Pragmatics 21:3.453-471 (2011)
International Pragmatics Association

RATER VARIATION IN THE ASSESSMENT OF SPEECH ACTS

Naoko Taguchi

Abstract

This study addresses variability among native speaker raters who evaluated pragmatic performance of learners of English as a foreign language. Using a five-point rating scale, four native English speakers of mixed cultural background (one African American, one Asian American, and two Australians) assessed the appropriateness of two types of speech acts (requests and opinions) produced by 48 Japanese EFL students. To explore norms and the reasoning behind the raters' assessment practice, individual introspective verbal interviews were conducted. Eight students' speech act productions (64 speech acts in total) were selected randomly, and the raters were asked to rate each speech act and then explain their rating decision. Interview data revealed similarities and differences in their use of pragmatic norms and social rules in evaluating appropriateness.

Keywords: Assessment; Pragmatic competence; Rater variation; Speech acts.

1. Introduction

Corresponding to the recent trend of internationalization and transculturalism, the concept of uniformed native-speakerness has been seriously challenged in second language (L2) teaching and assessment (Davies 2003). Researchers are increasingly cautious about using a native-speaker model as the basis for comparing L2 learners' performance or as the model of target behavior for learners to emulate. However, in the assessment of pragmatic abilities, native speaker norms, inferred through data collected from a sample of native-speaking participants, continuously serve as the criteria for evaluating L2 pragmatic behavior. In addition, native speakers are typically used to rate appropriateness of L2 pragmatic performance, and the literature is largely silent about raters' experience and attitudes that inevitably contribute to the norms they establish. Rater variation is rarely brought up due to the conventionalized practice of establishing interrater agreement.

This study challenges the current practice by addressing variability among native speaker raters when they evaluate L2 pragmatic performance. Using a five-point rating scale, four native speakers of English assessed appropriateness of speech acts of requests and opinions produced by 48 Japanese college students studying English. The raters formed a culturally mixed group: An Australian white male and female, an African-American male, and a Japanese-American female. To explore the norms and reasoning behind their assessment, individual retrospective interviews were conducted twice, each lasting about 40-50 minutes. Eight students' speech acts (a total 64 speech acts) were selected randomly, and the raters were asked to rate each speech act and then explain their rating decision. In addition to the interview data, norming sessions were recorded and analyzed. Data revealed similarities and differences among raters in their use of pragmatic norms and social rules in evaluating appropriateness. Variations among raters reflected their values and personal experience, and negotiation of these factors shaped the process of assessment.

2. Background

As situationally appropriate language use has become central to L2 communicative competence, much attention has been given to methods for examining and assessing pragmatic competence. Pragmatics refers to language use in relation to language users and language use situations (Levinson 1983; Mey 1993; Thomas 1995). Pragmatics is a study of “a dynamic process involving the negotiation of meaning between speaker and hearer, the context of utterance (physical, social, and linguistic) and the meaning potential of an utterance” (Thomas 1995: 2). In the field of L2 assessment, these definitions of pragmatics correspond with Bachman and Palmer's (1996) notion of pragmatic knowledge: The relationships between utterances, language users, and settings. Bachman and Palmer identified pragmatic competence with two sub-components: Functional knowledge (i.e., knowledge of conventions for performing language functions) and sociolinguistic knowledge (i.e., knowledge of appropriateness of these conventions in social contexts). Thus, pragmatic competence refers to the ability to evaluate contextual information and to perform language functions effectively and appropriately in social situations.

The definition of pragmatic competence has been incorporated into a number of assessment studies that operationalized pragmatic competence as a measurable construct and developed test instruments and tasks to elicit and examine the construct (e.g., Brown 2001; Cohen 1994; Cohen & Olshtain 1981; Hoffman-Hicks 1992; Hudson Detmer & Brown 1995; Liu 2007; Roever 2005; Sasaki 1998; Taguchi 2006; Walter 2007; Yamashita 1996, 2001; Yoshitake 1997). Among these studies, Hudson et al.'s (1995) test battery is probably most comprehensive in that it covers several prototypical testing methods by using the multi-method approach in assessing speech acts of requests, apologies, and refusals. Their test battery included six measures for assessing speech acts of L2 English learners: Oral and written discourse completion tests (DCT), baselines, multiple-choice tests, and two self-assessment measures. In DCTs and role-plays, participants were given a situational scenario that differed in contextual variables (i.e., power, social distance and degree of imposition) and produced a speech act according to the situation either orally or in writing. Native speaker raters rated each participant's speech act on a five-point rating scale for the following criteria: Ability to use the correct speech act, formulaic expressions, amount of speech used, information given, formality level, politeness level, and directness level. In multiple-choice tests, a situational scenario was displayed, followed by a list of speech act expressions that participants choose from. The self-assessment task involved participants evaluating their own speech act performance on a rating scale. Yamashita (1996) later adapted Hudson et al.'s test for learners of L2 Japanese.

More recently, Roever (2005) took a multi-construct approach and developed a web-based assessment battery measuring three pragmatic constructs: Comprehension of implicatures, comprehension of routines, and production of speech acts. The implicature section took a multiple-choice format to measure comprehension of two types of implicatures: Formulaic implicatures that were marked by conventional structures and idiosyncratic, non-conventional implicatures. The routine items also took a multiple-choice format and tested recognition of situational and functional routines. The speech act section had 12 written DCT items that elicited requests, refusals, and apologies, which were evaluated by native speaker raters based on a four-point scale, ranging from 'fail' to 'immaculately perfect.'

