Native and Non-Native Raters of L2 Speaking Performance

Valeriia Bogorevich

Northern Arizona University

Abstract

Rater variation in performance assessment can impact test-takers' scores and compromise assessments' fairness and validity (Crooks, Kane, & Cohen, 1996); therefore, it is important to investigate raters' scoring patterns in order to inform rater training. In this study, two groups of raters 23 native (North American) and 23 non-native (Russian) raters graded speech samples from Arabic (n = 25), Chinese (n = 25), and Russian (n = 25) L1 backgrounds. Raters' scores were examined using the Multi-Faceted Rasch Measurement using FACETS (Linacre, 2014) software to test group differences between native and non-native raters. The results revealed that there were no radical differences between native and non-native raters; however, the non-native raters showed more lenient grading patterns towards the students with whom their L1 matched.

Native and Non-Native Raters of L2 Speaking Performance

Background

Language testers have always been interested in rater variation that occurs in raters scoring L2 performance assessment (i.e., writing and speaking). Research has shown that raters differ in their scores for the same written essay (e.g., Barkaoui, 2007) or spoken sample (Orr, 2002).

One area of research on rater variability has addressed the possible group differences that might be caused by raters' native- or non-native-speaker status. Language testers have raised concerns that native and non-native raters may differ in terms of their understanding of certain aspects of rating, for example, cultural communication norms (e.g., Brown, 1995) or written rhetorical patterns (e.g., Kobayashi & Rinnert, 1996), which may cause differences in scores. Another argument that is given to support the prospective differences is that non-native raters can have very diverse backgrounds or come from an area with established English dialects. Such backgrounds of non-native raters can affect their ability to evaluate language performances. Studies comparing native and non-native raters have been done in writing assessment (e.g., Johnson & Lim, 2009; Shi, 2001), speech perception and pronunciation (Fayer & Krasinski, 1987; Kang, 2012; Saito & Shintani, 2016), and speaking assessment (Kim, 2009; Zhan & Elder, 2011). Some of the studies showed that non-native speakers are more severe (e.g., Zhang & Elder, 2014; Brown, 1995; Fayer & Krasinski, 1987; Kang, 2012) or that native speakers are more severe (e.g., Barnwell, 1989) whereas other studies showed no differences (Xi & Mollaun, 2009; Wei & Llosa, 2015).

Studies comparing native and non-native raters differ in their findings and contradict one another. One explanation is that the differences in the outcomes of the studies may be attributed to differences in rater populations and research designs. The studies that have compared native and non-native raters have been done in speaking or writing; with or without a rubric; grading mono or multi-lingual students; looking at English, Spanish, or Arabic; with or without rater training; involving naïve and experienced raters as well as teachers and non-teachers. Some studies (e.g., Zhang & Elder, 2011) showed that the quantitative difference could not be seen, but some differences can be uncovered using a qualitative approach. Zhang and Elder (2014) pointed out that native and non-native "raters may arrive at their judgments via somewhat different pathways and show different degrees of tolerance of breakdowns in relation to particular features of speech" (p. 318).

The differences that may occur when comparing *native* to *non-native* raters are also seen when comparing *native* to *native* raters grading speaking (e.g., Chalhoub-Deville 1995; Chalhoub-Deville & Wigglesworth, 2005). Chalhoub-Deville and Wigglesworth's study compared native speakers from four native speaker backgrounds. The results illustrated that the U.S. raters were most lenient, U.K. raters most severe, and Canadian and Australian raters were in-between. An interesting explanation for these results was offered in the area of educational measurement by Suto (2012) who stated that rater agreement or disagreement could depend on their "community of practice" or "school of thought." The author suggested that "it is likely that raters of equal experience and eminence would hold different understandings of what constitutes a good response and interpret the scoring criteria slightly differently, despite common training on those questions" (p. 23). Another research study also showed the discrepancies among native raters due to their L2 background, because they were heritage speakers, or communicated with

non-native speakers of a similar L1 background on a regular basis (Winke, Gass, and Myford, 2011).

The studies discussed above suggest that there may be a difference in raters that is not only driven by the native or non-native affiliation but also based on raters' familiarity and exposure to other people who speak similarly (or with a similar accent) due to the same L1 background. In support of this idea, some studies have suggested that raters' familiarity expressed through knowledge of test-takers' L1 would impact their ratings (e.g., Winke, Gass & Myford, 2011). In another study, Carey, Mannell, and Dunn (2011) looked at the impact of raters' residence in the examinees' country. Both studies revealed that familiarity and exposure affected raters' scores.

The current study focused on investigating the rating behavior of native and non-native raters in order to uncover the differences in rating patterns when scoring speaking performance by multilingual test-takers. A group of examinees with whom raters share the L1 was also included in order to examine another potential source of rater variability, which is L1 match of raters' and examinees.

