Part 2

Insights From Assessing Other Language Skills and Components

3 Pronunciation and Intelligibility in Assessing Spoken Fluency

Kevin Browne and Glenn Fulcher

Introduction

This chapter argues that any definition of the construct of fluency must include familiarity of the listener with the entire context of an utterance. This extends to pronunciation, the intelligibility of which is an interaction between the phonological content of the utterance and the familiarity of the listener with the second language (L2) speech produced by speakers from a specific first language (L1) background. This position recognizes that successful communication is not merely a matter of efficient cognitive processing on the part of the speaker. Fluency is as much about perception as it is about performance. This is a strong theoretical stance, which can be situated within an interactionist perspective on language use. Good theory generates specific predictions that may be empirically tested. If the listener is critical to the construct, we would expect to discover two facts. First, that variation in listener familiarity with L2 speech results in changes to scores on speaking tests. Secondly, that this variation is associated with estimates of intelligibility when the speaker is kept constant. In this chapter we describe a study that investigates these two predictions. We situate the findings in the context of language testing, where variation in familiarity among raters is a cause for concern.

The Fluency Construct

The construct of fluency is endemic in language teaching and applied linguistics research. Teachers feel especially relaxed in using the term to refer to a desirable quality of learner speech that approximates 'nativelike delivery' – or 'proficiency' in the broadest sense (Lennon, 1990). This comfortable assumption hides the fact that there is no single definition of 'nativelike' within a single language (Davies, 2004), and variation between languages is frequently considerable (Riazantseva, 2001). Early research by Fillmore (1979) and Brumfit (1984) provided a very broad definition of fluency, including 'filling time with talk' through automatized language production, selecting relevant content for context, and creating coherent utterances without becoming 'tongue tied'. Koponen and Riggenbach (2000) exposed the metaphorical nature of the fluency construct, characterizing speech as fluid, or flowing like a river: smooth and effortless in its passage from mind to articulation. The articulation includes pronunciation, which adds or subtracts from the perception of fluidity (Educational Testing Service, 2009) on the part of the listener.

The language of fluency definitions reveals what we have elsewhere called the 'janus-faced' nature of the construct (Fulcher, 2015: 60). Language testers often make the assumption that pronunciation is a simple 'on/off switch' for intelligibility (Fulcher, 2003: 25). But this assumption focuses too much upon the production of the individual speaker in relation to the acquisition of some standard, usually the notion of the 'native speaker'. It is the assumption that underlies the automated assessment of pronunciation in computer-based tests by matching performances on reductive task types such as sentence repetition (Van Moere, 2012) with preselected norms. The place of pronunciation in cognitive fluency models also treats phonological accuracy as merely the observational component of part of a speech-processing model such as that of Levelt (1989, 1999), so that measurements may be treated as surrogates for general L2 proficiency (Segalowitz, 2010: 76).

The reality is that pronunciation is variably problematic, depending on the familiarity of the listener with the L1 of the speaker. This realization is significant in the context of language assessment, where such familiarity becomes an important variable that impacts scores being assigned to speakers.

Defining intelligibility and familiarity

Familiarity shapes and facilitates speech processing. The intelligibility of speech is speaker–listener dependent (Riney *et al.*, 2005). Attention has been drawn to how differential rater familiarity with accent can affect test scores, posing a threat to both reliability and validity (e.g. Carey *et al.*, 2011; Winke *et al.*, 2013; Xi & Mollaun, 2009). Research into rater accent familiarity as a potential threat has tended to focus on listeners' shared L1 with the test takers (Kim, 2009; Xi & Mollaun, 2009), residency and employment in the country where the L1 of test takers is spoken (Carey *et al.*, 2011), and prior personal L2 study experiences (Winke *et al.*, 2013). In these studies the construct of familiarity was not carefully defined, but was inferred on the basis of different types and amounts of linguistic experiences a rater had with the L2 accent. A definition that can be extrapolated from these studies is that accent familiarity is a speech perception benefit developed through exposure and linguistic experience. Carey *et al.* (2011: 204) labelled it 'interlanguage phonology familiarity'.

Gass and Varonis (1984) released the earliest study of familiarity. They argued that four types of familiarity contribute to comprehension: familiarity with topic of discourse; familiarity with nonnative speech in general; familiarity with a particular nonnative accent; and familiarity with a particular nonnative speaker. Their study used 142 native-speaking university students as participants who listened to recordings of two male Japanese-English speakers and two male Arabic-English speakers completing three reading-aloud tasks: (1) reading a story; (2) reading a set of five 'related sentences' that pertained to the story although not included in the text; and (3) a set of 'unrelated sentences' with contexts or topics pertaining to 'real world knowledge'. The recordings were used to create 24 different 'tapes'. Each tape included first a reading of either the 'related' or 'unrelated' sentences. Next came a reading of the story, followed by the set of sentences not included prior to the story. The items were read by different combinations of speakers. Each listener was asked to complete transcription tasks of the related and unrelated sentences, and produce a short summary of the story as a measure of comprehension.

