Papers in Language Testing and Assessment, 2018, Volume 7, Issue 1, 32-60

Rater variability across examinees and rating criteria in paired speaking assessment

Soo Jung Youn
Northern Arizona University

This study investigates rater variability with regard to examinees' levels and rating criteria in paired speaking assessment. 12 raters completed rater training and scored 102 examinees' paired speaking performances using analytical rating criteria that reflect various features of paired speaking performance. The raters were fairly consistent in their overall ratings, but differed in their severity. The bias analyses using many-facet Rasch measurement revealed that a higher level of rater bias interaction was found for the rating criteria compared to those of the examinees' levels and the pairing type which reflects a level difference between two examinees. In particular, the most challenging rating category *Language Use* attracted significant bias interactions. However, the raters did not display more frequent bias interactions based on the interaction-specific rating categories, such as *Engaging with Interaction* and *Turn Organization*. Furthermore, the raters tended to reverse their severity patterns across the rating categories. In the rater and examinee bias interactions, the raters tended to show more frequent bias toward the low-level examinees. However, no significant rater bias was found based on the pairing type that consisted of high-level and low-level examinees. These findings have implications for rater training in paired speaking assessment.

Keywords: rater variability, paired speaking assessment, bias analysis, FACETS, rating criteria

One of the distinctive advantages of paired speaking assessment is its increased construct representation of speaking ability by eliciting various features of interactional competence (Taylor & Wigglesworth, 2009). In paired speaking assessment, an examinee can demonstrate various interactional skills while interacting with another examinee,

Email address for correspondence: Soo-Jung.Youn@nau.edu

such as initiating conversations, developing given topics, and providing listener support. At the same time, co-participants' contributions to paired speaking interaction are often inseparable as participants construct meaning together (Hall, 1995; Kramsch, 1986). These characteristics of paired speaking assessment present challenges to operationalizing and assessing interactional competence (e.g., McNamara, 1997; Chalhoub-Deville, 2003; Young, 2002). In response to these challenges, in-depth qualitative analyses of paired speaking interaction employing Conversation Analysis (CA) have enabled an improved understanding of the construct being measured (e.g., Galaczi, 2008, 2014; Ross & Kasper, 2013) and raters' perceptions of the interactional patterns found in paired speaking assessment (e.g., Ducasse & Brown, 2009; May, 2009, 2011). Compared to the qualitative studies on paired speaking assessment, what is relatively unknown to date is a quantitative account of ways in which raters maintain their severity across examinees' levels and rating criteria when scoring paired speaking performances, with the exception of a few notable studies (e.g., Wigglesworth, 1993). Thus, in this study I investigate rater variability across rating criteria, examinees' levels, and a pairing type determined by two examinees' level difference in paired speaking assessment using many-facet Rasch measurement. The literature review starts with the discussion of validity issues in paired speaking assessment, followed by current research on rater variability in speaking assessment.

Toward valid score interpretation in paired speaking assessment

A large body of paired speaking assessment research has focused on the qualitative analyses of paired speaking discourse to examine interactional patterns and the effect of pairing two examinees in eliciting paired speaking performance. For example, a recent study by Galaczi (2014) clearly illustrates the nature of paired speaking interaction. Galaczi reported descriptive interactional features representative of a varying level of interactional competence using CA. The interactional features that distinguished high-level examinees from low-level examinees include the degree of mutuality between examinees, as shown in the amount of turn-initiation questions and topic expansions, and listener support via backchannelling and confirmation of comprehension.

Considering various interactional features elicited in paired speaking assessment, one of the validity issues at stake is how raters perceive paired speaking interaction and employ rating criteria in awarding a score accurately (Taylor & Wigglesworth, 2009). Previous research on group and paired speaking assessment offers reasonable evidence of potential sources of rater disagreement and what influences scoring decisions. In the group speaking assessment context, raters' unstable performance has been already

reported. For example, Bonk and Ockey (2003) reported that raters had difficulty in awarding scores consistently for the *communication skills* category that included various interactional features, compared to *grammar* and *fluency* categories. Furthermore, social factors, such as group dynamics and interlocutors in group speaking tasks explained the large amount of error that was found when raters assigned scores for examinees (Van Moere, 2006). In the paired speaking assessment context, previous studies suggest that paired interactional patterns affect raters' scoring decisions. For example, May (2009) reported that raters had difficulty in reaching an agreement when paired interactional patterns were asymmetric, especially when one examinee deliberately dominates another examinee. May (2011) further investigated the features of paired speaking interaction that were salient to the raters by analyzing rater notes, stimulated verbal recalls, and rater discussions. The sources of rater disagreement included the lack of intelligibility of examinees' contributions, the degree to which examinees understood their partners' contributions, and the lack of examinees' responses during interaction. Further, the raters highly valued collaborative and authentic interaction, compared to asymmetric patterns of interaction. Ducasse and Brown (2009) reported that the presence of non-verbal interpersonal communication, such as gaze and body language, influenced rater judgments as well. Other studies suggest that proficiency differences between two examinees in a pair results in differing interactional patterns, which can be another source that influences rater judgements (e.g., Galaczi, 2008). Davis's (2009) study on the influence of interlocutor proficiency in paired speaking assessment reported that lower-level examinees paired with higher-level partners produced more words, although the interlocutor proficiency had no noticeable effect on examinees' ability estimates.

Taken together, the paired speaking assessment literature suggests potential factors that influence rater performance and thus require further research. Namely, the differing interactional patterns elicited in each pair can generate a context in which raters assign scores inconsistently for various features of paired speaking performance. Asymmetric paired speaking interaction might lead to inconsistency or differential severity in rater judgments or raters can face greater challenges toward comprehending the interactional features described in rating criteria. Thus, remaining questions include whether raters consistently and accurately connect the features of paired speaking performance conforming to rating criteria descriptions and whether raters display differential severity toward a specific pairing type. The raters' perceptions of paired speaking performances discussed in previous research which mainly took qualitative approaches need to be connected with quantitative accounts of rater performance, which is the focus of the current study. The next section discusses quantitative approaches to explicating the

sources of rater variability to review how raters interact with various facets in speaking assessment.

Rater variability in speaking assessment

Central to performance-based language assessment, observed scores heavily rely on rater judgments of examinees' performances in relation to descriptions in rating criteria. Naturally, the importance of rater training in performance assessment in ensuring valid score interpretation has long been recognized in the field of language testing. While rater training can increase rater consistency and accuracy, researchers have reported unexplained variability in trained raters' rating behaviors (McNamara, 1996). Among various rater variability issues, rater bias is characterized as the different degree to which raters interact with various factors involved in performance assessment, such as examinees' characteristics or certain rating criteria, apart from being different in terms of the degree of rater severity and leniency (Eckes, 2011; McNamara, 1996). Many-facet Rasch measurement (MFRM), particularly its bias analysis function, has been widely used in rater variability research. The bias analysis examines the extent to which a score provided by a particular rater for a particular rating category, for example, was higher or lower than expected, given the raters' severity and rating category's difficulty estimate (Myford & Wolfe, 2003). Wigglesworth (1993) has long argued that rater bias research is suitable for examining systematic patterns of rater behavior, reporting the reduced rater bias interactions with a specific rating category after rater training and feedback in speaking assessment. Since then, raters' bias information has been integrated into rater training to monitor rater performance (e.g., Elder, Barkhuizen, Knoch, & von Randow, 2007; Knoch, 2011). However, attempts to reduce rater bias through training have met with mixed results. For example, Knoch's (2011) longitudinal study reported that the effect of individual feedback during rater training varied across raters.

