The effects of rater's familiarity with test taker's L1 in assessing accentedness and comprehensibility of independent speaking tasks

Hwijung Lee
(Seoul National University)

Lee, Hwijung. 2017. The effects of rater's familiarity with test taker's L1 in assessing accentedness and comprehensibility of independent speaking tasks. *SNU Working Papers in English Linguistics and Language 15, 93-111*. This study aims to answer the question: does different degree of rater's familiarity with the test taker's L1 (Korean) affect how they assess accentedness and comprehensibility? Speech data was collected from four Korean undergraduate students and a total of 13 raters participated in the study. The raters are classified into three groups depending on the level of proficiency of test taker's L1 and years living in Korea: three of the raters are non-native speakers of Korean, five of them are bilingual speakers, and five are Korean native speakers. They were asked to rate the speech data of approximately 45 seconds for their accentedness and comprehensibility. The group of native speakers gave the lowest scores for accentedness and comprehensibility whereas the non-native speaker raters gave the highest scores for both. The bilingual rater scores fell in between the two groups: they were relatively lenient compared to the native speaker raters but harsh compared to the non-native speaker raters. There was a gradient-like result of the scores depending on the level of familiarity with the test taker's language. This study suggests that even a minimum level language familiarity can be a source of bias for there is a correlation with the severity in scoring. **(Seoul National University)**

Keywords: L1 familiarity, speaking assessment, bilingual raters, accentedness, comprehensibility,

1. Introduction

A recurrent issue in oral proficiency testing is rater bias and how to reduce it for fair and valid assessment of speech. Speaking construct is relatively difficult to operationalize. The distinct features that constitute speech are difficult to identify and assess individually since they are often closely linked to one another. If the listener cannot make out what is said due to a heavily accented pronunciation this may affect their

judgment of syntax, grammar or ideas as well. In this case, the pronunciation feature would act as a "first level hurdle" (Iwashita, Brown, Mcnamara, & O'Hagan, 2008). Even if the operationalization of speech measures is ideally accomplished, assessment and rating can be problematic if erroneous judgement results due to the rater's fault, hence the name *rater bias*. Rater effect could occur due to differences of individual severity, but this could also be related to rater's background such as their own L1 or their familiarity with the test's target language. For instance, one rater may be unable to comprehend speech because they lack the familiarity with the speaker's accent, whereas for another rater the speech may have been perfectly comprehensible. Then, how would a rater's familiarity with an accent play a role in their judgment of speaking tests?

This paper will begin by reviewing literature on accentedness and comprehensibility. Then, it will discuss several definitions of accent familiarity and explain how this study is a continuation as well as a complementation of this line of research. Lastly, it will present a preliminary study on the effects that the different levels rater's familiarity with test taker's L1 has in assessing their accentedness and comprehensibility.

2. Literature Review

2.1 Intelligibility, Accentedness, and Comprehensibility

Many scholars discuss the terms intelligibility, comprehensibility, and pronunciation in an attempt to analyze how disentangle them when researching oral proficiency assessment. Munro et al. (1999; 2006) clearly distinguish between foreign accent, comprehensibility, and intelligibility in their study. Their research on ten native speakers of Mandarin proves that the three measures are different and should not be confused with one another in rating scales. Intelligibility is the amount

of word-by-word transcription of an utterance. Comprehensibility is a measurement of how much the rater perceived they could understand. Foreign accent is the degree of native-like characteristics of the speech. This final definition could be a problematic way of assessing accentedness alone since it implies that an ideal educated native-speaker speech should be the norm. It is important to acknowledge that there are different accents and this paper is not concerned with how to "fix" them, but how different levels of familiarity play a role in rating them.

2.2 Defining Accent Familiarity

Scholars have different ways of judging rater's familiarity with a language. What Carey et al. (2011) call "interlanguage phonology familiarity" is related to (1) rater's self-identified exposure to the language and (2) the location where the rater is currently living in. Carey et al.'s study looks at three L1 groups (Chinese, Indian, and Korean) and find that candidates receive the highest pronunciation scores from the test centers located in their respective L1 country. The result of their study broadens the scope of what could be seen as accent familiarity. Bias could be related to not only the rater's knowledge of or proficiency of the test taker's L1, but their location as well. They did find that location plays an important role in rater's lenient scoring of pronunciation. Greater amounts of higher pronunciation scores were given to Korean test takers at the Korean test centers. However, self-identification and location alone are not sufficient factors that define language familiarity.

