Investigating Native Speakers' Intelligibility Ratings and Comments about Japanese EFL speakers' speech

Nobuhisa Hiraishi

Nagoya Gakuin University, Japan

0412

The Asian Conference on Language Learning 2013

Official Conference Proceedings 2013

Abstract

In this study, we investigated Japanese EFL (JEFL) speakers' intelligibility rated by two groups of native speakers of English, a peer group and a teacher group, and their comments on the speech samples. Reading and spontaneous speech samples were collected twice, pre and post study abroad from seven JEFL students who visited North America for about nine months. The raters were asked to evaluate the samples in terms of intelligibility, and also were asked to give a comment on each speech about what features seemed unclear or unnatural or some advice for further improvement. The results of the assessment revealed that the intelligibility of JEFL learners was confirmed to be improved. It also revealed that the means of the scores given by the peer raters were higher than the teachers' means. From the result, it seemed that the peer raters were more lenient to JEFL learners' speech. The terms occurring in the comments were divided into four categories for further analysis: phonetics, fluency, suprasegmentals, and grammar. The peer raters gave comments on fluency much more frequently than the teacher raters did. The comments were further investigated and separated into each score. Comments about phonetics were more frequently given to less intelligible speech, and the percentage fell as the intelligibility level rose. The frequency of comments for reading speech was confirmed to have a strong positive correlation, but it was not so strong for the spontaneous speech with the peer raters.

1. Introduction

One of the most important purposes of learning foreign language must be to make listeners understand what the speaker is saying in the foreign language. As babies start to learn how to communicate with others orally much earlier than how to read or write in their first language (L1), oral skills are apparently very important and crucial for human beings, so as for learners of English as the second language (L2) or as a foreign language (EFL). The very basic skill to make listeners understand what you are saying will be defined as *intelligibility* (Smith, 1992; Nelson, 2011). Unfortunately, it is often said that Japanese EFL speakers (JEFL) tend to carry a strong Japanese accent in speaking English. It is also said that it is rather difficult for non-Japanese native speakers to understand English utterances given by Japanese native speakers. However, what kind of features in their English utterances actually makes it difficult for native speakers of English to understand? Also what elements are native speakers concerned about when listening to non-native speakers' English utterances?

Mastering English pronunciation seems to be one of the most difficult targets for JEFL learners to achieve because it is very different from Japanese pronunciation. English has, for instance, about 19 vowels and 24 consonants (Ladefoged and Johnson, 2011). English also has about 67 consonant clusters (Takebayashi, 1996). On the other hand, Japanese has only five vowels and 26 consonants including consonant clusters. Not only are there differences in the pronunciation, but also there are several prominent differences in the suprasegmentals. For example, English is said to be a stress-timed language, and Japanese a mora-timed language (Trubetzkoy, 1958: as cited in Kubozono, 1998, p.4). In terms of syllable types, 56% of the syllables in English are said to be closed syllables, and 44% open syllables (Dauer, 1983). In contrast, about 90% of syllables in Japanese are open syllables (Kubozono, 1996). Also it is known that stress, duration, intensity (Fry, 1955), and pitch prominence (Bolinger, 1958) play important roles in English speech. On the other hand, pitch change is the only key for word emphasis, and neither duration nor intensity contributes to word emphasis in Japanese speech (Kubozono, 1998). What is more, Japanese has a unique accent type called *heiban-shiki accent*, which means *monotonous accent*. There are many more differences between English and Japanese in terms of phonetics and phonology.

It is known that L2 learners tend to transfer their L1 knowledge in the process of L2 acquisition (Ellis, 1997). So the transfer of those suprasegmental prominences in Japanese must contribute to their lack of intelligibility in speaking English. Although the differences in those features seem large and hard to adapt, not all the Japanese L1 speakers are unable to speak English proficiently or intelligibly. For example, many returnee children or returnee students use native-like English phonology, or nearly English phonology when speaking English. Also the students who have been in an English speaking country as exchange students seem to have become able to speak English more intelligibly when they come back from abroad. Many of them seem to have had favourable influences on their speech skills during their stay abroad. But does their intelligibility really improve? Will it be noticed and positively evaluated by English L1 speakers?

To answer those questions, we carried out an experiment in which we collected samples of reading and spontaneous speech utterances given in English by JEFL learners before and after staying in an English speaking country (U.S.A. or Canada) for about nine months. Also we collected evaluations of those utterances from two groups of native speakers of English, a university student group, and a teacher group. With the evaluation, we also collected comments from the raters about what features they were concerned about when evaluating JEFL speakers' utterances and what features the speaker might need to work on in the future in order to become an excellent speaker of English. We will discuss the findings and implications obtained from the comments.

2. Background and research questions

2.1. Foreign accent and intelligibility

Munro and Derwing (1995) investigated effects of foreign accent on intelligibility and comprehensibility. In their study, they found that accentedness did not have a significant correlation with either intelligibility or comprehensibility. The speakers who participated in this study were all native speakers of Mandarin. Listeners were all native English speakers who were undergraduate students at a Canadian university.

Derwing and Munro (1997) found that accentedness did not affect scores of intelligibility although about 40% of the listeners (ten out of 26) commented it was more difficult to understand the speech made at faster rate. In other words, although the speech was intelligible, some of the listeners needed to use some effort to understand what the speaker was saying. In this study, they used speech made by speakers with four different L1 backgrounds: Cantonese, Japanese, Spanish, and Polish. One-fourth of the speakers (12 out of 48) had taken the TOEFL test, and the mean score was 479. Those speakers represented upper range of proficiency in the sample. The speakers watched a short cartoon story, and after a while, they narrated the story. The listeners were all native English speakers who were born and raised in Canada.

Tajima, Port, and Dalby (1997) investigated if temporal correction influences the intelligibility of foreign-accented English. They recorded English short phrases spoken by Chinese L1 and English L1 speakers, and manipulated the duration of acoustic segments of those samples. They modified the durations in Chinese L1 speech samples to match the durations in English L1 speech samples, and vice versa. They found that the intelligibility of Chinese-accented utterances improved significantly after temporal correction, and the intelligibility of English L1 utterances declined after modification.

2.2. Effect of non-native speech rate on native listener comprehension

Anderson-Hsieh and Koehler (1988) investigated the effect of foreign accent and speaking rate in native speaker comprehension, and found that the comprehension scores of speech with a strong foreign accent at faster speech rate were lower than the speech at regular speech rate. The speakers' L1 was Chinese in this study, and they were all graduate students at an American university. A speaker read one of six passages at three different speech rates, fast, regular, and slow. Listeners were 224 undergraduate students at an American university.

Derwing and Munro (1997) found that English native speakers perceived strongly foreign-accented speech as too fast although the actual speech rate was not different.

Munro and Derwing (1998) found a different aspect about the speech rate from those studies mentioned above. They actually found that the non-native speech made at a slower rate was judged more accented and less comprehensible. Also they found that native English listeners preferred foreign-accented speech made at some speed, but not at a slower rate. The non-native speakers' L1 was Mandarin Chinese in this study.

From those studies, it was revealed that non-native speech with strong foreign accent that was made either too fast or too slow reduced the comprehensibility and forced English native speakers to make extra effort to understand what the speaker was saying.

2.3. Raters' familiarity with foreign accents and ethnicity

Gass and Varonis (1984) investigated the effect of familiarity on native speaker comprehension of non-native speaker speech with English language. They used speech samples made by two foreign speaker groups, Japanese native speakers and Arabic native speakers. Listeners were 142 native English-speaking undergraduate students at the University of Michigan. They found that familiarity with the topic, non-native speech in general, a particular non-native accent, and a particular non-native speaker facilitated listeners' comprehension.

Rubin (1992) also confirmed that undergraduate listeners who were willing to attend a class which was given by a non-native English-speaking teaching assistant (NNSTA) showed better understanding of Oriental speech than those who weren't. He also confirmed that the students who had more experience with NNSTA classes had better understanding.

Rubin and Smith (1990) investigated the effects of accent, ethnicity, and lecture topic on American undergraduate students' perception of non-native English-speaking teaching assistants. They collected two kinds of non-native English speech samples, highly accented and moderately accented Chinese-accented English. While they had 92 American undergraduate students listen to the speech samples, they projected a photograph of either Caucasian or Oriental/Asian at the front of the room as if they were the speaker. They found that with the Oriental/Asian photograph, the undergraduates did not pay much attention to the accentedness so that they did not distinguish the different levels of accentedness. On the contrary, with the photograph of Caucasian, they used accent as a basis for specifying ethnicity. They also found that American undergraduates tended to rate a highly accented instructor as a poor teacher. So in their study, it was found that American undergraduates tended to have a sort of prejudice that Oriental/Asian speaker would have strong accent and it would be rather difficult for them to understand what they were saying. In their study, they discussed how limited exposure to the international community influenced the undergraduates' judgements on accent and ethnicity. Rubin (1992) did a similar experiment using the speech spoken by a native English speaking American doctoral student who were born and brought up in Ohio, USA. The result revealed that the comprehensibility of the undergraduate students who had an Oriental person's photograph was worse than those who had a Caucasian person's photograph, and the speech with an Oriental person's photograph was perceived as more accented. He also studied about listeners' familiarity with NNSTA.

In the present study, the peer raters who participated in the experiment were visiting Japan for a month on a foreign exchange programme, so most of them had familiarity with the Japanese language and with the English spoken by Japanese native speakers. The teacher raters were more familiar with JEFL English utterances since they had been living in Japan for more than a decade and were having communication with JEFL learners in English frequently. So it was expected that the teacher raters' evaluations would be more lenient as a result of familiarity.

2.4. Rater group differences

Caban (2003) investigated the difference in non-native speech assessment by four different rater groups; English L1 MA students, Japanese L1 MA students, English L1 teachers, and Japanese L1 students at two language institutes in Hawaii. She used seven categories: fluency, grammar, pronunciation, compensation techniques, content of utterance, language appropriateness, and overall intelligibility. She found that English L1 MA students and English L1 teachers were more lenient in rating pronunciation. Japanese L1 MA students were more lenient in rating overall intelligibility. Japanese L1 students at language schools were more lenient in rating fluency and grammar.

