Perception and Pronunciation in Fluency Assessment

Kevin Browne and Glenn Fulcher

Introduction

This chapter argues that any definition of the construct of fluency must include the 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 perpective on language use in applied linguistics (and social sciences more generally). 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. Firstly, 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 linguistic research. Teachers feel especially relaxed in using the term to refer to a desirable quality of learner

speech that approximates “native-like delivery” – or “proficiency” in the broadest sense (Lennon, 1990). This comfortable assumption hides the fact that there is no single definition of “native-like” 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 automatised language production, the selection of 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 language of fluency definition reveals what we have elsewhere called the “janus-faced” nature of the construct (Fulcher, 2015). At one and the same time, fluency draws its meaning in part from the linguistic features of utterances, and from the perception and interpretation of those utterances on the part of a listener. However, a focus upon one or the other face of fluency has recently led to a radically different research emphasis.

A cognitive science perspective primarily concerns itself with observable elements of measurable speech performance. It is argued that variations in such observable elements are caused by an underlying construct of L2 cognitive proficiency, defined as the “efficiency of making word-meaning links” and “the functioning of attention-based mechanisms involved in more complex language processing” (Segalowitz, 2010: 76). The observable correlates of cognitive proficiency do not require explanation in and of themselves, but only in terms of variation in the causal construct; and the variation in the causal construct is discovered by variation in the observable elements. In this circular definition, it therefore follows that the observable elements themselves can “…serve as a stand-in measure of

general proficiency and L2 experience" (ibid.: 76). This argument provides the warrant required for the computer scoring of speech using temporal measures such as speech rate, mean speech run, or phonological accuracy, using reductive task types such as read-aloud, sentence repetition, and sentence building (Van Moere, 2012). Phonological accuracy in automated assessment is generally defined as the extent to which pronunciation matches a pre-selected native norm. Cognitive fluency models therefore treat phonological accuracy as the observational component of part of a speech processing model such as that of Levelt (1989; 1999). The measurement aim is supported by research through which "…it is hoped that it will be possible to identify a reasonably small set of cognitive processes that can be reliably associated with an equally reasonably small set of utterance fluency phenomena" (Segalowitz, 2010: 51).

Yet, Segalowitz (2010: 49) admits that "…listeners do not normally treat every pause and hesitation as evidence of dysfluency… implying that a certain amount of pausing and hesitation is acceptable and even expected in so-called fluent speech." A linguistic approach to fluency takes this as its starting point, rather than a post-hoc admission. Applied linguists, like cognitive scientists, wish to identify the observable surface elements in speech that define what we mean by fluency. They differ, however, in treating these observable elements as being in need of contextual interpretation (Fulcher, 1996). That is, the perception of fluency on the part of the listener and the context of utterance are just as important as the observable elements themselves. While it is true that a beginner in a language class may exhibit performance that is easily classifiable as a "dysfluency" because little of the language system has become automatised (Levelt, 1989: 2), there is a plateau after which low inference categories (such as number and length of pauses) become

irrelevant. A silence may be interpreted in many contexts (including a language test) as reflection, a way of expressing emotion, as a politeness marker, or as a turn-taking device, as a marker of suspense, or humour (e.g. Nakane, 2012). When the contextual communicative uses of the observable markers of fluency are taken into account at intermediate levels and higher, there is simply no one-to-one mapping with unobservable cognitive constructs.

There is therefore an interaction between speech phenomena, the intentionality of the speaker, the context, and the interpretation of the listener. Listeners can only understand if they are familiar with the entire context of the utterance, and the cultural constraints on production. This is the main reason why descriptors on a fluency rating scale usually refer both to the speaker's utterances, and the listener's interpretations (see Fulcher, 2003, for examples). Fluency is as much in the ear of the listener, as it is in the speech of the speaker. It is a function of the relationship between familiarity and context.

This applies equally to the role of pronunciation in the listener's perception of fluent speech. Language testers often make the assumption that pronunciation is a simple "on/off switch" for intelligibility (Fulcher, 2003: 25). But this is to make the same mistake committed by the cognitive scientists. 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." The reality is that pronunciation is variably problematic, depending on the familiarity of the listener with the L1 of the speaker. This is an important realization in the context of language assessment, where such familiarity becomes an important variable that impacts on scores being assigned to speakers. While this is of little concern to cognitive

scientists, for applied linguists and language testers the question of variable familiarity is a pressing matter.