As exemplified above, previous research has provided a rich array of options for tools and methods in assessing pragmatic competence. Among the studies, speech acts have been the most popular target of assessment. Typical practice has been to elicit speech acts via DCTs or role-plays and evaluate them on a rating scale using trained native speaker raters. Descriptions of bands in rating scales reflect pragmatic-specific aspects of language use, focusing on areas such as tone, clarity of intention, level of formality, directness, and politeness, and use of strategies and semantic moves used to support speech acts. In addition to pragmatics, the scales address general command of language use, including typicality of expressions, amount of speech, coherence and organization, grammar, and word choice. Rating could be holistic, noting general impression of learners' performance encompassing all dimensions listed above, or analytic, showing breakdowns in learners’ performance in each of the dimensions.

While these existing practices assume that raters conform to pre-determined criteria of appropriateness in speech act performance instilled in norming sessions, very few studies have examined actual processes involved in norming and rating. In particular, raters' perceptions of and orientation toward pragmatic appropriateness has rarely been addressed. As a result, many questions remain unanswered. For example, how do raters interpret and internalize descriptions of rating rubrics? Do they bring their own criteria in determining appropriateness of pragmatic behaviors? Do they prioritize one dimension of pragmatic appropriateness over others, and is there variation in their orientation? These questions are particularly relevant for pragmatics due to the nature of pragmatics. Pragmatics involves linguistic behaviors that are reflective of values and norms of a given culture and addresses a wide range of elements - forms, functions, contexts, social relationships, and cultural conventions. Given this complexity, it is conceivable that raters' background, cultural experience, and personality greatly influence the standards they use to judge appropriateness. There might be rater leniency or bias with respect to certain aspects of pragmatic competence, and the level of leniency or bias may vary across raters, depending on their norms and practice of social interaction.

Despite these possibilities, raters' characteristics and behaviors have rarely been taken up in pragmatic assessment. Walter (2007) is probably the only study that examined rater variation in pragmatic assessment. In his study, 42 learners of English participated in a baseline activity with a native English-speaking tester for 10 to 15 minutes. The activity included three oral pragmatic prompts: An assessment, a compliment, and a pre-sequence, which were embedded within three larger topic-discussions. The prompts were delivered spontaneously after each topic discussion. For instance, after discussing challenges of cultural adjustment, the tester delivered an assessment, e.g., speaking a foreign language is hard work, to which an assessment response (i.e., agreement or disagreement) was expected. Two raters, a native and non-native speaker of English, evaluated the baselines based on a four-point holistic rating scale. Dialogues between the raters as they resolved differences in scoring were recorded and analyzed. The results showed that the raters interpreted examinees' performance differently, leading to different scoring decisions. For instance, in the pragmatic target of compliment responses, the native speaker rater relied on his knowledge of normative patterns of compliment and compliment response in American English, while the non-native speaker considered L1 transfer as possible source of non-normative compliment response. The non-native speaker rater was also influenced by the examinee's fluency and clear pronunciation, leading him to give a higher score.

Although Walter (2007) is the only existing study that addressed rater variation in pragmatics, there are a number of studies in other areas of assessment that investigated raters' perspectives and orientations (e.g., Brown 2000, 2003, 2005; Ducassee & Brown 2009; Johnson & Lim 2009; May 2006, 2009; McNamara & Lumley 1997; O'Loughlin 1996; Polit & Murray 1996). Using introspective verbal protocols, these studies examined how raters' characteristics - gender, language background, experience, and competence - affected their evaluation of L2 oral interviews, writing samples, and paired dyads. After rating learner's performance, raters were asked to summarize their reasons for awarding the rating. A review of verbal reports revealed aspects of learners' performance that raters focused on (e.g., linguistic features, discourse management, rhetorical organization, and listening behaviors). A general consensus drawn from this body of literature is that, even after training, raters bring their own values and criteria in assessment, and they adhere to both criterion and non-criterion features.

These findings imply that analysis of rater perspectives in pragmatic assessment is a worthwhile investigation because it could reveal the precise dimensions of pragmatic competence that raters consider, along with background beliefs that raters subconsciously bring to the task of rating. These analyses could in turn help us better understand what pragmatic competence entails and fine tune the rating scale used to measure appropriateness, formality and tone involved in speech acts.

In addition, the degree and nature of rater variation revealed in these analyses could prompt a re-examination of the notion of uniformed native speakerness. A common practice in pragmatic assessment has been to use a group of native speaker raters to evaluate speech acts, assuming that the raters operate under the same standards and conform to the group norms established during the training session. However, in reality, native speakers do not form a unitary category. There are a variety of native speakers within any one language or culture. They can come from different regional, educational, and professional backgrounds, all which inevitably influence the manner in which they project politeness and the criteria they use to judge appropriateness of language behaviors. Hence, it is possible that native speaker raters from different backgrounds and experiences evaluate pragmatic performances differently. There might be great variation among native speakers on what an acceptable or unacceptable answer would be in pragmatic performance. Because previous research is largely silent about raters' experiences and attitudes that may influence the norms used by raters, future research that addresses these issues could add to the literature.

3. Purpose of the study

The purpose of the study was to investigate native speaker raters' orientation when assessing appropriateness of speech acts in L2 English. The study examines what raters focus on when rating speech acts and what variations manifest in their rating decisions.