Hsieh (2011) compared the judgement results of non-native speakers' speech samples on oral proficiency, accentedness, and comprehensibility rated by student and teacher raters. The speakers were international teaching assistants (ITA) in an American university. She adopted an English oral proficiency test, the Speak Proficiency English Assessment Kit (SPEAK), and the international ITAs responded to it. She found that student raters evaluated speech more globally while teacher raters did it more analytically. She also found that the student raters did not comment as much on intonation or stress patterns as the teacher raters did. She suggested that this was because the student raters were more likely to be "linguistically less sophisticated than the ESL teachers" (p.65). Also she found that the student raters tended to give harsher scores to the non-native speech. She discussed that this was because "the undergraduates were not familiar with the rating criteria for judging the examinees, thus, they sometimes made their rating decisions solely through their appraisal of whether they felt a particular examinee was qualified to be an ITA, a criterion not on the rating rubric", and it suggests that "undergraduates consider their personal feelings, perhaps even their fears, and their possible future experiences as students in ITA classes in judging ITA's speech" (p.64).

2.5. Research questions

The present study investigates the following research questions.

(1) Intelligibility scores in the pre and post study abroad speech:

- Does JEFL speakers' intelligibility improve as a result of study abroad?

(2) Tendency in comments for JEFL speech by English L1 peers and teachers:

- How do English L1 raters comment on JEFL speakers' utterances?
- And how do the ratings and comments by peers compare to those by teachers?

3. Method

In order to answer the research questions, JEFL learners' utterances (reading and spontaneous speech) were recorded before leaving and after coming back to Japan. After getting all the reading speech samples, four sentences out of 20 were selected and presented for native speakers' evaluation. The following sections provide more detailed description of the participants and procedures.

3.1. Participants

3.1.1. Speakers

The participants involved in this experiment were seven students who stayed either in the U.S. or the English speaking area of Canada for about nine months. All of them were in their second year at university when they left Japan in August 2011. There was one male and six females. All of them were majoring in English at a Japanese university. One of them had had a 5-month stay in Canada when she was a high school student. Another three had been abroad earlier on. It was the first time for the other three to go abroad.

3.1.2. Raters

3.1.2.1. Peer group

Ten American undergraduate students (peer raters) who were staying in Japan for one month while participating in an exchange programme were hired for the rating. There were eight females and two males. Four of them were in their teens; five were in their early 20s; and the other one in late 20s. They all had their primary and secondary educations in the U.S. Four of them declared that their Japanese proficiency was fair, and the other six said they could speak Japanese very little.

3.1.2.2. Teacher group

For comparison with those students' ratings, four American teachers who had been in Japan for more than a year and were teaching English at a Japanese university were hired for the assessment of students' utterances. All of them had their primary and secondary educations in the United States. They were all male. One of them was in his 30s, one in his 40s, one in his 50s, and one in his 60s. All of them had been living in Japan for more than a decade. Three of them indicated that they could speak Japanese well, and the other said his Japanese was fair. None of them reported of having any hearing problem.

3.2. Recordings

The recording of students' utterances was carried out in a small, quiet room on the campus. Students' utterances were recorded with a Sony PCM-D50 portable linear PCM digital recorder at a 22.05 kHz sampling rate with 16-bit sample size, through a stereo microphone ECM-MS907 with 90 degrees polar pattern. For comparison, 5 native English speakers' reading utterances were also recorded in their own study rooms at a university using a recorder Sony ICD-UX523 recorder, at 44.10 kHz sampling rate with 16-bit sample size. The omnidirectional stereo microphone embedded in the recorder was used. Those sound files were saved with the Waveform Audio File Format (WAV).

3.3. Speech materials

3.3.1. Reading passage

Students were given a copy of the reading passage when they came to the room for the recording. The passage was quoted from Shimaoka (2004). Students were told to read the passage several times before recording, silently or aloud, as preparation. No time limitation was given to the students for the preparation, but no one took more than 10 minutes. A few students read it aloud. Also they were told to ask for the correct pronunciation and the meaning of unknown words if there were any.

The English passage they read had a total 20 sentences with a total of 225 words with

300 syllables and 774 phonemes in the whole passage. The syllables were confirmed by referring to *Longman Dictionary of Contemporary English* (1987) and Oxford *Advanced Learner's Dictionary* (1995). The theme of the passage was about the four seasons in Japan. The mean number of words in a sentence was 11.3, with that of syllables 15.0, and that of phonemes 38.7. The maximum number of words per sentence was 20, that of syllables, 29, and that of phonemes, 74. The minimum number of words was five, and that of syllables nine, and that of phonemes 22. The standard deviation of the words was 3.87, with that of the syllables 4.94, and that of the phonemes 12.49.

3.3.2. Spontaneous speech

In the pre study abroad session, participants were asked to speak about themselves in English. They had been informed of this topic when their appointment was arranged, so they had time and opportunity to prepare what they would say during recording. Before being recorded in the room, they were given time to think about what they were going to speak about. The same procedure was employed in the post study abroad session. But at that time, an interview was given at the beginning of their session in Japanese, and they were asked about their school life, after-school activities, life at the dormitory or with the host family, and the places they visited in the country they stayed. They talked about their life or their experience in the country. Most of them had not prepared for the interview, but most of them were able to make a good speech.

3.4. Evaluation of JEFL learners' English utterances

3.4.1. Selection and manipulation of the speech sound data

Four sentences out of the 20 sentences of the reading passage were selected for the evaluation purpose. This cropping process was carried out with the software Audacity 1.3.14 Beta. The sentences selected were the 6th, 7th, 8th, and 12th sentences. To select the sentences, the first five and the last five sentences were first rejected. For the beginning part of the reading, many of the speakers seemed to start reading rather vigorously. But as the reading went on, their voice became less loud, and by the end quite a few of them seemed to have got worn out by reading English aloud. That is the reason why the first and the last five sentences were not used. Among the other ten sentences, the expressions that are not very familiar to Japanese were then excluded, such as *there* in "There they enjoy…" or *yet* in "Yet fall is one of the…" Those sound files were exported to the audio format MP3 so that the web survey software SurveyMonkey could load the sound files quickly enough for the rates to proceed smoothly. All of the sound files were amplified as to make the magnitude of the sound files fairly equal. The selected four sentences are presented in Figure 1.

Figure 1. Four selected sentences for the evaluation

No.6	In June, the rainy season begins.
No.7	The sky is overcast and we have very few sunny days for nearly a whole month.
No.8	Then summer comes with hot days and occasional showers.
No.12	In September, typhoons hit, causing damage to buildings and crops.

3.4.2. Software and rating scale

To facilitate collection of native speakers' evaluations, an internet survey programme called SurveyMonkey was employed. The sound files were mixed in random order. The same person's utterances were distanced from each other at the intervals of at least five files. The MP3 sound file was embedded in each question. Listeners were allowed to replay the sound file as many times as they liked. A 7-point Likert scale was employed. The labels on each scale were; *1- Not understandable. 2- Very difficult. 3- Difficult. 4- Understandable. 5- Easy. 6- Very easy. 7- Excellent*.

3.4.3. Evaluation of JEFL learners' intelligibility

For the teacher raters, written instructions were given on the rating site asking "How easy is it for you to understand the utterances?" Afterwards, we received some feedback from teacher raters that the question was rather too broad to give consistent judgements.

With the feedback received from teacher raters, the instructions for student raters were revised and some lines were added to give some ideas that we wanted the raters to have in mind while listening to the sound files, such as "If I were having a conversation with this speaker, I would find his/her English ___". Also on the scale, we used the word *guessing*, to give the idea that they might need to guess to understand what the speaker was saying. These lines were read aloud by the author to the raters before they started the session.

The raters were asked to write some comments on each speech file to indicate what parts of the utterance seemed unclear to them, and why they seemed unclear or unnatural. To the peer raters, a line was added in the instruction asking what features they thought should be improved by the speaker to become an excellent speaker of English. This supplement worked fairly well and we were able to get lots of valuable comments from student raters.

4. Results

4.1. Inter-rater reliability

Inter-rater reliabilities for both reading and spontaneous speech were tested by using Cronbach's coefficient alpha. The alpha for reading with 14 raters all together was .953, and the alpha for spontaneous speech was .900. These results indicate that the interrater reliability is high enough.

The alphas among the student raters and among the teacher raters were also calculated. The alpha among ten student raters was .928 for the reading speech, and .850 for the spontaneous speech. The alpha among four teacher raters was .892 for the reading, and .721 for the spontaneous speech. From these results, inter-rater reliabilities among ten student raters for both reading and spontaneous speech assessments were confirmed high enough to be reliable. Also the teachers' inter-rater reliability of the reading speech assessment was high enough to rely on, and that for the spontaneous speech was relatively high.

4.2. JEFL learners' speech intelligibility rated by English L1 peers and teachers

4.2.1. Scores for the reading speech

To make the mean robust, each learner's scores rated by the peer raters were truncated by cutting off the highest and lowest, so that each learner's mean was calculated as trimmed mean by applying remaining 8 raters' ratings. Therefore the sample size became 56 instead of 70 with both the pre and post study abroad readings. The trimmed mean of the pre study abroad reading speech was 4.77, and the standard deviation (SD) was 1.02, and the trimmed mean of post study abroad speech was 5.59 and the standard deviation was 0.73. Table 1 indicate the distribution of scores of the pre and post study abroad reading speech rated by peer raters. A dependent two-tailed *t*-test was carried out to confirm if the null hypothesis that the mean score of post study abroad reading speech was the same as that of the pre study abroad could be rejected. As the result, the null hypothesis was rejected (*df* =56, *t* =-7.7937, *p* =.0000 (*p* <.001)). So the difference of the post study abroad mean score and the pre study abroad mean score was confirmed to be statistically significant. Therefore, the intelligibility of JEFL speakers' reading speech was confirmed improved as a result of studying abroad by the America peer raters.

Table 1. Distribution of trimmed scores for reading speech rated by peer raters

Timing	Point	1	2	3	4	5	6	7	Total	M	SD
Pre	No. of scores	0	2	10	18	12	9	5	56	4.77	1.02
Post	No. of scores	0	0	1	13	18	13	11	56	5.59	0.73

Note: *df* =56, *t* =-7.7937, *p* =0000 (*p* <.001)

4.2.2. Scores for the reading speech rated by teacher raters

The distributions of scores of the reading speech rated by the teacher raters are available in Table 2. The sample size was 28 for each speech session. The mean score of the pre study abroad speech was 4.32, and the standard deviation was 1.06. The mean score of post study abroad speech was 4.86, and the standard deviation 1.01. The result of the dependent two-tailed t-test was *df* =28, *t* =-2.9480, *p* =.0065 (*p* <0.01). So the difference of the post study abroad mean score and the pre study abroad mean score of reading speech rated by teacher raters was confirmed to be statistically significant. Therefore, JEFL speakers' intelligibility in reading speech was confirmed improved as a result of studying abroad by the America teacher raters.