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; Xi & Mollaun, 2009; Winke *et al.*, 2013). 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), living and working in the country where the L1 of test takers is spoken (Carey *et al.*, 2011), and prior personal study 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) labelled it 'interlanguage phonology familiarity.'

Gass and Varonis (1984) is the earliest study of familiarity. They argued that four different types of familiarity contribute to comprehension: familiarity with topic of discourse; familiarity with non-native speech in general; familiarity with a particular non-native accent; familiarity with a particular non-native speaker. Their study used 142 native speaking university students as participants who listened to recordings of four male speakers (Japanese-English speakers n = 2; Arabic-English speakers n = 2) completing three reading tasks: (1) reading a story; (2) reading a set of five 'related sentences' that pertained to the

story though were not included in the text; (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 either a reading of 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 ($p < .05$) 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 non-native 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 ($p < .0001$) when 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 non-native speech in general were found to contribute to comprehensibility of non-native speakers, though 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 speaker error

instances 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).

Though Gass and Varonis did include the story summary for listener participants, no analyses or discussion of the summaries were included in the final paper that could support the claim that the different types of familiarity they examined contribute to comprehension, which would of necessity include the notions of locutionary or illocutionary force. While we do not wish to argue against the possibility that familiarity may contribute to comprehension and determining 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) suggested, the terms 'intelligibility,' 'comprehension' and 'interpretability' should be discussed and defined to avoid any confusion since these terms

have been applied in various ways and at times interchangeably (334). The definition of intelligibility this research adheres to is Field's (2005). It argues that intelligibility is determined by how the phonological content of a speaker is recognized by the listener. This version of intelligibility 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 upon intelligibility aspect of fluency, which directly results in score variation (Derwing *et al.*, 2004).

Research Questions

In order to investigate the role of intelligibility as a critical component of fluency within an 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 pronunciaiton 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, and also measure intelligibility success-rates. As a result, little is known about why score differences occur. The methodology was therefore designed to look at these two facets concurrently, in order to further elucidate the relationship between listener and speaker.

Participants

Eighty-seven ESL/EFL teachers and/or graduate students enrolled in applied linguistics or TESOL programs were recruited via email to participate as volunteer rater participants. Most (n = 73) were L1 English speakers and fourteen were L2 speakers (see Table 1).

Table 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

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.

The Test

A three-part test was constructed to measure intelligibility success rates of raters, and to observe how they scored the 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 in a quiet room with the use of headphones.

Part one 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 4-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 studied the Japanese language for one or more years (n = 38).

Part two 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 requested 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 based on the TOEFL iBT rating scales used to measure delivery (ETS, 2007: 44), which contains the notion of "fluidity" (see Table 2). Each recording was approximately 18 seconds in length. Raters could start, stop or replay the recording at their discretion. No visuals were provided so that raters had no additional information about the speakers that may lead to inferences that may impact on scores (e.g. gender, age, L1, nationality) (see Rubin, 1992). There are a number of limitations in the methodology. Firstly, raters completed the same test. 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 BKB sentence lists (Bench *et al.*, 1979), which were originally designed to measure the listening capabilities of children with varying degrees of sensorineural hearing loss (see Appendix A). 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. The choice to model this test on a sensorineural hearing loss measure was due to the hypothesis that accent familiarity directly affects how speech is processed. Like Bench *et al*.'s original tests this test was designed to measure differences in speech perception and processing with gap-fill transcription tasks. In Bench *et al*.'s tests clarity of speech was 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 eight to fifteen (see table 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 by utilizing the JACET 8000, a corpus of the 8000 most frequently used English words by Japanese speakers of English. Lexical complexity was restricted to the 3,000 most frequently used words in order to eliminate need to provide explanations of word meaning of pronunciation to speaker participants. As a result, each speaker was left to pronounce each word in a sentence as they thought fit.

Table 2. Pronunciation score descriptors

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, though 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.

Table 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.

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

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 they are 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 though contextually complex or unpredictable (see Table 4).

They were also designed to feature aspects of Japanese-English phonology that are known to be problematic both in production for the speakers and 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).

Table 4. The test sentences

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.

