Cognitive Processes of Native and Nonnative Teachers as Raters of L2 Speaking Performance

Valeriia Bogorevich

Northern Arizona University

Abstract

The present study used a mixed methods approach to investigate the prospective differences in how native English-speaking (NS) and nonnative English-speaking (NNS) teachers assess foreign students' speaking performance. The study analyzed teachers' scoring behavior when using a TOEFL iBT speaking rubric, which was used as an analytic one. The two groups of teachers were 5 American NS teachers and 5 Russian NNS teachers of English. All teachers scored the same 12 recordings answering an independent speaking task. Among speech samples, there were 4 Arabic speakers, 4 Chinese speakers, and 4 Russian speakers, which were pre-scored to determine variability. Teachers' analytic scores were analyzed using Multi-Faceted Rasch Measurement analysis, and the results showed that there is no difference between NNS and NS in terms of severity. The qualitative analysis of the think-aloud data and interview demonstrated that the raters exhibited several patterns of rating depending on their accent familiarity, focus while listening to examinees' performance and during the decision-making process, which caused score variability. These findings can be used by testing companies to study their raters' scoring behavior to individualize rater training in order to make exam ratings fair and raters interchangeable.

Cognitive Processes of Native and Nonnative Teachers as Raters of L2 Speaking Performance

Background

Because of the fact that performance assessment usually utilizes human raters to score test-takers using a scoring rubric, there are several ways in which subjectivity of rater judgments can cause variations. Raters bring in their own personal judgement standards that they are used to together with their own level of self-consistency. Moreover, differences in raters' practical rating experience with various test or with a specific test, raters' training, educational background and teaching experience add variance to the scores given by different raters to students with the same ability level (Lumely, 2005). Bachman, Lynch and Mason (1995) also emphasized that the potential sources of undesirable measurement error can be rater inconsistency, or bias towards the task or test-taker. There are multiple potential sources of rater variability including rater internal and external consistency, rater severity, quality of the rating scale, task demands, occasion of rating and interaction with other aspects of the rating process (Brown, 1995; Lumley & McNamara, 1995; Wigglesworth, 1993).

Research Questions

1. How did raters differ in their holistic and analytic score assignment?
2. Do specific patterns emerge in rater severity depending on whether raters belong to native or nonnative group?
3. What are the overall general strategies that raters utilized?

Methods

Participants

The raters in the study were 10 experienced teachers of English for speakers of other languages. Their age ranged from 26 to 43 (*M* = 31). Five of them were Russian, and the other five were American. All of the raters had an MA degree in Teaching English as a Second/Foreign Language or other relevant field such as (Applied) Linguistics or Translation. Their experience of teaching English as a second/foreign language ranged from 5 to 10 years (*M* = 7). Each teacher was provided with brief rater training as part of the study. According to the rater background questionnaire, NNS and NS had differences in terms of their familiarity with the test-takers' L1s used in the study. NNS raters had considerably lower familiarity scores with Arabic and Chinese L1 whereas NS raters had lower familiarity with Russian L1.

Table 1 describes raters' familiarity with examinees' L1s in terms of teaching students with those L1s, communicating with people in English for who those L1s are native languages,

Table 1

Familiarity of NNS and NS raters with accents in the study (group means)

	NNS	NS
	Teaching	
Arabic	1.0	3.8
Chinese	1.6	3.6
Russian	4.0	2.6
	Communication	
Arabic	2.4	3.8
Chinese	2.4	3.6
Russian	4.0	2.8
	Accent Familiarity	
Arabic	2.4	4.6
Chinese	2.8	4.6
Russian	5.0	3.4

Note: Teaching and communication were measured on a 1-4 scale (No, Little, Some, Extensive), Accent familiarity was measured on a 1-5 scale (Not familiar – Very familiar).

and raters' general impression of their familiarity with the accents that those people have in English. It can be seen that NNS were all very familiar with the accent that Russian speakers have in English, and had almost no familiarity with the English spoken by Arabic or Chinese people. There was one exception, one NNS rater taught in China for 3 months.

Examinees in the study were 24 Intensive English Program (IEP) students from three L1 backgrounds: Arabic (n=8), Chinese (n=8) and Russian (n=8); 9 of them were female and 15 male (Table 2). The recordings were not longer than 60 seconds.

Table 2
Total number of recordings for each part by score and L1

Part	L1	Score of 1	Score of 2	Score of 3	Score of 4	Total per L1
Training	Arabic	1F		1M	1M	3
	Chinese	1M	1M	1M		3
	Russian		1F		1F	2
Practice	Arabic		1M			1
	Chinese				1F	1
	Russian	1M		1F		2
Rating	Arabic	1M	1M	1F	1M	4
	Chinese	1M	1M	1F	1F	4
	Russian	1F	1M	1M	1M	4
Total per score		6	6	6	6	24

Note: M stands for male, F stands for female.

Instruments

An independent opinion prompt was used. The prompt asked students to express a preference for studying alone or in a group. TOEFL iBT independent speaking rubric was used by raters in this study. A questionnaire was used to collect the background information from all raters. The questionnaire was developed by the researcher for the purpose of the study utilizing parts of Language Experience Questionnaire (Harding, 2012) and Rater Language Background Questionnaire (Wei & Llosa, 2015). The questionnaire informed the researcher about the participants' academic, language learning and teaching background, previous rating background, and the level of raters' familiarity with test-takers' L1s.

Data Collection

Raters had a brief rater training, then scored the recordings followed by a brief interview. On average, each rater generated around 5000 words, with minimum around 3600 words and maximum 7300 words. The average time needed for each part is presented in Table 3.