Table 2. Distribution of trimmed scores for reading speech rated by teacher raters

Timing	Point	1	2	3	4	5	6	7	Total	M	SD
Pre	No. of scores	0	0	6	12	6	3	1	28	4.32	1.06
Post	No. of scores	0	0	2	8	12	4	2	28	4.86	1.01

Note: *df* =28, *t* =-2.9480, *p* =.0065 (*p* <.01)

4.2.3. Scores for the spontaneous speech rated by peer raters

The distributions of scores of spontaneous speech rated by the peer raters are shown in Table 3. As the highest and lowest was cut off at each learner's scores to make the mean robust, the sample size became 56 for both the pre and post study abroad speech ratings. The trimmed mean of the pre study abroad speech was 5.23 with a standard deviation of 0.66, and that of post study abroad speech was 5.64 with a standard deviation of 0.64. The result of dependent two-tailed t-test was *df* =56, *t* =-3.8227, *p*

=.0003 (*p* <0.001). So the means of scores for the pre and post study abroad rated by peer raters were confirmed to be statistically significant.

Table 3. Distribution of trimmed scores for spontaneous speech rated by peer raters

Timing	Point	1	2	3	4	5	6	7	Total	M	SD
Pre	No. of scores	0	0	3	12	14	23	4	56	5.23	0.66
Post	No. of scores	0	0	1	5	15	27	8	56	5.64	0.64

Note: *df* =56, *t* =-3.8227, *p* =.0003 (*p* <.001)

4.2.4. Scores for the spontaneous speech rated by teacher raters

The distributions of scores of the spontaneous speech rated by teacher raters are indicated in Table 4. The sample size was 28 for both the pre and post study abroad speech. The mean score of the pre study abroad speech was 4.50, and the standard deviation was 0.66. The mean score of post study abroad speech was 4.71, and the standard deviation was 0.86. The result of a dependent two-tailed t-test was *df* =28, *t* =-1.0301, *p* =.3121 (*p* >.05), Effect size *r* =.20. The probability was much larger than 0.05, and the effect size was small. As a result, the null hypothesis was not rejected so that the difference of the mean scores of spontaneous speech rated by teacher raters was not statistically significant.

Table 4. Distribution of trimmed scores for spontaneous speech rated by teacher raters

Timing	Point	1	2	3	4	5	6	7	Total	M	SD
Pre	No. of scores	0	0	3	12	9	4	0	28	4.50	0.66
Post	No. of scores	0	0	5	6	10	6	1	28	4.71	0.86

Note: *df* =28, *t* =-1.0301, *p* =.3121 (*p >.05*), Effect size *r* =.20 (small)

4.3. Raters' comments for the reading and spontaneous speech

In this section, the negative terms appeared in raters' comments are categorised and summarised. Depending on the area they were most closely related to the terms were divided into eight categories: phonetics, fluency, suprasegmentals, grammar, lexicon, confidence, comprehension, and topic. A table showing these categories and the terms is presented in Table 5.

The terms appearing in the comments on reading speech were summarised and categorised. Figure 7 presents the percentage of categorised negative terms by peer raters, and Figure 8 those by teacher raters for the reading speech. Terms referring phonetics appeared most frequently in both groups, comprising 68.6% of the terms used by peers, and 69.6% of those used by the teacher raters. Then suprasegmentals came second for the two groups, 41.4% for peers, and 55.4% for teachers. Grammar came third with peers (7.9%), and with teachers (12.5%).

Table 5. Table of categories of negative terms appearing in raters comments

Categories	Terms and ideas (examples)
Phonetics	"pronunciation", "articulation", "enunciation", "slur", "blending words", "some words were not clear", "phonetics", mention exact words, such as "nearly", "typhoon", "damage", etc., mention exact letters, such as "R", "L", "TH", "V", "B", etc.
Fluency	"stammer", "stutter", "stumble", "sentence sounds awkward", "pauses", "spaces", "hesitation", "speed", "pace", "slow", "too fast"
Suprasegmentals	"rhythm", "flow", "choppy", "staccato", "intonation", "tone", "stress"
Grammar	"grammar", "tense", "plural", "not pronouncing S at the end of words", "omit words"
Lexicon	"mix up words", "use wong words"
Confidence	"confident", "unsure"
Comprehension	"not understanding"
Topic	"topic", "content", mention about topic problems, such as "she was in Oregon, but talks about Hawaiian people?", "Where did she go? And she met an interesting guy?"

4.3.1. Negative comments for the reading speech

As can be seen in Figures 2 and 3, negative terms about phonetics were most frequently found in the comments for the reading speech by both the peer and teacher raters. In the peer group, 96 out of 140 comments for reading speech were about phonetics (68.6%). In the teacher group, 41 out of 56 comments for reading speech was about phonetics (73.2%). Negative terms about fluency were second frequently found in the peers' comments for 42 out of 140 speech samples (30.0%). On the other hand, teachers did not mention about fluency so much. It was found in the teachers' comments for 6 out of 56 reading speech (10.7%). Negative terms about suprasegmentals were found in teachers' comments as much as the terms about phonetics. It was found in the teachers' comments for 41 out of 56 reading speech (73.2%). On the other hand, the peer raters did not mention about suprasegmentals that much. It was found in the peers' comments for 21 out of 140 reading speech files (15.0%). Negative terms about grammar were fourth frequently found in the peers' comments, and third in the teachers' comments. In peers' comments, it was found for

11 out of 140 reading speech files (7.9%). In teachers' comments, it was found for seven out of 56 comments (12.5%).

As per the results seen in Figures 2 and 3, the four categories, phonetics, fluency, suprasegmentals, and grammar, looked salient, so those four categories were selected for further analysis.

Figure 2. Number of negative comments by the peer raters for the reading speech (n =140; multiple responses)

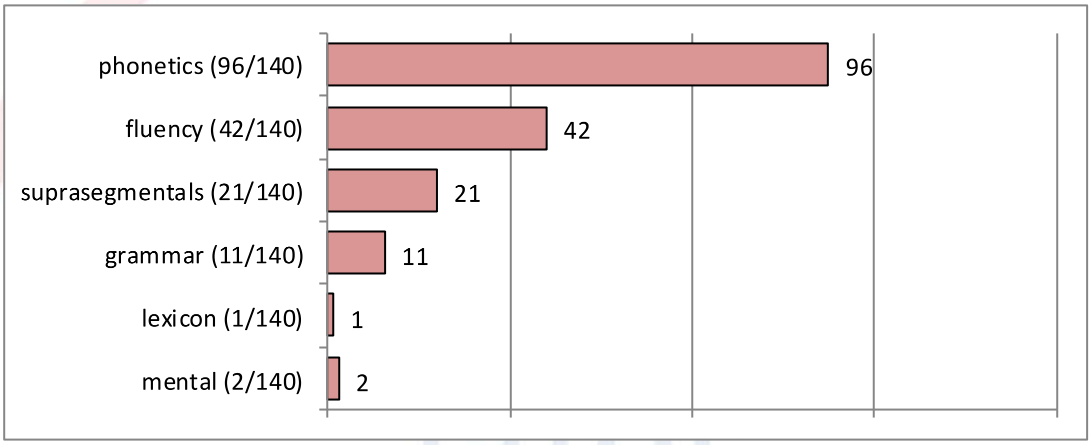
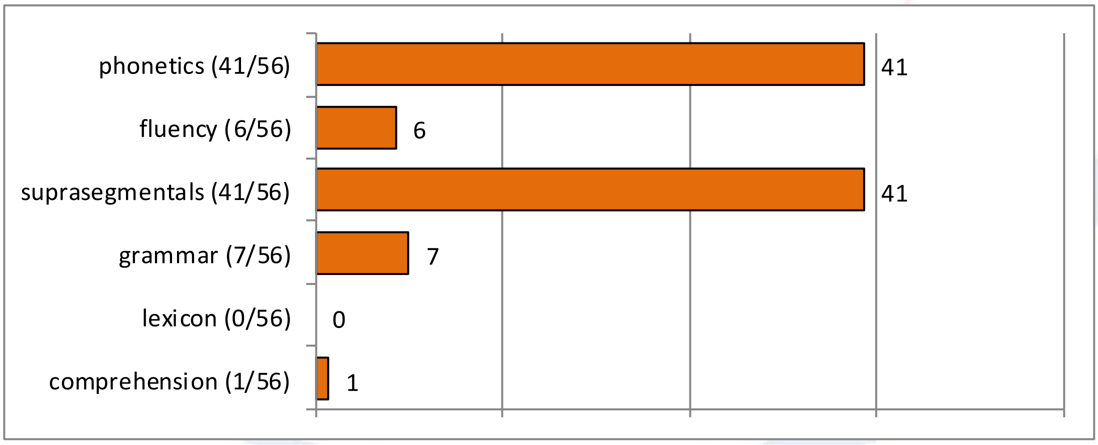


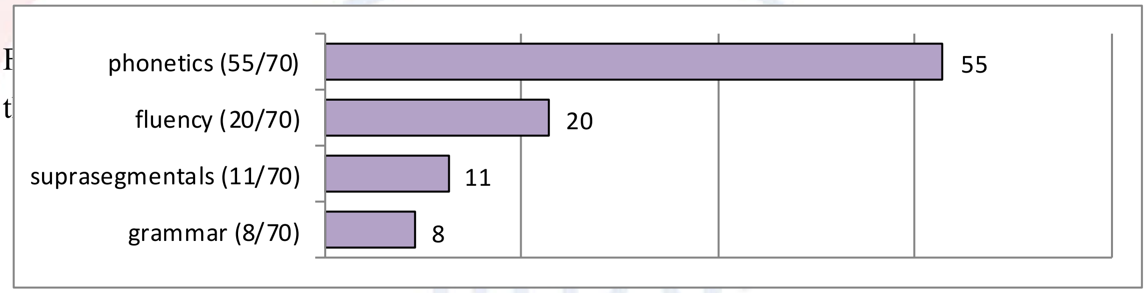
Figure 3. Number of negative comments by the teacher raters for the reading speech (n =56; multiple responses)