Various factors that influence rater variability have been researched in speaking assessment, such as examinees' gender and level (Eckes, 2005), rating experience (Isaacs & Thomson, 2013), rating context (Lumley & McNamara, 1995), task types and rating criteria (Wigglesworth, 1993), raters' first language (L1) (Kim, 2009), and raters' familiarity with examinees' L2 (Carey, Mannel, & Dunn, 2010; Winke, Gass, & Myford, 2013; Yan, 2014). Among them, the close relationship between rating criteria and rater variability needs to be noted. As established in the literature on rater training, well-developed rating criteria play a more critical role than raters' experience (Barkaoui, 2010; Knoch, 2009; Xi & Mollaun, 2011) and raters often internalize the performance level descriptors differently (e.g., Kim, 2015). Among various explanations offered toward

ways in which raters display variability toward rating criteria, Eckes's rater type hypothesis (2008, 2012) provides convincing accounts. Eckes argues that raters are quite fixed in their use of rating criteria, indicative of raters' distinct perceptions of the importance of rating criteria irrespective of their rating experience. Apart from Eckes's studies, which were mainly conducted in the writing assessment context, Cai (2015) reports that the rater type hypothesis is also applicable to speaking assessment. Cai explored how rater bias was related to rater types during the rating process in speaking assessment and reported that the raters' severity or lenience was explained by three rater types which were identified as form-oriented, balanced, and content-oriented, implicating the meaningful use of rater's weighting patterns for differential rater training.

In the rater variability literature, however, an emphasis on rater bias toward paired speaking performance is relatively underrepresented except for a few notable studies (e.g., Lumley & McNamara, 1995; Wigglesworth, 1993). Since raters heavily rely on rating criteria in making scoring decisions, it is an empirical question whether raters consistently and accurately award a score when using rating criteria that reflect the features of interactional competence. The close examination of how raters maintain severity based on rating criteria specifically designed to measure paired speaking interaction deserves more attention. These questions can be directly investigated using the bias analysis of MFRM.

Research questions

The literature review suggests that the potential sources of rater bias in paired speaking assessment include examinees' characteristics, such as their proficiency levels and the level difference between two examinees, and rating criteria. Given that the rater's ability to comply with the rating criteria descriptions and to separate examinees' abilities in paired speaking tests directly influences valid score interpretation, quantitative accounts of rater performance are necessary, which warrants the current study. Thus, the present study investigates rater variability across interaction-sensitive rating categories, 102 examinees' levels, and the pairing type determined by the level difference between two examinees using MFRM. The research questions addressed by the study are as follows:

1. In what ways do raters show bias across the rating criteria in paired speaking assessment?

2. In what ways do raters show bias across the examinees' performance levels?

3. In what ways do raters show bias across the pairing type that reflects two examinees' level difference?

Methods

Participants

Examinees

A total of 102 university-level English as Second Language (ESL) learners studying in North America voluntarily participated as examinees who responded to the researcher's call for research participants. The examinees received monetary incentives upon completion of the tasks. The examinees' TOEFL iBT® scores ranged from 65 to 111. Their L1s varied, the most common being Korean, Japanese, and Chinese. Of the 102 examinees, 70% were female and 30% were male. Almost an equal number of graduate and undergraduate students participated in this study.

Raters

A total of 12 raters, consisting of three males and nine females, scored the examinees' audio-recorded paired role-play performances. Seven raters were either English native speakers or bilingual English speakers. Five raters were non-native English speakers with advanced levels of academic English. Their L1s were Chinese, Japanese, Korean, and Vietnamese. The raters either held or were pursuing MA degrees in Second Language Studies at the time of the ratings. All raters had at least two years of experience in teaching English either at the university level or in a high school. Their previous rating experience was mostly done in classroom contexts or in-house speaking tests. Since the rater variability literature reported that raters' first language and prior rating experience usually have relatively little impact on rater performance (e.g., Barkaoui, 2010; Knoch, 2009; Xi & Mollaun, 2011), raters' varied first languages and a rather limited amount of rating experience were not considered as important in relation to the purpose of this study. During the rater recruitment and rater training, however, the researcher closely monitored how well the raters understood training materials and the quality of their practice ratings. Raters who did not meet these criteria were not included in this study.

Test instruments

Paired role-plays and role-play cards

The two role-plays involved two classmates working on a group project (a) to negotiate a mutually accommodating meeting time based on their weekly schedules and (b) to

decide a meeting mode (e.g., face-to-face vs. online) (see Appendix A). The target construct measured in these two role-plays was pragmatic competence in interaction, namely the ability to accomplish pragmatic actions (e.g., agreement, disagreement) jointly in organized sequences using linguistic and interactional resources. Unlike a closed role-play format that predetermines interactional outcomes, open role-plays were used to allow negotiation between the examinees (Kasper & Rose, 2002). In order to ensure some degree of authentic interaction and relatively standardized performances across the examinees, each examinee received a different role-play card, which was not shared with the partner. The same set of role-play cards was given to every pair of examinees. The role-play cards listed prompts and several contingencies related to the role-play situations (see Appendix A). For example, each examinee did not know each other's weekly schedules and which meeting mode option would be proposed by another examinee, as they were asked to express their own preferences on the meeting time and meeting option.

Rating criteria

The data-driven analytical rating criteria used in this study have been already validated using a validity argument framework. The conversation analysis performed on the varying level of examinees' paired role-play performances and the MFRM findings on rater performance were used as backing for three assumptions related to appropriate task design, rating criteria, and rater performance (see Youn, 2015, for a more comprehensive discussion) (see Appendix B). Table 1 summarizes the five categories reflecting various characteristics of the targeted construct. In order to assist raters' evidence-based scoring decisions, the concrete examples corresponding to a three-point scale for each category were integrated into rater training. Details regarding the rater training procedure are provided in the following sub-section.

Table 1. Descriptions of Five Rating Categories

Rating Category	Description
Content Delivery	• The ability to deliver a turn clearly and fluently
Language Use	• The ability to use various linguistic resources to agree and disagree appropriately
Sensitivity to Situation	• The ability to provide reasons and explanations when disagreeing and negotiating opinions in appropriate sequence organizations
Engaging with Interaction	• The ability to engage in conversations and to establish a shared understanding while interacting with a peer
Turn Organization	• The ability to take turns following turn-taking conventions in agreement and disagreement.

Procedure

Rater training

Each rater completed an individual training session and a post-rating interview administered by the researcher. Following the suggestions in the rater training literature (e.g., Lane & Stone, 2006), the training session focused on internalizing ways in which pragmatic actions are accomplished together in an organized sequence, establishing a direct relationship between the rating criteria descriptions and scores to ensure raters' evidence-based decisions, and using example role-play performances from varied levels. As a result, the training session consisted of three sequential steps. Firstly, the raters were familiarized with the paired role-play performances and ways in which various pragmatic action sequences in interaction (e.g., agreement, disagreement) were accomplished by the examinees together. The second training step (Norming) emphasized the raters' shared understanding of the rating criteria and the features that differentiated examinees' differing levels of performances in relation to the rating criteria descriptions. As reference materials for the raters, the transcriptions of example role-play performances that represent three different levels were used during the training. Additionally, during the norming phase, the differences among the five rating categories were explained. During the last sequence of rater training, the raters scored example performances and received immediate feedback from the researcher. Complex scoring cases were also discussed with the researcher.