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

Analyses

Facets 3.71, a Many Faceted Rasch Measurement (MFRM) software and SPSS (version 20) were used to analyse the test data. MFRM allows for multiple aspects, or facets, of a test to be examined together, and in the case of this study reveal a robust insight into raters' application of the rating scale and their abilities to transcribe utterances. Only data from the five L1 Japanese speakers was included in the MFRM analyses. This was designed to determine if rater accent familiarity differences resulted in significant score differences.

The pronunciation score and intelligibility success rates data were analysed separately in Facets due to the differences of tasks. Success was determined through accurate transcription per gapped item, and spelling errors were not penalized. Though Facets is capable of processing numerous types of data simultaneously, fit statistics for raters that serve as a type of quality control of measures (Green, 2013) were compromised when different tasks were analysed together. Linacre (personal communication) suggested that separate analyses were advisable.

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, 2012). The most informative and important piece of output from Facets analyses is the variable map. The variable map 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 in an equal-interval scale. Figure 1, which presents the Facets Variable Map for pronunciation scores, is separated into five vertical columns:

1. Column one displays the logit scale ranging from -7 to 2.
2. The second column displays the leniency of each rater from most (top) to least.
3. The third column shows the grouping facet revealing the severity and leniency of how each group scored the speakers.
4. Column four shows how each speaker performed with the most proficient at the top to the least proficient.
5. The fifth column shows the five-point rating scale used to score pronunciation. Each speaker participants' position in the fourth column is horizontal to their mean score on this rating scale.

Measr	+rater	+Familiarity Level	+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			

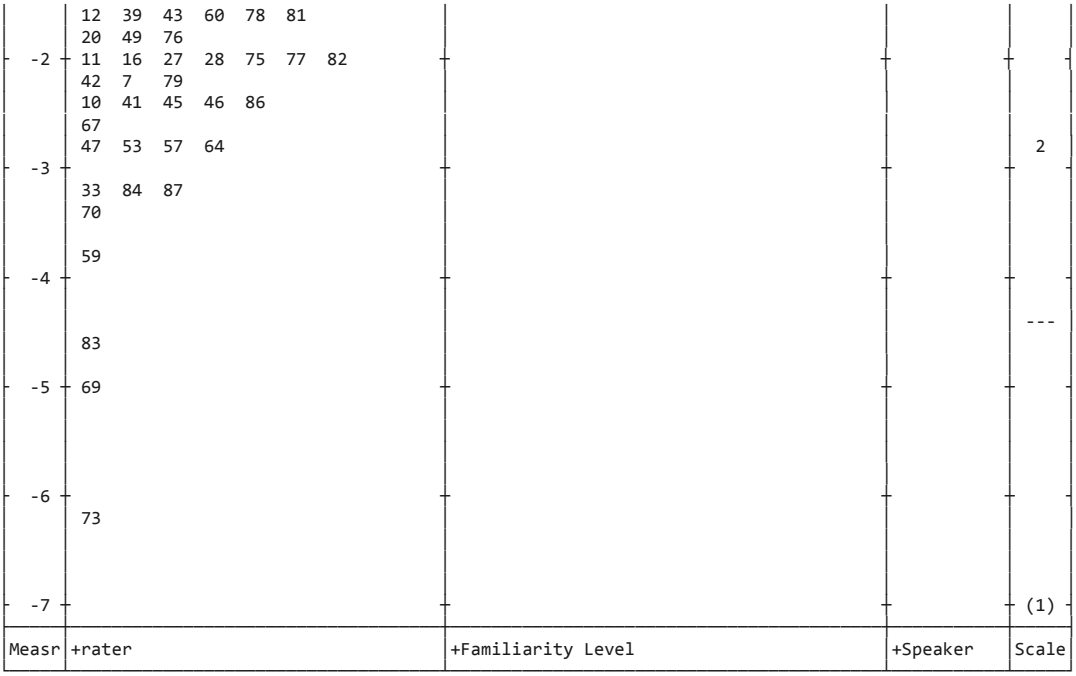

Figure 1. Facets Variable Map of Pronunciation Scores including Four Levels of Familiarity

The results revealed that all groups' In Fit and Out Fit statistics reflected acceptable values (0.5~1.5) (see Green, 2013: 219; and Table 5). The separation index reflects the difference between rater groups of 1.5 logits (see Table 6), particularly between the 'Very familiar' and 'No familiarity' groups. Reliability statistics are adequate.