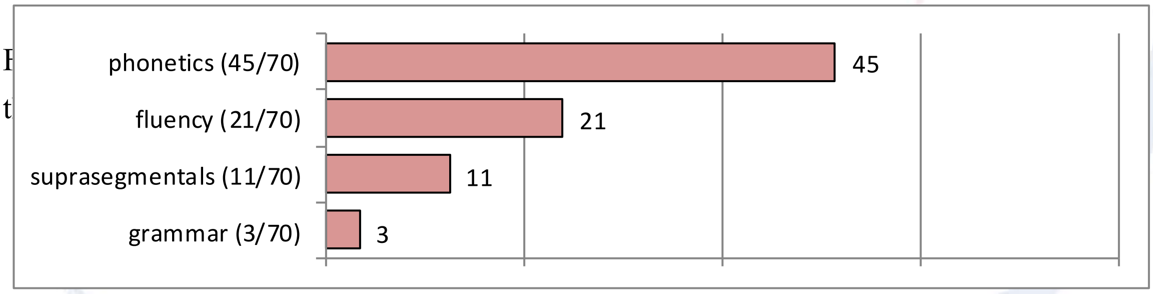
4.3.2. Negative comments given for the pre and post study abroad reading speech

Figures 4 and 5 present the numbers of comments in which the categorised negative terms were found in the peer raters' comments for the pre and post study abroad reading speech, and Figures 6 and 7 in the teacher raters' comments. As can be seen in the tables, comments on phonetics were the most remarked upon category in judging EFL speakers' intelligibility among the peer raters. Negative terms about phonetics were found in the peers' comments for 55 out of 70 speech samples of the pre study abroad reading speech (78.6%), and for 45 out of 70 speech samples of the post study abroad reading speech (64.3%). The numbers of negative comments on fluency was similar in the comments for both the pre and post study abroad speech. It was 20 for the pre, and 21 for the post study abroad reading speech. The percentages were 28.6% and 30.0% respectively. The numbers of negative comments on suprasegmentals was the third in both the peer and teacher raters' comments. It was 11 out of 70 speech samples in the comments for both the pre and post study abroad speech (25.7% each). Grammar was least mentioned for both the pre and post study abroad speech. It was eight for the pre study abroad speech (11.4%), and three for the post study abroad reading speech (4.3%).

F
t



F
t



The results of the correlation coefficient are r =.9821, R^2 =.9645, t =12.7592, and p <.001 (p =.0000). So it was confirmed that the comments for the pre and post reading speech by the peer raters a strong positive correlation.

Figures 6 and 7 present the numbers of negative comments found in the teacher raters' comments for the pre and post study abroad reading speech. The correlation coefficient between the number of negative comments for the pre and post study abroad reading speech was r =.9273, R^2 =.8599, t =6.0677, p <.001. So it was confirmed that the comments for the pre and post study abroad reading speech by teacher raters had a strong positive correlation.

The number of comments on phonetics was the largest in the comments for the both the pre and post study abroad speech, and it fell from 19 to 16 out of 28 comments (67.9% to 57.1%) for the post study abroad speech. The number of comments on suprasegmentals was second with 14 comments in the pre study abroad, and it fell to 11 (50.0% to 39.3%). The number of comments on fluency slightly fell from three to two (from 10.7% to 7.1%). The number of comments on grammar actually rose from two to six (7.1% to 21.4%).

Figure 6. Number of negative comments on the four categories by the teacher raters for the pre study abroad reading speech (n =28; multiple responses)

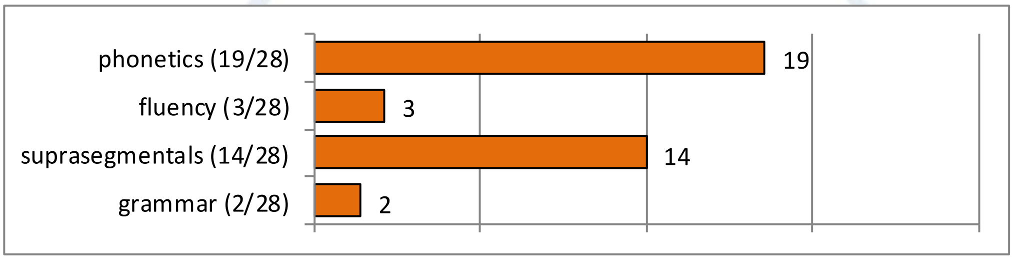
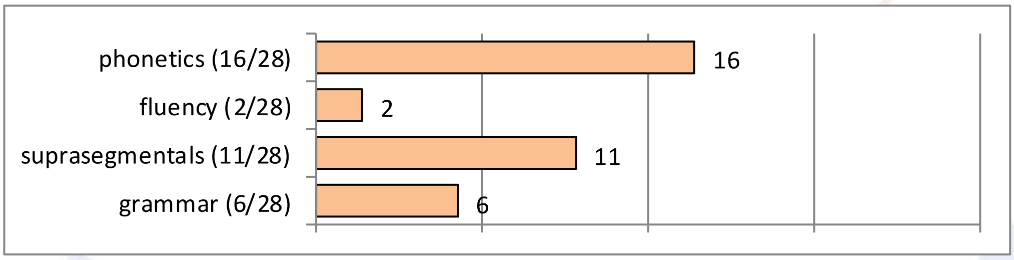


Figure 7. Number of negative comments on the four categories by the teacher raters for the post study abroad reading speech (n =28; multiple responses)



4.3.3. Negative comments for the spontaneous speech

The terms appearing in the comments on spontaneous speech were summarised and categorised in this section. Figure 8 presents the numbers of categorised negative terms given by peer raters, and Figure 9 by teacher raters respectively.

Terms about phonetics were most frequently used in both the peer and teacher raters' comments. In peer raters' comments, 52 out of 140 comments for spontaneous speech mentioned about phonetics (35.6%). Terms about grammar were also found most frequently in the peer raters' comments (56 comments, 35.6%). Terms about fluency came third (48 comments, 32.9%). Terms about suprasegmentals came fourth (9 comments, 6.2%). As can be seen in Figure 9, terms about fluency occurred much less frequently in the teacher raters' comments (6 out of 56 comments for spontaneous speech, 10.2%). Instead, negative terms about suprasegmentals were found more frequently (14 comments, 23.7%). Terms about grammar came third (13 comments, 22.0%).

The result of the correlation coefficient was r =.1572, R^2 =.0347, t =0.0347, and p >.05 (p =.3506). So it was confirmed that the comments for spontaneous speech by the peer and teacher raters had no correlation in terms of the four categories.

Figure 8. Number of negative comments by the peer raters for the spontaneous speech (n =140; multiple responses)

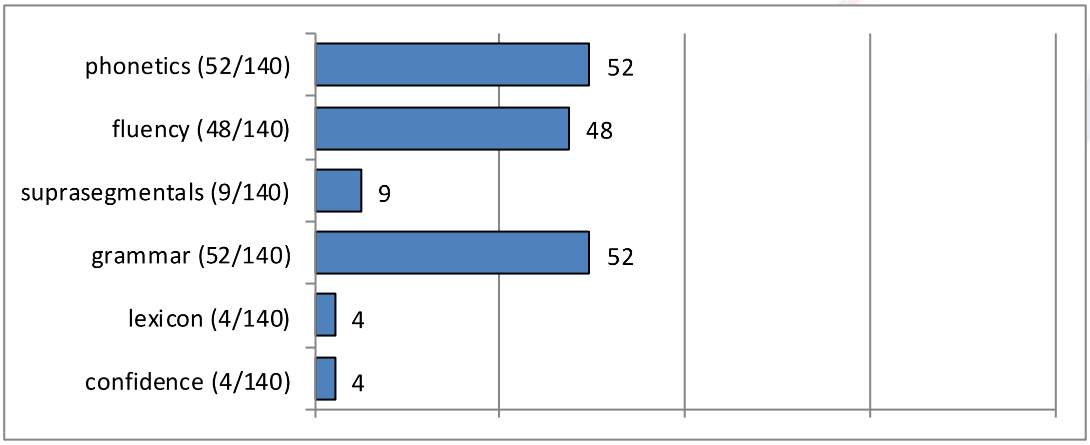
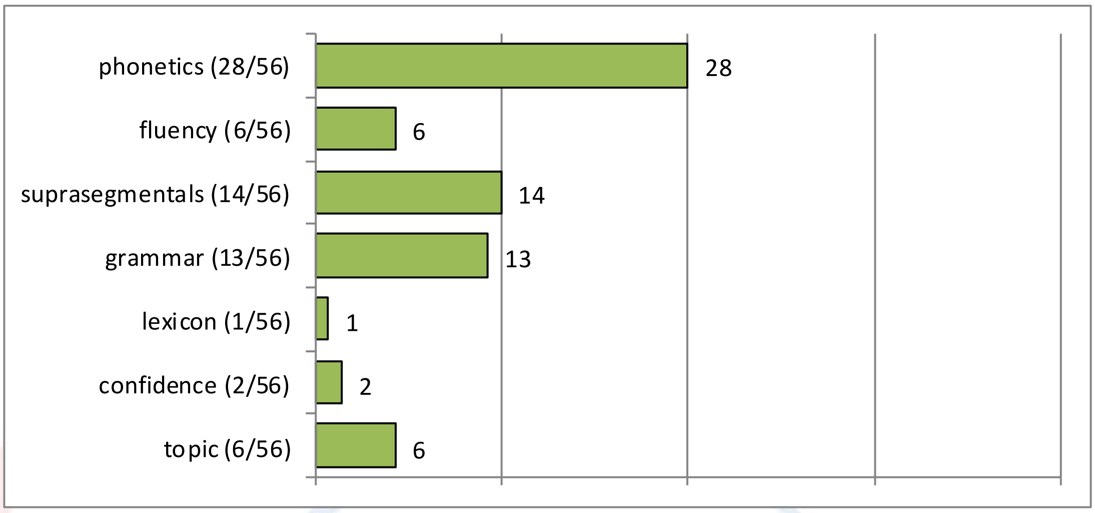


Figure 9. Number of negative comments by the teacher raters for the spontaneous speech (n =56; multiple responses)