Winke and Gass (2013) define familiarity of accent as (1) being native speakers of the test taker's L1, (2) having studied the test taker's L1 as an L2, and (3) extended exposure to speech. They divide the third factor into (a) having lived in the country where the L1 is spoken, (b) worked with or taught speakers of that L1, or (c) having grown up around L1 speakers of the target language. Their specific operationalization of

familiarity evidently shows that the second definition of familiarity Carey et al. addressed, location, covers only (a). Their first definition, which is self-identification, is in that sense more subjective.

2.3 Limitations in Previous Studies

One of the limitations that Winke and Gass (2013) point out with previous studies on the assessment of oral proficiency is the short length of speech data and unnatural rating procedure. Speech data that is as short as 4.5 to 10.5 seconds, as those used by Munro et al. (2006) could result in faulty judgement of accent and comprehensibility, especially without a carefully designed rubric. Also, assessing intelligibility by transcribing speech data word by word produces quantitatively significant data (Munro & Derwing, 1999). However, as Winke and Gass (2013) argue, rating processes of oral proficiency tests normally do not comprise of transcribing speech. Then this brings a question to mind, to what degree should speaking tests and speaking assessment procedure reproduce the real-life performance? For instance, Pronunciation has extensively analyzed in a word level or sub-word level, by each syllables (Iwashita et al., 2007). This approach is questionable in terms of real-life application because it is not an ordinary practice for listeners to break down speech word-by-word in naturalistic conversations. Transcription and overtly in-depth analysis of audio files may be unrepresentative of real-life communication.

Another limitation from previous studies is the relatively low number of data on L2 Korean raters. Vas amount of research has been conducted on test takers of various L1 groups and rater's familiarity with their L1 such as Chinese, Spanish, Indian, French etc (e.g. Zhang & Elder, 2011; Xi, 2011; Munro et al., 2006; Trofimovich & Isaacs, 2012; Winke et al., 2012; Winke & Gass, 2013). Winke et al. (2013) has found evidence of rater's L2 background as a source of bias in rating oral performance for Spanish and Chinese speaker. Although their study also covered Korean L1 test

takers, their rater group of Korean L2 raters consisted of a small number (n=11) resulting in lack of evidence in this case.

Furthermore, Winke et. al (2012) discuss that more studies must be done to reveal whether a certain *'tipping point'* of L2 experience exist in rater bias. Accent familiarity and exposure to a language are not features that can easily be categorized by the terminologies of native and nonnativeness. For instance, Xi (2011) studied raters from India and examined rater effect under regular training and specialized training based on benchmarks samples of Indian examinees only to see if training was the answer to reducing bias. The results imply that proper training is indeed helpful in reducing rater bias. However the limitation with this study is that they only study the raters that are native speakers of the test taker's L1, when in fact there are many bilingual or even multilingual raters as well.

3. Research Question

Having addressed the limitations that this study is concerned with, first my aim is to collect speech data that is sufficiently long enough for a valid assessment of accentedness and comprehensibility. In order to work with speech that is closer to real-life discourse they will be spontaneously produced but will be controlled with a test question. Also, to control possible problems that unguided ratings could have, I will implement a scoring rubric for each of the two criteria that I would like to look at. The main question I would like to focus on to guide this study is the following: Do the different degrees of rater's familiarity with the test taker's L1 (Korean) affect their assessment of accentedness and comprehensibility?