Gass and Varonis concluded that 'familiarity of topic' is the greatest contributor to comprehension of the four familiarity types researched (see also Kennedy & Trofimovich, 2008). This was determined by one-tailed *t*-tests comparing the pre- and post-text transcriptions of the related sentences. The results revealed a significant difference of means of errors ($p < 0.05$) for three of the four speakers (Gass & Varonis, 1984: 72). More errors were reported in the pre-story transcriptions of the 'related' sentences than in the post-story transcriptions, suggesting that native speakers are more capable of determining the content of nonnative speakers' utterances if they know the specific topic. Likewise, the 'unrelated' sentences determined to be comprised of 'real world knowledge' resulted in a significantly lower instance of errors ($p < 0.0001$) as compared to the 'related' sentences when they occurred in the pre-story position on the tapes.

Familiarity of speaker, familiarity of accent and familiarity of nonnative speech in general were found to contribute to the comprehensibility of nonnative speakers, although these findings were not based on any statistically significant differences in the data. Familiarity of accent was determined to positively affect transcription accuracy by observing instances of speaker error in the pre- and post-story positions. Greater accuracy was observed when listeners had encountered the same accent in the pre-story or story reading when transcribing the post-story sentences.

It can be argued that what Gass and Varonis discovered was that familiarity facilitates ‘intelligibility’ and not ‘comprehension’, according to the more useful definitions provided by Smith and Nelson (1985: 334). Smith and Nelson suggested the following interpretations of intelligibility, comprehension and interpretability:

- *intelligibility*: word/utterance recognition;
- *comprehensibility*: word/utterance meaning (locutionary force);
- *interpretability*: meaning behind word/utterance (illocutionary force).

Although Gass and Varonis did include the story summary for listener participants there was no analysis or discussion of the data to support the claim that the different types of familiarity they examined contribute to comprehension, which would include out of necessity the notions of locutionary or illocutionary force. While we do not wish to argue against the possibility that familiarity may contribute to comprehension and determination of meaning, Gass and Varonis’ findings can only be said to relate to intelligibility of word or utterance recognition, depending upon listener familiarity.

As Smith and Nelson (1985: 334) suggested, the terms ‘intelligibility’, ‘comprehension’ and ‘interpretability’ should be defined to avoid confusion, since these terms have been applied in various ways and at times interchangeably. The definition of intelligibility in this research follows Field (2005), as being how the phonological content of a speaker is recognized by the listener. This definition takes into account how the listener processes utterances, which we argue is a function of level of familiarity.

It is therefore theorized that increasing accent familiarity reduces the processing effort required for the phonological content of speech. Thus, raters with higher levels of familiarity are more likely to find speech intelligible, while lower levels of familiarity reduce intelligibility. Familiarity on the part of the listener is therefore the most important variable to impact the intelligibility aspect of fluency, which results directly in score variation (Derwing *et al.*, 2004).

Research questions

In order to investigate the role of intelligibility as a critical component of fluency within the argument that the construct exists as much within the listener as it does within the speaker, we formulated two research questions:

(1) How do raters’ familiarity levels with L2 English spoken by L1 speakers of Japanese affect pronunciation test scores?
(2) How do raters’ familiarity levels with L2 English spoken by L1 speakers of Japanese affect intelligibility success rates?

Methodology

No previous study of rater accent familiarity as a threat to test validity has simultaneously examined how raters score candidates on operational tests concurrently with rater intelligibility success rates. As a result, little is known about why score differences occur. The methodology therefore provides the means to investigate the relationship between these variables and identify potential impact on scores.

Participants

Eighty-seven ESL/EFL teachers and/or graduate students enrolled in applied linguistics or TESOL programmes were recruited via email to participate as volunteer rater participants. Most (*n* = 73) were L1 English speakers and 14 were L2 speakers (see Table 3.1).

Five first-year Japanese university students studying English as non-English majors at Tsukuba University (male: *n* = 1; female: *n* = 2), Waseda University (male: *n* = 2), and one American male from the Southern United States were recruited as the speaker participants. The students were enrolled in intermediate-level English courses at the time, and had studied English for six years prior to participating.

Table 3.1 Rater participants' home country list

United Kingdom	35
USA	34
Canada	7
South Africa	4
Japan	4
Australia	3
Brazil, France, Jamaica, Libya, Malta, Spain, St. Lucia, Sudan, Syria, Ukraine	1 (per country)
Total	87

The test

A three-part test was constructed to measure rater intelligibility success rates for comparison with the scores allocated to different speakers. Since participation was voluntary, the test was designed to be completed in less than 25 minutes. Rater participants required a computer connected to the internet and were recommended to complete the test with headphones in a quiet room.

Part 1 of the test included questions related to raters' professional, biographical and linguistic experiences. Questions focused on their L1(s), home country, country of residence at that time, ESL/EFL teaching and/or research experience, and familiarity with Japanese-English. Raters' familiarity with the accent was determined from responses to a four-level self-reporting scale. The scale and number of participants selecting each level was:

- **No familiarity** (*n* = 13).
- **Limited familiarity**: You have heard Japanese speakers of English but without regularity, and/or have not had Japanese students during the last two years (*n* = 32).
- **Some familiarity**: You have spent at least the last two years with students from Japan, have visited Japan and/or regularly watch TV or movies in Japanese (*n* = 4).
- **Very familiar**: You are a native speaker of Japanese, have lived in Japan for one or more years and/or have studied the Japanese language for one or more years (n = 38).