4. Methodology

4.1. *Test instrument*

This study developed a test instrument that assessed L2 English learners’ ability to produce speech acts. The notion of speech acts originates from Austin's (1962) claim that an utterance encodes a specific "act" or function that the speaker wants to achieve by producing the utterance. According to Austin, utterances have three kinds of meaning: Locutionary, illocutionary, and perlocutionary. The locutionary meaning refers to literal meaning of the utterance. If someone says "It's cold in here," the locutionary meaning is the low temperature of the room. The illocutionary meaning is the function of the utterance. The illocution of "It's cold in here" could be a request to shut the door. The last kind of meaning, perlocutionary meaning, is the intended effect of the utterance. If the listener of the utterance shuts the door, the perlocution of the utterance is observed. Thus, speech acts are purposeful: They have specific goals and are intended to have a specific effect on the listener (Clark 1979).

L2 learners’ speech acts were elicited through a computerized oral discourse completion test (ODCT). Participants read situational descriptions and produced speech acts accordingly. Two types of speech acts were elicited: Requests (*k*=4) and opinions (*k*=4). These types were both divided into two situation types: Low- and high-imposition situations. Low-imposition situations were informal situations in which the speaker addressed a person with the same power status. High-imposition situations were formal situations in which the speaker addressed a person with a higher power status. See Table 1 for the situations used in the ODCT.

The ODCT was computerized and administered to 48 Japanese students enrolled in the intensive ESL program in a university located in northern Japan. There were 16 males and 32 females, ranging in age from 18 to 21 with an average age of 18.3. They averaged 6.1 years of formal English education in Japan. Their entry TOEFL score was about 460 on average, ranging from 413 to 497. None of the students had studied abroad prior to data collection.

The ODCT was given individually three times over one academic year: Time 1 (April), Time 2 (July), and Time 3 (December). Students put on headsets with a microphone attached and read directions in English with Japanese translations. They were told to read each situational scenario and respond as if they were in a real situation and performing the given role. They had two practice items. Each item started with a situational scenario in English displayed on the computer screen. They read the scenario and prepare for the speech act. When they were ready, they clicked on the "continue" button and produced the speech act. The computer recorded their speech.

1. Low-imposition situations

Requests

1) You have a free writing task in class today, but you forgot to bring a pen. You want to borrow a pen from your friend in the class. He is sitting next to you. What do you say to him?

2) You and your friend are talking about your group presentation for tomorrow's class. Your friend said something about English class to you, but you didn't understand. What do you say to him?

Opinions

3) You are shopping with your friend. Your friend picked up a brown jacket and tried it on. You don't think he looks good in brown. He says, "What do you think?" What do you say to him?

4) Your friend asked you to check the first draft of her paper on the Japanese education system. The paper is well-written, but you think the introduction is too long. What do you say to him?

2. High-imposition situations

Requests

5) You have a small test in her class next Monday, but you have to go out of town that day because of your cousin's wedding. You want to take the test at some other time. What do you say to the professor?

6) Tomorrow is the due date of a paper for your history class. You caught a cold, and you've written only two pages so far. You want to ask for two extra days to finish. What do you say to the professor?

Opinions

7) Your professor gave you a mid-semester grade of C, but you don't think it's fair. You missed three classes, but you always turned in homework on time and got 80% on the test. You go to the professor's office to explain. What do you say?

8) You like the French professor, but she talks about French history most of the time and you are more interested in French pop culture. One day after class she says, "What do you think about the class?" What do you say to the professor?

Table 1: *Simplified situations used in the ODCT*

4.2. *Evaluation of speech acts*

Speech acts were evaluated on their overall appropriateness, which was defined as the ability to produce speech acts at the proper level of politeness, directness, and formality in the given situations. Appropriateness was assessed using a five-point rating scale ranging from 1 (very poor) to 5 (excellent). The sum of the ratings of the four low-imposition and four high-imposition speech acts were calculated.

5 Excellent
Almost perfectly appropriate and effective in the level of directness, politeness, and formality.

4 Good
Not perfect but adequately appropriate in the level of directness, politeness, and formality. Expressions are a little off from target-like, but pretty good.

3 Fair
Somewhat appropriate in the level of directness, politeness, and formality. Expressions are more direct or indirect than the situation requires.

2 Poor
Clearly inappropriate. Expressions sound almost rude or too demanding.

1 Very poor
Not sure if the target speech act is performed.

Table 2: *Rating scale*

Four native speakers of English evaluated the samples. They were asked to listen to the speech acts, along with the transcripts, and to indicate the rating based on the rating descriptions. Interrater reliability was *r*=.92. About 2% of the samples had two points off in rating. They were discussed in the follow-up meetings to reach a consensus. For the cases with one point off, the average score between the two raters was assigned as the final score.

4.3. *The raters*

The four raters formed a culturally -mixed group: An Australian white male and female, an African-American male, and a female Japanese-American. They had little background in Applied Linguistics or related fields and had limited experience in teaching English. Descriptions of the raters - Will, Erin, Nick, and Britney (pseudonyms) - are given below.

Will is an African-American male in his 30s. He is from a middle-class family with Jamaican parents. He graduated from a university in New York with a Bachelor of Arts in Political Science. He received a Masters of Management from a university in Arizona. He is married to a Japanese woman and has lived in Japan for six years. Language spoken at home is mostly English with occasional Japanese. He has taught English for six years in a college, a high school, and private language schools in Japan.