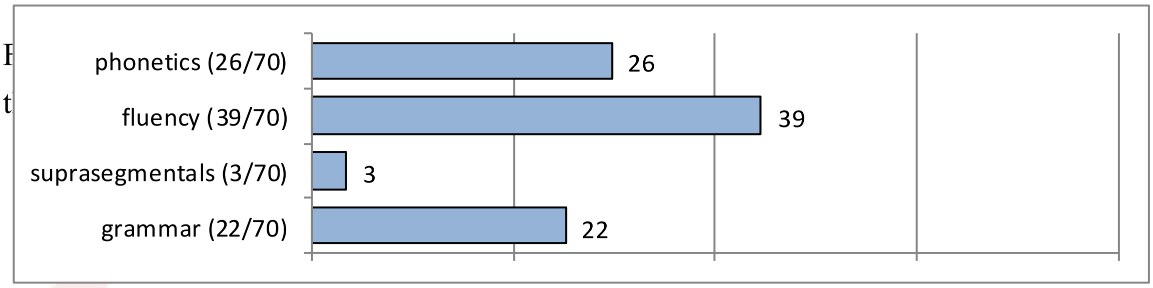


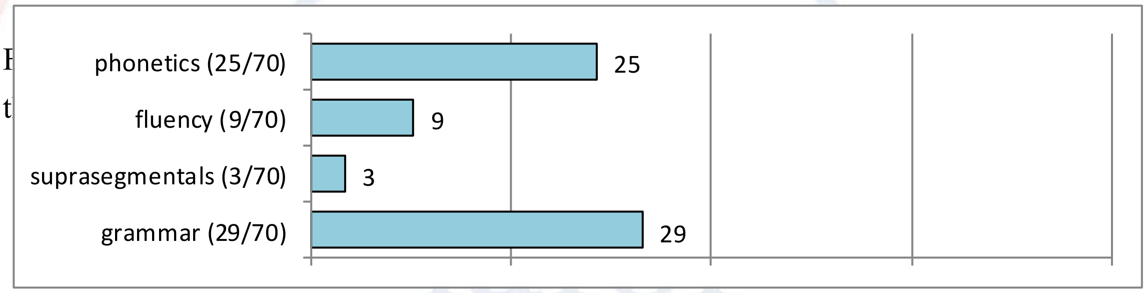
4.3.4. Negative comments for the pre and post study abroad spontaneous speech

Figures 10 and 11 present the numbers of the negative comments appearing in the peer raters' comments for the pre and post study abroad spontaneous speech. The sample size was 70. As can be seen in Figure 10, terms about fluency occurred most frequently for the pre study abroad spontaneous speech with 39 out of 70 comments (55.7%). Terms about phonetics came second with 26 comments (37.1%), and grammar third with 22 comments (31.4%). Comments on suprasegmentals were only three (4.3%). In the comments for post study abroad spontaneous speech, the number of comments about fluency dropped dramatically to nine (12.9%). On the other hand, the number of comments for grammar rose to 29 (41.4%). Terms about suprasegmentals remained the same in the comments for post study abroad (three comments, 4.3%).

The results of the correlation coefficient was r =-.2933, R^2 =.0860, t = 0.7515, p >.05 (p =.2794). So it was confirmed that the comments for the pre and post study abroad spontaneous speech by the peer raters had no correlation.

In Figures 12 and 13, the numbers of negative terms appearing in teacher raters' comments for the pre and post study abroad spontaneous speech was presented. Phonetics was the most frequently mentioned feature for both the pre and post study abroad speech, 14 out of 28 comments for the pre, and 12 for the post study abroad speech. The percentages were 50.0% and 42.9% respectively. Grammar came second with seven comments for the pre study abroad speech, and five comments for post study abroad speech. The percentages were 25.0% and 17.9% respectively. The number of negative comments on suprasegmentals fell from seven comments to four comments





(25% and 14.3% respectively). Comments on fluency were the least mentioned by teacher raters. The numbers of comments were four for the pre study abroad speech, and three for post study abroad speech (14.3% and 10.7% respectively).

The result of the correlation coefficient was *r* =.9817, R² = .9637, *t* =12.6298, *p* <.001 (*p* =.0000). So it was confirmed that the number of negative comments for the pre and post study abroad spontaneous speech had strong positive correlation.

Figure 12. Number of negative comments on the four categories by the teacher raters for the pre study abroad spontaneous speech (n =28; multiple responses)

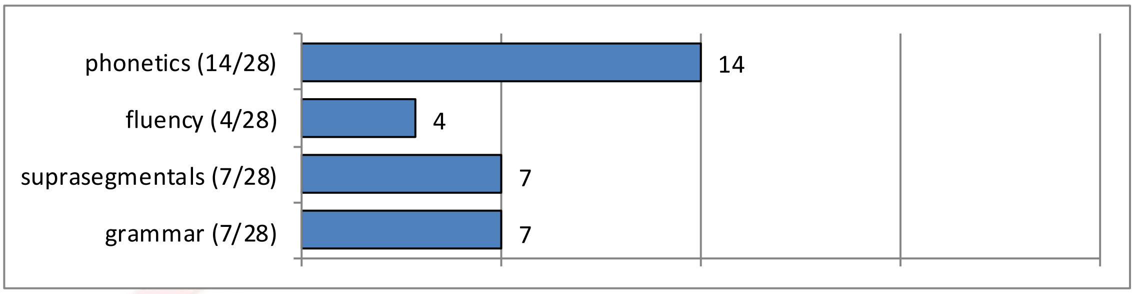
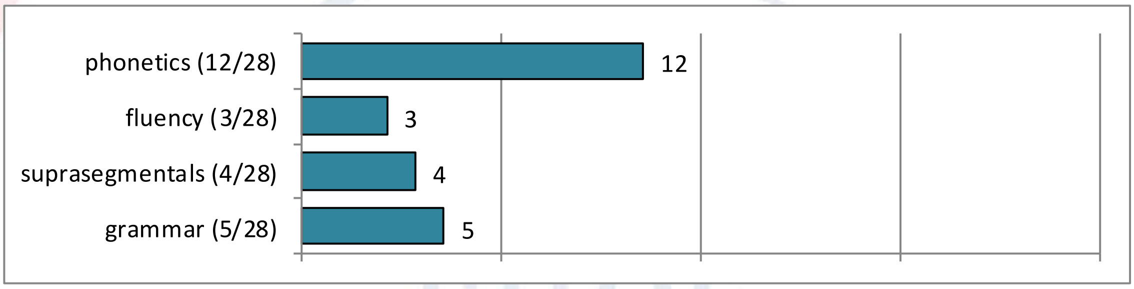


Figure 13. Number of negative comments on the four categories by the teacher raters for the post study abroad spontaneous speech (n =28; multiple responses)



4.4. Negative comments per score

In this section, the negative comments for reading speech are presented per score, per category, and per rater group. The scores 1 and 2 were disregarded in this section because the sample sizes of those scores were too small to analyse: the size of score 1 was n =1, and score 2 was n =2. So the results presented in this section are concerning the *scores 3, 4, 5, 6*, and *7*. Please note the no teacher raters gave any negative comments for the *score 7* speeches.

4.4.1. Negative comments per score given for the reading speech

The numbers of negative comments per score for reading speech given by the peer and teacher raters are shown in Figures 14 and 15 respectively. Negative comments about phonetics were most frequently found in the comments for the *score 3*, and the percentages got smaller as the intelligibility got better except for the comments for the *score 7*. Negative comments about fluency were found most frequently for the *score 5*, and the percentage got smaller as the intelligibility got either better or worse. The percentages of negative comments for suprasegmentals showed an inverted *W* shape.

In contrast, the percentages for grammar looked like a *W* shape. Negative comments about grammar were found for the *score 5* most frequently, and just a few or no negative comments were found in the comments for the other scores.

The correlation coefficients of the percentages of negative comments on the four categories per score were calculated and shown in Tables 6 and 7. As can be seen in Table 6, it was found that a positive correlation between fluency and grammar (r =.6666, p <.05), and a negative correlation between suprasegmentals and grammar (r =-.8709, p <.01) were confirmed to be statistically significant.

In the teachers' comments, negative comments on phonetics were most frequently found for the *score 3*, and tended to get smaller in percentage as the intelligibility got better. Comments on fluency were found most frequently for the *score 5*, and the percentage tended to get smaller as the intelligibility got either better or worse. Comments on suprasegmentals were found most frequently for the *score 4*, and second frequently found for the *score 6*. Comments on grammar were found for the *score 5* most frequently, and a few or no comments were found for the other scores.

Figures 16 and 17 show the polynomial line graphs for the peer sand teachers' comments for the reading speech per score respectively.

4.4.2. Negative comments per score given by the peer raters for the spontaneous speech

The numbers of negative comments per score about the four categories for the spontaneous speech given by both the peer and teacher raters were shown in Figures 18 and 19, and the correlation coefficients in Tables 8 and 9, and polynomial trend line graphs in Figures 20 and 21 respectively. As can be seen in Figure 18, the percentage of comments about phonetics given by the peer raters got smaller as the intelligibility got better. Comments about grammar showed similar tendency except for the comments for the *score 7*. The comments were about the plural and particles. As can be seen in Table 8, negative comments about grammar had positive correlations with phonetics and suprasegmentals, and a negative correlation with fluency. In the teacher raters' comments as shown in Figure 19, the percentage of the comments about phonetics were most frequently found for the *score 3*, and tended to get smaller in percentage as the intelligibility got better. Comments about fluency were most frequently found for the *score 6* and got smaller as the intelligibility got worse. Comments about suprasegmentals were found most frequently for the *scores 3* and *6*, and slightly less frequently for the *score 4*. Comments about grammar were most frequently found for the *score 5*, and then for the *score 3* and for the *score 6*. As can be seen in Table 9, negative comments on phonetics and fluency had a strong negative correlation.

