Monologic vs. Dialogic Assessment of Speech Act Performance: Role of Nonnative L2 Teachers' Professional Experience on Their Rating Criteria

Zia Tajeddin[1] & Iman Alizadeh[2]

[1]Corresponding author, Department of English Language and Literature, Allameh Tabataba'i University, zia_tajeddin@yahoo.com

[2]Department of English Language and Literature, Allameh Tabataba'i University, iman_alizadeh87@yahoo.com

Received: 10/09/2014 *Accepted: 09/03/2015*

Abstract

Few, if any, studies have investigated the effect of professional experience as a rater variable and type of assessment as a task variable on raters' criteria in the assessment of speech acts. This study aimed to explore the impact of nonnative teachers' professional experience on the use of criteria in monologic and dialogic assessment of 12 role-plays of 3 apology speech acts. To this end, 60 raters were divided into 2 subgroups of raters with under and over 5 years of professional experience and rated the role-plays monologically and dialogically. A content analysis of the raters' descriptions of the ratings showed 3 groups of criteria: the general criterion (appropriateness), pragmalinguistic criteria (linguistic features, L1 effect, paralinguistic features, directness, and adequacy), and sociopragmatic criteria (politeness, repair, truthfulness, promise, thanking, reasoning, personal trait formality, genuineness, and expression of apology). We also discovered that neither the more experienced nor the less experienced raters paid due attention to the sociopragmatic criteria in the monologic and dialogic ratings of pragmatic performances. Both groups of raters based their ratings primarily on the general criterion of appropriateness in the dialogic ratings. However, in the monologic ratings, the more experienced ones preferred pragmalinguistic criteria, and the less experienced ones opted for the appropriateness criterion. An analysis of the influence of the type of rating on the raters' application of criteria showed that the raters differed in the use of all the 3 groups of criteria in the monologic ratings, whereas in the dialogic ratings, their difference in the application of criteria narrowed down to the sociopragmatic criteria. The findings have implications for teacher education programs on pragmatic assessment, urge considerations for the role of teachers' experience in pragmatic assessment, and stress the inclusion of dialogic ratings in the assessment of speech acts for improving the quality of raters' assessments.

Keywords: Pragmatic Assessment; Professional Experience; Monologic Rating; Dialogic Rating; Apology Speech Act; Rating Criteria

1. Introduction

The significance of observing pragmatic norms in communication, due to serious consequences of failure to comply with proper norms of linguistic and social behavior, has stepped up research in several aspects of the area of pragmatics (Bardovi-Harlig, 2001; Billmyer & Varghese, 2000; Blum-Kulka et al. 1984; Brown & Levinson, 1987; Ervin-Tripp, 1976; Hudson, 2001; Kasper & Roever, 2005; Li, 2000; Matsumura, 2007; Rose & Kasper 2001; Taguch, 2010). According to McNamara and Roever (2006), studies on the assessment of learners' pragmatic competence, however, are in embryonic stages in comparison with studies on the teaching of pragmatic, and pragmatic assessment awaits further research and development. Rose and Kasper (2001) also point out that research on the assessment of pragmatic competence has gained less attention compared with the significant amount of research on pragmatic instruction. Studies conducted in the area of pragmatic assessment have examined rater's variability (Liu, 2006; Taguchi, 2011) and probed bias in the process of assessing learners' pragmatic performance (Youn, 2007). Few studies have also investigated rating criteria in the assessment of individuals' pragmatic performances (Alemi & Tajeddin, 2013; Tajeddin & Alemi, 2014). The paucity of research on pragmatic assessment comes as several studies have probed raters' variability (Brown, 1995; Eckes, 2005; Elder et al., 2007; Galloway, 1980; Lee, 2009; Wigglesworth, 1994) and bias (Kondo-Brown, 2002; Johnson & Lim, 2009; Schaefer, 2008; Wigglesworth, 1993; Zhang & Elder, 2011) in the assessment of learners' speaking or writing performances. A number of studies have also investigated raters' rating criteria in the assessment of L2 skills (e.g., Eckes, 2005).

Whereas the few existing studies have investigated raters' variability, bias, and, to a lesser extent, criteria in pragmatic assessment, it seems many other criteria and variables that may contribute to more effective ratings of speech acts have remained underexplored. To address the existing gaps in pragmatic assessment, the present study was designed to investigate the effect of nonnative English teachers' professional experience on the use of criteria in monologic and dialogic assessments of English language learners' apologetic performances.

2. Literature Review

The present study was an attempt to shed light on the effect of teachers' professional experience, as a rater variable, on their rating behavior in the assessment of learners' pragmatic performances. The study also aimed to discover the criteria that raters use in their monologic and dialogic assessments of learners' pragmatic performances. Related literature on raters' professional experience as well as raters' criteria along with studies conducted in these areas is reviewed in this part.

2.1. Professional Experience as a Rater Variable

Rater variables play a significant role in assessing learners' ability in the production of language that is appropriate, given the context of language use. The most frequent rater variables researched and discussed in the realm of performance assessment include the status of being native or nonnative (Brown, 1995; Lee, 2009), gender (Eckes, 2005), professional background (Liu, 2006), experience (Galloway 1980), expectations (Weigle 1998), and intensity of instruction (Elder et al., 2007). Rates' professional experience, as a rater variable, among other factors, is made up of their beliefs, perceptions, or expectations that subconsciously influence the way they approach a rating task (Barnwell, 1989; Shohamy, Gordon, & Kraemer, 1992). Studies delving into raters' behavior in performance assessments have reported significant degrees of variability originating from the characteristics of raters (e.g., Eckes, 2005; Engelhard & Myford, 2003; Schoonen, 2005). Eckes (2008) holds that the interaction between raters and rating criteria can be a source of rater variability. He argues that raters may vary in the degree to which they comply with the scoring rubric, the way they interpret criteria, and the degree to which their ratings are consistent across examinees, scoring criteria, and performance tasks. There is a paucity of research on the effect of raters' professional experience on pragmatic assessment. Some researchers, however, have examined the relationship between raters' background, including their experience, and test scores in the speaking assessment of English language learners (Bachman et al., 1995; Caban, 2003; Kim, 2009; Lumley & McNamara, 1995).

