Rater Variability in a Paired Speaking Task: A Mixed-methods Approach

Shi Chen

Northern Arizona University

Abstract

This exploratory mixed-methods study examined rater variability in a paired speaking task. Four female EFL raters from China participated in the study. All of them had at least one-year teaching experience. This study used a concurrent mixed-method approach. The raters graded 15 paired speaking tasks, which included 30 test takers. The speaking task was retrieved from an achievement test in the Program in Intensive English (PIE) at a major university in the United States. Test takers' performances were rated based on a 4-point scale which included 13 subcategories. The preliminary results of the factor analysis provide some validity evidence of the revised rating scale. The mixed-effect analysis confirmed that the raters had a significant effect on participants' ratings. Additionally, raters demonstrated more variance in the task completion and interaction model than the linguistic features model. The results of the study are able to inform rater training, quality control, and rating scale design.

Keywords: paired-speaking task, interactional competence, factor analysis, mixed-effects model

Rater Variability in a Paired Speaking Task: A Mixed-method Approach

Background

The purpose of this study was to find out the rater types exemplified in a paired speaking task. First, this study aimed to justify the revised rating scale by providing some validity evidence using exploratory factor analysis. If the rating scale provides some validity evidence, it could reduce some construct-irrelevant factors influence that a rating scale could have on the raters. Next, mixed-effects analyses were conducted to investigate whether raters had significant influence on the scores, and whether the participant was a random factor that accounted for much variance in the models. In addition, the ranking of the perceived importance of the rating scale was compared to the actual ratings. Finally, a preliminary coding scheme was established for the reference of coding raters' comments in the near future. Some interesting findings from the think-aloud sessions were discussed.

The study addressed the following research questions:

1. Are there any differences or similarities between the perceived importance of the rating scale and the operational ratings?
2. Is there any validity evidence for the paired-speaking task rating scale?
3. Was any rater variability demonstrated in ratings of the paired-speaking task?
4. How many types of raters can be identified?

However, to present the specific rater types of rating patterns, I still need to recruit more raters. Therefore, this current report only provides answers to part of the research questions.

Methods

Participants

Interviewee. The interviewee was an instructor who has been teaching listening and speaking for three years at Program of Intensive English (PIE). She has also been rating the paired speaking test for a few years.

Test takers. Test takers were 30 students who were level 4 students at PIE.

Raters. Raters were four female EFL teachers from China, and they had at least one year of English teaching experience.

Instruments and Data Collection Procedures

Semi-structured interview. A semi-structured interview was conducted to gather some comments about current rating scale used at PIE. One PIE instructor consented to take part in the interview. During the interview, questions in relation to rating scale, test, teaching, and rating decision making were asked.

The paired speaking task. One paired speaking task was selected from PIE archived data. The task was about opening a business enterprise.

Rating scale. The rating scale was developed based on the interview data. It included four categories: delivery, language use, task completion, and interaction. Raters not only rated the four large categories, but they also rated the subcategories underneath each category. In total, raters awarded scores based on 13 subcategories.

Questionnaires. The rater background questionnaire asked about demographic information and teaching experience. After raters finished their rater training sessions, they were asked to rank the importance of the rating scale categories that they perceived.

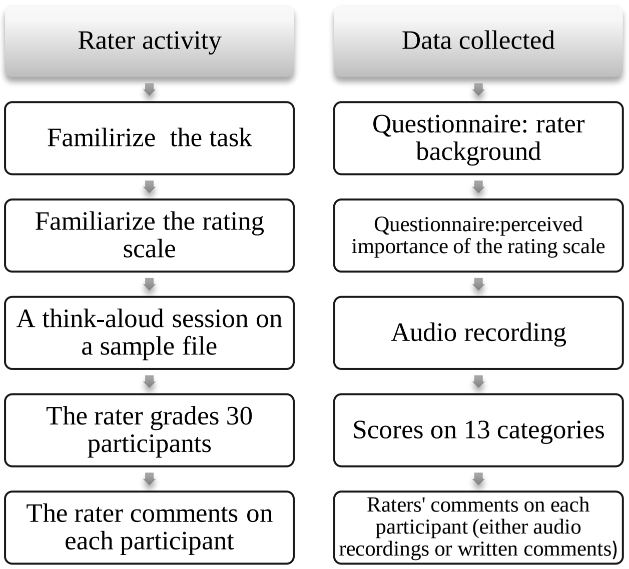

Rater activity
Familirize the task
Familiarize the rating scale
A think-aloud session on a sample file
The rater grades 30 participants
The rater comments on each participant
Data collected
Questionnaire: rater background
Questionnaire:perceived importance of the rating scale
Audio recording
Scores on 13 categories
Raters' comments on each participant (either audio recordings or written comments)

Figure 1. Sequence of rater activity and data collected.

