differences and rater severity, this finding is in line with the study by Shi (2001), which revealed strict native speakers and lenient EFL raters in writing assessment. It also lends further support to the study by Wen, Liu, and Jin (2005), which showed significant variation between native and non-native judges in assessing speaking. By contrast, this study is incompatible with Brown (1995), who found that NNES raters were harsher than NES raters in scoring pronunciation. Furthermore, similar to the current study, Lumley and McNamara (1995) confirmed the effect of rater training and claimed that it permitted the raters to re-organize a set of criteria for their ratings. However, in the interpretation of the compatibility of the results of this study with other findings, strong generalizations should not be made because other findings are focused on the assessment of linguistic skills such as speaking and wring rather than pragmatic production.

Generally, non-native raters in this study improved mostly in rating all refusal tasks. It shows that the workshop influenced the participants and their ratings got closer to those of natives. In effect, training minimized the differences with respect to rater variability and maximized the consistency among raters who were expected to focus on the selected rating criteria. This lends support to Knoch's (2007) study, which revealed that in terms of severity, training was successful in bringing the raters closer together in their ratings. Despite the effect of training on the closer correspondence between NS/NNS ratings, we need to be cautious about the function of NS rating as a benchmark. As the NS participants in this study were from various English-speaking countries, there might be certain degree of variation in their sociocultural norms, causing, in turn, variation in the benchmark rating per se.

Training sessions proved to be necessary for non-native raters in order to make them informed of the ILP rating benchmark to enhance their rating consistency. However, greater bias displayed by non-native teachers after training can be explained in four ways. The first possible reason is that due to their familiarity with various criteria for pragmatic appropriacy after receiving training, non-native raters became more analytical in their rating behavior. The move away from the more holistic to the more analytic rating can, therefore, be the effect of the application of various, albeit not necessarily similar criteria, in ratings and hence cause an increase in rater leniency/severity. The second reason is related to the nature of rating criteria in pragmatics. Largely different from the criteria for rating speaking and writing performance, the criteria for pragmatic appropriateness are largely sociopragmatic in nature and hence rooted in raters' long-established beliefs in social and cultural conventions. It follows that pragmatic training cannot easily change the perception of appropriateness. The third explanation for after-training rater bias is that familiarity with rating criteria is not equal to their application by raters. Despite training, as Haizhen's (2008) study shows, different raters may interpret and apply rating criteria differently. It seems that some of the raters in this study overused certain rating criteria with the effect of becoming too strict, e.g. rater 33, or underused them to practice rating different from the majority of the raters. The final reason is based on the observation in the literature that training does not bring about similar effects on participants. Variation in the degree of bias found in this study substantiates this observation.

7. Conclusion and Implications

Many studies have shed light on rater bias and criteria for assessing the performance of language skills (e.g. Eckes, 2005; Gamaroff, 2000). However, the interface between rater assessment and interlanguage pragmatics has remained largely unnoticed (for exceptions, see Liu & Xie, 2014, in this issue; Taguchi, 2011; Youn, 2007). The first objective of this study was to discover how native and non-native teachers rate L2 refusal production. Rater variability, which has been one of the biggest challenges for language assessment, diminished after the workshop, and more consistency was observed in non-native raters' judgments. In addition, changes in ratings across refusal situations show that pragmatic performance in certain situations caused more rating variation due to the complexity of variables involved, like power, imposition, and distance in different cultures. Generally, non-native raters in this study were more severe than native raters. However, their ratings across refusal situations approached those of native raters after participating in the rating workshop. In conclusion, as the findings show, the main contribution of rater training, which is to reduce rating error as measured against a benchmark and variation across raters, can be realized in the field of pragmatic performance assessment.

The second objective was to explore training effects on non-native teachers' bias in their rating of the speech act of refusal. Compared with its effect on rating accuracy, rater training for pragmatic assessment in this study proved to be less effective in decreasing rater bias. This suggests that rater training has differential effects on training participants; moreover, rating criteria, particularly in pragmatic assessment performance, may need continued rater training to change rating practice. It can also be concluded, in line with the common observation in rater training research (Lumley & McNamara, 1995; Wang, 2010) that bias can be decreased rather than eliminated.

Since rating criteria play a significant role in ILP assessment, the findings from this study have a number of implications. The first one is that there is a need for a pragmatic training program for non-native EFL teachers. Teachers usually apply, if any, different rating criteria to assess pragmatic performance. In fact, raters may have a different understanding of the construct being measured and such differences may have a direct influence on the ratings raters assign to test takers' performance in the testing context. Practically, while the scores non-native raters assigned to pragmatic students' performance before attending the workshop were not close to the natives', they became more accurate after attending the workshop since they got conscious of native sociopragmatic and pragmalinguistic norms for refusal. Therefore, it is necessary to make non-native teachers aware of the benchmark and to increase consistency among them. For this purpose, rater training should be implemented in teacher education programs to make a change in the assessment practice of teachers, and the decision makers need to take training programs into consideration for EFL raters. However, further studies need to be conducted on the rater training effect on rater bias to explore if pragmatic training will produce effective results. Finally, there is a need for longitudinal studies to investigate whether the effect of rater training will last over time.

References

Barnwell, D. (1989). 'Naïve' native speakers and judgments of oral proficiency in Spanish. *Language Testing*, *6*(2), 152-163.