2.2. Studies on the Role of Professional Experience in Assessment

Literature shows that there is no agreement on the way professional and novice raters do the task of assessing learners' language ability. Barnwell (1989), for example, discovered that in comparison with novice raters, professional ones are more tolerant of language errors. One reason for this can be the very fact that the experienced raters' exposure to a possibly wider range of linguistic abilities has enabled them to assess learners' performance more reliably. Brown (1995), on the other hand, found in a study on the effect of raters' experience on raters' assessment behavior that professional raters in her study were harsher in all areas of proficiency except pronunciation.

Galloway (1980) investigated the way students' communicative competence was evaluated by native high school Spanish teachers, nonnative high school Spanish teachers, nonteaching native speakers living in the U.S. with a fair-to-good command of English, and nonteaching native speakers living in Spain with no or poor command of English. The rating criteria used were the amount of communication, efforts to communicate, comprehensibility, paralanguage, and overall rating. The findings revealed that regarding specific aspects of speech, the nonteaching native speakers of Spanish living in the U.S. paid less attention to pronunciation than the other groups. It was also discovered that the nonnative high school Spanish teachers seemed to be more disturbed than the nonteaching native speakers of Spanish by the slow pace of the students' utterances. Based on the raters' comments on the performance of the students, it was revealed that the nonnative teachers focused primarily on grammatical accuracy, whereas the nonteaching native speakers focused mainly on the content or message. The findings indicated that teachers were more critical of students' grammatical abilities in comparison with ordinary people.

In a similar study, Caban (2003) examined the way Japanese L1 raters with and without ESL/EFL background, and English L1 raters assessed the Japanese students' English oral interviews. They rated the students' performances on the seven categories of fluency, grammar, pronunciation, comprehension techniques, content of utterance, language appropriateness, and overall intelligibility. The findings showed that the English L1 speaker raters were consistently more lenient in evaluating the pronunciation quality of Japanese-accented English than Japanese L1 raters. It was also discovered that the ESL/EFL-trained Japanese L1 raters rated pronunciation and grammar more harshly, but compensation techniques, language appropriateness, and overall intelligibility more leniently than the other groups. The Japanese L1 speakers also exercised more leniency in scoring fluency and grammar than the English L1 raters.

Kim (2009) also studied the way trained native versus nonnative English-speaking teachers scored an English-speaking test. Kim asked 12 native-speaking English teachers and 12 Korean English teachers to score 80 spoken responses from 10 Korean speakers to a computerized oral test. The raters received training with sample responses before the scoring. The results showed no total test score differences between these two groups of raters. Moreover, no positive or negative bias toward a particular task or task type was observed in the ratings of the two groups. Kim concluded that the way that the raters scored the performances might be the same; however, the reasons for the scoring might be different.

In a bid to investigate the development and maintenance of rating quality in writing assessment, Lim (2011) researched the rating behavior of new and experienced raters in a longitudinal study. The quality of the rating was operationalized in terms of raters' severity and consistency. The findings indicated that the novice and experienced raters did not always differ in terms of consistency and severity. It was also shown that the novice raters improved their ratings and learned to rate appropriately and relatively quickly.

2.3. Rating and Rating Criteria in Pragmatic Assessment

A significant variable in the area of assessment, in general, and pragmatic assessment, in particular, is rating variable or the type of rating. Most studies on pragmatic assessment have approached the task monologically, that is, a single examiner or rater completes the rating. It seems quite practical to ask teachers to rate the pragmatic tasks dialogically with other teachers or raters. A review of literature with the aim of spotting research in which modalities of assessment close to the dialogic assessment are used shows a study by Chau (2005). The study was on the assessment of language skills within the paradigm of collaborative assessment, which is the only framework relevant to the dialogic assessment used in the current study.

In addition, the issue of rating criteria has been introduced as an area of research in the studies on pragmatic assessment. Some researchers maintain that the way raters apply rating criteria for a performance can influence the fairness, accuracy, and interpretability of the assessment (Winke, Gass, & Myford, 2012). A review of literature on pragmatic assessment criteria, however, shows a few studies (e.g., Alemi & Tajeddn, 2013; Tajeddin & Alemi, 2014). Studies relevant to the discussion on rating and rating criteria are reviewed below.

In a bid to investigate the impact of collaborative assessment as a tool not only for evaluative purposes but also for learning purposes, Chau (2005) provided the L2 learners in her study with an opportunity to act as the teacher's partners in assessment and language learning. The tutor and the learners in the study collaborated to arrive at an agreed score for an assignment through discussion and negotiation in English. She found that, as a result of the collaboration with tutors, the students became active participants in the process of teaching and learning. It was also discovered that collaborative assessment contributed to the development of the learners' continual intellectual, experiential, and attitudinal growth. Chau affirmed that the advantages gained from this type of assessment came without spoiling the evaluative nature of the assessment.

Alemi and Tajeddin (2014) analyzed the descriptions of 51 native English teachers rating six different pragmatic situations for an apology discourse completion task (DCT) that were accompanied by an L2 learner's response to each situation. The content analysis of the raters' descriptions revealed five criteria they mostly applied in their rating: expression of apology, situation explanation, repair, offer, promise for future, and politeness.

In a separate study, Alemi and Tajeddn (2013) also investigated the ratings native English-speaking (NES) and nonnative English-speaking (NNES) teachers assigned to L2 refusal production and the criteria they applied in their ratings. They found the NES teachers applied 11 criteria in their pragmatic ratings: politeness, statement of alternative, postponing to another time, thanking, dishonesty, cultural problem, explanation/reasoning, irrelevancy of refusal, offer suitable consolation, statement of refusal, and brief apology. It was also discovered that the NNES teachers based their ratings on brief apology, irrelevancy of speech act, explanations, postponing to another time, statement of alternative, and politeness. Reasoning/explanation was the leading criterion in the assessment of the NES raters, and politeness was the main criterion for the NNES ratings. They also found that the NNES teachers were more lenient and divergent in their ratings. They concluded that there was a gap between the NES and NNES teachers in terms of rating criteria, strictness, and convergence in rating.