Table 3
Time needed for raters to finish each section (in minutes)

Rater	Training	Practice	Rating	Interview
NNS1	27	15	48	18
NNS2	30	17	57	8
NNS3	40	60	100	25
NNS4	23	18	46	13
NNS5	30	27	61	14
NS6	32	22	50	15
NS7	28	15	33	17
NS8	30	31	50	12
NS9	31	32	55	18
NS10	25	20	41	12

Note: Values are given in minutes, the time includes the researcher speaking, raters speaking and time for listening to recording.

Quantitative analysis. Many-Facet Rasch Measurement (MFRM) model was used to compare two groups of raters. Computer program FACETS, version 3.71.4 (Linacre, 2014) was used. The analysis was performed using the scores from 12 recordings graded during think-aloud rating. To match the variables in the study, three facets were specified in the model: Examinee (N=12); Rater (N=10); Criteria (N=3). The teacher group facet was entered as a dummy facet and anchored at zero, (N=2) native and nonnative. There were 480 valid responses used for estimation.

Qualitative analysis. The think-aloud protocols were transcribed and thematically coded using content analysis (Strauss & Corbin, 1998). An example of the coding scheme is provided. The codes and the relations between codes were generated from the data, and then examined to determine patterns. Transcriptions of each rater's think-aloud protocol were analyzed individually to identify themes and patterns and then compared among raters and rater groups to identify prospective variations between American and Russian teachers.

Results

RQ1: How did raters differ in their holistic and analytic score assignment?

Current TOEFL speaking rubrics are holistic, but they still have some analytic pattern that describes each part of the response namely delivery, content, and language use. Ultimately, the task of a rater is to assign a rank order number to each response on a holistic scale (e.g. from 0 to 4); however, each rater can potentially arrive to the same holistic score guided by different sub-score judgements. It is easy to give a score of 0 because it means that the task is not attempted or unrelated to the topic of the task. On the other hand, to get a score of 1 to 3 on TOEFL independent speaking task, two out of three judgements must fall into the same band; however, to give a score of 4, each sub-score must fall into the band 4. In other words, there can be only one combination for a score of 4 and 0, but several combinations for scores 1, 2 and 3. The most likely combinations for getting a score of 3 may vary according to these patterns: 3-3-3, 2-3-3, 3-2-3, 3-3-2, 4-3-3, 3-4-3, 3-3-4, 4-4-3, 3-4-4, 4-3-4. A smaller number of variance in the analytic scores exists for a score of two 2-2-2, 3-2-2, 2-3-2, 2-2-3, 1-2-2, 2-1-2, 2-2-1. The number of combinations for a score of one is smaller: 1-1-1, 1-1-2, 1-2-1, 2-1-1. This means that the holistic scores might be the same, but variation may occur in the analytic scores. Depending on the individual attributes of each rater as well as on what aspect of test-taker's performance they emphasize, raters can arrive at different scores as well as at the same score but using different paths or thought processes.

Figure 1 describes the use of scores for overall score and score for each criterion by group of raters. It can be seen that there is no clear systematic difference between rater groups, but some pattern by sub-criteria can be seen: All raters were more lenient with the delivery scores and less lenient with topic development and language use.

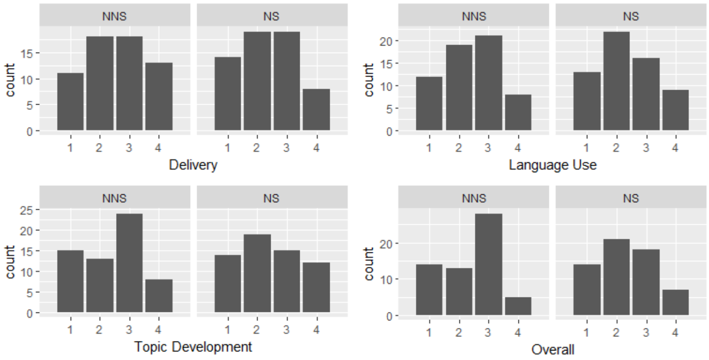


Figure 1. Score variations by rater group.

Figure 2 illustrates the overall scores and scores for each category given by each rater. It can be seen that all raters used the whole rating scale, which was expected as the recordings were selected from low to high oral proficiency.