Figure 14. No. of negative comments per score by the peer raters for the reading speech

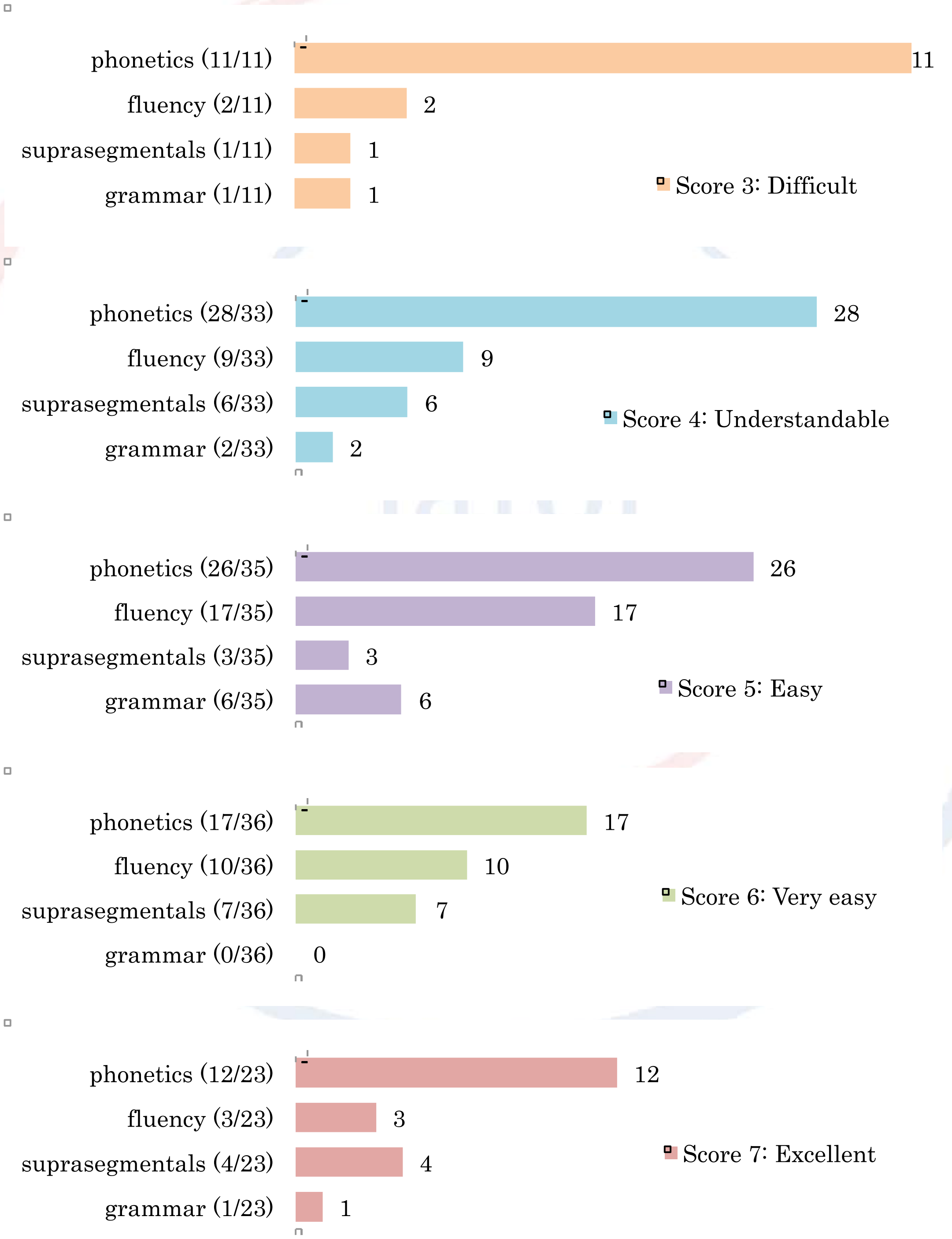


Figure 15. No. of negative comments per score by the teacher raters for the reading speech

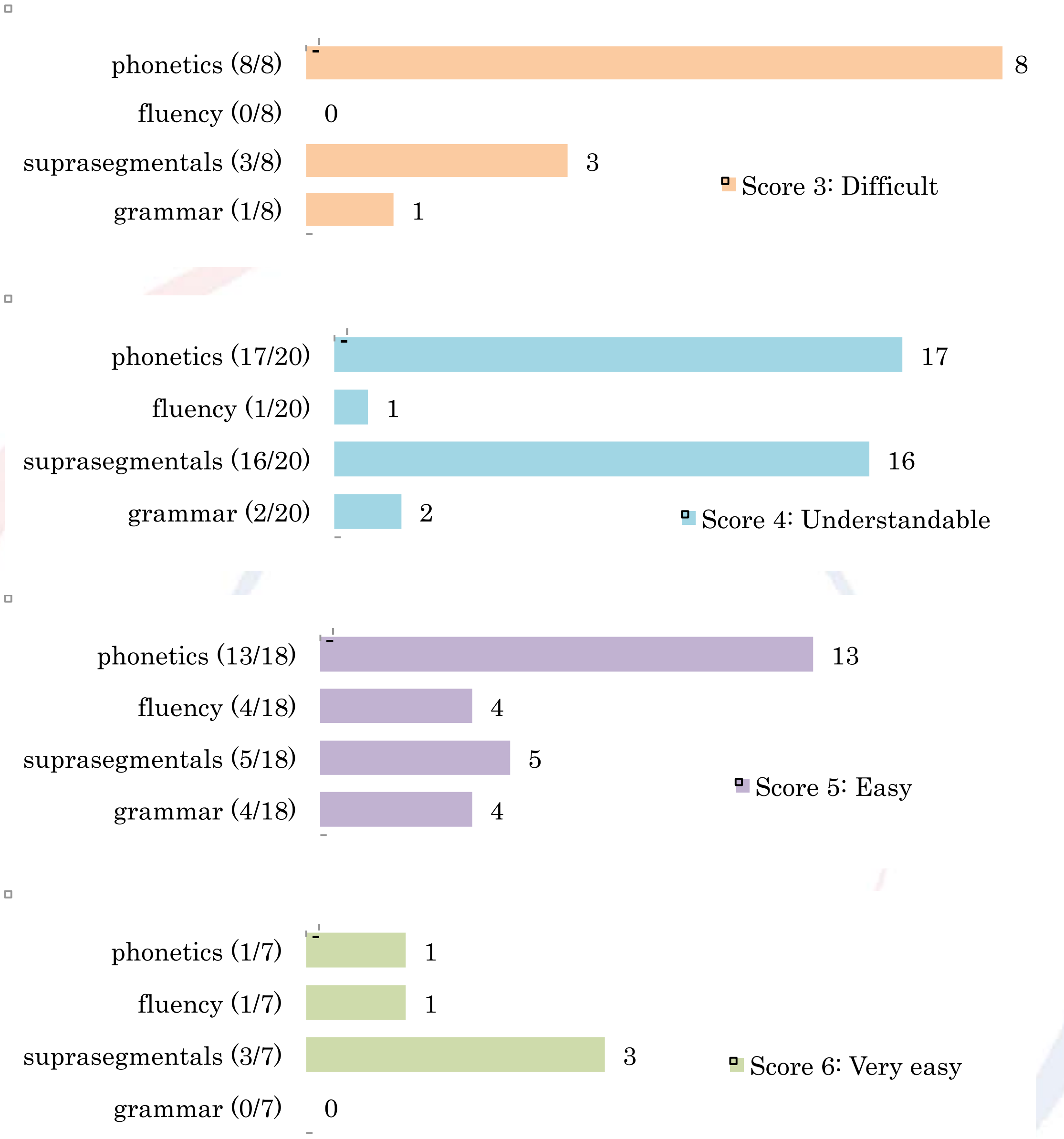


Table 6. Correlation coefficient of negative comments on the four categories per score by the peer raters for the reading speech

	phonetics	*fluency*	*supraseg.*	*grammar*
phonetics	1			
fluency	0.0526	1		
supraseg.	-0.6383	-0.4045	1	
grammar	0.5229	0.6666 *	-0.8709 **	1

Note: supraseg. = suprasegmentals; ** *p* <.01, * *p* <.05

Table 7. Correlation coefficient of negative comments on the four categories per score by the teacher raters for the reading speech

	phonetics	*fluency*	*supraseg.*	*grammar*
phonetics	1			
fluency	-0.5257	1		
supraseg.	0.1551	-0.4771	1	
grammar	0.6519	0.2950	-0.3475	1

Note: supraseg. = suprasegmentals

Figure 16. Polynomial trend line graph for the number of negative comments per score by the peer raters for the reading speech

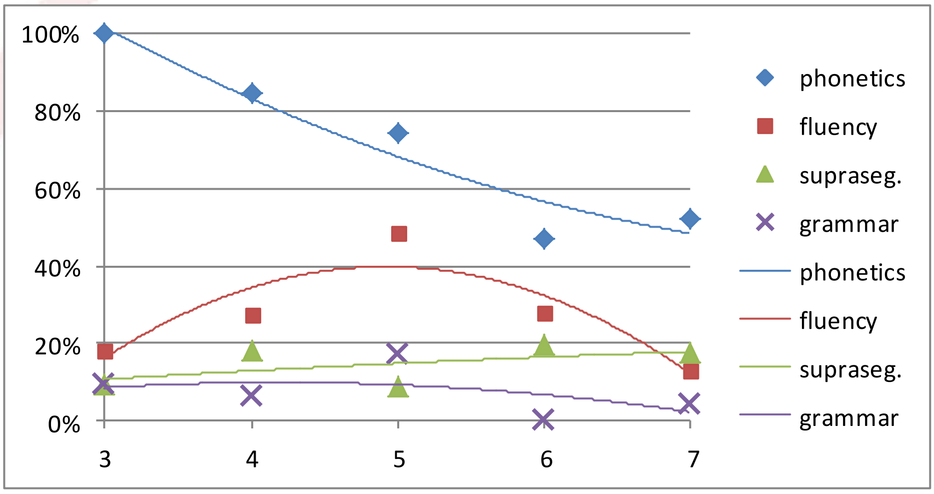


Figure 17. Polynomial trend line graph for the number of negative comments per score by the teacher raters for the reading speech

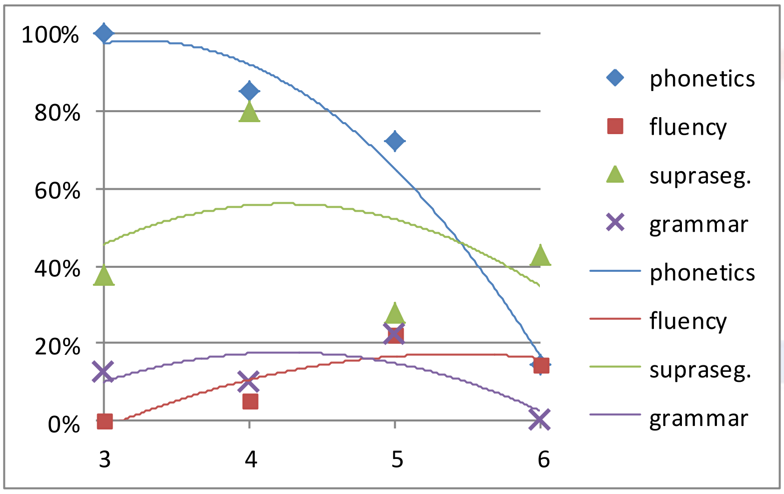


Figure 18. No. of negative comments per score by the peer raters for the spontaneous speech