Column three in Figure 1 shows that 'limited' and 'some' familiarity have nearly identical logit scores (0.03 and -0.09 respectively; also see Table 4); however, the 'very familiar' (0.43) and 'no familiarity' (-0.39) groups showed the greatest difference (.81). The results suggest four of the five speakers (all but Speaker B) would receive a one point higher score from the 'very familiar' raters than those with no familiarity (see Green, 2013, and McNamara, 1996, for more concerning Rasch for language testing).

Table 5. Pronunciation score Facets Rater Familiarity Level group measures

Familiarity Level	Measure	S.E.	InfitMS	InfitZ	OutfitMS	OutfitZ
No Familiarity	-0.38	0.22	0.91	-0.46	0.91	-0.46
Limited Familiarity	0.03	0.14	0.92	-0.66	0.96	-0.32
Some Familiarity	-0.09	0.38	0.71	-0.92	0.69	-0.97
Very Familiar	0.43	0.12	1.09	0.90	1.12	1.21

Model, Population: RMSE = .24, Adj (True) SD = .16, Separation = .68, Strata = 1.25, Reliability = .32.

Model, Sample: RMSE = .24, Adj (True) SD = .23, Separation = .98, Strata = 1.64, Reliability = .49.

Model, Fixed (all same): $\chi^2(3) = 12.3$, $p < .01$.

Model, Random (normal): $\chi^2(2) = 2.4$, $p = .30$.

Table 6. Pronunciation score Facets Rater separation and agreement measures

Model, Population: RMSE = .77, Adj (True) SD = 1.14, Separation = 1.49, Strata = 2.32, Reliability (not interrater) = .69.

Model, Sample: RMSE = .77, Adj (True) SD = 1.15, Separation = 1.50, Strata = 2.34, Reliability (not interrater) = .69.

Model, Fixed (all same): $\chi^2(86) = 246.1$, $p < .0001$.

Model, Random (normal): $\chi^2(85) = 72.2$, $p = .84$.

Interrater agreement opportunities = 6323, Exact agreements = 2399 or 37.9%, Expected = 2338.7 or 37.0%.

The Facets Variable Map for intelligibility scores is shown in Figure 2. The content of each column is as follows:

1. Column one displays the logit scale ranging from -4 to 6. With dichotomous data as in this intelligibility analysis the logit measures provide success probability calculations corresponding to item difficulty calibration and rater or rater group ability measurements.
2. How the individual raters performed in the intelligibility gap-fill exercises are shown in the second column. Raters' individual ability is reflected in their position on the map with the most able near the top of the column.
3. The third column reveals how rater groups performed completing the gap-fill exercises. As predicted, the 'Very Familiar' raters were the most successful completing the tasks and the 'No Familiarity' group had the least success.
4. The fourth column displays the items from most difficult (top) to least difficult (bottom). The items are identified first according to the speaker 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).

Measr	+rater Most Capable	+Familiarity Level Most Capable	+Item Most difficult
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		