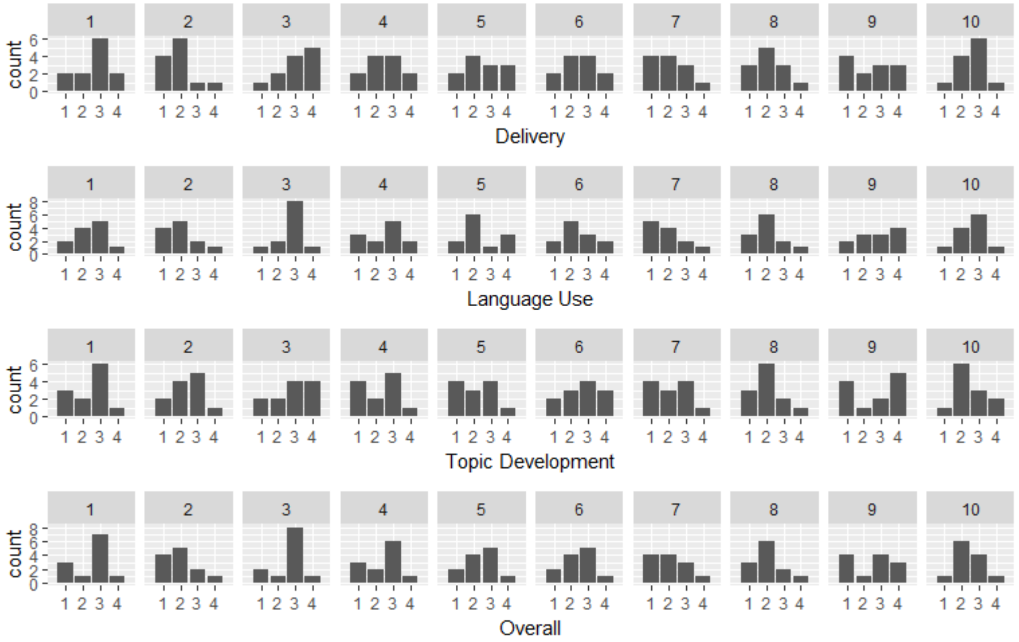


Figure 2. Score variation by each rater.

Figure 3 shows the descriptive statistic of how the two groups of raters varied in their holistic scores. Raters numbered 1 through 5 are NNS, and raters numbered 6 through 10 are NS. It was expected to see less variability in the rater's holistic scores as all of them received the same training. It can be seen that there was no variation for the first and the last recordings as those were absolutely clear examples of a score of 1 and 4.

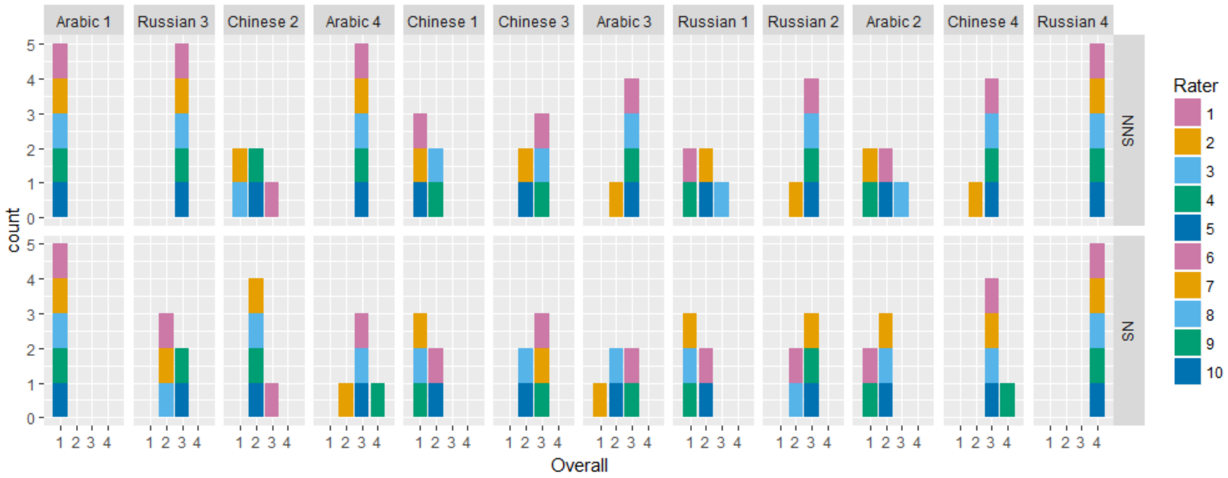
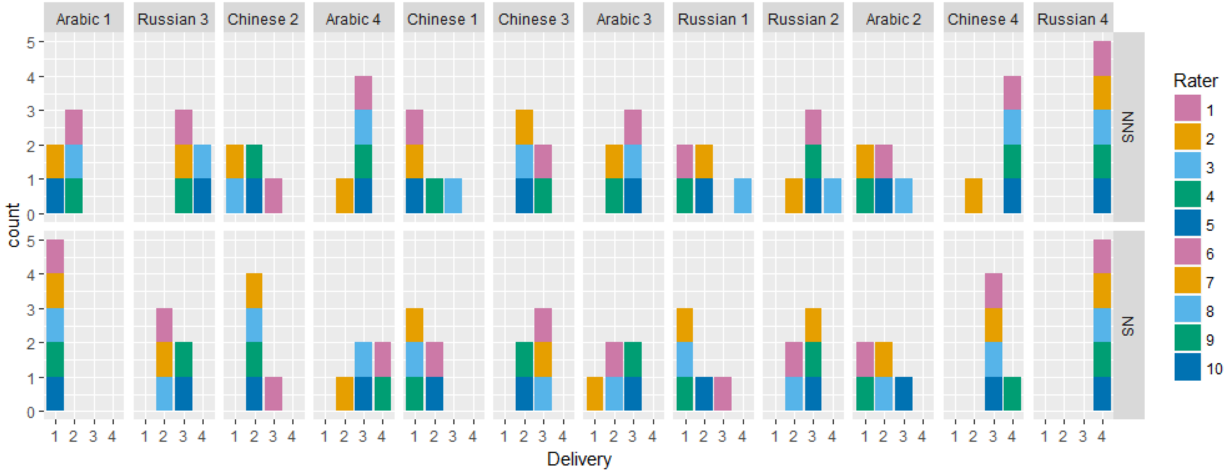
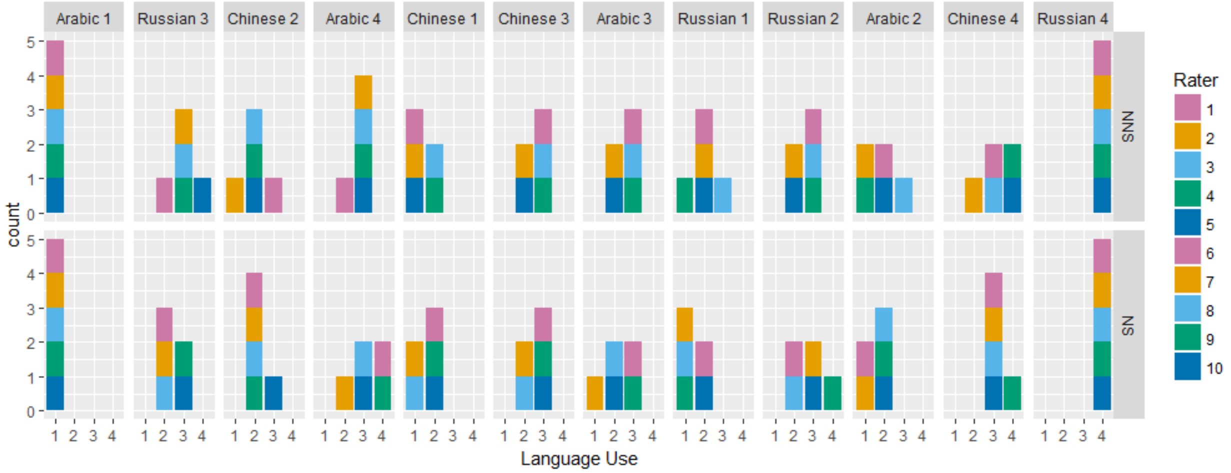


