Judgments of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary constructs?

Ying, Z., & Catherine, E. (2011). Judgments of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary constructs? *Language Testing*, *28* (1), 31-50. doi: 10.1177/0265532209360671

Background

The aim of this research is to investigate the possibility of differing orientations to the oral proficiency construct by native and non-native English speaking teacher assessors.
This research relates to two arguments:

(1) the role of native speaker (NS) norms in English language teaching and testing and in communication
(2) the tendency amongst non-native users of English to distrust local norms.

The ideal of English for EFL/ ESL learners is native speakers' English. Therefore, NS's conversations and texts were used as EFL/ ESL learners' text.
NSs are often used in test pilot studies, since in case they cannot show their performance in the test, the test must be unreasonable. Furthermore, NSs usually play a role of a rater of tests.

Recently, the idea which is English as a lingua franca is rapidly spreading.
Thus, the NS norm is put in question from the perspectives of (1) definition, (2) ownership, (3) social identity, (4) World Englishes and some other points.

A number of studies about the NS and non-native English speaker (NNES) raters' rating approach are published, but there are some ambiguities about those results.
No consensus regarding the effect of language background on rating behavior has yet been reached.

Methodology

The College English Test-Spoken English Test (CET-SET).

CET-SET is a large-scale nation-wide spoken English test, and it is used for Chinese college students to measure their oral English speaking proficiency.
This test is administrated in a face-to-face small group format. Three or four test takers and two examiners participate to this test. One rater concentrate on rating test takers' performance, and another examiner acts as an interviewer.

Sub-tests are fall into three categories;
(1) answering the questions raised by the examiner,
(2) making a presentation according to a prompt card,
(3) participating in a group discussion.

Raters gives 1-5 on each of the three rating categories;
(C1) Accuracy and Range of test-takers' language,
(C2) Size and Discourse Management,
(C3) Flexibility and Appropriacy.
The sub-scores are weighted (C1 x 1.2, C2 x 1.0, C3 x 0.8) and added up and the total is converted to a final grade for reporting purpose.
Total score= C1 x 1.2 + C2 x 1.0 + C3 x 0.8

Research Questions

1) Do NNES and NES rater groups differ in consistency and severity is unguided holistic rating of CET-SET performance?
2) How do raters from each group differ in defining the oral proficiency construct as manifest in CET-SET test performance?

RQ1 is set to see how a group of NNES and NES raters marked the spoken test with holistic scale which does not have any specific criteria.
RQ2 is set to analyze the rater group's construct of spoken English ability.

Participants

The characteristics of the raters are following table.

	NES	NNES	Sum
male	9	7	16
female	10	13	23
sum	19	20	39

All NNES raters had learned English at school and most of them educated in a teacher education university. Their teaching experiences are 2 to 35years.
10 NES raters teaching ESL or language-related subjects in university in Australia, and another 9 participants employed as EFL teacher in China. Six of them have a qualification as a rater of IELTS. Close to half of them have experienced as a rater of other kinds of oral tests.

Data elicitation

Each rater group assessed 10 CET-SET speech samples selected from a pool of officially rated videotapes to represent a wide range of proficiency levels. Each tape was approximately 20mins.
The assessment took the form five-point scale was used: that is 1 (very poor) to 5 (Excellent).
Raters did not use the official holistic scales.

Data analysis

As for RQ1, data were analyzed using many-facet Rasch measurement (MFRM) to see the differences in rater severity and consistency.
Each facet had a set of elements:

(1) 30 candidate elements in the Candidate facet,
(2) 39 rater elements in the Rater facet,
(3) 2 elements of NNES group and NES group in the Rater Group facet.

The first set encompasses fit analysis and a single rater-rest of the raters correlation, to provide information on intra- and inter-group consistency. The second set includes group severity measures at the individual and group level. The third one reported in logit estimates for each element of the Group facet along with the result of a chi-square test of the significance of differences.

For the analysis of the MFRM, Wright and Linacre's (1994) idea "mean square values of 0.6 to 1.4 as lower and upper limit quality control limits" was adopted.

RQ2 is used to explore group differences in defining the CET-SET oral proficiency construct, retrospective written comments on holistic rating were analyzed.
This analysis was based broadly on the method described by North and Schneider (1998), namely, 'the intuitive identification of key 'features' at different levels through rater discussion of performance samples ranked in consensus order'.

The feature of the oral proficiency construct that featured in raters' judgments was identified independently by individuals immediately after they had rated a performance sample.
The written comments given by each rater participant were transcribed, segmented into unit turns with a specific idea, and numbered.