Results

Quantitative Analysis: Descriptive Statistics, Factor Analysis and Mixed-effects Models

As for the first research question regarding the differences between perceived importance of the rating scale and operational ratings, in this report, only the first part of the question was answered, and the results are still not generalizable partly because only four raters were included. As shown in Table 1, the raters perceived “delivery” as the most important category, followed by language use and interaction. “Language use” was rated by the raters as the least important category.

Table 1

Perceived Importance of Rating Scale

Category	1=most important	2	3	4=least important
Delivery	50%	25%	0	25%
Language Use	25%	0%	25%	50%
Task Completion	0%	50%	25%	25%
Interaction	25%	25%	50%	0%

After performing a factor analysis, two factors were generated and were shown in Table 2. This table contains the loadings for each variable on each factor. According to the percentage of loadings, factor 1 was labeled as task completion and interaction and factor 2 was labeled as linguistic features.

Table 2

Pattern Matrix

	Factor	
	1=Task Completion and interaction	2=Linguistic features
rater_understand	-0.01	0.49
testtaker_understand	0.85	0.03
fluency	0.13	0.66
suprasegmental	0.08	0.43
respond_effective	0.49	0.37
vocab	0.04	0.86
grammar	-0.17	0.94
task_completion	0.57	0.13
detailed_evidence	0.68	0.16
reach_agreement	0.46	0.04
engagement	0.77	-0.04
authenticity	0.67	-0.08
ask_opinion	0.78	-0.11

Note. Extraction method: principle axis factoring. Rotation method: promax with Kaiser normalization.

After the two factors were generated, they were treated as two dependent variables for the latter analysis. Two mixed-effects models were performed in R Studio with raters treated as the fixed factor, and participants as the random factor. As shown in Table 3, raters' performances are significant different in task completion and interaction (factor 1), $p < 0.05$.

Table 3.

Factor 1 Task Completion and Interaction: Fixed Effects

	Estimate	Std. Error	df	t value	Pr(>|t|)
Intercept	0.96	0.19	81.32	5.20	0.00
rater	-0.36	0.05	69.00	-7.10	0.00

Table 4 displays the estimates of random effects and the variance accounting for the variability. The estimates of random effects refer to the unaccountable variance after the fixed effects have been accounted for. The random effect of participants accounts for 36% of the variability.

Table 4

Factor 1 Task completion and interaction: Random effects

Groups	Name	Variance	Std. Dev.
Participant	intercept	0.36	0.60
residual		0.36	0.60

Because SPSS doesn't automatically calculate variance of fixed effects, fixed effects predicted value and its associated Marginal and Conditional R^2 were calculated (Marginal R^2 =Fixed predicted variance/sum of all three variances). Marginal R^2 square indicates the effect size of the fixed factor, and Conditional R^2 ((fixed predicted variance + sum of random variances) / sum of all three variances) shows the effect size of both the fixed factor and the random factors. $R_{marginal}^2$ is associated with fixed effects, and Conditional R^2 is associated with the fixed effects and the random effects ($R_{marginal}^2$=0.22, $R_{conditional}^2$=0.61). Conditional R^2 is 61%, which means, the fixed effect of rater, combined with the random effects of participants, account for approximately 61% of the variance in task completion and interaction.

For the second model, linguistic features (factor 2) were treated as the dependent variable, as shown in Table 5, since *p* < 0.05, if we set the alpha level at 0.05, rater has a significant effect on factor 2, which is linguistic features.

Table 5

Factor 2 linguistic features: Fixed effects

	Estimate	Std. Error	df	t value	Pr(>|t|)
Intercept	0.95	0.20	85.60	4.74	0.00
rater	-0.36	0.06	59.00	-5.54	0.00

As shown in Table 6, the random effect of participant accounts for 13% of the variability.