Figure 3. Overall scores assigned to each recording by NNS and NS raters.

In terms of delivery, there was more rater variation than in terms of overall scores. Some variation appeared for the first recording as two Russian raters gave it a 2 for clear pronunciation. They mentioned that the recording was very short, and there was not enough speech to be graded, but they knew that the higher delivery grade would not change the overall score, so they wanted to acknowledge that even though the person said two sentences, his speech was intelligible. In terms of scores for language use and topic deliver, again, there was more rater variation; however, there was no systematic pattern between NNS and NS groups.





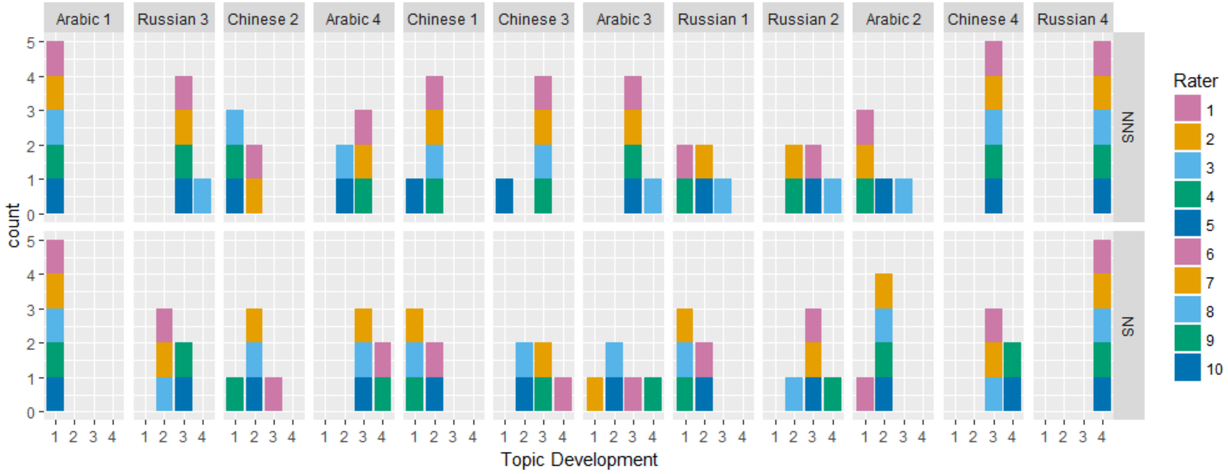
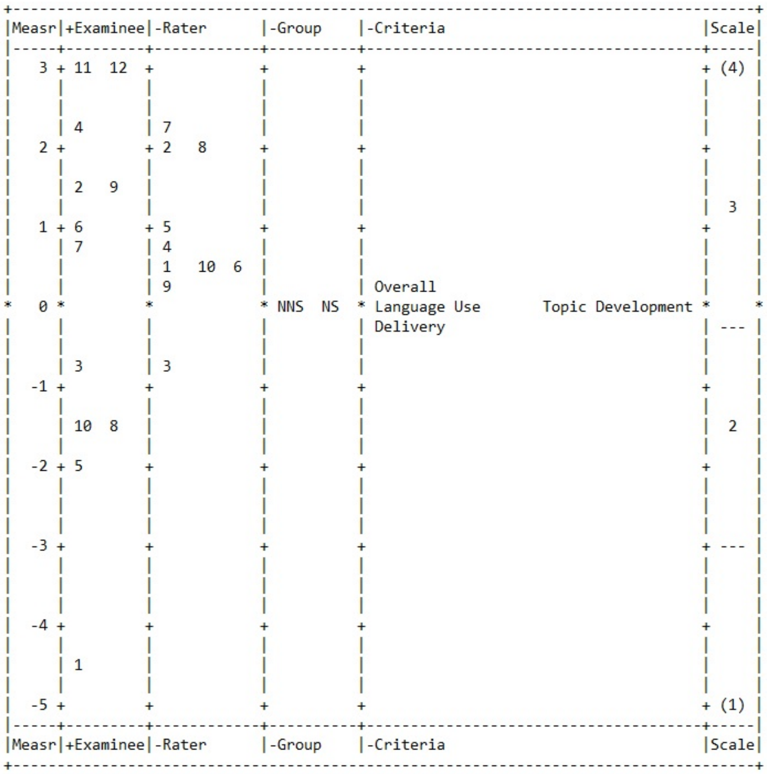


Figure 4. Scores assigned to each recording on each sub-category by NNS and NS raters.