Part 2 was divided into six sections, with one section for each speaker participant. Each section contained a recording of the speaker reading two sentences. The raters were asked to listen to each sentence and then complete an intelligibility gap-fill task by typing missing words from an incomplete transcript of the sentences on the screen. The native speaker was placed in first position. This was decided primarily to help the raters better understand the tasks they were asked to complete, and to serve as an 'easily intelligible' example of pronunciation to process. There were a total of 28 intelligibility gap-fill items in the test (24 spoken by the Japanese-English speakers; four spoken by the native speaker).

After completing the intelligibility task for one speaker, raters scored that speaker for pronunciation using a five-point scale adapted from the TOEFL iBT Speaking Scoring Rubric 'Delivery' sub-scale for the independent speaking tasks, which incorporates the notion of 'fluidity' (Educational Testing Service, 2009). The scale that the raters used in the current study is shown in Table 3.2. Each recording was approximately 18 seconds in length. Raters could start, stop or replay the recording at their discretion. No visuals were provided; raters had no additional information about the speakers that would lead to inferences that might impact scores (e.g. gender, age, L1, nationality) (see Rubin, 1992). There are a number of limitations in the methodology. First, raters completed test items in the same sequence. The survey website made randomizing the items prohibitive, as they were clustered according to speaker, so order effect could not be controlled. Secondly, the native speaker may have 'loomed over the study' (Isaacs & Thomson, 2013), but none of the raters reported the use of a native speaker example to have been problematic, and the data from the native speaker were not included in the analyses.

The sentences read by the speaker participants were adaptations of the Bamford–Kowal–Bench (BKB) sentence lists (Bench *et al.*, 1979), which were

Table 3.2 Pronunciation score descriptors used in the current study

5	Speech is generally clear and requires little or no listener effort. Only one listening required.
4	Speech is generally clear with some fluidity of expression, but it exhibits minor difficulties with pronunciation and may require some listener effort at times. Only one listening required.
3	Speech is clear at times, although it exhibits problems with pronunciation and so may require more listener effort. It was necessary to listen more than once before attempting to complete the gap fill.
2	Consistent pronunciation difficulties cause considerable listener effort throughout the sample. It was necessary to listen more than once before attempting to complete the gap fill.
1	Cannot comprehend at all.

Source: Adapted from the TOEFL iBT Speaking Scoring Rubric, Independent Tasks (Educational Testing Service, 2015: 189–190).

originally designed to measure the listening capabilities of children with varying degrees of sensorineural hearing loss. Sensorineural hearing loss is an affliction that affects how speech is processed. Regardless of the volume of the speech signal, sensorineural hearing loss affects the clarity of the acoustic signal the listener perceives. Like Bench *et al.*'s original tests, this test was designed to measure differences in speech perception and processing with gap-fill transcription tasks with clarity of speech determined through word identification accuracy.

The BKB test measures speech perception abilities using samples with pronunciation a ‘normal’ listener should find intelligible, whereas the test designed for the research described in this chapter measures speech perception using accented samples for which the rater participants had variable familiarity. The BKB sentences were standardized in length and lexical complexity and served to reflect natural speech of NS children aged 8–15 (see Table 3.3). The sentences designed for this study were also standardized in length and lexical complexity to represent the vocabularies of intermediate-level Japanese-English speakers. Lexical complexity was determined utilizing the JACET

Table 3.3 Examples of the original BKB sentences

An old woman was at home.
He dropped his money.
They broke all the eggs.
The kitchen window was clean.
The girl plays with the baby.

Source: Bench *et al.* (1979: 109)

8000, a corpus of the 8000 most frequently used English words by Japanese speakers of English. Lexical complexity was restricted to the 3000 most frequently used words in order to eliminate the need to provide explanations of word meaning or pronunciation to speaker participants. As a result, each speaker was left to pronounce each word in a sentence as they thought fit.

A unique aspect of the sentences designed for this instrument was the decision to intentionally construct them to have complex or unpredictable contexts. As previously discussed, Gass and Varonis (1984) argued that 'familiarity of context' was the most significant contributory type of familiarity to success in word/utterance identification tasks. This is because background knowledge of context helps the listener to successfully guess words or utterances that he or she is not able to otherwise identify. We judged that the use of sentences with complex or unpredictable contexts might effectively reduce the context familiarity benefit identified by Gass and Varonis, thus allowing us to see the impact of pronunciation alone on listener evaluation of intelligibility. The resulting sentences constructed for the test were not nonsensical; they were syntactically accurate although contextually complex or unpredictable (see Table 3.4).

The sentences were also designed to feature aspects of Japanese-English phonology that are known to be problematic both in production for the speakers and in distinction by unfamiliar listeners. Elements of problematic Japanese-English phonology incorporated in the test included /r/–/l/ distinction, the lax vowels /ɪ/, /ʊ/, /ʌ/ and /ə/, and the voiced dental fricative /ð/ (see Carruthers, 2006, for a complete discussion of pronunciation difficulties of Japanese speakers of English).

Part 3 of the test sought rater comments in order to gain additional insight into the raters' opinions of the research instrument and their experiences completing the test.