4. Methodology

4.1 Participants

All the participants, or test takers, of speech sample collection in this study were limited to Seoul National University undergraduate students in to limit them to a pool of higher education. By higher education, not necessarily assuming a high level of English proficiency, but rather test taker's familiarity with and capacity of handling academic context. A total of four participants volunteered for data collection of speech samples. They are Korean native speakers with no experience living abroad in an English-speaking country. Their age ranges from 20 to 26 (mean = 22.75), two of them are male and two are female. All of them had either a TEPS or TOEIC score obtained within the past two years (See Appendix A) and none of them had taken the TOEFL iBT before. Their test scores converted into TOEFL iBT scores shows that their English proficiency level ranges from 86 to 113. All of them listed English as their L2, except for one participant who listed English as their third language after Korean and Japanese. Their L1 use in everyday life including their academic setting is minimum 70% (mean = 86.75%). A questionnaire about their background information was given after they had finished the speaking task (See Appendix B).

4.2 Speech Data Collection

4.2.1 Instrument

The TOEFL iBT independent speaking test questions were used in order to elicit spontaneous speech data of maximum 45 seconds. The question presented to test takers was limited to opinion types. For the study two questions were given to each test taker. The first question was: "If friends from another country were going to spend time in your country, what city or place would you suggest they visit? Using details and examples, explain why". The second question was: "Some people enjoy taking risks and trying new things. Others are not adventurous; they are cautious and prefer to avoid danger. Which behavior do you think is better? Explain

why". In the background questionnaire, test takers rated their perceived level of difficulty for each of the two questions on a scale from one to ten (one being very easy and ten being very difficult). The two question were fairly balanced in terms of difficulty level: the first question received a 3.5 (standard deviation = 1) and the second question received a 6 (standard deviation = 1.83).

4.2.2 Procedure

Participants were given a short orientation on what they would be instructed to do prior to recording. The question, printed out on a paper, was handed out to them. Once they were done reading the question they were be given 15 seconds to brainstorm, without note-taking. Then they were timed for 45 seconds to answer the question, which was recorded. Once finished with the recordings they filled out a one page questionnaire about their background to check whether they qualified for the purpose of this study. The recording was done using a phone and the locations all varied, which resulted in noise for some audio files. Limitations related to this will be discussed in the later sections.

4.3 Raters

4.3.1 Subgroups

In this preliminary study a total of 13 undergraduate and graduate students from Seoul National University participated in speech rating. These untrained raters were categorized into three subgroups. I took into account three different factors to categorize them. These were (1) their L1, (2) the language in which they claimed highest proficiency in, and (3) years living in Korea.

The first group (n=3) consisted of raters that claimed either Korean was not their L1 or had lived for less than 6 years in total in Korea. There was an L1 English speaker, an L1 Ukrainian speaker, and only one of them claimed that Korean was their L1 although this rater lived only for four

years in Korea (born to three years old, and less than a year as an adult). This rater explained that they had chosen Korean as a L1 because it was the first language they had learned but claimed higher proficiency in English. All three raters of this first subgroup had maximum a year experience teaching English to Korean students or tutees. There were two male and one female rater in this group.

The second group of raters (n=5) all spoke Korean as their L1. This group claimed to be bilinguals, with proficiency levels similar to their L1 Korean or lived for more than 10 years in Korea. Only one of them claimed to have a higher proficiency level in English, but had experience living for 10 years in Korea, therefore was grouped with the other bilingual raters. Their experience teaching English to Koreans varied from none to more than five years. There were four female and one male rater for this group.

The third group of raters (n=5) were native Korean speakers with the highest proficiency level in Korean. One of them expressed to be equally proficient in English as well, identifying themselves as a bilingual but had lived in Korea for 18 years thus was grouped in this category. All of them lived in Korea for more than 18 years, and three of them had no experience living abroad at all. Their experience teaching English to Koreans also ranged from none to more than five years. There were three female and two male raters for this group.

It may also be important to point out that ten out of 13 of these raters are from the department of English language and literature whereas the three raters, all of them in the first group, belong in other departments.