Table 6

Factor 2 Linguistic Features: Random Effects

Groups	Name	Variance	Std. Dev.
Participant	Intercept	0.13	0.36
Residual		0.58	0.76

As mentioned above, the marginal R^2 and conditional R^2 were calculated. $R_{marginal}^2$ is associated with fixed effects, and Conditional R^2 is associated with the fixed effects and the random effects ($R_{marginal}^2$=0.13, $R_{conditional}^2$=0.22). Conditional R^2 is 22%, which means, the fixed effect of rater, combined with the random effects of participant, account for approximately 35% of the variance in task completion and interaction.

Qualitative Analysis: Think-aloud Sessions

There were several interesting findings that stood out from the think-aloud session. One rater mentioned that she intended to give similar scores to the participants for the participants in a pair, even if their performances might vary. This phenomenon was also observed by inspecting the scoring sheets of the other raters. Next, for a specific pair, one rater thought there were three people who participated in the paired speaking task. Even if I explained to her that the conversation flow between the two speakers were continuous, participants kept moving their microphones, she still insisted that the conversation sounded like three people participated in this task. In addition, for the sub-rating category, *reach agreement at the end of the conversation*, one rater pointed out that they agreed with the benefits of opening the business enterprise, however, they did not reach agreement at the end of the conversation in term of which business they wanted to open. Therefore, she could not decide whether this accounts as reaching an agreement at the end.

Another finding was that when I asked all four raters to comment on the overall impression of the participant's performances, three out of four raters mentioned that the female test taker had heavy accent, which is common among EFL raters. In addition, some raters guessed participants' nationalities in their think-aloud sessions. And they felt like they could relate more to the students who were from the same country, whereas, finding it harder to understand the speakers who were not from the same country as them.

Finally, one rater mentioned that if a test taker has a communication breakdown because he or she cannot understand their partners because of their pronunciation, she was not sure whom to blame and whether she should deduct test takers' points. Based on the think-alouds, a preliminary coding scheme was established for the reference of coding raters' comments in the near future (Table 7).

Table 7

Coding Scheme

Key words	Coding key
Accents related comments	ACC
The uses of grammar	GRA
Quality of fluency	FLU
The participant uses appropriate/inappropriate vocabularies	VOC
The participant reaches agreement regarding the business	AGG
Authenticity of the paired-speaking task	AUT
The participant makes good preparation for the task	PRE
The participant covers all four points in the task	COV
Raters find it hard to separate two participants' performances	SEP
Not using a score of 1 on the rating scale	RAT
Rates follow/understand the participant's turns	UND
Details and examples of the reasons given	EXA
The participant greets each other at the beginning of the conversation	GRE
The rater pinpoints participant's language background	LAN

Conclusion

RQ1: Are there any differences and similarities between the perceived importance of the rating scale and the operational ratings?

Regarding the perceived importance of the rating scale, the raters perceived "delivery" as the most important category, followed by language use and interaction. Raters perceived "language use" as the least important category. It is in line with the think-aloud session data that were collected during the rating training sessions. Out of four raters, three of them mentioned that the female participant in the sample file had heavy accent.

RQ2: Is there any validity evidence for the paired-speaking task rating scale?

After the factor analysis was performed, two factors were generated. According to the loadings on each factor, the first factor was labeled as task completion, and the second factor was labeled as linguistic features. Two factors do not greatly overlap and represent different constructs. If the rating scale does not provide some validity evidence, all the constructs measure in the rating scale would fall into one factor. Therefore, the factor analysis result offered some validity evidence of the rating scale.

RQ3: Are there any rater variability demonstrated in ratings of the paired-speaking task?

Two mixed-effects models were performed, and raters definitely have significant effect on the ratings. One possible reason might be raters tend to rate higher in task completion and interaction, whereas in linguistic resources, not a lot of participants can achieve high scores and perform very stable.

RQ4: How many rater types of raters can be identified?

This research questions cannot be answered for now because the ultimate goal of the research project is to determine rater types. However, there are insufficient number of raters. Therefore, this research question will be answered in the future once enough data are collected.

Relevance to PIE and Second Language Learning

Because each rater comes from different backgrounds, their judgement varies a lot. If we are able to identify rater types at PIE, it might be easier for us to tell whether a rater is severe or lenient in ratings for a specific category. For example, one rater might be lenient in language use, severe in task completion, but lenient in delivery. Additionally, it will also be beneficial for quality control of raters since we are able to identify different rater types. If so, PIE will be able to categorize instructors into different rater types, and thus having a better understanding of their rating habits. Last, this project redesigned the current rating scale for the paired speaking task. PIE might be able to use some parts of the redesigned rating scale.

