How Teacher–Raters Assess Learners' Pronunciation

Ayako Yokogawa
Tokyo University of Marine Science and Technology

Reference Data:

Yokogawa, A. (2015). How teacher–raters assess learners' pronunciation. In P. Clements, A. Krause, & H. Brown (Eds.), *JALT2014 Conference Proceedings*. Tokyo: JALT.

In this research I investigated how teacher–raters assess learners' pronunciation when the learners read a given passage aloud. The purposes of the research were (a) to identify the consistency of ratings by 19 English teachers on the pronunciation of 4 learners when they orally read a passage similar to the Read-a-Text-Aloud task in TOEIC Speaking Test and (b) to spot gaps between the ratings of native and nonnative English speaker teacher–raters, if there are any. The research was conducted as a preliminary study to explore valid and reliable ratings on learners' pronunciation by classroom teachers.

本研究では、学習者が与えられたパッセージの音読をする際の発音を、英語教師がどのように評価するかを精査する。研究の目的は、19人の英語教師によるTOEICスピーキングテストの音読問題に似たパッセージを4人の学習者が音読する際の発音に対する評価における一貫性、また英語を母語あるいは非母語とする教師との間で想定される評価における差異を明らかにすることである。本研究は、教師による妥当で信頼できる発音評価の可能性を探求するための予備研究として行われた。

As needs and demands for reliable and quantifiable assessment to evaluate English learners' speaking skills have grown in Japan, educational institutions and business entities have started to use speaking tests such as TOEIC Speaking Test (http://www.toeic.or.jp/sw/), Versant (http://www.versant.co.jp/), and BULATS (http://www.justycom.jp/cambridge02/bulats.html). These tests include a section that accesses test-takers' pronunciation and the assessment is supposed to be based on a valid and systematic scoring system. I taught a preparation course for the TOEIC Speaking Test for years and have wondered if classroom teachers could rate students' pronunciation in a reliable way so that students can improve their pronunciation and performance on the TOEIC Speaking Test, anticipating that there could be individual differences, preferences, or biases in teachers' ratings that can have a negative impact. Derwing and Munro (2005) pointed out that listeners' responses to an utterance may be influenced by their experience with accented speech or personal bias against particular accents or voices. Thus, ratings on speaker's utterance by listeners, including classroom teachers, can to a large extent be subjective.

When classroom teachers evaluate learners' pronunciation subjectively, their assessment of learners' pronunciation cannot be reliable. Assessment is important for language teachers, learners, and anyone involved in the process of language teaching and learning because valid assessment can reveal what has been learned and what needs to be learned. The effect of assessment on teaching and learning is called *washback* or *backwash*. Gates (1995) defined washback as "the influence of testing on teaching and learning" (p. 101). Washback can be either positive or negative. Bachman (1990) stated that positive washback occurs when the assessment used reflects the skills and content taught in the classroom. To generate positive washback in a classroom, teachers' assessment needs to be valid and reliable, not random and inconsistent, which is the main concern of the current research.

This study focused on assessment of pronunciation. In the research I investigated (a) how teacher–raters assess learners' pronunciation when the learners read a given passage aloud and (b) how consistent their ratings are. Numerical ratings with a scale of 0-3 were analyzed and compared against the actual ratings that the learners previously received on the TOEIC Speaking Test. I also examined gaps between ratings by native English speaker teachers and ratings by nonnative English speaker teachers to explore if there are any differences or tendencies in teachers' ratings depending on the raters' first language.

Research Questions

The following research questions guided the study:

1. Do classroom teachers rate learners' pronunciation similarly or differently and to what extent?
2. Do native speaker (NS) teachers and nonnative speaker (NNS) teachers rate learners' pronunciation similarly or differently and to what extent?

Method

For this research, 19 English teachers currently teaching in different institutions in and around Tokyo volunteered to rate learners' pronunciation. First, the 19 teachers, both NS (*n* = 9) and NNS (*n* = 10, all Japanese), were asked via email to listen to Benchmark 3 (high level) and Benchmark 1 (low level) sample answers of the Read-a-Text-Aloud task in the TOEIC Speaking Test provided by Educational Testing Service, the developer of the test, as an anchor for their assessment. The written script of the sample question was as follows:

> If you're shopping, sightseeing, and running around every minute, your vacation can seem like hard work. To avoid vacation stress, come to the Blue Valley Inn on beautiful Lake Mead. While staying at our inn, you'll breathe clean country air as you view spectacular sights. With its spacious rooms, swimming pool, and many outdoor activities, the Inn is the perfect place for a vacation you won't forget. (Educational Testing Service, 2010, p. 3)