Erin is a Japanese-American female in her 20s. She received a Bachelor's degree in journalism from a university in Connecticut. She has a Japanese mother who had her after she moved to the U.S.A. with her American husband who used to work for the military. They divorced when Erin was five. Her mother rarely spoke Japanese at home. Erin studied Japanese in high school and college. Her self-assessed Japanese ability is intermediate. She can do basic reading and writing in Japanese, but speaking is still difficult for her. Erin has lived in Japan for three months, teaching English in language schools.

Nick is an Australian, Caucasian male in his 30s. He grew up in an English-speaking family of high socio-economic status in a high-income area in Sydney. He earned a Bachelor's degree in law and Japanese in an Australian university. He worked as assistant executive in a travel agency and lived in Europe for nine months. He considers his Japanese reading comprehension as high-level because he can read employment contracts and leases in Japanese. He said that his spoken Japanese tends to be very simple because he hasn't had much speaking practice. He has lived in Japan for 10 months, teaching English in elementary schools and private language schools.

Britney is an Australian, Caucasian female in her 20s. She received a college certificate in fitness from an Australian university. She worked seven years in a gym in Australia and participated in summer camps for six months in the U.S.A. Two of her brothers are married to Japanese women, and she speaks Japanese with them sometimes. She studied Japanese in a private school for one year, and she is able to manage daily conversation in Japanese. She has lived in Japan for one year, teaching English in elementary schools and private language schools.

4.4. *Data collection and analysis*

Primary data for this study was individual interviews with raters. The interviews took a format of introspective verbal protocols to gain insight into the rating activity and raters' orientation toward pragmatic aspects. Sixty-four speech acts (32 requests and 32 opinions) were randomly selected from the samples. Raters evaluated each speech act using a five-point rating scale (Table 2) and then explained their rationale for each rating. Interviews were recorded and transcribed. In addition to the interview data, norming sessions were analyzed. There were two norming sessions, each lasting for two to three hours, in which the raters and the researcher together evaluated about 40 speech acts from the samples. Their conversations and discussions about rating decisions were recorded and transcribed.

Interview transcriptions were examined carefully to identify units of analysis and coding categories. Following Green (1997), a unit of analysis was defined as a single or several utterances within a single aspect of the event at the focus. Repetitions and elaborations were not recorded as new units. As shown in sample protocols below, one segment is divided into idea units with '/':

Low-imposition request (asking a friend for a pen)

> "Excuse me. I forgot my pen. Please give me a pen."
> Rater's comment:
> "Appropriateness score is 4. ‘Excuse me’ is overdone / and ‘Please give me a pen’ is kind of direct statement, instead of asking for a pen.”

High-imposition request (asking a teacher to reschedule a test)

> "Sorry, I forgot other appointment, so can I take a test some other time?"
> Rater's comment:
> "Appropriateness score is 3, not quite 4. / 'Sorry' is polite beginning. / There is a reason, but the reason is a bit too vague. / And request could be more positively worded. Saying 'could I' is better than 'can I.'"/

The next step was to code the idea units. Defining categories involves repeated data reduction as the researcher cycles through the data with rearranging and recoding multiple times. Following Green (1997), multiple categories were created at the discovery stage. These were then grouped by theme after more coding. This procedure was repeated until five dominant categories emerged. The following is the list of these categories with a sample of rater's comments.

(1) Amount of speech — Whether the amount of speech is appropriate
e.g., "Saying 'excuse me' is overdone."

(2) Clarity of intention — Whether the intention is communicated clearly
e.g., "The student said 'Could you borrow a pen?' Not clear what he meant."

(3) Directness — Whether expressions are at proper level of directness
e.g., "The request is a bit too direct. 'Could you please' could take it to 5, but 'Please change' is 3."

(4) Strategies — Whether the speech act contains supporting strategies
e.g., "The student says 'I promise I will do a better job.', which is a good thing, which is a teacher wants to hear."
e.g., "There is no clear reason stated to support the request. The reason is too vague."

(5) Politeness markers — Whether the speech act contains politeness markers
e.g., "There are no politeness markers, like 'please' or 'excuse me.'"

(6) Content — Whether the speech act contains valid content
e.g., "Asking a professor to change the whole date of the test sounds kind of inappropriate."

The researcher coded the full data set using these categories. Another coder independently coded 20% of the data. Interrater agreement was 85%. The disagreements in the double-coded data were discussed and resolved to arrive at the agreed set of ideas unit and coding categories. After the coding was complete, frequency of raters' comments on each category was tallied and compared across the four raters. See Table 3 for the data and data analysis methods used in this study.

Test data	Verbal report data	Verbal report analysis
Recordings and transcriptions of 64 speech acts produced by eight EFL learners in an oral DCT task	Four raters individually rate speech acts and verbalize their rating decisions	Transcription of verbal reports from four raters. Content analysis of transcription to develop coding grid.

Table 3: *Data collection and analysis*

5. Results

Table 4 displays interrater reliabilities (Pearson correlations) of the four raters based on the 64 speech acts used for the introspective verbal interviews. Out of the 64 speech acts, 18 were in complete agreement, 11 were two bands off, and the remaining 35 were one band off. Will and Britney had the lowest correlation, *r*=.65. Will and Erin, two Americans in the group, as well as Nick and Britney, two Australians, had the highest correlation, *r*=.80 or above, indicating that the same nationality might have contributed to the higher agreement in rating.