3. Purpose of the Study

As the raters in the current study assessed learners' performance of apology speech acts, their ratings were considered pragmatic performance assessment. The studies carried out in the area of pragmatic assessment have approached the assessment task monologically, that is, one particular examiner or rater conducted the rating (Alemi & Tajeddin, 2013; Liu, 2006; Plough, Briggs, van Bonn, & , 2010; Tajeddin & Alemi, 2014). Monologic rating embodies the whole range of activities and techniques a single rater deploys to come up with a score and/or criteria for rating a role-play. According to McNamara (1996), in performance assessment, the decisions of a single rater do not represent an appropriate estimate of a candidate's ability. Moreover, Knoch, Read, and von Randow (2007) argue that performance assessment is prone to various sources of bias and error that can threaten the quality of the assessment. Therefore, dialogic rating was used in the present study as a form of performance assessment to probe the raters' behavior in the application of criteria in pragmatic assessment. It seems quite practical to ask L2 teachers to rate pragmatic performances dialogically with other teachers or raters. Employing a dialogic mode of assessment has the potential for helping raters share their rating criteria with other raters, negotiate the criteria, and challenge other raters' criteria with the aim of coming up with the most appropriate criteria for assessment. Therefore, besides investigating raters' criteria in monologic ratings, the current study probed the raters' performance in dialogic ratings. Dialogic rating is defined in the current study as an alternative form of collaborative assessment in which two teachers or raters carry out the assessment task by negotiating the rating criteria. According to Dickinson (1988), collaborative assessment, in which one side is a student and the other is a teacher or another student, provides L2 learners with the opportunity to think through the task they are involved in and to reflect on their work. Chau (2005) specifies mutual goals, dynamic exchange of information, and role interdependence as key features of collaborative assessment. What makes dialogic assessment in the current study different from other forms of paired assessment, including collaborative assessment, is that both parties involved in the assessment are teachers.

Furthermore, Brown (1995) stresses that raters' decision-making comes from their experience. Raters' norms and beliefs have a significant role in their decision-making process and influence their behavior while making judgments. Therefore, it can be argued that raters' professional experience as a rater variable can play a significant role in the process of pragmatic assessment. A review of related literature, however, shows that the variable has apparently received little, if any, attention in the area of pragmatic assessment. Few studies, however, have probed the effect of teachers' experience on the assessment of learners' general language skills (Caban, 2003; Lim, 2009). The present study, as a result, aimed to delve into the effect of nonnative raters' professional experience on the use of criteria in dialogic and monologic pragmatic ratings. The study also sought to investigate whether raters' professional experience makes a change in the criteria they employ in monologic and dialogic ratings of the apology speech acts. The questions below were addressed in this study:

1. What criteria do nonnative English teachers with over and under 5 years of professional experience apply in dialogic and monologic ratings of learners' production of apology?
2. What are the similarities and differences in the criteria nonnative English teachers with over and under 5 years of professional experience use in dialogic and monologic ratings of learners' production of apology?

4. Method

The study benefited both qualitative and quantitative procedures for data collection and data analysis. As for the data collection process in the quantitative phase, the raters were asked to rate the learners' performances on a scale ranging from 1 (*the lowest*) to 5 (*the highest*) on a rating sheet. For the qualitative data collection, the raters were asked to attach their explanations and comments on the numerical ratings of the learners' performances to the rating sheet. In the data analysis phase, the study employed content analysis as a qualitative analytical method to discover the criteria underlying the raters' rating process. In the quantitative phase of data analysis, the study investigated the significance of the differences in the frequency of the criteria the raters used in the dialogic and monologic pragmatic assessment in terms of their professional experience.

4.1. Participants

A group of eight L2 learners and a group of 60 raters made up the participants. The learner participants were male English learners within the age range of 14-50. They all studied English as the L2 at Iran Language Institute (ILI) at the intermediate level of language proficiency. The learner participants were divided into four groups, each consisting of two interactants. The four groups of the L2 learners performed 12 role-plays and received no training or help while performing the roles. The 60 rater participants were male and female L2 teachers instructing English as the L2 in different language institutes in Iran. They were divided into two groups of raters with over 5 years of professional experience and the raters with under 5 years of professional experience in L2 teaching. They completed the monologic rating tasks individually on their own and the dialogic rating in cooperation with their colleagues.

4.2. Instrumentation

As many as 12 apology role-plays, that is, three role-plays performed by each of the four groups of the interlocutors based on three different apologetic situations and a pragmatic performance rating sheet were the instruments in the study. The reason for recording four role-plays for each of the situations was to provide the raters with different performances on the situations. By doing so, the raters could have a better grasp of the situations and consider different ranges of criteria in the rating process. Each of the three apology situations required one of the two learner participants in an apology speech act to perform a specific role. For the purpose of rating, the teachers were asked to rate the pragmatic performance of the interlocutor producing the apology speech act in each of the role-plays. The situations were designed with an eye to the social variables of power (P), social distance (D), and degree of imposition (I) proposed by Hudson, Detmer, and Brown (1995). These contextual variables characterized the situations because they have been found to play a decisive role in speech act realization patterns (Blum-Kulka, 1984; Brown, & Levinson, 1989). The situations and the way the social variables of P, D, and I interact in each of the situations are described below:

- **Situation 1:** *You are a student and you have asked for an appointment with your teacher. Despite being busy, the teacher accepts your request and agrees to receive you in his office for next Tuesday at 5 o'clock. When you go to his office on the fixed day, you notice that it is ten past 5 and the teacher is leaving the office. How would you as a student apologize for being late?* (+P, +D, +I)
- **Situation 2:** *You are a student at an advanced level of English proficiency. You have borrowed a book from one of your classmates, who is much younger than you, on Saturday and promised to give it back on Wednesday. Despite your promise, you forget to bring it over. How would you as a classmate apologize?* (= - D, -I)