RQ2: Do specific patterns emerge in rater severity depending on whether raters belong to native or nonnative group?

Figure 5 shows FACETS variable map, which shows that the groups did not differ in their severity. It can be seen that NNS3 was the most lenient rater, and the rest of the raters exhibited a more severe pattern of ratings. Raters NNS2, NS7 and NS8 were the most severe raters. In terms of scoring sub-rubric categories, delivery was the most leniently scored.



```
+-------------------------------------------------------------------------------------+
|Measr|+Examinee|-Rater      |-Group    |-Criteria                              |Scale|
|-----+---------+------------+----------+---------------------------------------+-----|
|   3 + 11  12  +            +          +                                       + (4) |
|     |         |            |          |                                       |     |
|     |         |            |          |                                       |     |
|     | 4       | 7          |          |                                       |     |
|   2 +         + 2   8      +          +                                       +     |
|     |         |            |          |                                       |     |
|     | 2   9   |            |          |                                       |     |
|     |         |            |          |                                       |  3  |
|   1 + 6       + 5          +          +                                       +     |
|     | 7       | 4          |          |                                       |     |
|     |         | 1   10  6  |          |                                       |     |
|     |         | 9          |          | Overall                               |     |
*   0 *         *            * NNS  NS  * Language Use        Topic Development *     *
|     |         |            |          | Delivery                              | --- |
|     |         |            |          |                                       |     |
|     | 3       | 3          |          |                                       |     |
|  -1 +         +            +          +                                       +     |
|     |         |            |          |                                       |     |
|     | 10  8   |            |          |                                       |  2  |
|     |         |            |          |                                       |     |
|  -2 + 5       +            +          +                                       +     |
|     |         |            |          |                                       |     |
|     |         |            |          |                                       |     |
|     |         |            |          |                                       |     |
|  -3 +         +            +          +                                       + --- |
|     |         |            |          |                                       |     |
|     |         |            |          |                                       |     |
|     |         |            |          |                                       |     |
|  -4 +         +            +          +                                       +     |
|     |         |            |          |                                       |     |
|     | 1       |            |          |                                       |     |
|     |         |            |          |                                       |     |
|  -5 +         +            +          +                                       + (1) |
|-----+---------+------------+----------+---------------------------------------+-----|
|Measr|+Examinee|-Rater      |-Group    |-Criteria                              |Scale|
+-------------------------------------------------------------------------------------+
```

Figure 5. FACETS variable map.

The interview questions provided raters' own feelings about their grading; however, most of the raters considered themselves either medium or lenient (Table 4). NS6 mentioned that she is harsh, but she was not sure about. One NNS rater noted that she is neither severe nor lenient, but she would show a more severe pattern when scoring examinees with bad pronunciation, Asian speakers specifically. Comparing the variable map and the raters' impressions, it can be seen that the position of NNS3 and her own perception of being too lenient matched.

Table 4

Raters' Perceived Severity Level

Rater	Perceived leniency/harshness
NNS1	Lenient
NNS2	Medium, not too severe, not too lenient
NNS3	Too lenient
NNS4	Depends, but harsh on Asian people because of pronunciation
NNS5	Lenient
NS6	Harsh
NS7	The happy medium
NS8	More lenient
NS9	More lenient
NS10	Not lenient at all

RQ3: What are the overall general strategies that raters utilized?

Raters approached the task from different perspectives. Most of the raters used the following pattern: Overall impression, delivery, language use, and topic development. Other raters started with language use and then comment on delivery or topic development. Some raters did not have a pattern, and decided on the spot which category is easier to score, so that they can start with it and postponed the more difficult score to the end.

Relevance to PIE and Second Language Learning

This study showed that the raters relied on different experiences in order to arrive at a score. Some raters had compassion to students and tried to suppress it, others had negative feelings towards Asian speakers and tried to overcome it. NNS raters had almost no familiarity with Arabic and Chinese speakers; therefore, they were not sure in the accuracy of their grades because their "ears were tired" from trying to understand the person. On the other hand, when the NNS raters had a match between their L1 and examinees' L1, there was hesitation again because "the Russian accent did not hurt their ears." Even though the analysis did not show any differences between the rater group, it might have been much more difficult for the raters to come up with the score correctly making scoring more time consuming. In addition, most of the raters were guided by their overall impression to this or that particular band, and raters were questioning whether it is a correct way of scoring, or they have to suppress their overall impression and analytically score each sub-score to be guided towards an overall score at the end. In terms of non-rubric comments, several raters mentioned that they are more liable to give a higher score to confident-sounding students rather than to students with softer voices because it is easier for the confident students to get their message across; therefore, their confidence makes them sound more proficient and makes the raters listen through mistakes. PIE can take these potential differences and hypothetical concerns into account when performing rater training sessions in order to address the possible covered hesitations raters might have.

References

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing, 12*, 238-257.

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

Lumley, T. (2005). *Assessing second language writing: The rater's perspective*. Frankfurt, Germany: Peter Lang.

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

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing*, 10, 305-335.

Appendix A
Rater Background Questionnaire

Adapted from Language Experience Questionnaire by Harding (2011) and Rater Language Background Questionnaire (Wei & Llosa, 2015)

Directions: Fill out the questionnaire to the best of your knowledge.

General.