	Will	Erin	Nick	Britney
Will	---	.80	.74	.65
Erin	---	---	.73	.70
Nick	---	---	---	.81
Britney	---	---	---	---

Table 4: *Interrater reliabilities of speech acts rating (k=64)*

Table 5 displays average ratings of the four raters. Will and Britney showed the largest discrepancy in their rating. Will tended to be more lenient with low-imposition speech acts than high-imposition speech acts, while the pattern was opposite for Britney: Compared with other three raters, her rating was more severe for low-imposition speech acts than it was for high-imposition acts.

	Will	Erin	Nick	Britney
Low-imposition speech acts				
Request (*k*=16)	3.68	3.56	3.56	3.38
Opinion (*k*=16)	4.31	4.06	4.25	3.94
High-imposition speech acts				
Request (*k*=16)	2.69	2.88	2.94	3.13
Opinion (*k*=16)	2.56	2.94	3.19	2.88

Note. The rating is based on a five-point scale.

Table 5: *Mean ratings of speech acts (k=64)*

The main purpose of this study was to investigate raters' orientation toward the evaluation of speech acts, in particular aspects of learners' performance that the raters focused on during rating and sources of variation in their rating decisions. Tables 6 through 9 present the areas of speech acts that the raters commented on while rating individual speech acts, as well as frequency of their comments on each area. As shown in Table 6, evaluation of low-imposition requests (i.e., asking a friend for a pen; asking a friend to repeat what he/she has said) yielded comments on two categories: Amount of speech and clarity. All four raters made reference to students' excessive verbosity in their requests and lowered their rating when students said more than necessary. Small, low-stake requests such as these require little face-work. As a result, the expressions used to perform these speech acts should be short, simple, and to the point, without much elaboration or lengthy explanations to alleviate the illocutionary force. Hence, the students' overuse of strategies and lengthy speech act were of concern for the raters.

	Will	Erin	Nick	Britney
Amount of speech	6	11	11	10
Clarity	3	2	2	1

Table 6: *Frequency of raters' comments, low-imposition requests (k=16)*

However, there was variation among raters in their level of tolerance toward lengthy low-imposition requests. Frequency of comment on this area was much greater for Erin, Nick, and Britney than for Will. The following interview excerpts illustrate this variation. See the raters' comments on a low-imposition request:

> Low-imposition request (asking a friend to repeat), Student #16
> "Ah, sorry, I beg your pardon? Please one more time, what you said."
> Britney:
> "Appropriateness is 3. To a close friend, apologizing and saying 'pardon' and 'please' are just too much. Overdone."
> Erin:
> "It's 4. It's a little overly polite. 'Beg your pardon' is a common way to say 'excuse me' or 'what did you say?' but 'I'm sorry' and 'Please beg your pardon' are a little too much.
> Nick:
> "For my perspective, 'Sorry I beg you pardon' is more than enough. Politeness and explanation more than required. To attempt to elaborate on that is unnecessary, long winded, so I'd say 3."

While these three raters gave a score of three or four because of the excessive verbosity, Will gave a perfect score of five on this speech act because the verbosity did not bother him as much. This excerpt below shows that Will was re-negotiating between his norms of appropriateness and those established by the group during the norming sessions.

> Will:
> "This is 5. 'Pardon? Say that again?' That's how I kind of think someone would say that, which, to me, OK, I'd say it again. Again, this is something I did agree to that, this over excessive 'please' and stuff might be too much, that's why I'd give 4, but I personally don't really think it's extreme thing. If I'm speaking to someone, maybe, if someone rolled that out of the tong - 'I'm sorry, pardon, could you say that again, please', I don't really think it's a problem. According to our rating, if I take out the sheet, that would be 4, because too many 'please'."

Below is another example of raters' variation. While Britney, Erin, and Nick gave a lower score of three because of the student's excessive use of "please" and "excuse me," Will gave a higher score of four indicating that he would say the same thing in the given situation.

> Low-imposition request (asking a friend for a pen), Student #44
> "Excuse me Ken, I forgot my pen, please borrow your pen, please."
> Britney:
> "Appropriateness is 3. There are so many "please" and "excuse me."
> Erin:
> "This is overly polite. For just borrowing a pen, it's too much, so I'd give 3 for appropriateness."

> Nick:
> “Appropriateness is high 3. It sounds strangely needy, so perhaps the person is desperate, so in that case second 'please' is justifiable, but just a hint of desperation for something seemingly trivial.”
> Will:
> “For appropriateness, I’d give 4. I think it's you're in class, a little explanation, “Oh, Ken, I forgot my pen. Could you lend me a pen?” That's what I would do, personally, so . . . “

Raters' variation was also observed in their rating of low-imposition opinions (i.e., expressing negative opinion about a friend's clothes and papers). See Table 7.

	Will	Erin	Nick	Britney
Directness	1	4	0	3
Strategies	3	5	21	7

Table 7: *Frequency of raters' comments, low-imposition opinions (k=16)*

As shown here, Nick made reference to the use of strategies far more often than other raters. His rating decision was based on whether or not the students gave a useful suggestion when criticizing his/her friend, while other raters did not seem to consider a suggestion as a necessary element. Below is an example of a speech act followed by Nick's comment.

> Low-imposition opinion (expressing opinion about a friend's clothes), Student #12:
> "It doesn't suit you, Jeff, I think."
> Nick:
> "Appropriateness is 3. Because no assuring comment, but no helpful suggestion either, which is more problematic for me, lack of helpful suggestion from a friend."