Next, the 19 teacher–raters listened to the sound files of four Japanese learners' oral rendition of a written text that was similar to the above speech script in length and linguistic complexity. The written script of the task was as follows:

> Thank you for calling Stalks Florists. Our office is currently closed. Our regular business hours are from 9 A.M. to 5 P.M., Monday through Thursday and on Saturday. For your convenience, we stay open until 8 P.M. on Fridays. If you leave a message, one of our florists will return your call as soon as possible. You can also order flower arrangements on the Internet at stalks.com. And remember, we offer free delivery in most areas of the country. We appreciate your patronage. (Educational Testing Service, 2008, p. 25)

Then, the 19 teacher–raters assessed the oral rendition by four learners regarding two elements: pronunciation (local features) and intonation and stress (global features) on a scale of 0-3 each. The teacher–raters were also invited to write comments about the learners and their own assessment if there was anything noticeable that they thought was worth sharing. Numerical ratings (0-3) and voluntary comments were collected via email and analyzed.

Learners' Profiles

The four learners were learning English in a foreign language institute in Tokyo, Japan, and were chosen from an advanced class of the TOEIC Speaking Test preparation course that I was then teaching. All of them had taken TOEIC several times and the TOEIC Speaking Test once and demonstrated a good level of English proficiency, as shown in Table 1.

Table 1. Learners' Results on TOEIC and TOEIC Speaking Test

Learner	TOEIC (Highest score: 990)	TOEIC Speaking Test (Highest score: 200)	Pronunciation on TOEIC Speaking Test (Highest score: 3)	Intonation/stress on TOEIC Speaking Test (Highest score: 3)
A	855	160	3	2
B	910	160	3	3
C	840	120	2	2
D	855	130	2	2

According to TOEIC Program Data and Analysis 2013 (The Institute for International Business Communication, 2014), the average score of Japanese test takers of the TOEIC Speaking Test is 123.4. Learners A and B demonstrated a strong speaking proficiency; Learners C and D were average. Learner A had lived in the US for 4 years when he was a child and returned to Japan at the age of nine. Learners B and C had never studied abroad. Learner D spent 1 month in Ireland participating in a school program.

Findings and Discussion

Comparison of Ratings: TOEIC Speaking Test Results Versus Teachers–Raters' Results

The TOEIC Speaking Test results and the teachers' assessments were similar for the most part (see Table 2). Learners C and D, in particular, were assessed by the teacher–raters just as the TOEIC Speaking Test raters had, so it seems that the raters involved in both assessments agreed that Learners C and D would be rated as at the medium level.

Table 2. TOEIC Speaking Test Results and Teacher–Raters' (*N* = 19) Results

Learner	Pronunciation Scale: 0-3		Intonation/stress Scale: 0-3	
	TOEIC-Speaking Test	Teachers' average ratings	TOEIC-Speaking Test	Teachers' average ratings
A	3	3.0	2	2.89
B	3	2.57	3	2.42
C	2	2.0	2	2.0
D	2	1.94	2	2.05

Learner B, however, demonstrated an interesting gap in the results. Learner B was rated lower by the teacher–raters than by the TOEIC Speaking Test raters (pronunciation: 2.57 versus 3, intonation/stress: 2.42 versus 3). Several possibilities could explain the gaps. The teacher–raters were possibly strict in assessing Learner B's pronunciation due to her strong Japanese accent, which made the raters hesitant to give her the highest score. Another possibility is that the TOEIC Speaking Test examines test-takers' *intelligibility* as a user of the English language, not their *native-likeness*, so Learner

B was rated as the highest because her speech was sufficient to deserve the highest level of intelligibility.

Gaps in Ratings: NS Teachers versus NNS Teachers

As shown in Table 3, there seemed to be a consensus among NS teachers and NNS teachers on the ratings of Learners A, C, and D but ratings for Learner B, again, demonstrated noticeable gaps.

Table 3. Average NS and NNS Teachers' Ratings (Scale 0-3)

	Pronunciation		Intonation/stress	
Learner	NS teachers' rating (*n* = 9)	NNS teachers' rating (*n* = 10)	NS teachers' rating (*n* = 9)	NNS teachers' rating (*n* = 10)
A	3.0	3.0	2.88	2.9
B	2.44	2.7	2.55	2.3
C	2.0	2.0	2.10	2.0
D	2.0	1.7	2.22	1.9

Both NS and NNS teachers rated Learner D's intonation/stress higher than her pronunciation. Interestingly, though, NS teachers rated Learner D higher than NNS teachers (pronunciation: 2 versus 1.7, intonation/stress: 2.22 versus 1.9). I speculate that NS teachers living in Japan may have become familiar with the Japanese-accented English spoken by Japanese people and thus tend to view the pronunciation of Japanese speakers of English as acceptable, whereas NNS teachers whose first language is Japanese may be more sensitive to and critical of Japanese-accented English.