4.3.2 Scoring Rubric

The scoring rubric used for a holistic rating of accentedness and comprehensibility was created for the purpose of this study (see Appendix C) and its content was adapted from two previous studies done on accent and comprehensibility (Trofimovich & Isaacs, 2012; Crowther

et al., 2015). A nine-point Likert-type scale has been adapted following Trofimovich and Isaacs's example. In their study they look at 19 different speech measures and found that accent was linked to the phonology domain, and comprehensibility was mostly linked to grammar and vocabulary. With that established, for my rubric I selected only some of the categories for accentedness which were segmental errors, word stress errors, intonation and rhythm. Comprehensibility was narrowed down to four categories related to grammar and vocabulary and they were lexical appropriateness, lexical richness, grammatical accuracy, and grammatical complexity. The examples and details given for each categories were directly adapted from Crowther et al.'s (2015) study. The definition of score points for one (lowest for poor performance) and for nine (highest for best performance) each are explained for each subcategories in the rubric. Eventually, the raters will give a single holistic score for accentedness and for comprehensibility, keeping those subcategories in mind.

4.3.3 Procedure

Since each test-taker answered two different questions there were a total of eight audio files. Using an audio editing program any noticeably high noise was reduced and the sound files were normalized.

The rating was done on computer using a google form survey that was sent individually to the raters, along with an attachment of the rubric and the audio files. Each rater was asked to give a holistic score for accentedness and comprehensibility for each audio file.

Once they were done rating, they had to fill out a background questionnaire online. They were asked to rate their understanding of the holistic rubrics for accentedness and comprehensibility from a scale of one to nine (maintaining the numbers consistent with the nine-point Likert-type scale they had used previously for audio ratings: 1- difficult to understand, 9- being very easy to understand). The accentedness rubric

received on average an acceptably high score of 7.69 (standard deviation = 1.84), and the comprehensibility rubric received 7.69 (standard deviation = 1.70) as well. The numbers show that raters had adequately understood the rubrics given for the rating task. Afterwards they were asked to fill out information about their L1, the language that they have the highest proficiency level in, years in Korea, experience teaching English to Koreans, and their major. Lastly, in order to qualitatively check for any possible off-rubric thinking I added a comments or suggestion section at the end of the questionnaire. Since this was not a requisite category, only six raters left comments.

5. Results of the Preliminary Study

Firstly, I looked at the mean scores that each test taker had received for accentedness and comprehensibility from all raters (Table 1). A pattern that is evident is that all test takers received a lower score on accentedness compared comprehensibility. However, these scores differences was not significant, not more than 1-point of difference. Test taker 4 received the highest average scores and test taker 1 received the lower scores. Coincidentally test taker 4 happened to have the highest proficiency level, but test taker 1 did not have the lowest proficiency level. A second pattern than can be observed is that, the test taker's score rankings for accentedness and for comprehensibility are identical (4>3>2>1). Possibly, there may not have been a drastic distinction made between accentedness and comprehensibility but this requires further study with more data.

Table 1. Mean scores for *accentedness* and *comprehensibility* of each test takers

Test taker	***Accentedness*** (n=13)	***Comprehensibility*** (n=13)
1	4.27 (SD=1.33)	4.35 (SD=1.65)
2	4.42 (SD=1.14)	4.85 (SD=1.26)
3	4.69 (SD=1.38)	5.08 (SD=1.38)
4	5.42 (SD=1.79)	6.15 (SD=1.57)

The mean scores for accentedness and comprehensibility of each test taker given by each rater subgroup is organized in Table 2 below. The standard deviation scores was also calculated to check for any statistically misleading results due to the interrater differences. Most of the standard deviation scores were within 0 to 1.5 range (with the exception of eight results which were over 1.5).

Table 2. Mean scores for *accentedness* and *comprehensibility* of each test takers by groups of rater's familiarity with test taker's L1

	Group (1) Non-native speakers of test taker's L1 (n=3)		**Group (2) Bilinguals of test taker's L1** (n=5)		**Group (3) Native speakers of test taker's L1** (n=5)	
Test taker	*Acc.*	*Comp.*	*Acc.*	*Comp.*	*Acc.*	*Comp.*
1	**3.17** (SD=1.79)	**2.83** (SD=0.41)	**4.50** (SD=1.78)	**4.30** (SD=1.49)	**4.70** (SD=1.57)	**5.30** (SD=1.64)
2	**4.33** (SD=0.82)	**4.17** (SD=1.17)	**4.30** (SD=1.49)	**4.70** (SD=1.06)	**4.60** (SD=0.97)	**5.40** (SD=1.35)
3	**3.67** (SD=1.03)	**4.00** (SD=0.89)	**4.80** (SD=1.48)	**4.90** (SD=1.45)	**5.20** (SD=1.23)	**5.90** (SD=1.1)
4	**5.33** (SD=1.37)	**5.50** (SD=1.87)	**5.10** (SD=1.76)	**6.00** (SD=1.76)	**5.80** (SD=2.10)	**6.70** (SD=1.06)