Tables 8 and 9 display frequency of raters' comments for high-imposition speech acts. Compared with low-imposition speech acts, high-imposition speech acts elicited a greater variety of comments, indicating that the raters were considering wide-ranging aspects of speech acts when scoring them.

	Will	Erin	Nick	Britney
Politeness markers	4	10	8	11
Directness	10	14	11	12
Strategies	11	7	8	1
Content	5	1	0	0

Table 8: *Frequency of raters' comments, high-imposition requests (k=16)*

	Will	Erin	Nick	Britney
Politeness markers	3	3	7	2
Directness	9	9	12	14
Strategies	10	9	10	8
Content	4	2	5	0

Table 9: *Frequency of raters' comments, high-imposition opinions (k=16)*

Rater variation for high-imposition requests was found in Will's evaluation patterns. He was less attentive to politeness markers (e.g., expressions of "please" and "excuse me") compared with other raters, but was more keen on the content of request - what the student asked for, rather than how he/she asked. The example below illustrates this point.

> High-imposition request (asking a professor to re-schedule a test), Student #19
> "Excuse me professor. I have test next Friday, but I have doctor's appointment too. I cannot go to the doctor another day. Please change the day of the test."
> Will:
> "He said 'Please change the day of the test.' Asking to change the date because of your doctor's appointment is a bit inappropriate."

As shown here, Will repeated the student's request but made no comment about the politeness marker "please" that appeared in the request. Instead, he commented solely on the appropriateness of requests' content. Will's priority on content over linguistic features in his rating decision contrasts with other raters' orientation toward linguistic politeness. See Erin's comment on the same student's speech act below. Here, she points out that the syntactic form used for the head act, "please + verb", is a direct statement, not a request, but politeness markers that come with the head act mitigate the directness.

> Erin:
> "I'll give 3 on this. He said 'Please change the date.' It's not asking, it's demanding, but at least he says 'excuse me' and 'please', so it's polite."

Like Erin, Britney mentioned that the request sounded too demanding because of the linguistic form used in the request (i.e., “Please” + verb). However, she gave a score of four because the form was not as direct as the form of “I want,” which is an expression of direct wants and wishes.

> Britney:
> “There is "excuse me" but the utterance "Please give me another two days" sounds a little demanding. But it’s not "I want," so I’d give it a 4.”

Among the four raters, Britney was most sensitive to the linguistic features of speech acts when determining appropriateness of high-imposition speech acts. As shown in Tables 8 and 9, she made frequent comments on politeness markers and the directness level of the request expressions, and she rarely verbalized semantic strategies or content of requests. This tendency was found in the norming session as well. See this excerpt of a discussion among raters that took place during the second norming session regarding a high-imposition opinion speech act:

> High-imposition opinion (expressing an opinion to a teacher about a grade)
> Student #23
> "Excuse me I have a question about my grade of C because I got the 80% of test, and I always spoke up in the class, but I don't know why it happened. I think it's unfair. So could you tell me what happened to my grade?"

1 Researcher: So Britney, why did you give a score of four on this?

2 Britney: Well I'm saying "excuse me" on a question. Ummm, it's very polite. "Could you tell me" things like that. So I gave it a 4. It's quite polite. It's not that demanding. The "unfair" comment wasn't very demanding.

3 Researcher: Others gave it a 3. What's your comment?

4 Erin: "I think it's unfair."

5 Will: Yeah the "unfair" thing. So she states her positive reasons. "I got 90%, I always spoke up in class, I don't know why it happened ..."

6 Nick: It's inappropriate choice of words.

7 Will: "So could you tell me what happened to my grade... So could you tell me why this happened…I don't know why it happened." I think it's kind of inappropriate to come to your professor that way.

8 Britney: But it's quite polite with "excuse me I have a question."

9 Will: Which is, why it's a 3 and not lower.

10 Erin: I think we already talked about that even though you use "I'm sorry" or "excuse me" it can still sound rude.

11 Britney: But so "could you."

In line 2, Britney said that she gave a score of four because the student used the politeness marker "please" and indirect request "could you". Other raters, however, gave a score of three because the student directly said to the professor that the grade was unfair (lines 4, 5 and 6). Nick noted the inappropriate choice of the word "unfair" (line 6). Following Nick, Erin said that the politeness markers do not compensate for the rudeness in the content of the speech (line 10). However, in lines 8 and 11, Britney still insisted that politeness markers were salient features of appropriateness.

Below is another example that illustrates Britney's focus on linguistic forms rather than on content. In line 4, she explained that she gave a score of four for this student's speech act because the expression "Please tell me" was polite. However, other raters (Nick and Erin) disagreed with Britney, claiming that the student did not provide adequate explanation to back up his complaint about the grade (lines 8 and 9).

> High-imposition opinion (expressing an opinion to a teacher about a grade)
> Student #24
> "Ah, Dr. Paker, I think, ah, why I got the C grade? I think I did my best, so please tell me why I got C, so . . . "

1 Nick: So Britney, why 4?

2 Britney: For grammar?

3 Nick: Ahh . . . appropriateness.

4 Britney: "Please tell me." I thought it was quite polite, "So please tell me..." "So please tell me why I got C."

5 Nick: And a few people gave a 3.

6 Erin: Yeah, he didn't give any explanation to back up himself, like, "Why

did I get the C. I didn't think I did my best, and please tell me why."
7 Nick: "Doing your best" isn't really . . .
8 Erin: Really. "I did my best teaching you so I should get an A." (laugh) Yeah it didn't mention about the 80% on the test, the speaking up in class, missing class and skipping homework. Just, "I think I did my best." It didn't give enough substance, I don't think.
9 Nick: So Britney, if there is no explanation it goes down. It's a three.