Test administration

An individual meeting, which took about 20 minutes, was held between the researcher and the two examinees to obtain consent forms and administer the two role-plays. In terms of planning time, there was no time limit. Most of the examinees did not take more than 5 minutes for planning, except for lower-level examinees who needed more time for the researcher to explain instructions. During the tasks, the researcher remained silent and did not interfere with the performance. The duration of each role-play interaction time for each pair ranged from 1 to 2 minutes. Since the same set of role-play cards was given to all examinees, the role-play interaction time was relatively similar across the pairs. The paired role-play performances were audio-recorded. Both role-plays were completed with the same partner.

In order to pair 102 examinees systematically, each examinee was assigned a proficiency of High, Mid, or Low, based on their standardized proficiency test scores. The High category was given to those whose TOEFL iBT® scores were higher than 90. The Mid category was given to those whose TOEFL iBT® scores ranged between 75 and 90. The

rest was categorized as Low. For those who only took IELTS tests, converted TOEFL iBT® scores were used. For the scores obtained more than two years ago at the time of data collection or borderline cases, examinees' length of time living in the USA and their ESL program placement test results were considered in determining the examinees' current proficiency levels. Each examinee was randomly paired with another examinee based on their proficiency levels, which resulted in six pairing types: High-High, Mid-Mid, Low-Low, High-Mid, Mid-Low, High-Low.

Rating

Upon rater training, each rater was given a packet that included an audio CD of examinees' performances, excel sheets to enter their scores, and the rating criteria. The raters were asked to provide an individual score for each rating category on each examinee's performance. Each rater completed the ratings alone and was asked to score all examinees' performances on one role-play situation first and then move on to the next role-play situation. For practicality, 12 raters scored a different number of performances, ranging from 30% to 65%, rather than of a fully-crossed rating design (Schumacker, 1999). In order to link the subsets of the data from each rater, all 12 raters scored 30 examinees' two role-play performances of varied proficiency levels, which ensured connectivity between the subsets of the rating from each rater for MFRM (Linacre, 2012c). Depending on the rater, it took each rater approximately 10 to 20 hours to score the paired role-play performances.

Data analysis

MFRM was conducted using the computer program FACETS, version 3.70.1 (Linacre, 2012a). FACETS calibrates various facets on the same equal-interval scale, the logit scale. A partial credit model (Masters, 1982), which assumes the rating scale for each criterion is modeled to have its own category structure, was used to fit the data. Four facets (examinee, rater, task, rating criteria) were considered in MRFM to examine the main effect of each facet and to answer Research Questions 1 and 2. The examinees were allowed to float as they are the objects of measurement in this study; the other facets were centered. Apart from the four facets, a dummy facet of *pairing type* was added in a model statement for FACETS for the sole purpose of bias analysis between the raters and the pairing type to answer Research Question 3. In FACETS, a dummy facet is used for investigating interaction only rather than main measurement effect. In other words, the dummy facet is not considered in computing the probability of any examinee responding to any assessment task for any rating category threshold for any rater. The dummy facet is anchored at zero logits, meaning that it is not used for estimating measures (Linacre,

2012b, 2013). The pairing type facet indicates six different pairing types based on the two examinees' proficiency levels (i.e., High-High (n = 20), High-Mid (n = 30), High-Low (n = 12), Mid-Mid (n = 14), Mid-Low (n = 14), Low-Low (n = 12)). Since the pairing type was a dummy facet, the six pairing types did not influence estimating the overall measures: they were only used to investigate whether raters maintained their level of severity across certain pairing types. Lastly, three separate bias analyses were conducted to examine the interactions involving the rater facet with three facets (examinee, rating criteria, pairing type). For the bias analyses, a bias size, which calculates the size of bias measure in terms of the logit scale (Linacre, 2013), and its t statistics are reported in the results section.

Results

The MFRM analysis revealed that the raters differed in terms of severity while maintaining internal consistency and the rating criteria functioned relatively well. However, the bias analyses indicated that some raters displayed unexpected rating patterns based on particular rating categories, low-level examinees, and the pairing types that involve low-level examinees.

Four-facet Rasch analysis

Figure 1 presents a FACETS variable map that locates the four facets (examinee, rater, role-play task, and rating criteria) on the logit scale in the first column. All measures of the four facets are positioned on the logit scale, which enables interpreting the results in a single framework of reference. The elements in each column are located in different positions, indicating they differ in terms of abilities for the examinee facet (second column) and severities for the rater facet (third column), for example. The *pairing type* was a dummy facet anchored at zero mean, as it is not used for estimating the measures. This means that the elements of the pairing type facet (i.e., six pairing types) are listed in a row at zero logits, which does not really provide informative value in a variable map. Thus, this dummy facet was intentionally excluded from the variable map, which is possible by changing the model statement properties in a FACETS specification file (Linacre, 2013). Table 2 shows summary statistics for the four facets.

```
+-------------------------------------------------------------------------------------------------+
|Measr|+Examinee   |-Rater                   |-Task       |-Criteria   | S.1 | S.2 | S.3 | S.4 | S.5 |
|-----+------------+-------------------------+------------+------------+-----+-----+-----+-----+-----|
|   7 + *          +                         +            +            + (3) + (3) + (3) + (3) + (3) |
|     |            |                         |            |            |     |     |     |     |     |
|     | *          |                         |            |            |     |     |     |     |     |
|     |            |                         |            |            |     |     |     |     |     |
|     |            |                         |            |            |     |     |     |     |     |
|   6 +            +                         +            +            +     +     +     +     +     |
|     |            |                         |            |            |     |     |     |     |     |
|     |            |                         |            |            |     |     |     |     |     |
|     |            |                         |            |            |     |     |     |     |     |
|     | *          |                         |            |            |     |     |     |     |     |
|   5 + *          +                         +            +            +     +     +     +     +     |
|     |            |                         |            |            |     |     |     |     |     |
|     | *          |                         |            |            |     |     |     |     |     |
|     | **         |                         |            |            |     |     |     |     |     |
|     | ***        |                         |            |            |     |     |     |     |     |
|   4 + *          +                         +            +            +     +     +     +     +     |
|     | *          |                         |            |            |     |     |     |     |     |
|     |            |                         |            |            |     |     |     |     |     |
|     | **         |                         |            |            |     |     |     |     |     |
|     | ***        |                         |            |            |     |     |     |     |     |
|   3 + ****       +                         +            +            +     +     +     +     +     |
|     | ******     |                         |            |            |     |     |     |     |     |
|     | ******     |                         |            |            |     |     |     |     |     |
|     | *****      |                         |            |            |     |     |     |     |     |
|     | ***        |                         |            |            |     |     |     |     |     |
|   2 + ******     +                         +            +            +     + --- +     +     +     |
|     | ********** | R1                      |            |            | --- |     | --- |     | --- |
|     | *          |                         |            |            |     |     |     | --- |     |
|     | ****       | R2                      |            |            |     |     |     |     |     |
|     | *******    |                         |            |            |     |     |     |     |     |
|   1 + **         + R12                     +            + Language   +     +     +     +     +     |
|     | ***        |                         |            | Content    |     |     |     |     |     |
|     | ******     |                         |            |            |     |     |     |     |     |
|     | **         |                         |            |            |     |     |     |     |     |
|     | *          | R9                      |            |            |     |     |     |     |     |
*   0 * **         * R10  R3   R4   R6   R8  * Mode  Time *            *  2  *  2  *  2  *  2  *  2  *
|     |            |                         |            |            |     |     |     |     |     |
|     | **         |                         |            | Turn       |     |     |     |     |     |
|     | **         |                         |            | Situation  |     |     |     |     |     |
|     | **         | R5                      |            | Engaging   |     |     |     |     |     |
|  -1 + **         +                         +            +            +     +     +     +     +     |
|     | ****       |                         |            |            |     |     |     |     |     |
|     | **         |                         |            |            |     |     |     |     |     |
|     | **         | R11  R7                 |            |            |     |     |     | --- |     |
|     | *          |                         |            |            | --- |     | --- |     | --- |
|  -2 +            +                         +            +            + (1) + (1) + (1) + (1) + (1) |
|-----+------------+-------------------------+------------+------------+-----+-----+-----+-----+-----|
|Measr| * = 1      |-Rater                   |-Task       |-Criteria   | S.1 | S.2 | S.3 | S.4 | S.5 |
+-------------------------------------------------------------------------------------------------+
```