Graph 1 and 2 below provides visual aid for data from Table 2. Noticeably there were differences among rater subgroups on their severity of ratings. The green line, which represent group 1, non-native speakers of test taker's L1, is located lowest in the graph meaning that

they had a tendency to give the most lenient scores compared to the yellow line located highest in the graph, representing group 3 native speakers of the test taker's L1. The blue line, for group 2 of bilingual speakers, is located in the middle of the lines for group 1 and group 3. This shows that the bilingual rater group were less lenient compared to the native speakers of Korean yet they were relatively harsher compared to non-native speakers of Korean. This pattern was true for both accentedness and comprehensibility.

Figure 1.

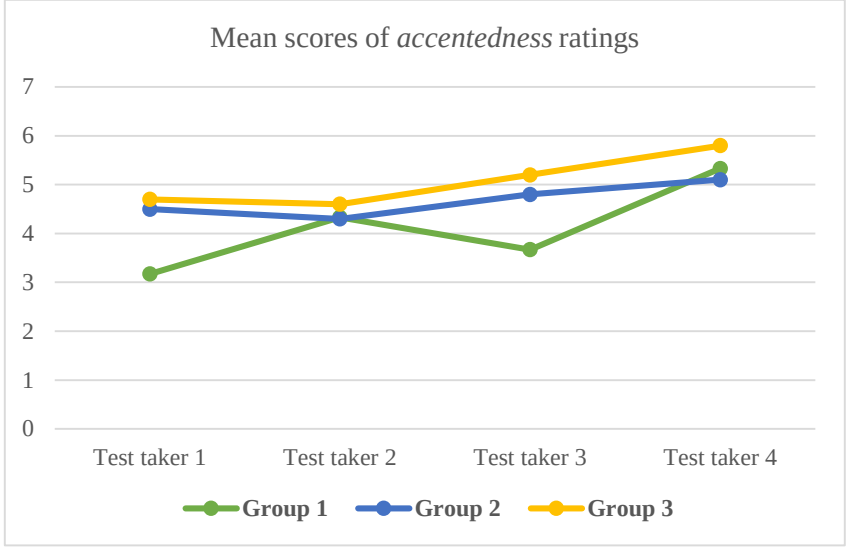
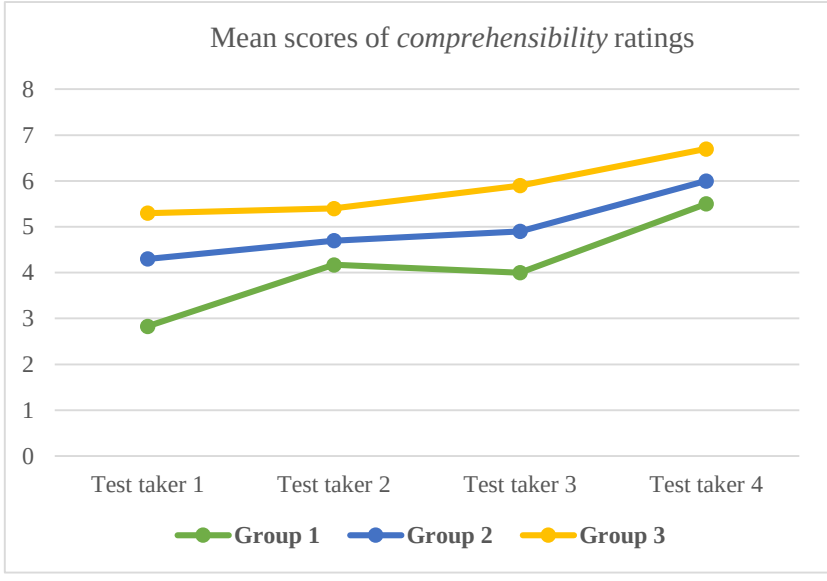


Figure 2.