- **Situation 3:** *You are a student in an advanced course and you are supposed to have a presentation on a subject which has been assigned to you. You, however, forget to do the assignment and prepare your power points for the presentation. Your teacher calls you for the presentation. How would you as a student apologize to the teacher for not being ready?* (+P, +D, +1)

The pragmatic performance rating sheet, employed to collect the data on P, *the* raters' rating performances of the apology role-plays, consisted of three parts: (1) An instruction part where the raters were instructed how to rate the performances of the interlocutor producing the apology speech act, (2) a five-point Likert scale ranging from 1 (*the weakest*) to 5 (*the best*) used by the raters for scoring the interlocutors' performance, and (3) a comments section for the raters to show their rating criteria and descriptions for assessing each of the role-plays.

4.3. Data Collection Procedure

The data were collected in two phases: (1) recording the role-plays performed by the learners, and (2) ratings by the raters. The method employed was recording the open role-plays as a modality of spoken interaction. As there were four role-plays for each of the three different apology situations, a total of 12 role-plays were recorded. Each role-play lasted between 1.5 and 2.5 min, and the role-plays were built into as many as 12 video clips. The video clips showed a description of the apology situations based on which the role-plays were performed. The clips were, subsequently, recorded onto some CDs and were submitted to the raters. In the next stage, 60 raters were asked to rate the role-plays. The rating process consisted of two phases. In the first phase, the raters were asked to indicate their ratings of the apologetic performance of the learners in the role-plays on the rating sheet monologically. The raters were also asked to explain their reasons for giving a participant a particular score. In the second phase of the rating process, the teachers were invited to come together and rate the role-plays dialogically, that is, in pairs and in association with another teacher.

4.4. Data Analysis

A content analysis method was employed to explore the criteria the nonnative raters with different years of experience used in the ratings of the apology speech acts. To do so, a careful analysis of the descriptions was conducted based on the apology strategy frameworks proposed by Olshtain and Cohen (1983), Holmes (1990), and Bergman and Kasper (1993). According to Olshtain and Cohen (1983), the production of an apology speech act involves two general and three situation-specific strategies. They further explain that using an Illocutionary Force Indicating Device (IFID) and expressing responsibility are the general strategies, and explanation, offer of repair, and promise of forbearance are the situation-specific strategies for producing apology. Holmes (1990) introduced an apology strategy framework including four categories. Bergman and Kasper (1993) also distinguished seven different apology strategies.

Despite being proposed in the frameworks for producing apology speech acts, the strategies helped the researchers identify the notes in the raters' description that referred to the learners' failure or success to comply with these strategies while performing the speech act.

5. Results

A content analysis of the descriptive comments of the raters' monologic and dialogic ratings of the role-plays led to the discovery of 16 criteria shown in Table 1:

Table 1. *Criteria Discovered in Dialogic and Monologic Assessments by Raters*

Criteria	Definition	Example
Linguistic features	All factors pertaining to grammar, vocabulary, structure, use of tenses, and so forth	There are many grammatical mistakes in the conversation.
Politeness	Clues that somehow relate to issues like degree of imposition, social dominance, distance, and the application of terms like *polite, rude,* or *impolite*	Shaking hands with teacher does not seem polite for students.
Genuineness	Clues that show how naturally or artificially an apology speech act is produced	The apology was not natural.
Paralinguistic features	The way the speaker expresses his or her apology, that is to say, directly or indirectly	
Directness	The way the speaker expresses his or her apology, that is to say, directly or indirectly	The student uses the direct form of "excuse me" for his apology.
Expression of apology	Explicit apology forms such as *excuse me, sorry,* and *regret*	The participant expressed his apology.
Formality	The extent to which the speakers observe the norms of formality and/or informality in their productions of speech acts: being straightforward and too bookish, formal, and friendly are the constituents of this criterion	His apology is too formal.

Criteria	Definition	Example
Appropriateness	All cases that the raters evaluated the interactants' performances as "acceptable," "proper," "good," and the like	His apology was acceptable.
Repair	The effort made by the speaker to make for the fault on his part	He promises by saying I will bring the book next session.
Truthfulness	Features that indicate that an interactant is telling the truth and is speaking honestly, including cases that imply an interactant is not talking sincerely	It is obvious that he is lying.
Promise	An effort through which the speaker gives his or her word for doing or avoiding an act in the future	He promises to speak more quietly.
Thanking	Cases in which an interactant expresses gratitude or acknowledgement in a conversation which includes apology	He thanked his professor.
Reasoning and explanation	Justification by explaining the cause of the wrong action on the part of the speaker which was beyond his or her control	The examinee explains the reason behind his coming late.
Adequacy	The amount or size of the information that an interactant uses for performing an apology speech act	His apology is too short.
L1 effect	Traces of an interactant's L1 that are noticed in a conversation involving an apology speech act	It seems he is speaking Farsi [Persian].
Personal trait	Features in the behavior or personality of an interactant that raters think were influential in his or her success or failure to produce the apology	He is too shy.