Results

RQ1

Table 1. Group measurement report

Group	Obsvd Average	Fair–M Average	Measure	Model S.E.	Infit		Outfit		Estim. Discrm.	Corr. PtBis
					MnSq	ZStd	MnSq	ZStd		
NNES	6.4	6.35	−0.41	0.04	0.91	−1.6	0.89	−1.7	1.16	0.58
NES	6.4	6.34	−0.41	0.04	1.07	1.1	1.14	2.1	0.84	0.54
Mean	6.4	6.35	−0.41	0.04	0.99	−0.2	1.02	0.2		0.56
S.D.	0.0	0.00	0.00	0.00	0.08	1.4	0.12	2.0		0.02

Separation: 0.00; Strata: 0.33; Reliability: 0.00
Fixed (all same) chi-square: 0.0; d.f.: 1; Significance (probability): 0.97

Column 4

Column 6

Column 11

The column 6 shows the infit mean-square index. This result indicates the intra-group consistency of both the NES and NNES raters.

The column 11 shows the single rater-rest of the raters' correlation of the two rater groups. This result implies a high inter-group consistency.

The column 4 reveals that the NNES and the NES rater groups are equally severe in their holistic rating of the CET-SET. There is no significant difference in severity between the two rater groups.

RQ2

Figure 1 (p.40) shows the percentage of mentions of each language proficiency category by rater group.

This finding shows that

Linguistic Resources was mentioned more often by the NNES raters as a relevant factor in their judgments than by the NES raters;

The NES raters mentioned all the other categories more frequently than the NNES did.

This result indicates that the NES raters focused on a wider range of abilities in judging test taker' oral test performance than did the NNES raters.

The chi-square test was used to see the different frequencies of mentions of the category between the two rater groups. The table 2 (p.41) shows the result. There are no significant differences for the category of fluency and content, on the other hand, significant differences were observed in other 5 categories.

NNES raters gives more comment for linguistic resources category than NES raters, in contrast, NES gives many comments for every categories.

Table 2. Categories showing significant differences in number of mentions by rater group

Category		Frequency of mentions		p	Direction of difference
		NNES	NES		
1	Fluency	101	156		NNES < NES
2	Content	174	234		NNES < NES
3	Linguistic Resources	332	296	*	NNES > NES
4	Interaction	65	139	*	NNES < NES
5	Demeanor	18	49	*	NNES < NES
6	Compensation Strategy	4	19	*	NNES < NES
7	Other General Comments	19	42	*	NNES < NES

*$p < 0.05$

Table 3 shows the result of subcategories difference between the two rater groups. Significant differences are observed within every subcategories.

Table 3. Subcategories showing significant differences in number of mentions by rater group

Category	Chi-square	p	Subcategory	Directionality
Content	88.59	0.00*	*Idea	NNES < NES
			*Relevancy to the Topic	NNES < NES
			*Completeness of Response	NNES > NES
			*Appropriateness of Response to Context	NNES > NES
			*Content (global)	NNES < NES
Linguistic resources	30.26	0.00*	*Vocabulary	NNES < NES
			*General Linguistic Resources	NNES > NES
Interaction	33.538	0.00*	*Interaction (global)	NNES < NES
			*Participation in the Group	NNES > NES
			*Impact on Listener (intelligibility)	NNES < NES
			*Conversation Management	NNES < NES
			*Other Interactive Communication Skills	NNES > NES

*$p < 0.05$

Interesting result appears in the subcategory of vocabulary. Although NNES raters gave more comments about the linguistic resources than NES did, NES gives comments more frequently in respect of subcategory of vocabulary.

Discussion

The answer to the RQ1 is that there are no differences in consistency and severity between two rater groups.

The authors conclude that “it may not matter whether native or non-native speakers are used as raters

at least as far as the likelihood of being assigned to poor, good or excellent categories is concerned."
The differences of the tendency between two rater groups were observed in the category of the written comments.
The NES raters' intensive focus on the Interaction and Compensation Strategy categories, on the other hand, the NNES raters focused more on test takers' underlying language ability as manifested through task performance.
The differences of their rating orientation may be caused by their various language experiences.

Conclusion and implication

There are no differences in scoring decision, but a number of quantitative and qualitative differences were observed in the process of written justifications.
The native/ non-native dichotomy is not meaningful in that raters from their different background.
Different orientation for rating may be due to their different social, cultural and educational experience deriving from the fact that English is the first language used at home or in the wider society.
The limitations of this study are; convenience sampling, test context and the component of this study.

Additional comment

This study is taken place in limited way and specific situation; however, the result shows the coordinate scoring decision between non-native English speaker raters and native English speaker raters. In Japan, native English speakers are mythicized that they are eminent English teacher, but this result might demolish the myth. As the result of RQ2 shows the NNES gave more comments about linguistic resources. Personally, this tendency may be observed in raters of English teachers of Japanese. Paying attention to other facets of test takers' performance is quite important for raters to give a helpful feedback to test takers.
Three facets were set to analyze the differences in rater severity and consistency. MFRM is used to observe the interacting factors in data, and Rasch model has the feature of test-free person measurement and sample-free item calibration; therefore, three facets were used to analyze the differences in rater severity and consistency.