Table 3.4 The test sentences developed for the current study

Speaker 1	They had a tiny day.
	The old soaps are dirty.
Speaker 2	They are paying some bread.
	The play had nine rooms.
Speaker 3	The institution organism was wet.
	The dog made an angry reader.
Speaker 4	The ladder is across the door.
	He cut his skill.
Speaker 5	The union cut some onions.
	She sensed with her knife.
Speaker 6	Mine took the money.
	The matches lie on the infant.

Analyses

Facets 3.71 Many Faceted Rasch Measurement (MFRM) software (Linacre, 2013) and SPSS (Version 20) were used to analyze the test data. MFRM allows for multiple aspects facets of a test to be examined together and, in the case of this study, to investigate raters' intelligibility scores and their abilities to transcribe utterances. Only data from the five L1 Japanese speakers were included in the MFRM analyses. This was designed to determine whether rater accent familiarity differences resulted in significant score differences. The pronunciation score and intelligibility success rates data were analyzed separately (as recommended by Linacre, personal communication) due to the differences of tasks, as fit statistics were compromised when the different tasks were analyzed together.

Two facets (the raters and speakers) and one grouping facet (raters' familiarity level with Japanese-English) were examined. The intelligibility data were also analysed examining two facets – the raters and the items – again with familiarity level as a grouping facet.

Findings and Discussion

MFRM analyses of the pronunciation scores yielded results supporting previous findings that raters' familiarity with speakers' accents can have a significant effect on oral proficiency scores (e.g. Carey *et al.*, 2011; Winke *et al.*, 2011, 2013). The most informative and important piece of output from Facets analyses is the variable map, which summarizes the key information of each facet and grouping facet into one figure. The scale utilizes measurements in terms of 'logits' that reflect probability estimates on an equal-interval scale. Figure 3.1, which presents the Facets variable map for pronunciation scores, is separated into five vertical columns:

(1) Column 1 displays the logit scale ranging from –7 to 2. The scale provides a reference for measurements of all other columns. The measure 0 represents even likelihood, or 50–50 odds of prediction.
(2) The second column displays the leniency of each rater from most (top) to least.
(3) The third column shows the grouping facet revealing that the 'very familiar' group of raters were most lenient in scoring pronunciation.
(4) Column 4 shows the ability measures of each speaker-participant. The most proficient was speaker E shown at the top.
(5) The fifth column shows the five-point rating scale used to score pronunciation. Each speaker participant's position in the fourth column is horizontal to their mean score on this rating scale.

```
|Measr|+Rater (Most lenient - top)          |+Familiarity Level (Most lenient - top)  |+Speaker   |Scale|
|-----+-------------------------------------+-----------------------------------------+-----------+-----|
+   2 +                                     +                                         +           + (5) +
|     |                                     |                                         |           |     |
|     |                                     |                                         | Speaker E | --- |
|     | 14  48                              |                                         |           |     |
+   1 + 50  63  65                          +                                         + Speaker C +     +
|     |                                     |                                         |           |     |
|     | 25  3                               |                                         |           |     |
|     | 23  61                              | Very Familiar                           |           |     |
|     | 19  2   24  26  37  62  85          |                                         |           |  3  |
+   0 + 17  55                              + Limited Familiarity  Some Familiarity   +           +     +
|     |                                     |                                         |           |     |
|     | 13  18  34  40  6   66  74  9       | No Familiarity                          | Speaker B |     |
|     | 38  54                              |                                         |           |     |
|     | 1   15  21  22  29  31  44  52  8   |                                         | Speaker D |     |
+  -1 + 35  36  56  68  80                  +                                         +           +     +
|     | 58                                  |                                         | Speaker F | --- |
|     | 30  32  4   5   51  71  72          |                                         |           |     |
|     | 12  39  43  60  78  81              |                                         |           |     |
|     | 20  49  76                          |                                         |           |     |
+  -2 + 11  16  27  28  75  77  82          +                                         +           +     +
|     | 42  7   79                          |                                         |           |     |
|     | 10  41  45  46  86                  |                                         |           |     |
|     | 67                                  |                                         |           |     |
|     | 47  53  57  64                      |                                         |           |  2  |
+  -3 +                                     +                                         +           +     +
|     | 33  84  87                          |                                         |           |     |
|     | 70                                  |                                         |           |     |
|     |                                     |                                         |           |     |
|     | 59                                  |                                         |           |     |
+  -4 +                                     +                                         +           +     +
|     |                                     |                                         |           | --- |
|     | 83                                  |                                         |           |     |
+  -5 + 69                                  +                                         +           +     +
|     |                                     |                                         |           |     |
|     |                                     |                                         |           |     |
|     |                                     |                                         |           |     |
+  -6 +                                     +                                         +           +     +
|     | 73                                  |                                         |           |     |
|     |                                     |                                         |           |     |
|     |                                     |                                         |           |     |
+  -7 +                                     +                                         +           + (1) +
|-----+-------------------------------------+-----------------------------------------+-----------+-----|
|Measr|+Rater (Most severe -bottom)         |+Familiarity Level (Most severe -bottom) |+Speaker   |Scale|
```

Figure 3.1 Facets variable map of pronunciation scores including four levels of familiarity

The range of rater severity shown in Column 2 (7.39 logits) is wider than the spread of speaker-participants' ability in Column 4 (2.7). This indicates that the individual differences of rater severity were high. A closer examination of rater performance by familiarity level provided in Table 3.5 indicates that, as familiarity level increases, so do scores and rater leniency. Pearson's chi-square indicates significant differences of pronunciation scores between the four groups ($\chi^2(3) = 12.3$, $p = 0.01$).