Figure 1. FACETS variable map

Note: The *pairing type* facet is intentionally excluded in the variable map; S.1: Content delivery, S.2: Language use, S.3: Sensitivity to situation, S.4: Engaging with interaction, S.5: Turn organization

Examinees

In terms of the examinee facet, the substantial variability within examinees' ability logit values ranging from -1.76 to 6.62 was found, confirmed by the following statistics. A separation index of 4.17 and the corresponding reliability of examinees' level differences was 0.95, indicating that the two role-plays functioned reliably in separating the 102 examinees' varying abilities. The significant chi-square statistic (χ^2=2773.5, *d.f.* = 101, p <

.01) also rejects the null hypothesis that each examinee is equal in terms of ability. The fit statistics were also acceptable. Applying a generally accepted infit values range from 0.5 and 1.5 (Linacre, 2013), two examinees were slightly below the lower-limit (i.e., overfit) and three examinees were over the upper-control limit (i.e., misfit or underfit).

Table 2. Summary Statistics for MFRM analysis

Statistics	Examinees	Raters	Tasks	Rating Criteria
M measure	1.70	0.00	0.00	0.00
M SE	0.36	0.10	0.04	0.06
χ^2	2773.3*	1135.0*	9.2*	687.7*
df	101	11	1	4
Separation index	4.17	9.61	2.86	12.92
Separation reliability	0.95	0.99	0.89	0.99

*$p < .01$

Raters

The third column in Figure 1 compares 12 raters in terms of their degree of severity and Table 3 summarizes the rater measurement report. The raters were not identical in terms of their level of severity, supported by the separation index of 9.61 and the corresponding reliability of rater severity differences estimated at 0.99. The significant chi-square statistic (χ^2=1135.0, *d.f.* = 11, $p < .01$) also rejects the null hypothesis that each rater is equal in terms of severity. Yet, it is worth noting that the spread in severity of the majority of the raters (8 out of 12, 67%) falls within ±1 logit from the mean, which is typical for language performance assessment. Applying a narrower infit values range from 0.7 to 1.3 to the raters (Bond & Fox, 2007; McNamara, 1996) since rater performance is important in this study, the raters' infit values were all acceptable, indicating no raters showed unexpected rating behaviors or too little variation.

Table 3. Measurement Report for Raters

Raters	Severity Logit	Model Error	Infit Mean Square
1	1.75	0.08	1.01
2	1.35	0.08	0.79
12	0.91	0.11	1.09
9	0.13	0.12	1.10
6	0.04	0.08	0.88
10	-0.02	0.12	1.08
3	-0.02	0.08	1.00
4	-0.03	0.08	0.92

8	-0.09	0.12	1.13
5	-0.78	0.09	1.14
11	-1.60	0.14	1.12
7	-1.64	0.14	0.99
M	0.00	0.10	1.02
SD	1.03	0.02	0.11

Paired role-play tasks

The paired role-plays were almost similar in terms of difficulty (0.08, -0.08, respectively), corresponding to the two elements near zero in the fourth column of Figure 1. The infit values for both role-plays (0.99, 0.99, respectively) were acceptable as well, which satisfies the assumption of a unidimensional measurement model (Bond & Fox, 2007).

Rating criteria

As seen in the difficulty logits in Table 4, the five categories differed in terms of difficulty, confirmed by the separation index of 12.92 with the corresponding reliability of 0.99 and the significant chi-square statistic (χ^2=687.7, *d.f.* = 4, $p < .01$). As the partial credit model was used, the sixth to the last columns in Figure 1 respectively present how the three-point scales for five rating criteria were utilized. There was relatively little use of the 1 score band for all criteria. The most challenging category was *Language Use* and the least challenging category was *Engaging with Interaction*. Generally speaking, the interaction-specific categories (*Engaging with Interaction*, *Turn Organization*) ranked as relatively less difficult compared to those categories concerning the features in the examinee's individual utterance (*Content Delivery*, *Language Use*). The infit values of the criteria fell within the acceptable range from 0.7 to 1.3.

Table 4. Measurement Report for Rating Criteria

Rating Criteria	Difficulty Logit	Model Error	Infit Mean Square
Language Use	0.96	0.06	0.90
Content Delivery	0.79	0.06	0.95
Turn Organization	-0.47	0.06	1.10
Sensitivity to Situations	-0.56	0.06	1.05
Engaging with Interaction	-0.72	0.07	0.97
M	0.00	0.06	0.99
SD	0.81	0.00	0.08

Bias analyses

Three separate bias analyses were conducted: (a) Rater x Rating Criteria, (b) Rater x Examinee, and (c) Rater x Pairing Type. Table 5 summarizes the bias measures and the standardized fit statistics (i.e., *t* statistics), which report the statistical significance test for the size of the bias between the raters and the elements of particular facets. All of the absolute *t* values either equal to or greater than 2 were statistically significant at the .05 level. Of the three bias analyses, more frequent bias interactions were found for Rater x Rating criteria (53%), compared to Rater x Examinee (11%) and Rater x Pairing type (20%). More detailed information from each bias analysis is presented below.

Table 5. Summary Statistics for Bias Analyses

Statistics	Rater x Criteria	Rater x Examinee	Rater x Pairing Type
Total number of interaction	60	576[a]	72
Minimum bias measure	-1.25	-3.84	-1.34
Maximum bias measure	2.39	3.84	1.25
Absolute *t* values ≥ 2[b]	32 (53%)	63 (11%)	15 (20%)
Minimum *t* value	-7.08 (*d.f.* = 129)	-4.04 (*d.f.* = 9)	-6.70 (*d.f.* = 99)
Maximum *t* value	6.82 (*d.f.* = 131)	4.67 (*d.f.* = 9)	4.54 (*d.f.* = 99)
M	0.03	-0.05	-0.03
SD	2.93	1.24	1.84

[a] The mixed rating design resulted in 576 interactions between raters and examinees
[b] *t* values were statistically significant at $p < .05$.