6. Discussion and Further Study

Although the small sample size of my preliminary study can only lead to premature discussion, the results indeed prove to have some significance. In line with previous studies (Winke et al, 2012; Winke & Gass, 2013; Carey et al., 2010) there is certain bias that occurs due to rater's familiarity with the test taker's L1. One of the implications this study has for further studies is that a rater's language familiarity should be examined as a gradient. Rather than it being based on the concrete fact of whether they are natives or non-natives, a deeper investigation for each rater's background should be done through interviews or questionnaires. Many factors can be considered for this such as their L1, the language in which they have the highest proficiency level in, their experience living in a country of the target language, experience teaching

or working with the target language, etc. This study shows the possible association between rater's familiarities of the test taker's L1 and their leniency in scoring their accentedness and comprehensibility. The more familiar and experienced with Korean language raters were they would give the highest scores showing lenient scoring, but the more unfamiliar they were they would give the harshest or lowest scores.

There are several limitations to point out from this preliminary study. Firstly, there is the issue of sample size. There is still a need to collect large scale data for Korean test takers and raters with different levels of familiarity with the Korean language. Moreover, the data size for each subgroups were not well balanced. Overall the participants were small in sample size. Further study would require more test takers as well as more raters, balanced in numbers for each subgroup. Secondly, comments from the rater's questionnaire reflect on poor audio quality and noise. The technical problems may have interfered in their ratings of comprehensibility. Thirdly, the rater subgroup's independent variable should have been controlled only for their L1, the language with highest proficiency level, and their experience living in Korea. However, there was another factor that may have accounted for rating differences, which is that all three raters from the first group (non-native speakers of Korean) did not belong in the department of English language and literature whereas all the other raters did. This means that regardless of their language proficiency or experience in Korea, they may have been less familiar in the field of linguistics and may have had more trouble with the terminologies of the scoring rubric. In fact, one of the three raters of the first subgroup gave a very low score (four points) for their understanding of both rubrics. Even though I was dealing with untrained and nonprofessional raters, their familiarity with the rubric is a factor that could have been controlled by limiting raters to the same English language and literature major.

7. Conclusion

Supplementing previous research, my preliminary study looks at the rater effects when assessing speech accentedness and comprehensibility depending on the degree of familiarity that raters have with the test takers' L1. What discriminates this study is that it had been done on only Korean test takers. Also, I divided the rater groups based on minimal differences in language familiarity. This included a group of bilingual raters that claimed they were equally proficient in both the test taker's L1 and the target language. The preliminary study results were positive in that there was a gradient-like relationship between the level of familiarity and the severity in scoring. This suggests that the different degrees in language familiarity could be a possible source of rater bias.

References

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*(2), 201-219.

Crowther, D., Trofimovich, P., Saito, K., & Isaacs, T. (2015). Second Language Comprehensibility Revisited: Investigating the Effects of Learner Background. *TESOL Quarterly,49*(4), 814-837.

Derwing, T. M., & Munro, M. J. (2005). Second Language Accent and Pronunciation Teaching: A Research-Based Approach. *TESOL Quarterly,39*(3), 379.

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

Isaacs, T. (2008). Towards Defining a Valid Assessment Criterion of Pronunciation Proficiency in Non-Native English-Speaking Graduate Students. *Canadian Modern Language Review,64*(4), 555-580.