Figure 3.2 shows the correlation between level of familiarity and pronunciation score. Although shared variance is a modest 15%, in a speaking test such an influence may make a significant impact on an individual score.

The Facets variable map for intelligibility scores is shown in Figure 3.3. The content of each column is as follows:

Table 3.5 Pronunciation score: Facets rater familiarity level group measures

Familiarity level	*Total score*	*Obs. Av.*	*Measure in logits*	*Model SE*	*Infit MnSq*	*ZStd*
No	144	2.22	–0.38	0.22	0.91	–0.4
Some	48	2.4	–0.09	0.38	0.71	–0.9
Limited	388	2.47	0.03	0.14	0.92	–0.6
Very	526	2.77	0.43	0.12	1.09	0.9
Mean	276.5	2.46	0	0.21	0.91	–0.3
SD	190	0.2	0.29	0.1	0.14	0.7

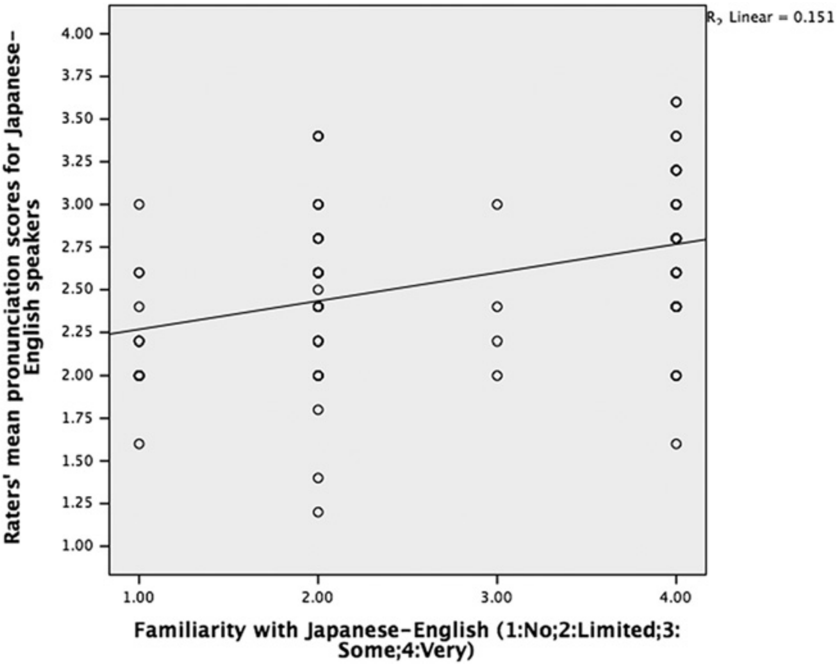
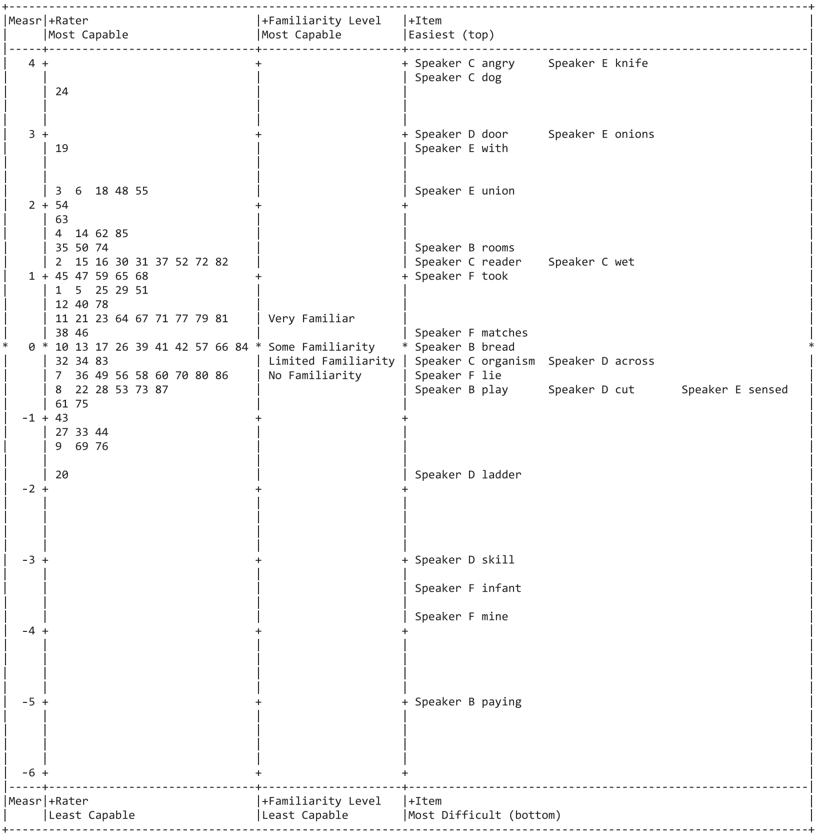


Figure 3.2 Scatterplot showing the correlation between accent-familiarity level and pronunciation scores

(1) Column 1 displays the logit scale ranging from –4 to 6.