1. Age: _______
2. Gender: Male/Female
3. In which country were you born? ___________________________
4. Have you ever lived in another country for more than 3 months? __Yes __No. If no, skip #6–8.
5. Where?____
6. For how long?____
7. For what purpose?____
8. Educational background (fill out those that apply):

Undergraduate degree in ____
Certificate in ____
Master’s degree in ____
Doctoral degree in ____
Other ____

Languages.

1. What is your native language/mother tongue? _______________________
2. Other languages spoken:

Additional language 1 _____________________
Additional language 2 _____________________
Additional language 3 _____________________
Additional language 4 _____________________

3. Please rate your ability to use these languages (low/intermediate/advanced/almost native).

Additional language 1 _____________________
Additional language 2 _____________________
Additional language 3 _____________________
Additional language 4 _____________________

4. Is English your native language/mother tongue? __Yes __ No. If yes, skip #5, 6, 7.
5. For how long have you studied English? ___________
6. Where did you study English? Select all that apply.

__ Kindergarten __ Primary school __ Secondary school __ College/university __ Other (please specify)

7. Have you studied English abroad? __Yes, __No. If no, skip #8
8. Where? ____
9. Please rate your ability to use English in academic settings by checking the appropriate level in the table below.

	Low	Intermediate	Advanced	Almost Native
Listening				
Speaking				
Reading				
Writing				

Teaching experience.

1. For how many years in total have you taught English?_________
2. In what countries have you taught? ____________________________________

   ____
3. Students from what countries have you had in your classroom?____
4. Describe how much experience do you have in teaching nonnative speakers for whom these are native languages?

	No	Little	Some	Extensive
Arabic				
Chinese				
Russian				

Rating experience.

1. Have you scored any standardized language tests before? ___Yes ___ No. If no, skip # 2, 3, 4, 5.

2. If yes, what is/are the test(s) that you scored? ___________________________________
____

3. What are the language skills that you scored? (Select all that apply):
__ Speaking __ Reading __ Listening __ Writing

4. Did you receive any formal training as a rater? __ Yes __ No. If no, skip #5.

5. Briefly, describe the training that you received____
____

Familiarity with accents spoken by nonnative English speakers.

1. How often do you speak English to people for whom English is not a native language?
__ Never __ Rarely __ Sometimes __ Often

2. In general, how familiar are you with English spoken with the following accents? (Please highlight one number for each accent)

	Not familiar				Very familiar
Arabic accent	1	2	3	4	5
Chinese accent	1	2	3	4	5
Russian accent	1	2	3	4	5

3. How much experience do you have listening/talking to English spoken by people for who these languages are native?

	No	Little	Some	Extensive
Arabic				
Chinese				
Russian				

Appendix B
Speaking Prompts

Answer a Question #1
Prepare: 1 minute; Speak: 1 minute

Preparing: Read the following question and then prepare your answer. You may take notes on this paper. Your response will be scored according to:

- Development of ideas
- Pronunciation
- Grammar and vocabulary

Question: You have an exam next week. Do you want to study alone or in a group? Include reasons and examples to support your answer.

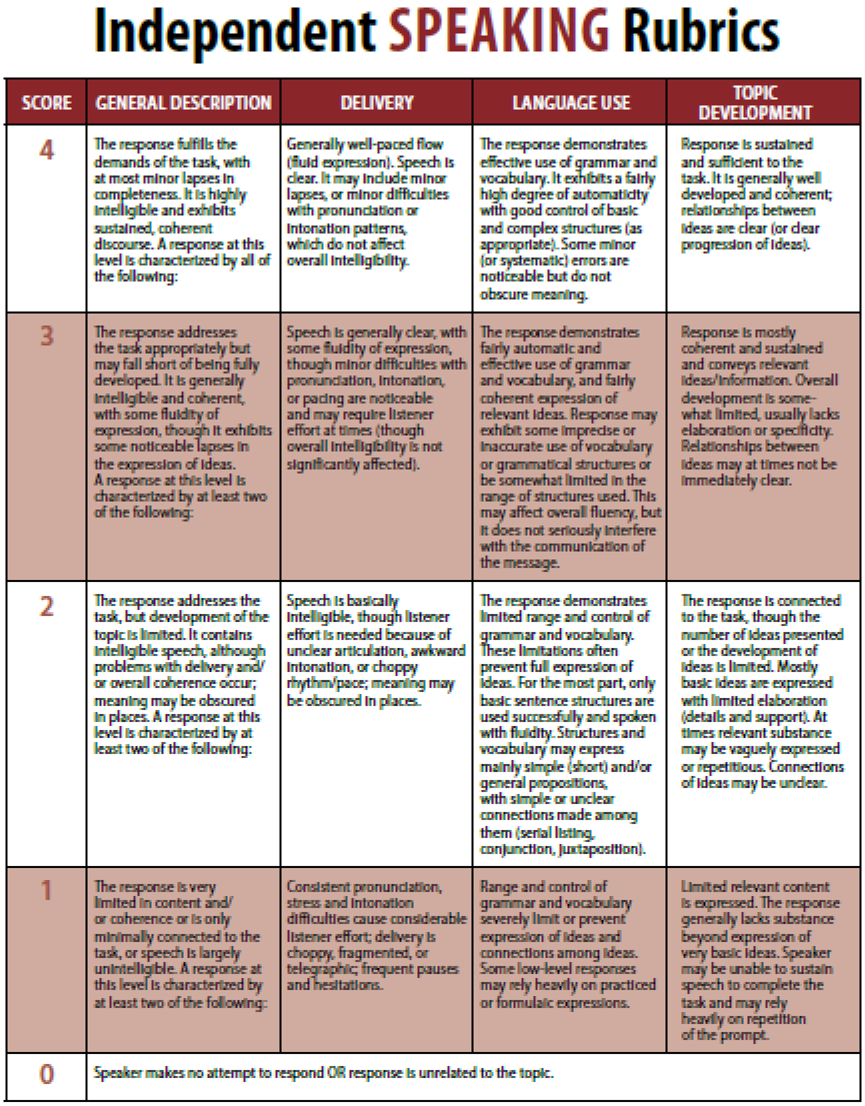
Answer a Question #2
Prepare: 1 minute; Speak: 1 minute