Rater and rating criteria

Table 6 reports the bias measures from the bias calibration report for Rater x Rating criteria. The bias measures greater than zero indicate observed scores were higher than expected based on the model (i.e., lenient ratings), while estimates smaller than zero indicate observed scores were lower than expected (i.e., harsher ratings) (Eckes, 2011). In Table 6, statistically significant bias measures, based on *t*-statistics that fall outside the acceptable limits (i.e., -2 to +2), are noted in bold. The positive t-values indicate more lenient ratings than expected while the negative t-values indicate harsher ratings than expected. Of 12 raters, Rater 7, the most lenient rater in general, was particularly more lenient on *Sensitivity to Situation* than expected (bias measure = 2.39, t = 3.24). Interestingly, some raters alternated between more lenient ratings on one rating category and harsher ratings on another category. For example, Rater 1 showed leniency on *Content Delivery* (bias measure = 1.18, t = 6.82) and *Language Use* (bias measure = 0.38, t = 2.12), but displayed severity on *Sensitivity to Situation* (bias measure = -0.63, t = -3.65) and

Engaging with Interaction (bias measure = -0.05, t = -4.79). Of the significant bias interactions, the raters noticeably showed unexpected ratings (9 out of 60 interactions, 15%) for *Language Use*, the most challenging category in general. In particular, the bias measures from several raters (Raters 5, 6, and 10) were noticeably large, which indicates that they were harsher than expected in general when scoring *Language Use*. Further, some raters (Raters 5, 7, and 11) were particularly more lenient for *Sensitivity to Situation*, which resulted in the relatively large averaged bias measure (0.37). With regard to the interaction-specific rating categories, the averaged bias measure was relatively small (0.10, 0.00 respectively) for *Engaging with Interaction* and *Turn Organization*.

Table 6. Bias Analysis between Rater and Rating Criteria: Bias Measures

Rating Criteria	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	Mean bias measure
Content Delivery	**1.18**	**0.51**	-0.21	0.28	**-0.70**	0.14	**-0.54**	-0.32	-0.17	**-1.07**	**-0.92**	0.23	-0.13
Language Use	**0.38**	**0.39**	**0.43**	0.33	**-1.25**	**-0.82**	0.12	**0.77**	**0.63**	**-1.09**	**1.04**	0.04	0.08
Sensitivity to Situation	**-0.63**	**-0.85**	**0.71**	**-0.47**	**1.53**	-0.19	**2.39**	-0.21	-0.28	0.32	**1.65**	0.50	0.37
Engaging with Interaction	**-0.81**	**0.37**	**-0.45**	-0.28	**1.93**	0.16	-0.35	0.09	0.38	**1.06**	-0.42	-0.48	0.10
Turn Organization	-0.05	**-0.40**	**-0.39**	0.05	**0.45**	**0.89**	-0.30	-0.32	-0.52	**1.44**	-0.55	-0.26	0.00

Note: Bias measures in bold are statistically significant at the .05 level (i.e., absolute *t*-values either equal to or greater than 2)

Rater and examinee

Figure 2 plots the Rater x Examinee interaction using the t-values for all 12 raters. The x-axis indicates examinees' ability logit values with higher-level examinees toward the right side. The line plots the 102 examinees' ability logits. The significant t-values are noted with x and non-significant interactions are noted with the • sign. The bias patterns are quite clear in Figure 2. The x signs were more frequent toward the left side of the x-axis where the lower-level examinees are located. However, the significant positive t-values were rarely shown on the right side (i.e., higher-level examinees) of Figure 2. This indicates that the raters were likely stable in maintaining their level of severity when scoring the higher-level examinees' performances. Table 7 further summarizes the significant bias interactions for all 12 raters across three levels of examinee's ability logits range. The most frequent bias interactions were shown for the lower-level examinees (29

out of 192 interactions, 15%), and the frequency of significant bias decreased for the higher-level examinees.

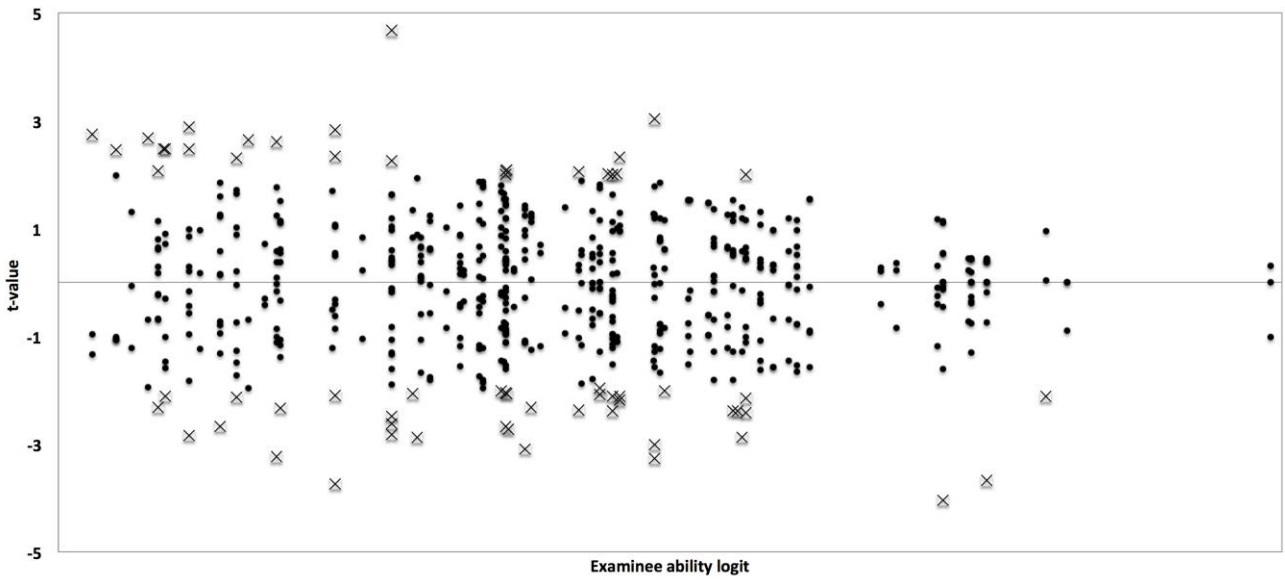

Figure 2. Bias analysis between rater and examinee

Table 7. Bias Interactions Pattern for Examinee Groups

	Low-level group	Mid-level group	High-level group
Total number of *t*-values	192	195	189
t-values equal or above +2	15 (7.8%)	9 (4.6%)	1 (0.5%)
t-values equal or below -2	14 (7.2%)	15 (7.6%)	9 (4.8%)
Mean of bias measures	-0.01	0.03	0.09

Rater and pairing type

Table 8 shows the bias measures for the Rater x Pairing type interaction. Statistically significant bias measures, based on t-statistics that fall outside the acceptable limits (i.e., -2 to +2), are noted in bold. Of 12 raters, five raters (Raters 1, 3, 8, 11, and 12) did not display any significant bias interactions based on the pairing type, confirmed by the relatively small bias measures. Further, it is noteworthy that the raters showed different severity patterns based on rating criteria and pairing types. For example, two raters (e.g., Raters 1 and 3) failed to keep their level of severity for the rating criteria, but they maintained the severity levels for the pairing type (i.e., relatively small bias measures in Table 8). On the other hand, Rater 4, who rarely showed the significant bias interaction

with the rating criteria, displayed unexpected ratings (i.e., relatively large bias measures) for the pairing types. The averaged bias measures for Mid-Mid and Low-Low were relatively small (0.00, -0.03, respectively). However, the largest averaged bias measure (0.14) was found for High-High, particularly due to two raters (Raters 7 and 9) whose bias measures were somewhat large (1.09, 1.25, respectively). This means that Raters 7 and 9 showed more noticeably lenient ratings than expected toward the High-High pairing type. Of the significant bias interactions, the Low-Low (5 out of 15, 33%) and the Low-Mid (4 out of 15, 27%) pairing types were common. It is noteworthy that the bias measures were relatively small, with no significant bias interaction for the Low-High pairing type.