(2) The second column shows how the individual raters performed in the intelligibility gap-fill exercises. Raters' individual abilities are reflected in their position on the map with the highest scoring raters at the top.

(3) The third column reveals how rater groups performed. As predicted, the 'very familiar' raters were the most successful and the 'no familiarity' group the least successful at completing the tasks.



```
+----------------------------------------------------------------------------------------------------------------------+
|Measr|+Rater                         |+Familiarity Level   |+Item                                                     |
|     |Most Capable                   |Most Capable         |Easiest (top)                                             |
|-----+-------------------------------+---------------------+----------------------------------------------------------|
|   4 +                               +                     + Speaker C angry     Speaker E knife                      |
|     |                               |                     | Speaker C dog                                            |
|     | 24                            |                     |                                                          |
|     |                               |                     |                                                          |
|     |                               |                     |                                                          |
|   3 +                               +                     + Speaker D door      Speaker E onions                     |
|     | 19                            |                     | Speaker E with                                           |
|     |                               |                     |                                                          |
|     |                               |                     |                                                          |
|     | 3  6  18 48 55                |                     | Speaker E union                                          |
|   2 + 54                            +                     +                                                          |
|     | 63                            |                     |                                                          |
|     | 4  14 62 85                   |                     |                                                          |
|     | 35 50 74                      |                     | Speaker B rooms                                          |
|     | 2  15 16 30 31 37 52 72 82    |                     | Speaker C reader    Speaker C wet                        |
|   1 + 45 47 59 65 68                +                     + Speaker F took                                           |
|     | 1  5  25 29 51                |                     |                                                          |
|     | 12 40 78                      |                     |                                                          |
|     | 11 21 23 64 67 71 77 79 81    | Very Familiar       |                                                          |
|     | 38 46                         |                     | Speaker F matches                                        |
*   0 * 10 13 17 26 39 41 42 57 66 84 * Some Familiarity    * Speaker B bread                                          *
|     | 32 34 83                      | Limited Familiarity | Speaker C organism  Speaker D across                     |
|     | 7  36 49 56 58 60 70 80 86    | No Familiarity      | Speaker F lie                                            |
|     | 8  22 28 53 73 87             |                     | Speaker B play      Speaker D cut       Speaker E sensed |
|     | 61 75                         |                     |                                                          |
|  -1 + 43                            +                     +                                                          |
|     | 27 33 44                      |                     |                                                          |
|     | 9  69 76                      |                     |                                                          |
|     |                               |                     |                                                          |
|     | 20                            |                     | Speaker D ladder                                         |
|  -2 +                               +                     +                                                          |
|     |                               |                     |                                                          |
|     |                               |                     |                                                          |
|     |                               |                     |                                                          |
|  -3 +                               +                     + Speaker D skill                                          |
|     |                               |                     |                                                          |
|     |                               |                     | Speaker F infant                                         |
|     |                               |                     |                                                          |
|     |                               |                     | Speaker F mine                                           |
|  -4 +                               +                     +                                                          |
|     |                               |                     |                                                          |
|     |                               |                     |                                                          |
|     |                               |                     |                                                          |
|  -5 +                               +                     + Speaker B paying                                         |
|     |                               |                     |                                                          |
|     |                               |                     |                                                          |
|     |                               |                     |                                                          |
|  -6 +                               +                     +                                                          |
|-----+-------------------------------+---------------------+----------------------------------------------------------|
|Measr|+Rater                         |+Familiarity Level   |+Item                                                     |
|     |Least Capable                  |Least Capable        |Most Difficult (bottom)                                   |
+----------------------------------------------------------------------------------------------------------------------+
```

Figure 3.3 Facets variable map of intelligibility gap-fill outcomes including four levels of familiarity

(4) The fourth column displays the items from easiest (top) to most difficult (bottom). The items are identified first according to the speaker from whose recording they originated, and the target word. The column reveals that all five speakers produced items that were both easier (with logit scores above zero) and more difficult (with negative logit scores).

The most important results in Column 3 show that the more familiar raters are with Japanese English the more capable they are at transcribing the speakers' utterances. Table 3.6 shows that as familiarity with Japanese-English increases, so does observed intelligibility. Raters 'very familiar' with Japanese-English were 20% more successful than the raters with 'no familiarity'. Figure 3.4 shows that the correlation of the two variables share 31% variance, which indicates a potentially large impact of familiarity on intelligibility.