Preparing: Read the following question and then prepare your answer. You may take notes on this paper. Your response will be scored according to:

- Development of ideas
- Pronunciation
- Grammar and vocabulary

Question: There are different ways to teach students. Some universities have large classes with many students. Other universities have small classes. Which of these classrooms is better for learning? Use specific examples to support your answer.

Appendix C
TOEFL Independent Speaking Rubric



Independent SPEAKING Rubrics

SCORE	GENERAL DESCRIPTION	DELIVERY	LANGUAGE USE	TOPIC DEVELOPMENT
4	The response fulfills the demands of the task, with at most minor lapses in completeness. It is highly intelligible and exhibits sustained, coherent discourse. A response at this level is characterized by all of the following:	Generally well-paced flow (fluid expression). Speech is clear. It may include minor lapses, or minor difficulties with pronunciation or intonation patterns, which do not affect overall intelligibility.	The response demonstrates effective use of grammar and vocabulary. It exhibits a fairly high degree of automaticity with good control of basic and complex structures (as appropriate). Some minor (or systematic) errors are noticeable but do not obscure meaning.	Response is sustained and sufficient to the task. It is generally well developed and coherent; relationships between ideas are clear (or clear progression of ideas).
3	The response addresses the task appropriately but may fall short of being fully developed. It is generally intelligible and coherent, with some fluidity of expression, though it exhibits some noticeable lapses in the expression of ideas. A response at this level is characterized by at least two of the following:	Speech is generally clear, with some fluidity of expression, though minor difficulties with pronunciation, intonation, or pacing are noticeable and may require listener effort at times (though overall intelligibility is not significantly affected).	The response demonstrates fairly automatic and effective use of grammar and vocabulary, and fairly coherent expression of relevant ideas. Response may exhibit some imprecise or inaccurate use of vocabulary or grammatical structures or be somewhat limited in the range of structures used. This may affect overall fluency, but it does not seriously interfere with the communication of the message.	Response is mostly coherent and sustained and conveys relevant ideas/information. Overall development is somewhat limited, usually lacks elaboration or specificity. Relationships between ideas may at times not be immediately clear.
2	The response addresses the task, but development of the topic is limited. It contains intelligible speech, although problems with delivery and/ or overall coherence occur; meaning may be obscured in places. A response at this level is characterized by at least two of the following:	Speech is basically intelligible, though listener effort is needed because of unclear articulation, awkward intonation, or choppy rhythm/pace; meaning may be obscured in places.	The response demonstrates limited range and control of grammar and vocabulary. These limitations often prevent full expression of ideas. For the most part, only basic sentence structures are used successfully and spoken with fluidity. Structures and vocabulary may express mainly simple (short) and/or general propositions, with simple or unclear connections made among them (serial listing, conjunction, juxtaposition).	The response is connected to the task, though the number of ideas presented or the development of ideas is limited. Mostly basic ideas are expressed with limited elaboration (details and support). At times relevant substance may be vaguely expressed or repetitious. Connections of ideas may be unclear.
1	The response is very limited in content and/ or coherence or is only minimally connected to the task, or speech is largely unintelligible. A response at this level is characterized by at least two of the following:	Consistent pronunciation, stress and intonation difficulties cause considerable listener effort; delivery is choppy, fragmented, or telegraphic; frequent pauses and hesitations.	Range and control of grammar and vocabulary severely limit or prevent expression of ideas and connections among ideas. Some low-level responses may rely heavily on practiced or formulaic expressions.	Limited relevant content is expressed. The response generally lacks substance beyond expression of very basic ideas. Speaker may be unable to sustain speech to complete the task and may rely heavily on repetition of the prompt.
0	Speaker makes no attempt to respond OR response is unrelated to the topic.			

Appendix D
Think-Aloud Protocol Script

Hi! How are you doing today? Thank you for agreeing to participate in my research study. Today we will spend about 1.5 hours. Our session will have five parts:

1. Rubric and task familiarization
2. Training
3. Practice
4. Rating
5. Interview

Let's begin with the first one!

Rubric and task familiarization

You will be scoring speech recordings from English learners in response to this prompt. Students had 1 minute to prepare and 1 minute to speak. The question was "…". Now you can take time to read through the prompt. You can make comments and ask questions.

Here is the rubric which will be used to score the recordings. It assesses delivery, topic development, and language use. It has also a general description of overall performance. The possible scores can vary from 0 to 4. As you can see, 0 means that there is no response or response is not on the topic. Scores from 1 to 4 describe the quality of appropriate responses. Scores for each category might be the same or might fall into different bands.

Now you can take time to read through the rubric paying attention to each criteria in each score band. You can make comments and ask questions (give time needed, approximate length 5-10 min).

Ok, now let's review the rubric together. I'll explain the salient features for each category.

First the salient features that distinguish 3 and 4 are that the score of 4 must have all three elements in its band, and three however, should have two elements in its band and one in a band lower or higher.