Learner B demonstrated intriguing gaps. Although NS teachers rated Learner B as more or less 2.5 for pronunciation (2.44) and intonation/stress (2.55), NNS teachers gave better ratings on pronunciation (2.7) than they did on intonation/stress (2.3). One possible explanation is that NNS teachers were less aware of, or less able to detect, the flaws in Learner B's pronunciation than NS teachers were. If the speculation holds true, it is possible that NNS Japanese teachers are able to detect Japanese-accented intonation in English more easily than mispronounced words. Another possibility is that NNS have demonstrated the tendency to rate accents more harshly than they rate intelligibility, as some of the previous research findings (e.g., Munro, Derwing, & Morton, 2006) have shown.

Table 4 shows the distributions of NS and NNS teachers' ratings on pronunciation (local features). NS teacher–raters seemed to be more critical about local features than NNS teacher–raters: Learner B received 3 (high level) from four NS teacher–raters and from seven NNS teacher–raters, and Learners C and D received 3 from no NS teacher–raters but they received 3 from three NNS teacher–raters. In other words, NS teacher–raters were more generous about global features, as previously mentioned in the discussion about Table 3.

Table 4. Distribution of NS and NNS Teachers' Ratings of Pronunciation

Learner	NS teachers' ratings (*n* = 9)			NNS teachers' ratings (*n* = 10)		
	3	2	1	3	2	1
A	9	0	0	10	0	0
B	4	5	0	7	3	0
C	0	9	0	2	6	2
D	0	9	0	1	7	2

In addition, Learners C and D were rated as 2 (medium level) by all NS teachers, but ratings by NNS teachers were not unified. For example, two NNS teachers rated Learner C as 3 (high level) and two rated him as 1 (low level), but no NS teachers rated him as either 3 or 1. This inconsistency could have been because the NNS teachers had not yet established their own criteria to assess learners' pronunciation. Possibly, they would have been able to rate Learner C more consistently if they had received formal instructions as raters for the task and become familiar with how to apply the benchmark answers in their assessment process.

Table 5 shows the distribution of NS and NNS teachers' ratings on intonation/stress (global features). Both NS teachers and NNS teachers rated learners' intonation/stress more similarly than they did pronunciation, except in the case of Learner B. As discussed in the section about Table 3, NNS teachers were stricter on Learner B's intonation/stress (high: 3, medium: 7) than NS teachers were (high 5, medium: 4).

Table 5. Distribution of NS and NNS Teachers' Ratings of Intonation/Stress

Learner	NS Teachers' Rating (*n* = 9)			NNS Teachers' Rating (*n* = 10)		
	3	2	1	3	2	1
A	8	1	0	9	1	0
B	5	4	0	3	7	0
C	1	8	0	1	8	1
D	2	7	0	1	7	2

Comments from Teacher–Raters

Some of the 19 teacher–raters commented on the learners and their own assessment voluntarily as follows:

> "In terms of intelligibility and appropriateness, they are all good enough to be easily understood." (NNS rater 1)

> "All speakers are mostly intelligible. While some speakers experienced challenges pronouncing some words from the text, this did not interfere with understanding their message." (NS rater 1)

> "Learner A gave the listener a feeling that he was forced to record this script. Learner B's rhythm was above average. Learner C has a strong Asian accent. He's understandable, but not pleasant to listen to. Learner D has a strong nasalization and a strong Asian accent on some words: See-Saw rhythm." (NS rater 2)

> "I think Learner B deserves 2, not 3, on TOEIC Speaking Test." (NSS rater 2)

As seen in the above comments, there were individual differences in interpretation of the learners' oral rendition. Some raters put emphasis on holistic intelligibility; others paid closer attention to specific phonetic features. NNS rater 1 and NS rater 1 used the term *intelligibility/intelligible*. Munro and Derwing (1995) defined intelligibility as "the extent to which an utterance is actually understood" (p. 291). NNS rater 1 and NS rater 1 seem to have valued intelligibility because they understood what was read aloud, although NS rater 1 mentioned some flaws in the speakers' pronunciation.