Research Questions

RQ1: To what extent do NS and NNS raters differ in terms of consistency and severity of their analytic scores?

RQ2: To what extent do NS and NNS raters differ in terms of scoring examinees by L1?

Method

Participants

Raters. The raters in the study were comprised of 23 experienced Russian EFL/ESL teachers as non-native speaking raters (NNS) and 23 experienced North American EFL/ESL teachers as native speaking raters (NS) (Table 1).

Table 1

Rater Demographic Information

	NS	NNS
Number	23	23
Age	*M* = 34, *SD* = 10*	*M* = 30, *SD* = 5
Gender	10 males and 13 females	4 males and 19 females
Teaching experience	*M* = 8.55, *SD* = 6.57	*M* = 7.78, *SD* = 5.41

Note. **M* = 32, *SD* = 6 without the oldest participant (71 years old).

Examinees. This study used 99 speech recordings in response to two independent speaking prompts (see Instruments) from a semi-direct speaking test. The recordings included Chinese (*n* = 33), Arabic (*n* = 33) and Russian (*n* = 33) speakers (Table 2). In terms of gender, there were more male than female recordings (50 males and 25 females). There were 5 female and 20 male recordings for Arabic speakers, 9 females and 16 males for the Chinese group, and 11 female and 14 male speech samples in the Russian group. The recordings from Arabic and Chinese L1 backgrounds were obtained from an archived database from an administration of a placement test at an IEP in the United States, while Russian L1 recordings were collected from an IEP located in Russia using the same administration process of the speaking task.

Table 2

Total Number of Recordings by Each L1

Procedure	Arabic	Chinese	Russian	Total
Training and Calibration	8	8	8	24
Individual Rating	25	25	25	75
Total per L1	33	33	33	99

Instruments

Speaking prompts. Two speaking prompts were used to obtain speakers' speech samples. Task 1 was an opinion task asking an alternative question about how a person prefers to study for an exam (i.e., alone or in a group). Task 2 was another opinion task asking an alternative question about what size of classes is better for students (i.e., big or small).

Rating rubric. The raters used the TOEFL iBT independent rubric in this study. The rubric was chosen because it represents a common speaking rubric with four rating sections including General Description, Delivery, Language Use, and Topic Development.

Results

RQ1: To what extent do NS and NNS raters differ in terms of consistency and severity of their analytic scores?

First, the NS and NNS raters are compared in terms of their ability to maintain their internal consistency. Regarding rater self-consistency, there were 5 misfitting NS raters and 4 NNS raters as well as 7 overfitting NS raters and 4 overfitting NNS raters. In other words, there were almost the same number of NS and NNS raters who exhibited erratic rating patterns, and there were more NS raters who showed overly-consistent rating patterns. In terms of self-

consistent raters, there were 11 NS and 15 NNS whose infit statistics were within the targeted 1.2 and 0.8 range. Overall, the mean infit square for both groups were close to each other M = .99, SD = .27 for the NS group and M = 1.01, SD = .20. Thus, there was no significant group difference in self-consistency: t = -0.31, df = 44, p = 0.758025 coupled with minimal Cohen's d = 0.09. Overall, the NS and the NNS rater groups exhibited similar internal consistency patterns.

Second, the NS and NNS raters were compared in terms of their statistical severity measures based on the Facets measurement report. There were 15 NS and 13 NNS raters placed above the average severity level of 0, while 8 NS raters and 10 NNS raters exercised more lenient rating patterns. The mean severity logits for both groups were close to each other with M = .12, SD = .55 for the NS group and M = -.12, SD = .78 for the NNS group. Therefore, there were no significant group differences in severity: t = 1.15, df = 44, p = 0.256356; however, Cohen's d = 0.34 showed a small effect size. In other words, although the NS raters showed a tendency to provide more severe ratings, there were no statistically significant differences regarding the overall severity of the NS and NNS groups of raters.

RQ2: To what extent do NS and NNS raters differ in terms of scoring examinees by L1?

To compare whether the NS and NNS groups rated each examinee L1 group differently, three separate Facets analyses were conducted, one for each examinee L1. The statistical information from Facets output files about rater performance is presented in Table 12 for Arabic L1, Table 13 for Chinese L1, and Table 14 for Russian L1. Rater groups' consistency and severity measures were compared across examinee L1s.

First, looking at infit statistics of rater groups for each examinee L1, it can be seen that neither NS nor NNS raters exceeded the targeted 1.2 - 0.8 values. Infit measures for the NS

raters were *M* = .95, *SD* = .26 for Arabic L1, *M* = .96, *SD* = .33 for Chinese L1, and *M* = 1.04, *SD* = .62 for Russian L1. For the NNS raters, the infit measures were *M* = 1.04, *SD* = .33 for Arabic L1, *M* = 1.03, *SD* = .38 for Chinese L1, and *M* = .99, *SD* = .51. It can be concluded that, on average, both rater groups exhibited similar internal consistency across examinee L1 groups viz. both rater groups scored each examinee L1 group consistently.