Iwashita, N., Brown, A., Mcnamara, T., & O'Hagan, S. (2008). Assessed

Levels of Second Language Speaking Proficiency: How Distinct? *Applied Linguistics,29*(1), 24-49.
Munro, M. J., & Derwing, T. M. (1999). Foreign Accent, Comprehensibility, and Intelligibility in the Speech of Second Language Learners. *Language Learning,49*, 285-310.
Munro, M. J., Derwing, T. M., & Morton, S. L. (2006). The Mutual Intelligibility Of L2 Speech. *Studies in Second Language Acquisition,28*(01).
Trofimovich, P., & Isaacs, T. (2012). Disentangling accent from comprehensibility. *Bilingualism: Language and Cognition,15*(04), 905-916.
Winke, P., Gass, S., & Myford, C. (2012). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing,30*(2), 231-252.
Winke, P., & Gass, S. (2013). The Influence of Second Language Experience and Accent Familiarity on Oral Proficiency Rating: A Qualitative Investigation. *TESOL Quarterly,47*(4), 762-789.
Xi, X., & Mollaun, P. (2011). Using Raters From India to Score a Large-Scale Speaking Test. *Language Learning,61*(4), 1222-1255.
Zhang, Y., & Elder, C. (2011). Judgments of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary constructs? *Language Testing,28*(1), 31-50.

Hwijung Lee
Hwijung94@snu.ac.kr

Appendix A

Data collected from the background questionnaire of test takers (participants)

Test taker	English test (score)	Converted TOEFL iBT score	Languages in order of proficiency	Test taker's perceived percentage of L1 use
1	TEPS(737)	98	Korean, English	95%
2	TOEIC(885)	103	Korean, English, Chinese	97%
3	TEPS(615)	86	Korean, English, Russian	85%
4	TOEIC(940)	113	Korean, Japanese, English	70%

Appendix B

Questionnaire for test takers to check whether they qualify for the study.

1. Department/major: _____
2. Age: _____
3. Gender: _____
4. List the languages you speak in order of proficiency:
   i. _____
   ii. _____
   iii. _____
   iv. _____
5. (If answer to 4-i is Korean) What percentage of use does your mother language occupy out of all the languages that you speak? _____%
6. Did you take an English test within the past two years?

_____

A. If yes, which test? _____

B. What is your score? _____

7. (If answer to 1-A is not TOEFL iBT) Have you taken the TOEFL iBT test before? _____

8. Have you lived or studied abroad? _____

   A. If yes, where? _____

   B. For how long? _____

   1 (very easy) ←----------→ (very hard) 10

9. On a scale of 1-10, how difficult was speaking task 1? _____

10. On a scale of 1-10, how difficult was speaking task 2? _____

Appendix C

Holistic scoring rubric given to raters for their assessment.

Accentedness	
Segmental errors (1 = frequent, 9 = infrequent or absent)	*dat* instead of *that, pin* instead of *pen, 'ouse* instead of *house, supray* instead of *spray*
Word stress errors (1 = frequent, 9 = infrequent or absent)	*com-pu-TER* instead of *com-PU-ter,* or absence of discernible stress giving all syllables equal prominence such as *com-pu-ter*
Intonation (1 = unnatural, 9 = natural)	Appropriate pitch moves that occur in native speech such as rising tones for yes/no questions (*Will you be home tomorrow?*↑) or falling tones at the end of statements (*Yeah, I'll stay at home*↓)
Rhythm (1= unnatural, 9= natural)	Appropriate difference in stress/emphasis. For instance *They RAN to the STORE.* Content words *ran* and *store* are stressed more than *they, to,* and *the* which are function words

Comprehensibility	
Lexical appropriateness (1 = many inappropriate words used, 9 = consistently uses appropriate vocabulary)	Speaker's choice of words to accomplish the task. Poor lexical choices include incorrect, inappropriate, and non-English words (ex: *A man and a woman bumped into each other on a walkside)*
Lexical richness (1 = few, only simple words used, 9 = varied vocabulary)	Sophistication of the vocabulary used (*The girl arrived home her dog was happy she arrived* home ↔ *The girl arrived home to find her dog overjoyed at her return*)
Grammatical accuracy (1 = poor grammar accuracy, 9 = excellent grammar accuracy)	Defined as the number of grammar errors, for instance word order error (*What you are doing?*), morphology (*She go to school every day*), and agreement (*I will stay there for five day*)
Grammatical complexity (1 = simple grammar, 9 = elaborate grammar)	Sophistication of the grammar. Grammatical complexity is low if the speaker uses simple, coordinated structure without embedded clauses or subordination (*The man wore a black hat and enjoyed his coffee* ↔ *The man that was wearing a black hat was enjoying his coffee*)