Table 8. Bias Analysis between Rater and Pairing Type: Bias Measures

Pairing Type	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	Mean bias measure
High-High	-0.01	-0.14	-0.10	**0.90**	**-1.15**	0.35	1.09	-0.62	**1.25**	0.13	-0.11	0.09	0.14
Mid-Mid	-0.36	-0.16	-0.06	**0.48**	-0.21	0.22	-0.26	0.14	0.29	-0.05	-0.13	0.13	0.00
Low-Low	0.13	**0.49**	-0.19	**-0.69**	**0.65**	0.15	-0.06	0.18	**-0.86**	**-0.73**	0.29	0.28	-0.03
Low-Mid	0.12	-0.07	0.12	**-1.34**	**0.88**	**-0.44**	**0.82**	0.46	-0.34	0.23	0.31	-0.18	0.05
Low-High	0.25	-0.28	0.12	0.29	-0.15	0.15	-0.68	-0.19	0.38	0.00	-0.29	-0.24	-0.05
Mid-High	-0.01	0.13	0.05	**0.56**	-0.27	-0.16	**-0.74**	-0.48	0.18	0.36	-0.50	-0.04	-0.08

Note: Bias measures in bold are statistically significant at the .05 level (i.e., absolute *t*-values either equal to or greater than 2)

Discussion

Although the raters maintained internal consistency, the bias analyses indicated that the raters' severity patterns were not consistent across the rating criteria, the examinees' levels, and the pairing types. Comparing the ratio of significant bias interactions to the total number of interactions, a greater amount of rater variability was found for Rater x Rating Criteria (52%), compared to Rater x Pairing Type (20%) and Rater x Examinee (11%). Similar findings were reported elsewhere; for example, Eckes (2005, 2012) also reported that raters commonly failed to maintain their level of severity across rating criteria, compared to examinees and assessment tasks. These findings, in turn, further confirm that rating criteria critically impact on raters' decision-making processes.

First of all, the *Rater* x *Rating Criteria* bias analysis provides an answer for the first research question of whether the raters maintained their level of severity when using the interaction-focused rating criteria. The distinct difficulty estimates and stable infit

statistics were reported for the rating criteria, although the raters noticeably underused the score of 1 on all rating criteria. Nonetheless, considering the greater amount of rater variability found toward the rating criteria in general, the raters maintained their level of severity when assigning scores on the interaction-specific rating categories compared to other rating categories and the averaged bias measures were not noticeably larger than other criteria. These findings indicate that the raters were relatively consistent in assigning the scores for the *Engaging with Interaction* and *Turn Organization* categories. Taken together, these findings support the facilitative effect of rater training in this study, which focused on ways in which the raters can distinguish the varying levels of the two categories that directly tap into interactional competence. In fact, the source of rater variability based on the rating criteria was found in other rating categories. For example, the averaged bias measures value was somewhat large (0.37) for *Sensitivity to Situation*. This means that some raters (e.g., Raters 5, 7, and 11) showed more leniency than their averaged severity when scoring *Sensitivity to Situation*. Further, nine out of 12 raters displayed inconsistent severity patterns when scoring *Language Use*, which measures the ability of using appropriate linguistic expressions in making an agreement and disagreement. Of these nine raters, six raters showed more leniency than their averaged severity. Thus, it can be speculated that the severe raters might have compensated for their severity, especially when providing a score on the challenging rating category. Another potential reason might be that the rater training that emphasized interaction-related rating categories might have influenced the raters to rely more on their own intuition when scoring *Language Use* and *Sensitivity to Situation*. Finally, raters' tendency to reverse their scoring patterns as they award scores on the rating categories was observed. In other words, when the raters, such as Rater 1, were more lenient than expected on *Content Delivery* and *Language Use*, they were more severe on *Sensitivity to Situation* and *Engaging with Interaction*, and vice versa. Given this, it is speculated that the raters compensated for the degree of severity as they moved from one rating category to another. Rater's tendency to alternate between harsher ratings on one category and more lenient ratings on another category was also reported in two rater bias studies in writing assessment (Eckes, 2012; Schaefer, 2008). Such intricate patterns of rater variability based on the rating criteria call for more studies on raters' scoring decision processes involved in operationalizing rating criteria descriptions.

Turning to the second research question of whether the raters displayed unexpected ratings related to particular examinees' levels, interesting bias patterns were found. The *Rater* x *Examinee* bias analysis indicated that more frequent biases were found on the lower-level examinees. With regard to this finding, previous research on rater performance across examinees' levels offers potential reasons. For example, Yan (2014)

reported a higher degree of rater agreement on high-level examinees than on low-level examinees when scoring monologic speaking performances. In Kondo-Brown's (2002) rater bias research on L2 Japanese writing assessment, raters frequently gave unexpected ratings for low-level examinees' performances that consisted of extremely short passages. May (2011) also reported raters' perceptions of the importance of the intelligibility of candidate contributions in scoring paired speaking interaction. Thus, the lack of clarity in delivery and language use in low-level examinees' performances might have caused inconsistent rater judgments when assigning a score for a particular rating category. A second possibility is that the raters might have used differential rating strategies depending on examinees' levels. Kuiken and Vedder (2014) reported differing expectations and distinct types of language features that raters attended to depending on examinees' levels, such as more discourse-level features for high-level examinees and more emphasis on grammar for low-level examinees. It is possible that the raters in this study might have focused on the linguistic features (i.e., *Language Use* category) more often when rating low-level examinees to compensate for their limited abilities to sustain the conversation, which resulted in more biased ratings.

Regarding the third research question of inconsistent rater severity related to a particular pairing type, the *Rater* x *Pairing Type* bias analysis indicated that the raters were most stable when scoring the Mid-Mid pairing type and no significant rater bias was shown for the Low-High pairing type, confirmed by the relatively small averaged bias measures. The largest averaged bias measures (0.14) were found for High-High pairings. Looking at the bias measures more closely, a few raters (Raters 7 and 9), whose bias measures were larger than 1 (i.e., noticeably more lenient ratings than expected), contributed to such overall lenient scoring pattern found in the High-High pairing type. Significant bias interactions were found for the Low-Low or Low-Mid pairing types. The general characteristics of low-level examinees' paired speaking interaction include a low degree of mutuality established between the speakers and minimal topic development (Galaczi, 2014), which raters themselves usually pay attention to during scoring in paired speaking assessment (May, 2011). Thus, it is possible that each examinee's contribution to sustain the conversation might be quite noticeable in the Low-High pairs, which could have assisted the raters to consistently maintain a degree of severity in scoring each examinee's performance. In contrast, two examinees' performances from the Low-Low or Low-Mid pairs might not be easily separable and their limited language competence can present more challenges for the raters, resulting in more frequent bias interactions. Further, the noticeable underuse of score 1 on all rating criteria may indicate the greater uncertainty in raters' decision-making for the lower-level examinees. While this study's findings are limited to generalized systematic sources of rater variation, the results from both *Rater* x

Examinee and *Rater* x *Pairing Type* bias interactions indicate that the raters' inconsistent ratings tended to occur when scoring the lower-level examinees' performances. The lower-level examinees' paired speaking performances likely consist of unclear utterances, a shorter turn length and lengthy pauses, which can result in miscommunication during paired speaking interaction. Such characteristics in combination might have presented more challenges to the raters in terms of awarding scores consistently in accordance with the rating criteria descriptions.