NS rater 2 commented on each speaker's global features in detail. NNS rater 2 stated that Learner B should have been rated lower on the TOEIC Speaking Test. It is interesting to see that NS rater 2 commented on Learner B positively, but NNS rater 2 commented on the same speaker negatively, which reflects teacher–raters' subjectivity in perception of speakers' utterances.

Suggestions for Future Research

Following are possible ways to develop this research.

1. More learner subjects with various backgrounds and proficiency levels can provide more data. Because the research was

conducted as a preliminary study, it was important for the researcher to limit the number of variables to consider. Therefore, the subject–learners were selected from one class that the researcher was then teaching, and no huge gaps among learners' proficiency levels were anticipated. If subject–learners are chosen from various environments, there will be more varied results to be examined in depth.

2. If NS teacher–raters living overseas rate the same four learners, they can provide interesting data. As mentioned above, there is a possibility that NS teachers living in Japan for some time have become accustomed to Japanese-accented English. NS teachers who have never lived in Japan or are unfamiliar with Japanese-accented English might rate learners' pronunciation differently, which might provide an interesting comparison.
3. Teacher–raters' backgrounds could be analyzed and compared to their ratings. Teacher–raters are not free from preferences or biases that come from their backgrounds, such as nationality, regional accent, and first language. This may influence them when assessing learners' pronunciation. It can reveal interesting results if teachers' ratings and their backgrounds are analyzed from several aspects. For example, one anticipated result is that American NS teachers would rate learners who have studied in the UK lower than would British NS teachers.

Conclusion

This study was an attempt to investigate research questions on (a) the consistency of teachers-raters' assessment on learners' pronunciation, and (b) gaps between the ratings by NS teacher–raters and NNS teacher–raters. The 19 teacher–raters, both NS and NNS, rated learners' pronunciation consistently for the most part but their ratings were inconsistent in some aspects. NS teacher–raters were more critical of pronunciation (local features) but less critical of intonation/stress (global features) than were NNS teacher–raters.

I would like to point out that there is a limit to the validity of a direct comparison of the TOEIC Speaking Test results and teacher–raters' results. Actual speech scripts of the TOEIC Speaking Test questions are unavailable for reproduction because all test items are not disclosed to the public. Therefore, it is impossible to compare the ratings of the TOEIC Speaking Test and those of teacher–raters on the same speech script. For this preliminary research, however, the learners' actual ratings on the TOEIC Speaking Test served as a useful comparison to the ratings by the teachers involved.

Pedagogical Implications

This small study has shown that classroom teachers rated learners' pronunciation similarly for the most part, but their ratings were inconsistent in some aspects. For valid and reliable ratings on learners' pronunciation by classroom teachers, there are things that can be done. First and foremost, teachers should keep in mind that their assessment on learners' pronunciation can be affected by individual differences and may not be the ultimate judgment for learners. Also, both NS teachers and NNS teachers need to be aware that they can have different perspectives on learners' pronunciation. This could provide learners with opportunities to see room for improvement from various angles. Finally, because pronunciation assessment can vary depending on what is assessed (e.g., intelligibility versus native-likeness) and how the assessment is conducted (e.g., in this research on a scale of 0-3), teachers should be aware of how they assess learners' pronunciation and let the learners know what they want the learners to accomplish and how it will be evaluated. Such an environment for assessment will provide learners with positive washback.

Bio Data

Ayako Yokogawa is a designated associate professor at Tokyo University of Marine Science and Technology. Her primary research interests are in the area of assessment and faculty development. <ayokog0@kaiyodai.ac.jp>

References

Bachman, L. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Derwing, T. M., & Munro, M. J. (2005). Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly*, *39*, 379-397.

Educational Testing Service. (2008). *TOEIC test official preparation book, vol. 3*. Tokyo: The Institute for International Business Communication.

Educational Testing Service. (2010). *TOEIC speaking and writing tests*. Tokyo: The Institute for International Business Communication.

Gates, S. (1995). Exploiting washback from standardized tests. In J. D. Brown & S. O. Yamashita (Eds.), *Language testing in Japan* (pp. 101-106). Tokyo: JALT.

Munro, M. J., & Derwing, T. M. (1995). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and Speech*, *38*, 289-309.

Munro, M. J., Derwing, T. M., & Morton, S. L. (2006). The mutual intelligibility of L2 speech. *Studies in Second Language Acquisition*, *28*, 111-131.

The Institute for International Business Communication. (2014). TOEIC program data and analysis 2014. Retrieved from http://www.toeic.or.jp/library/toeic_data/toeic/pdf/data/DAA.pdf