The second question of the study addressed the similarities/differences in the criteria employed by the nonnative L2 teachers in the monologic and dialogic pragmatic ratings of the apology speech acts. Details on the frequency and percentage of the criteria in the pragmatic ratings of the speech acts are given in

Table 2. As shown, “appropriateness” with the frequency of 889 is the most preferred criterion, and “thanking” with the frequency of 29 is the most underrepresented criterion in the ratings by the raters:

Table 2. *Frequency and Percentage of Criteria in Both Dialogic and Monologic Ratings*

Criteria	Frequency	Percentage
Linguistic features	739	19.1
Politeness	353	9.1
Directness	60	1.5
Paralinguistic features	393	10.1
Genuineness	226	5.8
Expression of apology	415	10.7
Formality	119	3.1
Appropriateness	889	23.0
Repair	56	1.4
Truthfulness	123	3.2
Promise	88	2.3
Thanking	29	.7
Reasoning and explanation	178	4.6
Adequacy	122	3.2
L1 effect	32	.8
Personal trait	51	1.3
Total	3819	100.0

To see how the raters deployed the criteria in their monolgic and dialogic ratings of the three apology speech acts in terms of their teaching experience, the differences or similarities in the frequency and percentage of the criteria in each of the two modalities of pragmatic rating were calculated. Table 3 shows that “appropriateness” with the frequency of 326 was the most preferred criterion in the ratings of the less experienced raters, whereas the more experienced raters gave priority to “linguistic features” by applying it 265 times in their monologic ratings:

Table 3. *Frequency of Criteria in Monologic Ratings by Raters With +/-5 Years of Experience*

Criteria	Experience	Frequency	Percentage
Linguistic features	Under 5	194	42.3
	Over 5	265	57.7
Politeness	Under 5	96	42.5
	Over 5	130	57.5
Directness	Under 5	22	41.5
	Over 5	31	58.5
Paralinguistic features	Under 5	117	45.9
	Over 5	138	54.1

Genuineness	Under 5	71	46.4
	Over 5	82	53.6
Expression of apology	Under 5	114	43.2
	Over 5	150	56.8
Formality	Under 5	62	71.3
	Over 5	25	28.7
Appropriateness	Under 5	326	56.2
	Over 5	254	43.8
Repair	Under 5	19	54.3
	Over 5	16	45.7
Truthfulness	Under 5	49	58.3
	Over 5	35	41.7
Promise	Under 5	32	49.2
	Over 5	33	50.8
Thanking	Under 5	10	71.4
	Over 5	4	28.6
Reasoning	Under 5	61	48.0
	Over 5	66	52.0
Adequacy	Under 5	29	42.0
	Over 5	40	58.0
L1 effect	Under 5	6	60.0
	Over 5	4	40.0
Personal trait	Under 5	10	40.0
	Over 5	15	60.0

A chi-square test was used to investigate whether the difference between the frequencies of the criteria used in the monologic ratings of the teachers with over and under 5 years of experience was significant. Table 6 displays the results of the tests. As Table 4 shows, except for “appropriateness” (X^2 = 8.93, *df* = 1, *p* = 0.003), “politeness” (X^2 = 5.11, *df* = 1, *p* = 0.024), “expression of apology” (X^2 = 4.90, *df* = 1, *p* = 0.027), “formality” (X^2 = 8.93, *df* = 1, *p* = 0.000), and “linguistic features” (X^2 = 10.98, *df* = 1, *p* = 0.001), the frequency of other criteria in the monologic ratings of the more and the less experienced raters was not statistically different. That is to say, the raters’ professional experience only played a significant role in the choice of the abovementioned criteria:

Table 4. *Chi-Square Results for Criteria in Monologic Ratings by Raters*

Criteria	Chi-Square	*df*	*Sig.*
Linguistic features	10.98	1	.001
Politeness	5.11	1	.024
Directness	1.52	1	.216
Paralinguistic features	1.72	1	.188
Genuineness	.79	1	.374
Expression of apology	4.90	1	.027
Formality	8.93	1	.000
Appropriateness	8.93	1	.003
Repair	.25	1	.612
Truthfulness	2.33	1	.127
Promise	.01	1	.901
Thanking	2.57	1	.109
Reasoning	.19	1	.657
Adequacy	1.75	1	.185
L1 effect	.40	1	.527
Personal trait	1.00	1	.317

The study also aimed to delve into the dialogic ratings of the apology speech acts by two the groups of raters. The statistics with regard to this purpose of the study are given in Table 5. As shown, both the less and the more experienced raters assigned the highest significance to the “appropriateness” criterion (the frequencies of 160 and 148, respectively) in their dialogic ratings of the speech act:

Table 5. *Frequency of Criteria in Dialogic Ratings by Raters With +/-5 Years of Experience*

Criteria	Experience	Frequency	Percentage
Linguistic features	Under 5	140	50.0
	Over 5	140	50.0
Politeness	Under 5	46	36.2
	Over 5	81	63.8
Directness	Under 5	0	0
	Over 5	7	100.0
Paralinguistic features	Under 5	64	46.4
	Over 5	74	53.6
Genuineness	Under 5	38	48.7
	Over 5	40	51.3
Expression of apology	Under 5	57	37.7
	Over 5	94	62.3
Formality	Under 5	25	83.3
	Over 5	5	16.7
Appropriateness	Under 5	160	51.9

	Over 5	148	48.1
Repair	Under 5	10	47.6
	Over 5	11	52.4
Truthfulness	Under 5	26	53.1
	Over 5	23	46.9
Promise	Under 5	9	39.1
	Over 5	14	60.9
Thanking	Under 5	5	45.5
	Over 5	6	54.5
Reasoning	Under 5	27	52.9
	Over 5	24	47.1
Adequacy	Under 5	25	47.2
	Over 5	28	52.8
L1 effect	Under 5	6	42.9
	Over 5	8	57.1
Personal trait	Under 5	18	69.2
	Over 5	8	30.8

The chi-square statistics are given below to show how significantly the use of the criteria by teachers with over and under 5 years of experience differed in their dialogic rating of the apology speech acts. As Table 6 shows, except for “politeness” ($X^2 = 9.646$, $df = 1$, $p = 0.002$), “expression of apology” ($X^2 = 9.06$, $df = 1$, $p = 0.003$), “formality” ($X^2 = 13.33$, $df = 1$, $p = 0.000$), and marginally “personal trait” ($X^2 = 3.84$, $df = 1$, $p = 0.050$), the frequency of other criteria in the dialogic ratings of the teachers with over and under 5 years of experience in L2 teaching was not statistically different:

Table 6. *Chi-Square Results for Criteria in Dialogic Ratings by Raters*

Criteria	Chi-Square	*df*	*Sig.*
Linguistic features	.00	1	1.000
Politeness	9.64	1	.002
Directness	—		—
Paralinguistic features	.72	1	.395
Genuineness	.05	1	.821
Expression of apology	9.06	1	.003
Formality	13.33	1	.000
Appropriateness	.46	1	.494

Repair	.04	1	.827
Truthfulness	.18	1	.668
Promise	1.08	1	.297
Thanking	.09	1	.763
Reasoning	.17	1	.674
Adequacy	.17	1	.680
L1 effect	.28	1	.593
Personal trait	3.84	1	.050

6. Discussion

One of the purposes of this study was to explore the type and frequency of the criteria the nonnative raters with over and under 5 years of professional experience apply in the monologic and dialogic ratings of three apology speech acts. A content analysis of the raters' descriptions of the reasons for their ratings led to the discovery of 16 criteria categorized in three groups of: (1) the general (appropriateness), (2) pragmalinguistic (linguistic features, directness, paralinguistic features, adequacy, L1 effect), and (3) socipragmatic (genuineness, expression of apology, formality, politeness, repair, truthfulness, promise, thanking, reasoning, and personal trait). We discovered that in the monologic ratings, the "appropriateness" criterion topped the list in the ratings of the less experienced raters and the more experienced ones preferred the "linguistic features" criterion. We also discovered that both groups of raters attached less significance to the sociopragmatic criteria in comparison to the two other groups of criteria in the monologic ratings, whereas social norms are viewed as one of the indispensable building blocks of a pragmatic performance (Thomas, 1995). Leech (1983) also divides pragmatics into two parts of sociopragmatics and pragmalinguistics, with the former focusing on socially appropriate language use and the latter involving linguistic strategies for expressing speech acts. The raters' failure to pay due attention to the sociopragmatic criteria is arguably a fundamental flaw in their monologic ratings, indicating the ineffectiveness of the raters' professional experience in pushing them toward considering social factors in the ratings. One reason behind the underrepresentation of social norms in the ratings of both groups of raters in the monologic ratings can be the nonnative raters' unawareness of the construct of pragmatic assessment that stresses both linguistic and social dimensions of pragmatic competence. Apparently, due to the recency of research on pragmatic assessment in the Iranian context (Alemi & Tajeddin, 2013; Tajeddin, Alemi, 2014), and the resultant unfamiliarity of the Iranian teachers with the tenets of pragmatic assessment, the raters based their ratings upon the holistic assessment of general language skills. Therefore, it can be argued that because of the absence of such an assessment in the Iranian context, the raters relied on their experience in the assessment of L2 skills while rating pragmatic tasks.

Another line of argument is that in the monologic ratings, the more experienced raters mainly based their assessments on the pragmalinguistic criteria and the less experienced ones preferred the general criterion of “appropriateness.” This shows that the more experienced raters performed better in the ratings as they tapped at least part of the construct they were supposed to rate. The raters with under 5 years of experience, however, mainly focused on the general criterion of “appropriateness,” which is practically neither pragmalinguistic nor sociopragmatic, making their rating distant from the construct of pragmatic assessment.

Results of the investigations into the dialogic rating of the apology speech acts by the less and the more experienced raters showed that both groups assigned the highest significance to the “appropriateness” criterion. This finding is in line with the findings by Lim (2011) who concluded that novice raters’ ratings are not always different from those of experienced ones. Lim also discussed that novice raters’ rating quality improves relatively quickly and reaches acceptable limits. The reason behind the nonnative raters’ propensity for the general criterion of “appropriateness” can be the nature of the dialogic assessment together with the expertise of the raters in the field. Dialogic assessment requires negotiations between two raters to reach an agreement on the competency of a learner in performing a pragmatic task. Negotiating an agreement on general criteria in the process of assessment is easier than working on detailed context-specific ones. As a result, the raters in this study preferred to arrive at a compromise on the learners’ ability in performing the apologetic speech acts by describing the interlocutors’ performances as “good,” “bad,” “acceptable,” and the like. One of the most likely causes for such a compromise on only the general criterion of “appropriateness” can be the raters’ unfamiliarity with the tenets of pragmatic competency and lack of expertise in the field of pragmatic assessment. Discussing the details of the interlocutors’ pragmatic performance requires a mastery over pragmatic competence and assessment. As the nonnative raters lacked the required knowledge of pragmatic competence and mastery in the pragmatic assessment, they accordingly decided to base their assessments on the general criteria such as “bad” and “good” that are more easily negotiable.

Moreover, the nonnative raters’ relied on pragmalinguistic and general criteria, and little attention was paied to social norms and conventions, whereas assessing pragmatic performances dialogically and mmonologically indicates that the raters actually assessed part of the construct they were supposed to rate that jeopardizes the validity of their assessments. As social and linguistic factors along with other contextual factors constitute the construct of pragmatic assessment, it can be argued that ignoring each of these factors, while assessing pragmatic tasks, leads to an invalid assessment. The findings on the raters' criteria correspond to the findings by Taguchi (2010) where she asserted that the raters carried out pragmatic assessments holistically and based their ratings on a number of general features.

This study also aimed to discover the criteria making a difference in the monologic and dialogic ratings of the nonnative teachers. The analyses of the frequencies of the criteria used in the monologic rating showed that, except for "appropriateness," "politeness," "expression of apology," "formality," and "linguistic features," the frequency of other criteria in the monologic ratings of the more and the less experienced raters was not statistically different. It follows that the raters' professional experience played a significant role in the choice of these criteria in the monologic ratings of the speech acts. A scrutiny of the five criteria shows that they represent the three categories of criteria in the study: the general (appropriateness), pragmalinguistc (linguistic features), and sociopragmatic (politeness, expression of apology, formality). This finding also shows that the less experienced raters attached more significance to the appropriateness criterion, whereas the more experienced ones were more inclined to the pragmalinguistic factors. The frequency of the sociopragmatic criteria was low in the ratings of both groups of raters. The more experienced raters' preference for the pragmalinguistic criteria, despite the low representation of socipragmatic criteria in their ratings, is an indication that the raters' experience had an effect on their move toward more pragmatically oriented assessment.