Second, severity of NS and NNS raters was compared per examinee L1. Based on the severity mean for the NS group (*M* = .07, *SD* = .71) and the NNS group (*M* = -.07, *SD* = .77), there were no radical differences between the NS and NNS raters scoring Arabic L1 students. Moreover, there were no differences between these groups rating Chinese L1 students (NS: *M* = .06, *SD* = .84 and NNS: *M* = -.06, *SD* = .96). Although there were no differences, in both cases, the NNS group can be described as a more lenient one. Furthermore, more difference can be seen between the rater groups when they rated Russian L1 students. The NNS raters exhibited a more lenient scoring pattern (*M* = -.27, *SD* = 1.02) than the NS group (*M* = .27, *SD* = .58), which was statistically significant (*t* = 2.16, *df* = 44, *p* = .036308, Cohen's *d* = 0.65). Overall, based on the severity measures, the NS rater group was more severe across L1s. There were no significant differences between the NS and NNS rater groups for Arabic and Chinese L1s, but the NNS rater group was significantly more lenient when rating Russian L1 examinees who share the same L1.

Relevance to PIE

The PIE can use the results of the study when considering if NS and NNS raters may exhibit differences in rating L2 speaking performance on achievement and placement/exit tests. The present study provides backing to the fact that proficient and experienced NNS can exhibit severity and consistency levels comparable to NS raters, which means that NNS can be used for

scoring speaking exams. Even though there were detected L1 match differences, the study still argues for the inclusion of NNS as raters, but suggests providing more specific training for NNS raters who can be prone to exhibit some degree of positive bias. The special training can include more comprehensive guidelines for NNS raters about how to approach scoring examinees with shared L1.

References

Barkaoui, K. (2007). Participants, texts, and processes in ESL/EFL essay tests: A narrative review of the literature. *Canadian Modern Language Review*, *64*, 99–134.

Barnwell, D. (1989). 'Naive' native speakers and judgments of oral proficiency in Spanish. *Language Testing*, *6*, 152-163.

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

Carey, M. D., Mannell, R. H., & Dunn, P. K. (2011). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing*, *28*, 201–219. doi:10.1177/0265532210393704

Chalhoub-Deville, M. & Wigglesworth, G. (2005). Rater judgment and English language speaking proficiency. *World Englishes*, *24*, 383–391.

Chalhoub-Deville, M. (1995). Deriving oral assessment scales across different tests and rater groups. *Language Testing*, *12*, 16–33.

Crooks, T. J., Kane, M. T., & Cohen, A. S. (1996). Threats to the valid use of assessment. *Assessment in Education*, *3*, 265-285.

Fayer, J. M. & Krasinski, E. (1987). Native and non-native judgments of intelligibility and irritation. *Language Learning*, *37*, 313–326..

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

Kang, O. (2012). Impact of rater characteristics and prosodic features of speaker accentedness on ratings of international teaching assistants' oral performance. *Language Assessment Quarterly, 9*, 249-269.

Kim, Y. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing*, *26*, 187–217. doi:10.1177/0265532208101010

Kobayashi, H. and Rinnert, C. (1996). Factors affecting composition evaluation in an EFL context: cultural rhetorical pattern and readers' background. *Language Learning 46*, 397–437.

Linacre, J. M. (2014) Facets computer program for many-facet Rasch measurement, version 3.71.4. Beaverton, Oregon: Winsteps.com

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

Saito, K., & Shintani, N. (2016). Do native speakers of North American and Singapore English differentially perceive second language comprehensibility? *TESOL Quarterly, 50*, 421-446.

Shi, L. (2001). Native- and non-native-speaking EFL teachers' evaluation of Chinese students' English writing. *Language Testing*, *18*, 303–325.

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

Wei, J., & Llosa, L. (2015). Investigating Differences Between American and Indian Raters in Assessing TOEFL iBT Speaking Tasks. *Language Assessment Quarterly, 12*, 283-304.

Winke, P., Gass, S., & Myford, C. (2011). The relationship between raters' prior language study and the evaluation of foreign language speech samples. *ETS Research Report Series*, 2, i-67.

Xi, X., & Mollaun, P. (2009). How Do Raters From India Perform in Scoring the TOEFL iBT™ Speaking Section and What Kind of Training Helps?. *ETS Research Report Series*, 2009 2, i-37.

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*, 31–50. doi:10.1177/0265532209360671

Zhang, Y., & Elder, C. (2014). Investigating native and non-native English-speaking teacher raters' judgments of oral proficiency in the College English Test-Spoken English Test (CET-SET). *Assessment in Education: Principles, Policy & Practice*, *21*, 306-325.