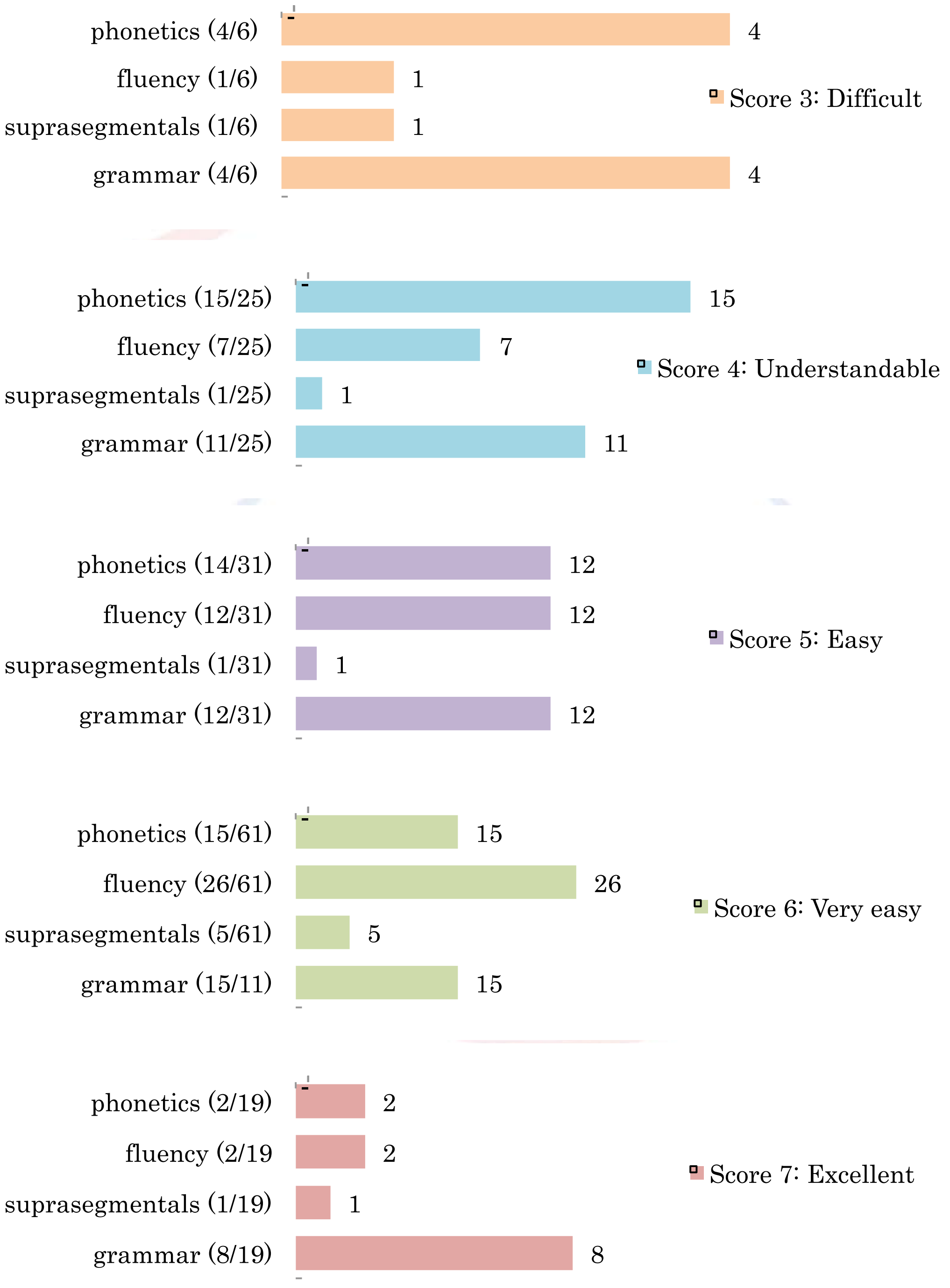


Figure 19. No. of negative comments per score by the teacher raters for the spontaneous speech

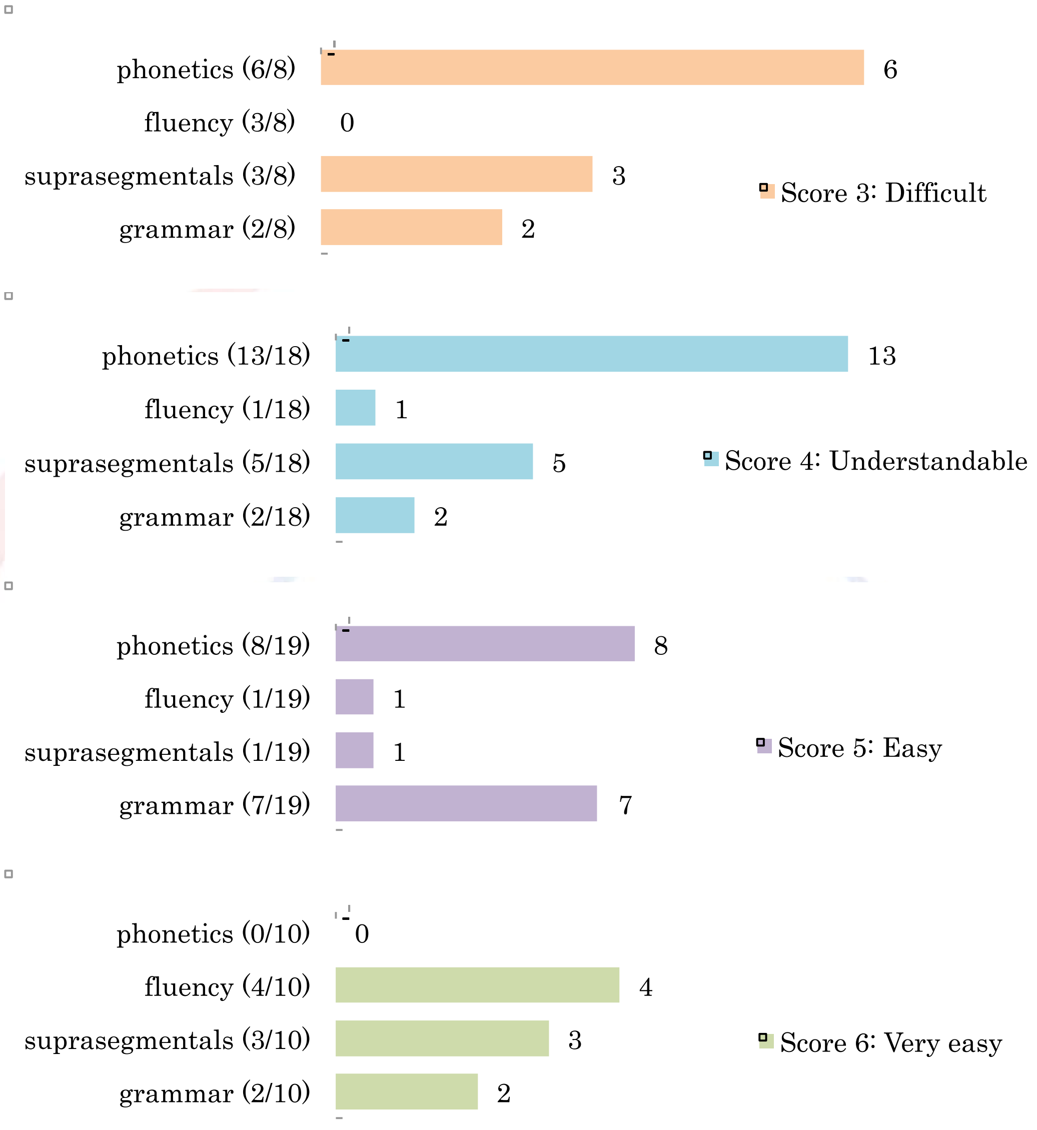


Table 8. Correlation coefficient of negative comments on the four categories per score by the peer raters for the spontaneous speech

	phonetics	*fluency*	*supraseg.*	*grammar*
phonetics	1			
fluency	-0.0200	1		
supraseg.	0.4557	-0.3327	1	
grammar	0.6769 *	-0.6785 *	0.6654 *	1

Note: supraseg. = suprasegmentals; * *p* <.05

Table 9. Correlation coefficient of negative comments on the four categories per score by the teacher raters for the spontaneous speech

	phonetics		*fluency*	*supraseg.*	*grammar*
phonetics	1				
fluency	-0.9267	***	1		
supraseg.	0.1937		0.1368	1	
grammar	-0.1521		-0.2117	-0.6627	1

Note: supraseg. = suprasegmentals; *** *p* <.001

Figure 20. Polynomial trend line graph for the number of negative comments per score by the peer raters for the spontaneous speech

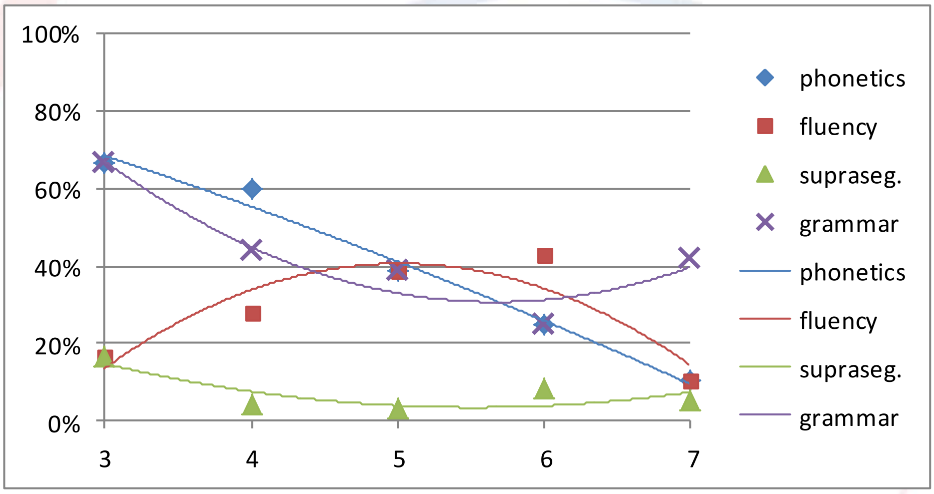
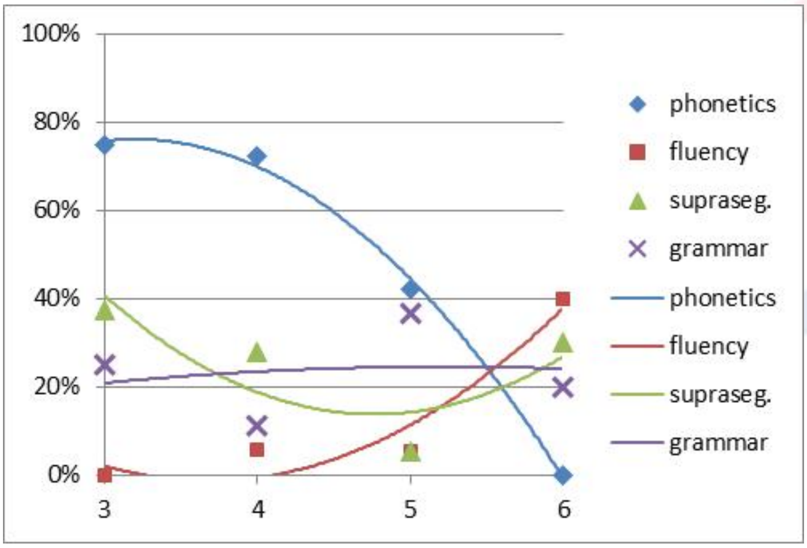