As shown in the examples above, raters' priorities sometimes conflicted. The main difficulty in the rating process was to reconcile multiple aspects of a complex speech act production and decide on one score. The raters demonstrated different approaches to this process. Some raters assessed speech samples holistically by weighing all the elements, and comparing and contrasting them to form a general impression of learners' performance, while others took a primary-trait assessment approach and focused on one specific dimension of performance to decide on a score. As shown below, Nick exemplified the former approach by making reference to multiple features in his evaluation decisions: Politeness markers (e.g., "excuse me"), use of address term with title, use of softener "in fact," directness level of the head act, sufficiency of reason, and use of positive politeness strategies (i.e., "I'll promise I will turn in an excellent work.").

> High-imposition request (asking a teacher for an extension of an assignment)
> Student #44
> "Excuse me professor, ah . . . please… my, my work haven't finished yet, so, could you, could you give me more time? Ah, I firmly promise to finish more excellent work."
> Nick:
> "Appropriateness is 4 because there is a lot of polite language, such as 'excuse me' and as an introduction to the professor using a title, and 'in fact' is a softener, in this case. And there is no problem with directness. But the reason is vague, why the work hasn't finished yet. 'Could you give me more time?' is a polite request, and I think the promising to do an excellent work with more time is a nice touch."

Erin, on the other hand, acknowledged the polite language use (i.e., "excuse me" and "could you"), but she prioritized sufficient amount of reason over linguistic politeness. She gave a score of two because the speech act was lacking an adequate justification for the request.

> Erin:
> "Appropriateness is 2. Even though she says 'excuse me' and 'could you give me more time' it doesn't give a reason why. If there is no specific reason, as a professor, I wouldn't give students any time, if there is no reason."

See below for another example. Nick attended a variety of aspects of the speech act (i.e., the politeness marker "I'm sorry," the vague reason for the request, and the syntactic form of the request) to arrive at a score of three. Although Erin gave the same score, she based her decision solely on the syntactic form of the request "Can I take the test another time?"

High-imposition request (asking a teacher to reschedule a test)
Student #44
“Sorry, I forgot, ah, other appointment at your, your test class, so can I take a test some another time?”
Nick:
“Appropriateness is 3. Not quite 4. 'Sorry' is polite beginning, and there is a reason, but the reason is a bit too vague. And request could be more politely worded. ‘Please, could I take’ is better than ‘can I’”.
Erin:
“Appropriateness is 3 because it’s giving a request by saying ‘Can I take the test another time?’”

In summary, these interview excerpts illustrate that rater variation often appears when raters prioritize aspects of performance differently when assigning one score. When making a decision, raters often place the situation at hand in the context of their own personal experience, verbalizing how they would feel or what utterance they would find useful if they were in the same situation.

6. Summary and conclusion

Introspective verbal protocols revealed divergent focus of the four raters when evaluating appropriateness of speech acts. Some raters were more focused on linguistics forms such as the directness level of expressions or the use of politeness markers, while others based their scoring decision on non-linguistics aspects such as the use of positive/negative politeness strategies and semantic moves as well as the content of speech. Yet other raters still incorporated additional, unique features that they felt were salient into the evaluation criteria (e.g., whether or not a student provided a useful suggestion when criticizing his/her friend). Even when focused on the same dimension, the raters differed in their degree of acceptance. For instance, all raters considered excessive verbosity as problematic in a small request, but they had different criteria for determining how much is too much. These variations revealed in the data suggest that the raters consider a variety of dimensions when they evaluate speech acts. Raters also base some of their assessment decisions in their own personal experiences.

The present findings have several implications. First, the study found that native speaker raters do not form a unitary category. They can vary widely in their perceptions and interpretations of appropriateness, politeness, and formality in pragmatic performance because they come from cultures that have very different community norms for social interaction and communicative events. This variation is not solely due to speakers’ regional backgrounds, having more to do with their own individual frames of reference: As shown in this study, raters from the same geographical area (i.e., Australia and North America) still revealed different orientations in judging appropriateness. All four raters differed in what they considered to be salient features of pragmatic appropriateness.

Second, the rater variation found in the present data tells us which dimensions of pragmatic performance raters actually heed and how they reach their rating decisions. As was evident in the excerpts of conversations among raters, it seems that native speakers’ norms are constantly changing through discussion and negotiation during the process of rating. Test norms can be open and emergent during the process of rating itself. Hence, the pragmatic norms in assessment might be refined by analyzing negotiation among the raters. Raters' spontaneous comments and feedback that emerge during the evaluation process could cyclically feed into the development of more fine-tuned rating scales and test specifications.

Third, with regard to research methodology, the present study triangulated data and combined the analysis of verbal protocols with analysis of dialogues among raters during their norming sessions. However, the inherent subjectivity of the analysis of verbal protocols, in terms of deciding on idea units and the coding of the protocols, is an area of concern. It is possible that raters subconsciously tailored their comments to meet the perceived expectations of the researcher and did not produce a report that reflected their response to the original performance. In order to increase the validity of the data analysis and the conclusions drawn in the analysis, verbal protocols could be supplemented with interviews with raters after the production of verbal reports.