Further examining the *Rater* x *Criteria* and *Rater* x *Pairing Type* bias interactions, the raters displayed distinct bias patterns. For example, Rater 4 displayed unexpected ratings (i.e., relatively large bias measures) for the pairing types, while maintaining expected scorings (i.e., relatively small bias measures) for the rating criteria. However, an opposite bias pattern was observed for Rater 1, who showed unexpected ratings for rating criteria, but relatively expected ratings for the pairing type. In other words, we can infer that raters might display distinct patterns of unexpected ratings based on different factors (e.g., rating criteria, examinees' levels) involved in performance assessment. Such findings might be the result of the interaction between various factors involved in the study. Nonetheless, the distinct rater bias patterns shown in this study might be indicative of raters' different perception of factors involved in performance assessment. The extent to which the raters utilized different strategies cannot be directly inferred in this study, calling for future research on distinct rater types and rating behaviors in the paired speaking assessment context.

Taken together, this study's findings hold the following implications for rater variability and rater training in paired speaking assessment. Raters might experience more difficulties with scoring low-level examinees and borderline cases from adjacent-level examinees' performances. Thus, specific strategies to comply with the rating criteria descriptions specific to low-levels and ways to deal with unintelligible utterances from low-level performances can be emphasized during rater training. Furthermore, raters will likely have difficulties with scoring adjacent-level performances rather than pairs with distinct level differences. To meet this challenge, more research on explicating key features that distinguish adjacent-level performances will be needed for more accurate and consistent scoring of borderline cases, which can be integrated into the development of construct-specific rating criteria and rater training methods. Moving forward, the remaining question concerns the potential of utilizing rater bias patterns for training raters in paired speaking assessment. Eliminating rater variability entirely might not be an ideal goal. Instead, rater bias information can be further integrated into rater training

to examine how well rater training works or to explicate the sources of persistent rater variability.

The limitations of this study should be noted. First of all, the small number of the raters limits the generalizability of the findings. Future research with a larger pool of raters to confirm the findings from this study is needed. Additionally, while systematic patterns of rater bias were revealed, supporting findings from previous research, the present study does not directly provide a comprehensive explanation of rater behaviors in paired speaking assessment. Such limitations in this study warrant future research to explicate the systematic sources of rater variability and to develop a valid rater training method for paired speaking assessment, such as mixed-methods approaches to a relationship between raters' thought process using think-aloud protocol (e.g., Baker, 2012; Barkaoui, 2011; Suto, 2012; Suto & Greatorex, 2008) and resulting rater performance. Lastly, the generalizability of the interaction-sensitive rating criteria used in this study is still in question, especially for scoring low-level examinees. More explicit accounts of each rater's cognitive process in utilizing rating criteria at different proficiency levels are necessary. Additional research on explicating the generalizable features of paired speaking performance at various levels and the development of rating criteria that can be applicable to other paired speaking tasks deserves further attention.

Conclusion

The present study examined the rater variability issue in paired speaking assessment. Although exploratory in nature, the findings provide various implications for understanding the nature of rater variability and ways to connect it to rater training in paired speaking assessment in particular and performance assessment in general. The findings suggest that the raters had difficulty in scoring low-level examinees' performances and that rater performance was more variable in scoring the most challenging rating category. However, the raters were not necessarily more biased in assessing the interaction-specific rating categories and the pairing type with a distinct level difference. While the direct effect of rater training cannot be confirmed in this study, rater training conforming to the nature of targeted constructs might have a facilitative effect on ensuring rater performance. Nonetheless, the rating process in speaking assessment can be quite unique and complex. A variety of features inherent in examinees' speaking performances, such as pronunciation, intelligibility, and mutuality, are heeded by raters, which can impact upon their intuitive perceptions of spoken performances and decision-making processes. Extending the preliminary findings from this study, future research on rater cognition which strives to understand raters' mental representations

and rating styles (Myford, 2012) will further enlighten rating behaviors, ultimately contributing to more valid score interpretations in paired speaking assessment.

References

Baker, B. A. (2012). Individual differences in rater decision-making style: An exploratory mixed-methods study. *Language Assessment Quarterly*, *9*, 225–248.

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly*, *7*, 54–74.

Barkaoui, K. (2011). Think-aloud protocols in research on essay rating: An empirical study of their veridicality and reactivity. *Language Testing*, *28*, 51–75.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model* (2nd ed.). New York: Routledge.

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the L2 group oral discussion tasks. *Language Testing*, *20*, 89–110.

Cai, H. (2015). Weight-based classification of raters and rater cognition in an EFL speaking test. *Language Assessment Quarterly*, *12*, 262–282.

Carey, M. D., Mannell, R. H., & Dunn, P. K. (2010). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing*, *28*, 201–219.

Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing*, *20*, 369–383.

Davis, L. (2009). The influence of interlocutor proficiency in a paired oral assessment. *Language Testing*, *26*, 367–396.

Ducasse, A. M., & Brown, A. (2009). Assessing paired orals: Raters' orientation to interaction. *Language Testing*, *26*, 423–443.

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

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

Eckes, T. (2011). *Introduction to many-facet Rasch measurement*. Frankfurt am Main: Peter Lang.

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

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

Galaczi, E. (2008). Peer-peer interaction in a speaking test: The case of the First Certificate in English examination. *Language Assessment Quarterly*, *2*, 89–119.

Galaczi, E. (2014). Interactional competence across proficiency levels: How do learners manage interaction in paired speaking tests? *Applied Linguistics*, *35*, 553–574.

Hall, J. K. (1995). (Re)creating our worlds with words: A sociohistorical perspective of face-to-face interaction. *Applied Linguistics, 16*, 206–232.

Isaacs, T., & Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly*, *10*, 135–159.

Kasper, G., & Rose, K. R. (2002). *Pragmatic development in a second language*. Malden: Blackwell.

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

Kim, H. J. (2015). A qualitative analysis of rater behavior on an L2 speaking assessment. *Language Assessment Quarterly*, *12*, 239–261.

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing*, *26*, 275–304.

Knoch, U. (2011). Investigating the effectiveness of individual feedback to rating behavior-a longitudinal study. *Language Testing*, *28*, 179–200.

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

Kramsch, C. (1986). From language proficiency to interactional competence. *The Modern Language Journal, 70*, 366–372.

Kuiken, F., & Vedder, I. (2014). Rating written performance: What do raters do and why? *Language Testing*, *31*, 329–348.

Lane, S., & Stone, C. A. (2006). Performance assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 387–431). Westport, CT: American Council on Education.

Linacre, J. M. (2012a). Facets Rasch measurement computer program, version 3.70.1. Chicago: Winsteps.com.

Linacre, J. M. (2012b, January). Many-facet Rasch measurement: Facets Tutorial 3. Estimation and interactions. Retrieved from http://www.winsteps.com/a/ftutorial3.pdf

Linacre, J. M. (2012c). Many-facet Rasch measurement: Facets Tutorial 4. Anchoring. Retrieved from http://www.winsteps.com/a/ftutorial4.pdf

Linacre, J. M. (2013). *A user's guide to FACETS: Rasch-Model Computer Programs*. Chicago: Winsteps.com.

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

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174.

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

May, L. (2011). Interactional competence in a paired speaking test: Features salient to raters. *Language Assessment Quarterly*, *8*, 127–145.

McNamara, T. (1996). *Measuring second language performance*. New York: Addison Wesley Longman.

McNamara, T. (1997). 'Interaction' in second language performance assessment: whose performance? *Applied Linguistics, 18*, 446–466.