Table 3.6 Facets intelligibility familiarity level measurements

Familiarity level	*Total score*	*Total count*	*Obs. Av.*	*Measure in logits*	*Model SE*	*Infit MnSq*	*ZStd*
No	156	312	0.50	–0.41	0.16	0.99	0.0
Limited	435	768	0.57	–0.13	0.10	0.91	–1.7
Some	59	96	0.61	0.07	0.30	0.84	–1.0
Very	634	912	0.70	0.46	0.10	1.08	1.4
Mean	321.0	522	0.59	0.00	0.17	0.96	–0.4
SD	227.4	331	0.07	0.32	0.08	0.09	1.2

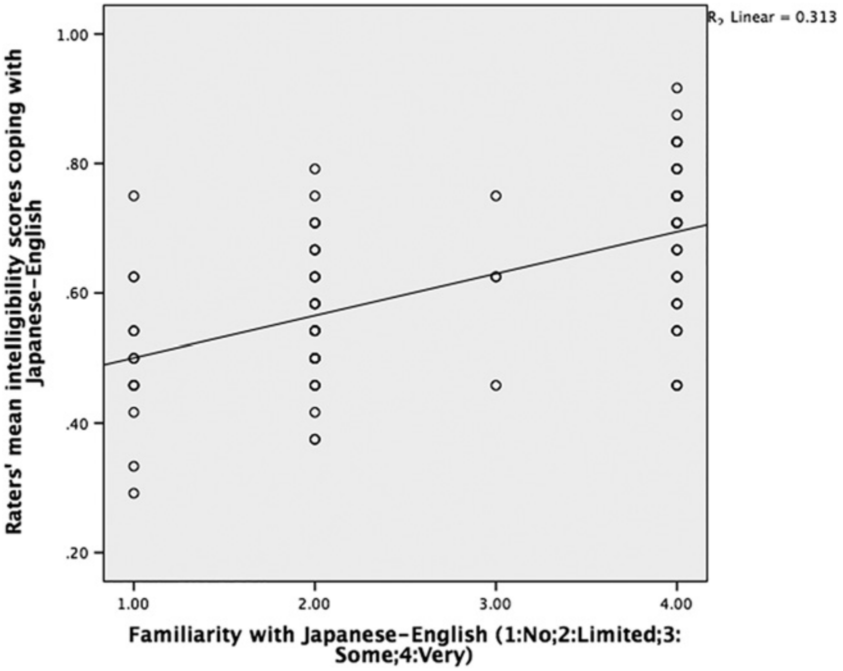


Figure 3.4 Scatterplot showing the correlation between accent-familiarity and intelligibility

Conclusion

We have argued that an understanding of fluency, and the place of pronunciation within a model of fluency, must take into account the listener. The study reported in this chapter addresses the two empirical correlates of the theoretical stance taken. The findings show that both pronunciation test scores and intelligibility vary as a function of listener familiarity. While the current study focuses on pronunciation as one component of fluency, the study supports the theoretical stance that the construct of fluency more generally, and intelligibility more specifically, is situated as much within the listener as the speaker. Perhaps the reason for the listener being ignored in recent cognitive research is the absence of the listener from models of cognitive processing, such as that of Levelt, where it is argued that there are two major parts to speech processing:

> … a semantic system which 'map[s] the conceptualization one intends to express onto some linear, relational pattern of lexical items' and a phonological system which 'prepare[s] a pattern of articulatory gestures whose execution can be recognized by an interlocutor as the expression of … the underlying conceptualization'. (Levelt, 1999: 86)

A speech-processing model of this kind is typically represented as a flowchart. It therefore represents a 'software-solution' to the problem of mind and language. Taken literally, the interlocutor is relegated to the role of a passive recipient of the speaker's output, for which the speaker is completely responsible.

This is a convenient place to be if one wishes to use automated speech assessment systems, as the construct does not involve a listener, and the use of monologic and semi-direct tasks is rendered unproblematic. It could also be argued that listener variability is little more than error, which is eliminated by the removal of variable human raters in automated assessment (Bernstein *et al.*, 2010). However, if listeners are part of the construct, it would seem unreasonable to eliminate them from the equation completely. Language, after all, is a tool for human communication, and so it makes a difference who you are talking to, the context in which you are talking, and the purpose of the communication.

What this research does not do is identify a 'familiarity threshold' that might be recommended for a particular type of speaking test. What it does do is to argue that familiarity is inevitably part of the construct, and to problematize the relationship between familiarity, intelligibility and test scores for the purposes of assessing speaking. This is likely to be of particular importance in contexts where single raters are asked to rate the L2 speech of test takers drawn from a large variety of L1 backgrounds. This situation is common in large-scale L2 testing, where at present there is no attempt to match raters with speakers on the basis of rater familiarity with accented L2 pronunciation from the L1. The issue for high-stakes speaking assessment is the principle that construct-irrelevant facets of a test should be a matter of indifference to the test taker. The principle implies that the test taker should get a similar score (given random error) no matter which rater is randomly selected from the universe of raters available for selection. We normally refer to this as the generalizability of the score across facets of the test (see Schoonen, 2012).

The discovery that the construct resides in the listener as much as in the speaker therefore leads to a dilemma: should familiarity be controlled in order to retain generalizability and the principle of equal treatment, or should familiarity be allowed to vary (as at present) as it is construct relevant? The problem is that although we have argued that familiarity is construct relevant, scores vary with familiarity. Unless it is possible to specify the level of familiarity that would be expected in the target domain to which test scores are intended to predict performance, it would seem reasonable to expect at least a minimum level of familiarity. This is certainly the case in large-scale tests that are used for a variety of decision-making purposes. Achieving familiarity may be obtained in one of two ways: first, by using a measure of familiarity such as the one used in this study to match raters with test takers; and secondly, by providing accent familiarity training to raters across the range of L1s represented in the test taker population at large. Further research is also required into the levels of rater familiarity required for there to be no impact on scores from intelligibility. Such research may need to have wider scales of familiarity than that used in this research, and have a much larger *n*-size for each L1 population, in order to maximize reliability. A larger study may be able to identify a plateau on the scale, which could then be used in conjunction with rater training to select raters for use with test takers from specific L1 backgrounds.