The analyses of the differences between the frequencies of the criteria used in the dialogic ratings of the less and the more experienced teachers shows that, except for "politeness," "expression of apology," and "formality" criteria, the frequency of other criteria was not statistically different. The three criteria, making a difference in the dialogic assessments of the two groups of raters, belong to the sociopragmatic category. It was also found that the more experienced raters considered "politeness" and "expression of apology" more significant than the less experienced raters who preferred the "formality" criterion. This can be mainly due to differences in the raters' perceptions of politeness, expression of apology, and formality. Along the same line of discussion, Eckes (2008) found that raters have different views on the significance of scoring criteria. He further discusses that, in his study, the raters attached different levels of significance to fluency, train of thought, completeness, description, structure, argumentation, syntax, vocabulary, and correctness in scoring the examinees' writing performance. Moreover, he argued that the raters' background variables partially accounted for the differences. Among the many factors that can have a role in the formation of the raters' perceptions is their experience. Roever (2001) underscores evaluator's built-in biases in the process of evaluation, arguing that evaluators' perceptions introduce highly subjective factors that make their evaluations more or less inaccurate or biased. Moreover, as politeness has been proved to be one of the overriding criteria in the studies on pragmatic assessment (Tajeddin & Alemi, 2014), one can claim that the more experienced raters in the current study had a more accurate perception of the sociopragmatic criteria by attaching higher significance to this criterion in comparison to the less experienced teachers who preferred the formality criterion.

It was also found that unlike the divergence of the raters' ideas on the general criterion of "appropriateness" as well as on pragmalinguistic criteria in the monologic ratings, their differences in the dialogic assessments narrowed down to only sociopragmatic ones. Such a change shows that the sociopragmatic criteria were the ones on which the two groups of raters had come to significantly different conclusions in the dialogic ratings that can be mainly due to their professional experience. This finding is in line with the findings reported by Barnwell (1989) who discovered that in comparison to novice raters, professional raters are more tolerant of language errors and focus on factors other than linguistic ones. Barnwell discusses that one reason for such a difference can be the very fact that the experience raters' exposure to the widest possible ranges of linguistic ability enables them to assess learners' performance more reliably. Another important line of argument for the finding is that the dialogic ratings could remove raters' differences in the application of the general criterion of "appropriateness" and pragmalinguistic criteria, but not the sociopragmatic criteria. One of the reasons behind this can be the nature of the criteria on which the raters compromised, and the raters' expertise in the assessment of general L2 skills and pragmatic performances of interlocutors. Reaching an agreement on the general criterion of "appropriateness," due to the nature of the criterion, seems to be a matter of common sense and does not require an in-depth analysis of the performances assessed. Therefore, both groups of raters could have arrived at an agreement on the appropriateness criterion by referring to their common sense. Moreover, the convergence of the ideas of the two groups of raters on pragmalinguistic criteria and the divergence of their views on the sociopragmatic criteria are apparently a matter of the raters' knowledge and mastery in the field. The two groups of raters had enough knowledge of the assessment of the linguistic aspects of the learners' performances; therefore, they could bring up cogent arguments in their discussions to reach an agreement on the criteria. However, it seems that the two groups of raters had different views on the sociopragmatic criteria in their dialogic ratings because they had different levels of knowledge of pragmatic competence and lacked necessary experience in the assessment of social aspects of the construct under assessment. Besides, it seems when assessing the pragmatic performance dialogically, the raters reiterated the stances they had in the monologic ratings because they did not have the knowledge to negotiate the sociopragmatic criteria and hence the differences they had in the monologic assessment remained unresolved. Due to the raters' ignorance of the magnitude of social norms in the pragmatic assessment, they did not know the significance of different criteria while rating the pragmatic performances. Therefore, they could not reach an agreement on the intuitive criteria they have brought up for rating the social dimension of pragmatic assessment.

7. Conclusion and Implications

Three groups of the appropriateness criterion, pragmalinguistic criteria, and sociopragmalinguistc criteria were discovered in a content analysis of the descriptions provided by the raters with over and under 5 years of professional experience in the monologic and dialogic ratings of three apology speech acts. The study revealed that the less experienced raters preferred "appropriateness" in both monologic and dialogic ratings. The more experienced raters, on the other hand, preferred pragmaliguistic criteria in the monologic ratings, and the appropriateness criterion in the dialogic ratings. It can be concluded that neither the more experienced nor the less experienced raters could rate the learners' pragmatic performance appropriately, as they both mostly focused on the general criterion of appropriateness, and failed to apply social factors in their assessments. As social factors constitute one of the dimensions of the pragmatic competence, the raters' failure to consider these factors can lead to the invalidity of the assessment. To avoid invalidity in the assessment of pragmatic tasks, raters are strongly advised to consider the full picture of pragmatics by involving both pragmalinguistic and sociopragmatic factors.

The results also showed that both groups of raters had enough familiarity with the pragmalinguistc factors. They, however, seemed to lack required knowledge in the social dimensions of pragmatic competence and assessment that necessitates the training of the raters in these aspects. Through the organization of pragmatic training courses and pragmatic assessment workshops, the raters, irrespective of their previous experience, can gain enough knowledge of pragmatic competence as well as pragmatic assessment, which, in turn, will improve the accuracy and validity of their assessments.

We also discovered that the two groups of raters differed in the application of the all the three groups of criteria in their monologic assessments. They, however, were found to have differences only in the application of sociopragmatic criteria in the dialogic ratings. This shows that the dialogic assessment helped both groups of raters compromise on pragmalinguistic criteria and attach almost the same level of significance to the nonspecific criterion of appropriateness. According to the findings, the two groups of raters apparently did not have the necessary knowledge of the social aspects of pragmatics to negotiate the sociopragmatic criteria. Raters could have reached an agreement on the sociopragmatic criteria if they had known pragmatics well. If the raters had been familiar with the social dimensions of pragmatic assessment, they could have had the chance to find common grounds while considering different aspects of the construct under assessment. In the current study, neither the more nor the less experienced raters had sufficient experience in rating the social aspects of the interlocutors' performance and were apparently unaware of the significance of these aspects in pragmatic assessment. The conclusion urges the implication that the raters should be instructed how to conduct pragmatic assessment by considering areas other than linguistic ones.