Barritt, L., Stock, P. L., & Clark, F. (1986). Researching practice: Evaluating student essays. *College Composition & Communication*, *37*, 315-327.

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing*, *12*(1), 1-15.

Caban, H. L. (2003). Rater group bias in the speaking assessment of four L1 Japanese ESL students. *Second Language Studies*, *21*(1), 1-44.

Cason, G. J., & Cason, C. L. (1984). A deterministic theory of clinical performance rating. *Evaluation and the Health Professions, 7,* 221-247.

Congdon, P. J., & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement, 37*(2), 163-178.

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.

Eckes, T. (2012). Operational rater types in writing assessment: Linking rater cognition to rater behavior. *Language Assessment Quarterly, 9*(3), 270-292.

Elder, C., Barkhuizen, G., Knoch, U., & von Randow, J. V. (2007). Evaluating rater response to an online training program for L2 writing assessment. *Language Testing*, *24*(1), 37-64.

Engelhard, G. Jr. (1992). The measurement of writing ability with a many faceted Rash model. *Applied Measurement in Education, 5*(3), 171-191.

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.

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

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

Gass, S.M., & Houck, N. (1999) *Interlanguage refusals: A cross-cultural study of Japanese-English.* Berlin: Mounton.

Haizhen, W. (2008). A study on raters' interpretation and application of the rating criteria in TEM4-Oral. *Theory and Practice of Foreign Languages Teaching 2*, 33-39.

Hsieh, C.-N. (2011). Rater effects in ITA testing: ESL teachers' versus American undergraduates' judgments of accentedness, comprehensibility, and oral proficiency. *Spaan Fellow Working Papers in Second or Foreign Language Assessment, 9*, 47-74.

Huot, B. (1990). Reliability, validity and holistic scoring: What we know and what we need to know. *College Composition & Communication, 41*, 201-213.

Johnson, J. S., & Lim, G. S. (2009). The influence of rater language background on writing performance assessment. *Language Testing*, *26*(4), 485-505.

Kachru, B. (1992a). World Englishes: Approaches, issues and resources. *Language Teaching, 2,* 1-14.

Kachru, B. (1992b). *The other tongue: English across cultures.* Champaign, IL: University of Illinois Press.

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.

Kirkpatrick, A. (2006). Which model of English: Native-Speaker, nativized, or lingua franca? R. Rubdy & M. Saraceni (eds.), *English in the world: Global rules, global roles*. London: Continuum.

Knoch, U. (2007). *Diagnostic writing assessment: The development and validation of a writing scale* (Unpublished doctoral dissertation). The University of Auckland, New Zealand. Retrieved from https://researchspace.auckland.ac.nz/

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

Linacre, J.M. (1989). *Many-facet Rasch measurement*. Chicago, IL: MESA Press

Linacre J. M., Wright B. D. (1996) Guttman-style item location maps. *Rasch Measurement Transactions*, *10*(2), 492-493.

Liu, J, & Xie, L. (2014, forthcoming). Examining rater effects in a WDCT pragmatics test. *Iranian Journal of Language Testing, 4*(1).

LoCastro, V. (2000). Evidence of accommodation to L2 pragmatic norms in peer review tasks of Japanese learners of English. *JALT Journal, 22*(2), 245-270.

LoCastro, V. (2012). *Pragmatics for language educators: A sociocultural perspective*. New York: Routledge.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to raters? *Language Testing*, *19*(3), 246-276.

Lumley, T. (2005). *Assessing second language writing: The rater's perspective.* Frankfurt am Main: Peter Lang.

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

Lunz, M. E., & Stahl, J. A. (1990). Judge consistency and severity across grading periods. E*valuation and the Health Professional, 13*(4), 425-444.

McIntyre, P. N. (1993). *The importance and effectiveness of moderation training on the reliability of teachers' assessment of ESL writing samples* (Unpublished master's thesis). University of Melbourne.

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.

Nation, I. S. P., & Macalister, J. (2010). *Language curriculum design*. New York: Routledge.

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

Seidlhofer, B., & Jenkins, J. (2003). English as a lingua franca and the politics of property. In C. Mair (ed.), *The politics of English as a world language*. Amsterdam: Rodopi.

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

Sugita, Y. (2011). Differences in raters' severity, consistency and biased interactions between trained and untrained raters in the context of a task-based writing performance test. *Proceedings of the 16th Conference of Pan-Pacific Association of Applied Linguistics.*

Taguchi, N. (2011). Rater variation in the assessment of speech acts. *Pragmatics*, *21*(2), 453-471.

Tajeddin, Z., & Mohammad Bagheri, M. (2012). The status of pragmatic awareness and instruction among Iranian EFL teachers. Unpublished manuscript.

Wang, B. (2010). On rater agreement and rater training. *English Language Teaching, 3*(1), 108-112.

Weigle, S. C. (1994a). *Effects of training on raters of English as a second language compositions: Quantitative and qualitative approaches* (Unpublished doctoral dissertation). University of California, Los Angeles.

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

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

Wen, Q, Liu, X., & Jin, L. (2005). Native and nonnative judgements of Chinese learners' English public speaking ability. *Foreign Language Teaching and Research, 37*(5), 337-342.

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing, 10*(3), 305-335.

Wigglesworth, G. (1993). Second language performance testing: The Ontario test of ESL as an example. *Language Testing, 4*, 28-47.

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