References

Brooks, L. (2009). Interacting in pairs in a test of oral proficiency: Co-constructing a better performance. *Language Testing*, *26*, 341–366.

Davis, L. E. (2012). *Rater expertise in a second language speaking assessment: The influence of training and experience* (Unpublished doctoral dissertation). University of Hawai'i at Manoa, Hawaii.

Ducasse, A. M. (2010). *Interaction in paired oral proficiency assessment in Spanish: Rater and candidate input into evidence based scale development and construct definition*. Frankfurt: Peter Lang.

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

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

Gorsuch, R. L. (1983). Factor analysis. 2nd. *Hillsdale, NJ: LEA*.

Han, Q. (2016). Rater cognition in L2 speaking assessment: a review of the literature. *Working Papers in Applied Linguistics and TESOL*. doi: 10.7916/D8MS45DH

Hashemnezhad, H. (2015). Qualitative content analysis research: A review article. *Journal of ELT and Applied Linguistics*, *3*(1).

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher*, *33*(7), 14-26.

Knoch, U., Fairbairn, J., & Huisman, A. (2016). An evaluation of an online rater training program for the speaking and writing sub-tests of the Aptis test.

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

May, L. (2011). Interactional competence in a paired speaking test: Features salient to raters. *Language Assessment Quarterly*, 8(2), 127-145.

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

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

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

Nakatsuhara, F. (2004). *An investigation into conversational styles in paired speaking tests*. Unpublished master's dissertation, University of Essex, Essex, United Kingdom.

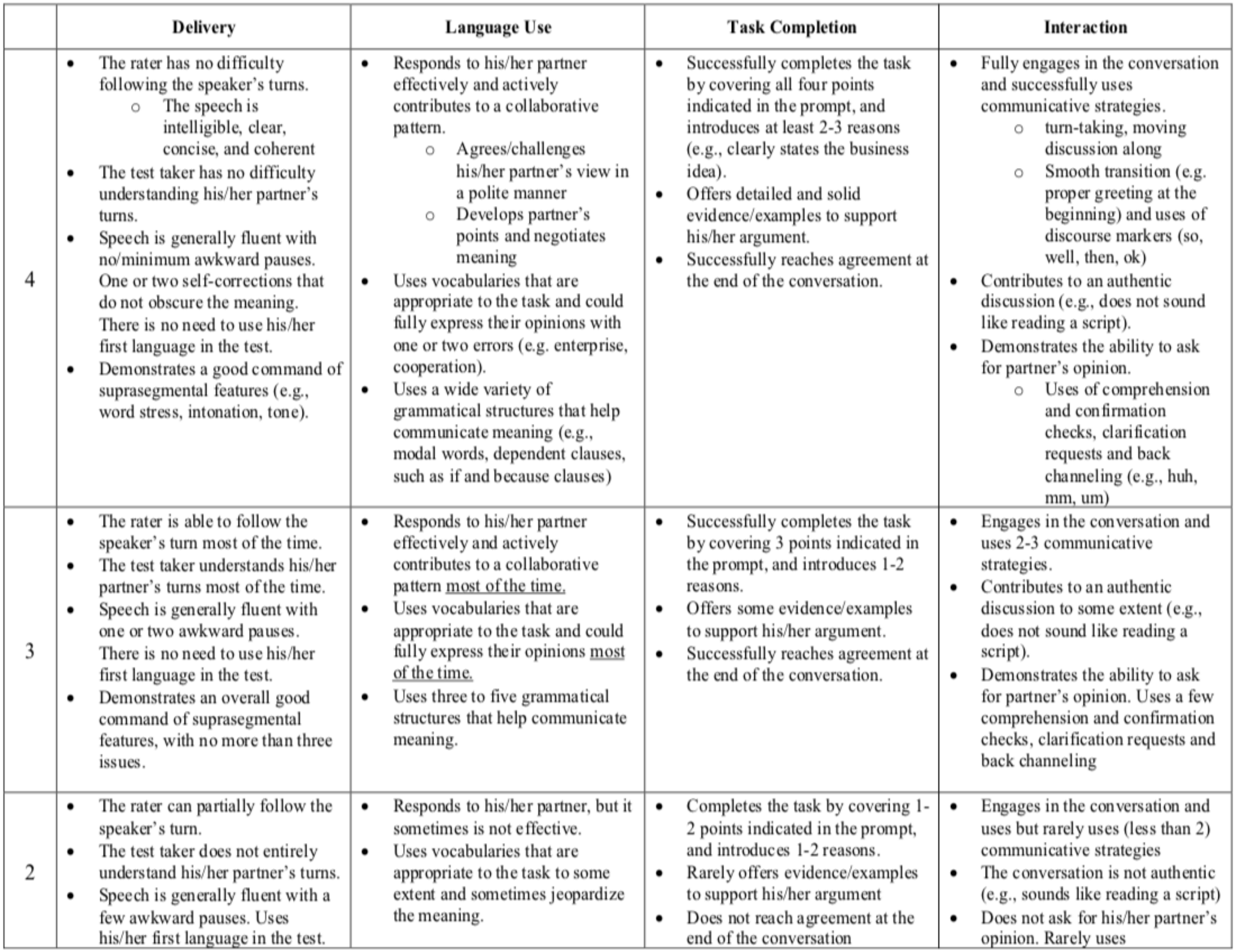
Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4,* 133-142.