Figure 21. Polynomial trend line graph for the number of negative comments per score by the teacher raters for the spontaneous speech



5. Discussion

In the present study, it was found that the mean scores given by the peer raters were more lenient for both reading and spontaneous speech than the mean scores by the teacher raters. Gass and Varonis (1984) stated that raters' familiarity with the topic, the utterances given by any non-native speakers, the utterances given by the same nationality as the speaker, and the speaker himself/herself facilitated the raters' comprehensibility. In the case of present study, the teacher raters were more familiar with English utterances given by Japanese L1 speakers than the peer raters. In Perlmutter (1989), she adopted a 5-point scale for 21 American undergraduate students to rate the overall intelligibility of the speech samples spoken by 24 international graduate students whose average score on TOEFL was 561 with the range from 490 to 627. The mean score of intelligibility was 2.71 and 3.00 for the pre and post-training speech. She stated that "the upper end of the rating scale was not used by the listener-subjects very often" (p.520). Considering the TOEFL scores of the speakers, the ratings by the American undergraduate students seem rather severe. This suggested that the peer raters' evaluation would have been rather harsh. As was mentioned in chapter 2, Hsieh (2011) stated that "undergraduates were not familiar with the rating criteria for judging the examinees, thus, they sometimes made their rating decisions solely through their appraisal of whether they felt a particular examinee was qualified to be an ITA, a criterion not on the rating rubric" (p.64).

The result obtained in the present study was opposed to their results however. One of the reasons for the peer raters to have been lenient to JEFL speech could be because the peer raters were visiting Japan to learn Japanese and experience Japan. Their age and attitude toward Japan and Japanese culture should be considered also. Their decisions to participate in an exchange programme and visit Japan would not have been easily made. Rather, it could have been a big decision. So their favourable attitude and respect toward Japan and Japanese culture, or their state of actually being in Japan might have influenced their ratings in a positive way.

We would like to discuss the raters' comments as well. The peer raters gave comments about fluency more frequently than the teacher raters, while the teachers concerned more about suprasegmentals. Hsieh (2011)'s statement about the undergraduates' attitude indicates that teachers were more sophisticated linguistically, and able to distinguish the problems in the non-native speech. Teachers in general are aware of the criteria in rating examinees' performance since that is one of the main works in their profession. On the other hand, students are not so sophisticated linguistically. That could be one of the reasons why the peer raters did not mention suprasegmentals, but fluency. We also think it to be possible that the teacher raters might have ignored about JEFL learners' fluency intentionally. Since speaking English fluently is very difficult for JEFL learners, and also because the teachers had been in Japan for more than a decade teaching English to them, they might have thought that JEFL learners would not be able to speak English fluently and/or would give many pauses in the utterances. That could be one of the reasons why they were able to pay attention to suprasegmentals if they ignored fluency.

Also we found that some of the peer raters gave negative comments for the *score 7* speech while the teachers did not. It seems that the teacher raters seem to have acknowledged the *score 4* "*understandable*" as the mid-point and set it as the benchmark. And therefore they seem to have acknowledged the *score 7 "excellent"* as the ceiling which should not be easily given to the speeches which were not really excellent. In fact, the number of overall speech samples that achieved *score 7* was merely four out of 112 speech samples in the teacher raters' ratings (3%). On the other hand, the number was 42 out of 280 overall speech samples in the peer raters' ratings (15%). Also not a few peer raters gave negative comments to the speech samples that achieved *score 7,* that was 28 out of 42 samples (66.6%). One of the reasons why they gave negative comments for the *score 7* speech samples could be because they were L2 learners themselves and visiting Japan to learn the language. Another reason could be observable in the instruction of the rating software we applied. In the description of the comment box, we wrote: "can you tell/describe what features you think should be improved by the speaker to become an excellent speaker of English". As they themselves were L2 learners and some of them, or perhaps all of them, might have been struggling with Japanese language, they might have felt empathy for the speakers and gave extra advice. In fact, many of the negative comments had particles together with the negative terms, such as "minor errors", "a bit slowly", or "could word on the r sound a little more", etc. For the speakers, their advice will be highly beneficial, but we wonder if it was relevant for L2 speech assessment. It might be better to change the description in the instruction, or should we instruct the raters orally in advance of the rating session. We did give oral instruction in the present experiment, but did not instruct about the benchmark. So this matter needs reconsidering.

6. Conclusion

In this study, it was confirmed by English L1 peer and teacher rater groups that nine-month study abroad experience improved JEFL learners' oral intelligibility. We applied two types of speech tasks, reading and spontaneous speech, and two rater groups evaluated the JEFL learners' speech intelligibility and gave comments for each speech file. The results revealed various things. The peer raters were found to be more lenient toward JEFL learners' intelligibility than the teacher raters. Phonetics was found to be the major factor of less intelligible speech. The peer raters gave comments on fluency more frequently than the teachers did. On the other hand, the teacher raters gave comments about suprasegmentals more frequently than the peer raters did. Both fluency and suprasegmentals were not mentioned very much in the less intelligible speech.

The result that English L1 raters were concerned more about phonetics could be considered for a pedagogical suggestion. At high school in Japan, the regulation that English lessons are basically to be given in English was enforced on the 1st of April in 2013. This regulation has been controversial, and many teachers seem to be uncertain whether they are able to run a lesson only in English. However, this implies that oral proficiency in English has become more important, and will be much more important in the future. The importance of practice in English pronunciation should be reconsidered nationwide.

References

Anderson-Hsieh, J. & Koehler, K. (1988). The effect of foreign accent and speaking rate on native speaker comprehension. *Language Learning*, 38 (4), 561-613.

Bolinger, D. L. (1958). A theory of pitch accent in English. *Word*, 14, 109-149.

Caban, H. L. (2003). Rater group bias in the speaking assessment of four L1 Japanese ESL students. *Second Language Studies*, 21 (2), 1-44. Retrieved from http://www. hawaii.edu/sls/sls/wp-content/uploads/2011/06/Caban.pdf

Cummins, R. A. & Gullone, E. (2000). Why we should not use 5-point Likert scales: The case for subjective quality of life measurement. *Proceeding, Quality of Second International Conference on Quality of Life in Cities* (National University of Singapore, Singapore), 74–93.

Dauer, R. M. (1983). Stress-timing and syllable-timing reanalized. *Journal of Phonetics.*

11 (1), 51-62.

Derwing, T. M. & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition*, 19 (1), 1-16.

Derwing, T. M., Rossiter, M. J., Munro, M. J., & Thomson, R. I. (2004). Second language fluency: Judgments on different tasks. *Language Learning*, 54 (4), 655-679.

Ellis, R. (1997). *Second language acquisition.* Oxford: Oxford University Press.

Fry, D.B. (1955). Duration and intensity as physical correlates of linguistic stress. *The Journal of the Acoustical Society of America,* 27 (4), 765-768.

Garland, R. (1991). The mid-point on a rating-scale: Is it desirable? *Marketing Bulletin*,

2, 66-70.

Gass, S. & Varonis, E. M. (1984). The effect of familiarity on the comprehensibility of

nonnative speech. *Language Learning*, 34 (1), 65-87.

Hsieh, C. N. (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. Retrieved from http://www.cambridgemichigan.org/sites/default/files/resources/SpaanPapers/Spaan_V9_Hsieh.pdf

Kubozono, H. & Ota, S. (1998). *Onin Kozo to Akusento* [The structure of Phonology and

Accent]. Tokyo: Kenkyusha Shuppan.

Ladeforged, P. & Johnson, K. (2011). *A course in Phonetics* (6th Edn.). Boston, USA: Wadsworth.

Longman Dictionary of Contemporary English (2nd ed.). (1987). London: Longman Group UK Limited.

Munro, M. J. & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, 45 (1), 73-97.

Munro, M. J. & Derwing, T. M. (1998). The effects of speaking rate on listener evaluation of native and foreign-accented speech. *Language Learning*, 48 (2), 159-182.

Nelson, C. L. (2011). *Intelligibility in world Englishes.: Theory and application.* New York and London: Routledge.

Oxford Advanced Learner's Dictionary (5th ed.). (1995). Oxford: Oxford University

Press.

Perlmutter, M. (1989). Intelligibility rating of L2 speech pre- and postintervention. *Perceptual and Motor Skills,* 68, 515-521.

Rubin, D. L. (1992). Nonlanguage factors affecting undergraduates' judgments of nonnative English-speaking teaching assistants. *Research in High Education*, 33 (4), 511-531.

Rubin, D. L. & Smith, K. A. (1990). Effects of accent, ethnicity, and lecture topic on undergraduates' perceptions of nonnative English-speaking teaching assistants. *International Journal of Intercultural Relations*, 14 (3), 337-353.

Shimaoka, T. (2004). *Nihon-go kara super-native no eigo he* [From Japanese to super-native English]. Tokyo: Soutakusha syuppan.

Smith, L. E. (1992). Spread of English and issue of intelligibility. In Kachru, B. B. (Ed.)

The other tongue: English across cultures (pp. 75-90). Urbana: University of Illinois Press.

Tajima, K., Port, R., & Dalby, J. (1997). Effects of temporal correction on intelligibility

of foreign-accented English. *Journal of Phonetics,* 25 (1), 1-14.

Takebayashi, S. (1996). *Eigo Onsei-gaku* [English phonetics]. Tokyo: Kenkyusha.