Finally, this study is limited in that it used only four raters. With such a small number of raters, it is difficult to detect the norms that the raters have chosen in their task of rating. As a result, future research should include more raters in order to yield more reliable findings.

References

Austin, J.L. (1962) *How to do things with words.* Cambridge: Harvard University Press.

Bachman, Lyle, and Adrian Palmer (1996) *Language Testing in Practice: Designing and Developing Useful Language Tests.* Oxford: Oxford University Press.

Brown, Annie (2000) An investigation of rater's orientation in awarding scores in the IELTS interview. In R. Tulloch (ed.), *IELTS Research Report* 3: 49-84.

Brown, Annie (2003) Interviewer variation and the co-construction of speaking proficiency. *Language Testing* 20: 1-25.

Brown, Annie (2005) *Interviewer Variability in Oral Proficiency Interviews*. Frankfurt am Main: Peter Lang.

Brown, James D. (2001) Six types of pragmatics tests in two different contexts. In K. Rose & G. Kasper (eds.), *Pragmatics in Language Teaching*. Cambridge: Cambridge University Press, pp. 301-325.

Clark, H.H. (1979) Responding to indirect speech acts. *Cognitive Psychology* 11*:* 430-477.

Cohen, Andrew (1994) *Assessing Language Ability in the Classroom.* Rowley, MS: Newbury House.

Cohen, Andrew, and Elite Olshtain (1981) Developing a measure of socio-cultural competence: The case of apology. *Language Learning* 31: 113-134.

Davies, Alan (2003) The native speaker in applied linguistics. In A. Davies & K. Elder (eds.), *Handbook of Applied Linguistics*. New York: Blackwell, pp. 431-450.

Ducassee, Ana M. (2009) Assessing paired orals: Raters' orientation to interaction. *Language Testing* 26: 423-443.

Green, A. (1998) *Verbal Protocol Analysis in Language Testing Research: A Handbook (vol.5).* Cambridge: Cambridge University Press.

Hoffman-Hicks, Sheila (1992) Linguistic and pragmatic competence: Their relationship in the overall competence of the language learner. In L.F. Bouton & Y. Kachru (eds.), *Pragmatics and Language Learning Monograph Series Vol.3*. Urbana-Champaign, IL: University of Illinois, pp. 66-80.

Hudson, Thom, Emily Detmer, and James D. Brown (1995) *Developing Prototypic Measures of Cross-Cultural Pragmatics* (Technical Report No.7). Honolulu, HI: University of Hawai'I at Manoa, Second Language Teaching & Curriculum Center.

Johnson, Jeff, and Gad Lim (2009) The influence of rater language background on writing performance assessment. *Language Testing* 26: 485-505.

Levinson, Stephen (1983) *Pragmatics*. Cambridge: Cambridge University Press.

Liu, Jianda (2006) *Measuring Interlanguage Pragmatic Knowledge of EFL Learners*. Frankfurt: Peter Lang.

May, Lyn (2007) Interaction in a paired speaking test: The rater's perspective. Ph.D. dissertation: The University of Melbourne.

May, Lyn (2009) Co-constructed interaction in a paired speaking test: The rater's perspective. *Language Testing* 26: 397-421.

McNamara, Tim, and Tom Lumley (1997) The effect of interlocutor and assessment mode variables in overseas assessments of speaking skills in occupational settings. *Language Testing* 14: 140-156.

Mey, Jacob (1993) *Pragmatics: An Introduction.* Oxford: Blackwell.

Roever, Carston (2005) *Testing EFL Pragmatics*. Frankfurt: Peter Lang.

O'Loughlin, Kieran (2002) The impact of gender in oral proficiency testing. *Language Testing* 19: 169-192.

Poliitt, Alastair, and Neil Murray (1996) What do raters really pay attention to? In M. Milanovic & N. Saville (eds.), *Performance Testing Cognition and Assessment: Selected papers from the 15th Language Testing Research Colloquium*. Cambridge: Cambridge University Press, pp. 74-91.

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

Thomas, Jenny (1995) *Meaning in Interaction: An Introduction to Pragmatics*. London: Longman.

Sasaki, Miyuki (1998) Investigating EFL students' production of speech acts: A comparison of production questionnaires and role-plays. *Journal of Pragmatics* 30: 457-484.

Walter, Scott (2007) A conversation-analytic hermeneutic rating protocol to assess L2 oral pragmatic competence. *Language Testing* 24: 155-183.

Yamashita, Sayoko (1996) *Six Measures of JSL Pragmatics* (Technical report #14). Honolulu, HI: University of Hawai'i at Manoa, Second Language Teaching & Curriculum Center.

Yamashita, Sayoko (2001) Using pictures for research in pragmatics-eliciting pragmatic strategies by picture response tests. In T. Hudson & J.D. Brown (eds.), *A Focus on Language Test Development: Expanding the Language Proficiency Construct Across a Variety of Tests*. (Technical Report #21). University of Hawaii, Second
language Teaching and Curriculum Center, pp. 35-56.

Yoshitake, S. (1997) Measuring interlanguage pragmatic competence of Japanese students of English as a foreign language: A multi-test framework evaluation. Unpublished Ph.D. thesis, Columbia Pacific University, Novata, CA.