Orr, M. (2002). The FCE speaking test: Using rater reports to help interpret test scores. *System*, *30*(2), 143-154.

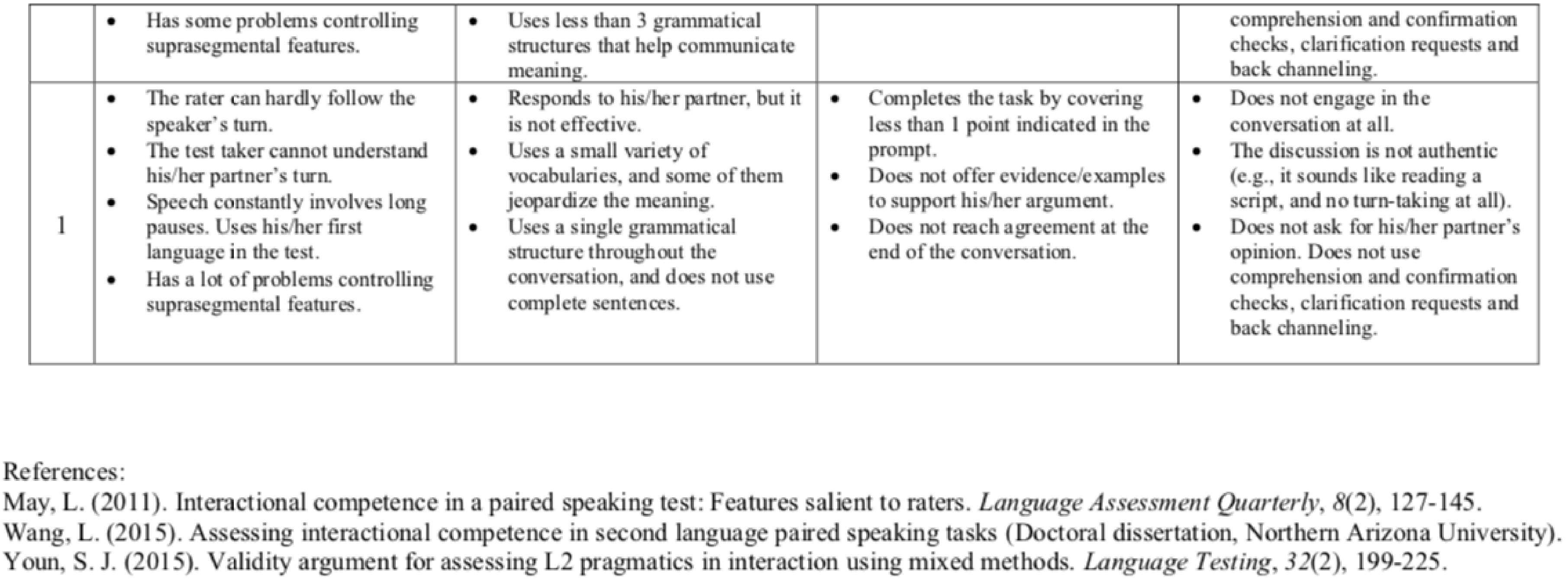
Richards, K. (2003). *Qualitative inquiry in TESOL*. Springer.