The salience of test method facets in score variance has always been one of the main considerations in investigating the fairness of decision making. It becomes even more problematic when the variance is construct relevant, but potentially random depending on how raters are selected. This paper problematizes the issue of potentially unfair construct-relevant variance, and points the way forward to potential remedies and future research.

References

Bench, J., Kowal, A. and Bamford, J. (1979) The BKB (Bamford–Kowal–Bench) sentence lists for partially hearing children. *British Journal of Audiology* 13, 108–112.

Bernstein, J., Van Moere, A. and Cheng, J. (2010) Validating automated speaking tests. *Language Testing* 27 (3), 355–377.

Brumfit, C. (1984) *Communicative Methodology in Language Teaching: The Roles of Fluency and Accuracy*. Cambridge: Cambridge University Press.

Carey, M.D., Mannell, R.H. and Dunn, P.K. (2011) Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing* 28 (2), 201–219.

Carruthers, S.W. (2006) Pronunciation difficulties of Japanese speakers of English: Predictions based on a contrastive analysis. *HPU TESL Working Paper Series* 4, 17–23.

Davies, A. (2004) The native speaker in applied linguistics. In A. Davies and C. Elder (eds) *The Handbook of Applied Linguistics* (pp. 431–450). London: Blackwell.

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

Educational Testing Service (2009) *The Official Guide to the TOEFL Test* (3rd edn). New York: McGraw-Hill.

Field, J. (2005) Intelligibility and the listener: The role of lexical stress. *TESOL Quarterly* 39 (3), 399–423.

Fillmore, C.J. (1979) On fluency. In C.J. Fillmore, D. Kempler and S.-Y. Wang (eds) *Individual Differences in Language Ability and Language Behaviour* (pp. 85–101). New York: Academic Press.

Fulcher, G. (2003) *Testing Second Language Speaking*. Harlow: Longman.

Fulcher, G. (2015) *Re-examining Language Testing: A Philosophical and Social Inquiry*. London and New York: Routledge.

Gass, S. and Varonis, E.M. (1984) The effect of familiarity on the comprehensibility of nonnative speech. *Language Learning* 34 (1), 65–87.

Isaacs, T. and Thomson, R.I. (2013) Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly* 10 (2), 135–159.

Kennedy, S. and Trofimovich, P. (2008) Intelligibility, comprehensibility, and accentedness of L2 speech: The role of listener experience and semantic context. *Canadian Modern Language Review* 64, 459–490.

Kim, Y.-H. (2009) An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing* 26 (2), 187–217.

Koponen, M. and Riggenbach, H. (2000) Overview: Varying perspectives on fluency. In H. Riggenbach (ed.) *Perspectives on Fluency* (pp. 5–24). Ann Arbor, MI: University of Michigan Press.

Lennon, P. (1990) Investigating fluency in EFL: A quantitative approach. *Language Learning* 40 (3), 387–417.

Levelt, W. (1989) *Speaking: From Intention to Articulation*. Cambridge, MA: MIT Press.

Levelt, W. (1999) Producing spoken language: A blueprint of the speaker. In C. Brown and P. Hagoort (eds) *The Neurocognition of Language* (pp. 83–122). Oxford: Oxford University Press.

Linacre, J.M. (2013) Facets Rasch measurement software (Version 3.71) [computer software]. Chicago, IL: WINSTEPS.com.

Riazantseva, A. (2001) Second language proficiency and pausing: A study of Russian speakers of English. *Studies in Second Language Acquisition* 23 (4), 497–526.

Riney, T.J., Takagi, N. and Inutsuka, K. (2005) Phonetic parameters and perceptual judgments of accent in English by American and Japanese listeners. *TESOL Quarterly* 39 (3), 441–466.

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

Schoonen, R. (2012) The generalizability of scores from language tests. In G. Fulcher and F. Davidson (eds) *The Routledge Handbook of Language Testing* (pp. 363–377). London and New York: Routledge.

Segalowitz, N. (2010) *Cognitive Bases of Second Language Fluency*. New York: Routledge.

Smith, L.E. and Nelson, C.L. (1985) International intelligibility of English: Directions and resources. *World Englishes* 4 (3), 333–342.

Van Moere, A. (2012) A psycholinguistic approach to oral language assessment. *Language Testing* 29 (2), 325–344.

Winke, P., Gass, S. and Myford, C. (2011) *The Relationship Between Raters' Prior Language Study and the Evaluation of Foreign Language Speech Samples*. Princeton, NJ: Educational Testing Service.

Winke, P., Gass, S. and Myford, C. (2013) Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing* 30 (2), 231–252.

Xi, X. and Mollaun, P. (2009) How do raters from India perform in scoring the TOEFL iBT speaking section and what kind of training helps? TOEFL iBT Research Report No. RR-09-31. Princeton, NJ: Educational Testing Service.