And according to the rubric, 4 is a fluent, clear, intelligible answer which might be a bit flawed. It is a well-developed answer, with clear and connected ideas. Grammar and vocabulary are good, but might be a bit flawed which does not obscure the meaning.

3 is not as easy to understand as 4 and it might require some listener effort. It's grammar, vocabulary and topic development are good but a bit limited.

2 is more difficult to understand and it requires listener effort. Grammar and vocabulary affect expression of ideas in a negative way. Topic development is basic, not elaborated, vague, repetitive with unclear or not connected ideas.

1 can have a lot of pauses, hesitations and pronunciation mistakes and it needs a lot of listener effort. Its grammar is severely limited. The ideas are very basic and maybe repeating the prompt, using memorized expressions and be repetitive.

We will not have any recordings with the score of 0 because zero means no answer.

So, 4 is the best, 1 is the worst and 3 and 2 are in the middle. You can look at the descriptions of 2 and 3 in order to find how you would differentiate them. (Give time)

Elicit answer:

Now we are finished. Let's move on to training.

Training

Now we will have the training. I'll play 8 one-minute recordings overall.

These recordings are from a placement test, so students did not study this topic in a classroom and their ideas are on the spot ideas.

You will hear the recording once from the beginning to the end. Then, you can listen to the recording again for as many times as you want and pause it if needed.

After you are comfortable with the recording, I will tell you what score it was assigned.

Then, using the rubric, you will express your opinion why this recording was given this particular score.

When you are giving scores, you will be thinking aloud. What I mean by "talk aloud" is that I want you to say out loud everything that you would say to yourself silently while you think. Just act as if you were alone in the room speaking to yourself. Please provide as thorough a justification as possible.

Do you have any questions? Let's start.

Recording #1. Let's listen. You may take notes if you wish.

Now you can listen again and pause if needed.

This recording was given a score of 2. Why do you think it was given this score?

The score that I provide to you is the overall score for the recording, but I would ask you to try predict what scores it was possibly given for each section: delivery language use and topic development.

Other possible additional probes:

You said "…" what do you mean by that?

You said "…" what exactly do you mean?

Can you give an example?

Why do you think so?

Ok, can you tell me more?

And?

So?

So, you are saying that …. is ….?

So, you want to say that …. is … ?

So, you mean that …. is …?

So, what you are saying is ….?

Ok. Do you have anything else to add?

Let's move on to recording #2.

The same pattern with the rest of the recordings.

Practice

Now that we have had training, we will have practice. I'll play 4 more recordings.

You will listen to each recording once from the beginning to the end and verbalize your thoughts using the rubric.

Then, you can listen to the recording again as many times as you want and pause it if needed. You will continue verbalizing your thoughts. Then, you will arrive at a final score for each category.

After that, I will tell you what score this recording was assigned, and we will discuss if your rating is the same or different from that score and why.

Do you have any questions? Let's start.

Recording #1. Let's listen. You may take notes if you wish.

What are your thoughts about the recording based on the rubric?

Now you can listen again and pause if needed.

Continue verbalizing your thoughts.

This recording was given a score of 4.

Does it differ from your scores?

Why do you think it is the same?

Why do you think it is different?

Other possible additional probes: the same as before.

Ok. Do you have anything else to add?

Let's move on to recording #2.

Rating

Now that we have had practice, we will have rating. I'll play 12 more recordings.

You will listen to each recording once from the beginning to the end and verbalize your thoughts using the rubric.

Then, you can listen to the recording again as many times as you want and pause it if needed. You will continue verbalizing your thoughts. Then, you will arrive at a final score for each category.

This time I will not tell you what score it was assigned. Then I will ask if it was easy or difficult for you to grade this recording and why.

Do you have any questions? Let's start.

Recording #1. Let's listen. You may take notes if you wish.

What are your thoughts about the recording based on the rubric?

Now you can listen again and pause if needed.

Continue verbalizing your thoughts.

Was it easy or hard for you to grade this recording? Why?

Other possible additional probes: the same as before.

Ok. Do you have anything else to add?

Let's move on to recording #2.

Thank you for your input! I really appreciate it! Let's move on to the short interview.

Appendix E
Interview Script

1. How was your rating experience? (prompt to outline experiences, concerns, or difficulties).
   Other probes:
   Why are you saying it was hard/easy?
   You said … why do you think it was hard?
2. Did you have any specific test-takers which were hard/easy to rate? Why?
   Other probes:
   Can you give an example?
   You said …., what did you mean?
   Can you expand on that?
   So, your opinion is …? (pause to elicit continuation)
3. Do you consider yourself a severe (harsh) or a lenient (liberal) rater? Why?
   Other probes:
   You said that you …, could you explain?
   Are you always a severe/lenient rater?
   You said … can you give an example?
   What do you mean by …?
   Can you give more details?
   So, you mean …? (pause to elicit continuation)
4. Do you think you were harsher/more lenient on some test-takers? Why?
   Other probes:
   When you say … you mean….? (pause to elicit continuation)
   Why do you think so?
   Can you tell me more?
   So, you are saying …? (pause to elicit continuation)
5. Looking at the rubric, do you think that each sub-category is equally important? Why?
   Other probes:
   So, you want to say …? (pause to elicit continuation)
   Why do think so?
   Any examples?
6. Do you think you are harsher/more lenient on some sub-categories? Why?
   Other probes:
   So, you are saying …? (pause to elicit continuation)
   So, why do you think that … is the most important?
   So, why do you believe that … is #1 for you?
7. Do you think you have any specific pattern of rating? What do you do first, second, etc.?
   Other probes:
   Why do you listen for … first?
   Do you take notes?