Appendix B

Rating Scale for The Paired Speaking Test



	Delivery	Language Use	Task Completion	Interaction
4	• The rater has no difficulty following the speaker's turns. ○ The speech is intelligible, clear, concise, and coherent • The test taker has no difficulty understanding his/her partner's turns. • Speech is generally fluent with no/minimum awkward pauses. One or two self-corrections that do not obscure the meaning. There is no need to use his/her first language in the test. • Demonstrates a good command of suprasegmental features (e.g., word stress, intonation, tone).	• Responds to his/her partner effectively and actively contributes to a collaborative pattern. ○ Agrees/challenges his/her partner's view in a polite manner ○ Develops partner's points and negotiates meaning • Uses vocabularies that are appropriate to the task and could fully express their opinions with one or two errors (e.g. enterprise, cooperation). • Uses a wide variety of grammatical structures that help communicate meaning (e.g., modal words, dependent clauses, such as if and because clauses)	• Successfully completes the task by covering all four points indicated in the prompt, and introduces at least 2-3 reasons (e.g., clearly states the business idea). • Offers detailed and solid evidence/examples to support his/her argument. • Successfully reaches agreement at the end of the conversation.	• Fully engages in the conversation and successfully uses communicative strategies. ○ turn-taking, moving discussion along ○ Smooth transition (e.g. proper greeting at the beginning) and uses of discourse markers (so, well, then, ok) • Contributes to an authentic discussion (e.g., does not sound like reading a script). • Demonstrates the ability to ask for partner's opinion. ○ Uses of comprehension and confirmation checks, clarification requests and back channeling (e.g., huh, mm, um)
3	• The rater is able to follow the speaker's turn most of the time. • The test taker understands his/her partner's turns most of the time. • Speech is generally fluent with one or two awkward pauses. There is no need to use his/her first language in the test. • Demonstrates an overall good command of suprasegmental features, with no more than three issues.	• Responds to his/her partner effectively and actively contributes to a collaborative pattern most of the time. • Uses vocabularies that are appropriate to the task and could fully express their opinions most of the time. • Uses three to five grammatical structures that help communicate meaning.	• Successfully completes the task by covering 3 points indicated in the prompt, and introduces 1-2 reasons. • Offers some evidence/examples to support his/her argument. • Successfully reaches agreement at the end of the conversation.	• Engages in the conversation and uses 2-3 communicative strategies. • Contributes to an authentic discussion to some extent (e.g., does not sound like reading a script). • Demonstrates the ability to ask for partner's opinion. Uses a few comprehension and confirmation checks, clarification requests and back channeling
2	• The rater can partially follow the speaker's turn. • The test taker does not entirely understand his/her partner's turns. • Speech is generally fluent with a few awkward pauses. Uses his/her first language in the test.	• Responds to his/her partner, but it sometimes is not effective. • Uses vocabularies that are appropriate to the task to some extent and sometimes jeopardize the meaning.	• Completes the task by covering 1-2 points indicated in the prompt, and introduces 1-2 reasons. • Rarely offers evidence/examples to support his/her argument • Does not reach agreement at the end of the conversation	• Engages in the conversation and uses but rarely uses (less than 2) communicative strategies • The conversation is not authentic (e.g., sounds like reading a script) • Does not ask for his/her partner's opinion. Rarely uses



	• Has some problems controlling suprasegmental features.	• Uses less than 3 grammatical structures that help communicate meaning.		comprehension and confirmation checks, clarification requests and back channeling.
1	• The rater can hardly follow the speaker's turn. • The test taker cannot understand his/her partner's turn. • Speech constantly involves long pauses. Uses his/her first language in the test. • Has a lot of problems controlling suprasegmental features.	• Responds to his/her partner, but it is not effective. • Uses a small variety of vocabularies, and some of them jeopardize the meaning. • Uses a single grammatical structure throughout the conversation, and does not use complete sentences.	• Completes the task by covering less than 1 point indicated in the prompt. • Does not offer evidence/examples to support his/her argument. • Does not reach agreement at the end of the conversation.	• Does not engage in the conversation at all. • The discussion is not authentic (e.g., it sounds like reading a script, and no turn-taking at all). • Does not ask for his/her partner's opinion. Does not use comprehension and confirmation checks, clarification requests and back channeling.

References:

May, L. (2011). Interactional competence in a paired speaking test: Features salient to raters. *Language Assessment Quarterly*, *8*(2), 127-145.

Wang, L. (2015). Assessing interactional competence in second language paired speaking tasks (Doctoral dissertation, Northern Arizona University).

Youn, S. J. (2015). Validity argument for assessing L2 pragmatics in interaction using mixed methods. *Language Testing*, *32*(2), 199-225.