Myford, C. M. (2012). Rater cognition research: Some possible directions for the future. *Educational Measurement: Issues and Practice*, *31*, 48–49.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part 1. *Journal of Applied Measurement*, *4*, 386–422.

Ross, S., & Kasper, G. (Eds.). (2013). *Assessing second language pragmatics*. Basingstoke, UK: Palgrave Macmillan.

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

Schumacker, R. E. (1999). Many-facet Rasch analysis with crossed, nested, and mixed designs. *Journal of Outcome Measurement*, *3*, 323–338.

Suto, I. (2012). A critical review of some qualitative research methods used to explore rater cognition. *Educational Measurement: Issues and Practice*, *31*, 21–30.

Suto, W. M. I., & Greatorex, J. (2008). What goes through an examiner's mind? Using verbal protocols to gain insights into the GCSE marking process. *British Educational Research Journal*, *34*, 213–233.

Taylor, L., & Wigglesworth, G. (2009). Are two heads better than one? Pair work in L2 assessment contexts. *Language Testing*, *26*, 325–339.

Van Moere, A. (2006). Validity evidence in a group oral test. *Language Testing*, *23*, 411–440.

Winke, P., Gass, S., & Myford, C. (2013). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing*, *30*, 231–252.

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

Xi, X., & Mollaun, P. (2011). Using raters from India to score a large-scale speaking test. *Language Learning*, *61*, 1222–1255.

Yan, X. (2014). An examination of rater performance on a local oral English proficiency test: A mixed-methods approach. *Language Testing*, *31*, 501–527.

Youn, S. J. (2015). Validity argument for assessing L2 pragmatics in interaction using mixed methods. *Language Testing*, *32*, 199–225.

Young, R. F. (2002). Discourse approaches to oral language assessment. *Annual Review of Applied Linguistics*, *22*, 243–262.

Appendix A: Role-play task and cards

Situation: After class, you're going to talk with your classmate who is doing a class project (article presentation) regarding **when and how your group members will meet** to discuss the project. The third member (Tom) is absent today in class. Your presentation is next Friday.

Task: You will receive role-play cards that describe what you are going to tell your classmate. Please have a conversation with your classmate naturally.

Role-play Card (Meeting time)

Jessie
Jessie
1. **Look at your schedule.** Respond to a Phoenix' question.
Jessie
Jessie
2. You need to leave soon since you have another class soon. So, whether you found a good time or not, **suggest** asking the third member (Tom)'s opinion to make a final decision.
Jessie
3. Respond what Phoenix says

Phoenix
1. As approaching to Jessie, start a conversation about an upcoming class project (article presentation). **Suggest** discussing an appropriate meeting time. Propose one available time slot based on your schedule.
Phoenix
Phoenix
2. Respond to Jessie's time availability **based on your own schedule**.
Phoenix
Phoenix
3. Respond what Jessie says

Jessie's Schedule

Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
9am-1pm: Classes	Part-time Work (10am-5pm)	9am-1pm: Classes	Part-time Work (10am-5pm)	9am-1pm: Classes		Part-time Work (2-9pm)

Phoenix' Schedule

Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
10am-3pm: Classes	No class	10am-1pm: Classes	10am-3pm: Classes	Meeting with an advisor at 2pm	BBQ party with friends at 5pm	

Role-play Card (Discussion mode)

Jessie	Phoenix
1. Move the discussion to a discussion mode. **Suggest** discussing how you will meet all together to discuss a project. Propose one option between **face-to-face discussion** and **online discussion (e.g., chatting)** that you personally prefer.	
Jessie	**Phoenix**
	1. Respond to what Jessie proposes. **Choose one option that you prefer** and express your own opinion.
Jessie	**Phoenix**
2. Respond to Phoenix' opinion.	2. Respond to Jessie's opinion. **Suggest** that you want to ask the third group member (Tom) who is absent today to make a final decision about how members will meet.
Jessie	**Phoenix**
3. Wrap up the conversation	3. Wrap up the conversation

Appendix B: Rating Criteria

Score	Content Delivery	Language Use	Sensitivity to Situation	Engaging with Interaction	Turn Organization
3	• Clear, concise, fluent • Smooth topic initiations with transitional markers (i.e., smooth turn initiation) **Rating Phoenix**: *asking time for a meeting* & Phoenix' responses to Jessie's questions **Rating Jessie**: *asking how to meet* & Jessie's responses to Phoenix' questions ***Note***: Who initiates 'Asking Tom for a final decision' is not a crucial rating point, but focus more on delivery of follow-up contents.	• Pragmatically appropriate linguistic expressions (**bi-clausal, conditional, past progressive tense**: I was thinking, I don't think I can; **modal verbs**: would, could, might) • Good control of grammar and vocabulary that doesn't obscure meaning **Focus**: **asking questions, expressing different opinions** and **suggestions** ***Note***: No need to heavily rely on elaborated complex structures, but diverse grammatical structures for pragmatic meaning need to be observed for '3'.	• Consistent evidence of awareness and sensitivity to situations exists in an appropriate sequence **Examples**: what is needed for a team project (e.g., time negotiation, back up time slots for Tom), accounts for disagreement, explanations (at least brief) for time and meeting mode preference, pay attention to classmate's opinions ***Note***: Although not all examples need to be observed, a substantial amount of evidence needs to be observed for '3'	• A next turn shows understandings of a previous turn throughout the interaction (i.e., shared understanding) • Evidence of engaging with conversation exists (e.g., clarification questions, backchannel, acknowledgement tokens) ***Note***: Non-verbal cues also serve as acknowledgement, so no need to heavily rely on the amount of discourse markers.	• Complete adjacency pairs (e.g., question & answer) • Interactionally fluid **without awkward pauses** or **abrupt overlap** (especially between disagreement) ***Note***: Interactionally meaningful pauses include those before refusal and between disagreements ***Note***: Even with the elaborated language use ('3' in Language Use), this may not necessarily be done properly with a pause (esp. disagreement). Then, '3' in Turn Organization may not necessarily be awarded.
2	• Generally smooth, but **occasionally unclear** (which confuse a classmate), or **unnecessarily wordy** • **Abrupt** topic initiation (in terms of contents) • **Unclear transitional cues** (e.g., unclear intonation and stress)	• Able to use modal verbs in **mono-clausal** (e.g., could, can, might), but doesn't use various grammatical structures for pragmatic meaning • Linguistic expressions are occasionally inaccurate and a bit limited that sometimes obscure meaning	• Inconsistent evidence of awareness and sensitivity to situations (e.g., provide accounts for opinions, but do not necessarily handle the disagreement situation properly)	• Some evidence of engaging with the conversation, but not consistent (e.g., literally read the role-play card), • A next turn does not sometime show understanding of a previous turn	• Some turns are **delayed** and a next turn is **absent** in adjacency pairs (e.g., absence of answers) • Sometimes **abruptly cutoff** previous turns
1	• Delivery is choppy, fragmented, and minimal (due to a lack of language competence)	• Expressions sound **abrupt** or **not polite enough** (e.g., I'm busy, I can't) • Linguistic expressions are inaccurate and quite limited that obscure meaning	• Little evidence of situational sensitivity (e.g., absence of providing accounts for *disagreements in particular*, handle disagreement awkwardly)	• Noticeable absence of discourse markers • Evidence of not achieving a shared understanding	• Noticeably abrupt overlap or no pauses between disagreements and refusal • **Noticeably long pauses** or **noticeable cutoff** between turns