The findings also showed that the dialogic rating can melt down the differences between raters while considering the nonpragmatic criterion and pragmalinguisitic criteria. Such findings motivate the implication that dialogic assessment can help convergence of raters' idea in the assessment of domains with which they enough familiarity. It can be concluded that raters should gain sufficient knowledge of pragmatic competence and assessment for the emergence of an agreement on the application of sociopragmatic criteria in dialogic assessment. The implication is that in order for the raters to benefit from the advantages of dialogic assessment in pragmatic ratings, training programs and courses should be organized to improve and raise raters' knowledge and awareness of pragmatic competence and pragmatic assessment.

The present study had two limitations: The first was that both the raters and learners were from one language background, so the results cannot be generalized to raters and learners of other backgrounds. The second was that the study focused only on the raters' experience, so other rater variables that might have some influence on their performance were not considered in the study.

References

Alemi, M., & Tajeddin, Z. (2013). Pragmatic rating of L2 refusal: Criteria of native and nonnative English teachers. *TESL Canada Journal, 30*(7), 65-83.

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

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

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

Bergman, M. L., & Kasper, G. (1993). Perception and performance in native and nonnative apology. In G. Kasper & S. Blum-Kulka. (Eds.), *Interlanguage pragmatics* (pp. 82-107*)*. New York: Oxford University Press.

Billmyer, K., & Varghese, M., (2000). Investigating instrument-based pragmatic variability: Effects of enhancing discourse completion tests. *Applied Linguistics, 21*(4), 517-552.

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

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

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

Caban, H. L. (2003). Rater group bias in the speaking assessment of four L1 Japanese ESL students. *Working Papers in Second Language Studies, 21*(3), 1-44.

Chau, J. (2005). Effects of collaborative assessment on language development and learning. *The Language Learning Journal*, *32*(1), 27-37.

Cohen, A., & Olshtain, E. (1981). Developing a measure of sociocultural competence: The case of apology. *Language Learning, 31*(1), 113-134.

Cohen, A. D., & Shively, R. L. (2007). Acquisition of requests and apologies in Spanish and French: Impact of study abroad and strategy-building intervention. *The Modern Language Journal, 91*(2), 189-212.

Dickinson, L. (1988). Collaborative assessment: An interim account. In H. Holec (Ed.), *Autonomy and self-directed learning: Present fields of application* (pp. 121-128). Strasbourg, France: Council of Europe.

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. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, *25*(2), 155-185.

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

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: College Entrance Examination Board.

Ervin-Tripp, S. (1976). Is Sybil there? The structure of some American English directives. *Language in society, 5*(1), 25-66.

Galloway, V. B. (1980). Perceptions of the communicative efforts of American students of Spanish. *Modern Language Journal, 64*(4), 428-433.

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

Hudson, T. (2001). Indicators for pragmatic instruction. In K. R. Rose & G. Kasper (Eds.), *Pragmatics in language teaching* (pp. 283-300). Cambridge: Cambridge University.

Hudson,T., Detmer, E., & Brown, J., D. (1995). *Developing prototypic measures of cross-cultural pragmatics*. Honolulu, Hawai'i: University of Hawaii, Second Language Teaching and Curriculum Center.

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

Kasper, G., & Roever, C. (2005). Pragmatics in second language learning. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 317-334). Mahwah, NJ: Lawrence Erlbaum Associates.

Kim, Y-H. (2009). An investigation into native and nonnative teachers' judgments of oral English performance: A mixed methods approach. *Language Testing, 26*(2), 187-217.

Knoch, U., Read, J., & von Randow, J. (2007). Retraining 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 L2 writing performance. *Language Testing, 19*(1), 3-31.

Lee, H. K. (2009). Native and nonnative rater behavior in grading Korean students' English essays. *Asia-Pacific Education Review, 10*(3), 387-397.

Leech, G. (1983). *Principles of pragmatics*. London: Longman.

Li, D. (2000). The pragmatics of making requests in the L2 workplace: A case study of language socialization. *The Canadian Modern Language Review, 57*(1), 58-87.

Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing*, *28*(4), 543-560.

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

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

Matsumura, S. (2007). Exploring the aftereffects of study abroad on interlanguage pragmatic development. *Intercultural Pragmatics, 4*(2), 167-192.

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

McNamara, T., & Roever, C. (2006). *Language testing: The social dimension.* Malden, MA & Oxford: Blackwell.

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

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.

Roever, C. (2001). *A Web-based test of interlanguage pragmalinguistic knowledge: Speech acts, routines, and implicatures*. Unpublished doctoral dissertation, University of Hawai'i, Honolulu, Hawai'i.

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

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

Shohamy, E., Gordon, C., & Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *Modern Language Journal*, *76*(1), 27-33.

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing, 22*(1), 1-30.

Taguchi, N. (2006). Analysis of appropriateness in a speech act of request in L2 English. *Pragmatics, 16*(4), 513-535.

Taguchi, N. (2010). Longitudinal studies in interlanguage pragmatics. In A. Trosborg (Ed.), *Pragmatics across languages and cultures* (pp. 333-361). Berlin: Mouton de Gruyter.

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

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

Thomas, J. (1995). *Meaning in interaction: An introduction to pragmatics*. London: Longman.

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

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

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. (1994). Patterns of rater behaviour in the assessment of an oral interaction test. *Australian Review of Applied Linguistics*, *17*(2), 77-103.

Winke, P., Gass, S., & Myford, C. (2012). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing, 30*(2), 231-252.

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