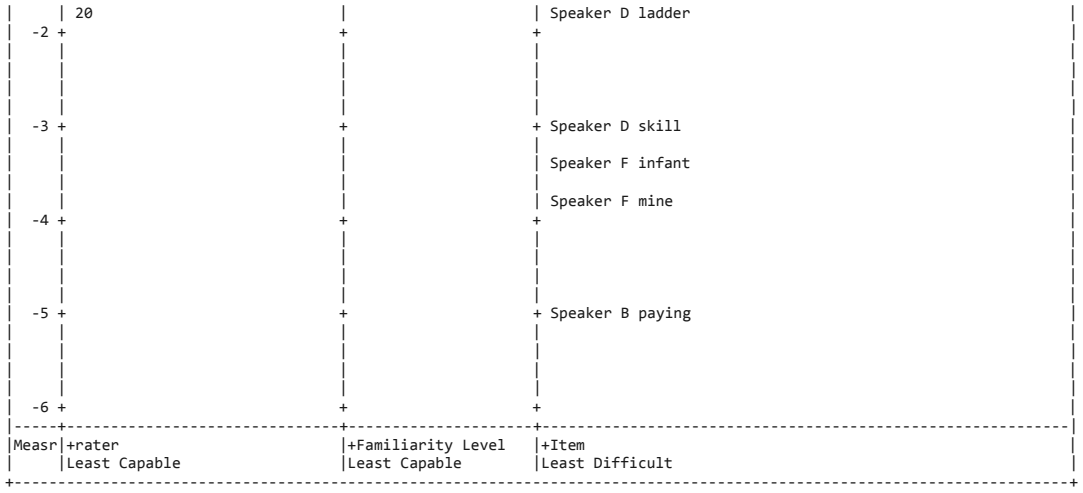

Figure 2. Facets Variable Map of Intelligibility Gap-fill Outcomes Including Four Levels of Familiarity

The MFRM results demonstrate that raters with more familiarity with Japanese-English experience greater intelligibility success-rates than raters with less familiarity. Raters very familiar with Japanese-English were 20% more successful in their ability to find the speakers intelligible than the raters with no familiarity (see Table 7).

Table 7. Facets Intelligibility Familiarity Level Measurements

Familiarity Level	Total Score	Total Count	Obsvd Average	Fair Average	Measure	S.E.	Infit MS	InfitZ	OutfitMS	OutfitZ
No Familiarity	156	312	0.50	0.51	-0.41	0.16	0.99	0.00	1.04	1.02
Limited Familiarity	435	768	0.57	0.58	-0.13	0.10	0.91	-1.70	0.91	-0.40
Some Familiarity	59	96	0.61	0.63	0.07	0.30	0.84	-1.00	0.56	-0.80
Very Familiar	634	912	0.70	0.71	0.46	0.10	1.08	1.40	1.29	1.20
Model, Population: RMSE .19 Adj (True) S.D. .26 Separation 1.39 Strata 2.19 Reliability .66										
Model, Sample: RMSE .19 Adj (True) S.D. .32 Separation 1.71 Strata 2.61 Reliability .74										
Model, Fixed (all same): $\chi^2(3) = 26.6$, $p < .00$.										
Model, Random (normal) $\chi^2(2) = 2.7$ $p = .26$.										

Kendall's Tau-b tests (nonparametric rank correlation) were also conducted using SPSS to determine if level of familiarity and intelligibility success for each item were statistically

dependent. The results revealed additional details concerning how the rater groups coped with each item. Familiarity was positively linked to the intelligibility success-rates of all items; 10 of the 24 items revealed significant intelligibility success-rate differences (see Table 8). This supports the interpretation that a positive correlation exists between familiarity level and intelligibility success.

Table 8. Kendal's tau-b results of significant intelligibility items

Item	T^b
Speaker B play	.433***
Speaker B rooms	.346***
Speaker D ladder	.548***
Speaker D across	.449***
Speaker D door	.278*
Speaker E union	.242*
Speaker E sensed	.255**
Speaker F mine	.198*
Speaker F matches	.394***
Speaker F lie	.255**

Note. ***$p \leq .001$, **$p \leq .01$, *$p \leq .05$.

Conclusion

We have argued that a linguistic 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 flow chart. It therefore represents a "software-solution" to the problem of mind and language. Taken literally, within a strong cognitive approach 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) whichever 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 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. Firstly, by using a measure of familiarity such as the one used in this study to match raters with test takers.

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 maximise 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

Alderson, C. & Bachman, L. (2004) Foreword in Luoma, S. *Assessing Speaking*. Cambridge: Cambridge University Press.

Bachman, L.F. and Savignon, S.J. (1986) The evaluation of communicative language proficiency: A critique of the ACTFL Oral Interview. *Modern Language Journal* 70, 380-390.

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. (1990) *Principles of Language Testing*. Oxford. Basil Blackwell.

Davies, A. (2003) *The Native Speaker: Myth and Reality.* Multilingual Matters Ltd. UK.

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. M., Rossiter, M. J. and Munro, M. J. (2002) Teaching native speakers to listen to foreign-accented speech. *Journal of Multilingual and Multicultural Development 23* (4), 245-259.

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

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. (1996) Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing* 13 (2), 208–238.

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.

Green, R. (2013) *Statistical analyses for language testers*. London: Palgrave Macmillan.

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.

Kachru, B. B. (1997) World Englishes and English-using communities. *Annual Review of Applied Linguistics* 17, 66-87.

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

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.

Lantolf, J.P. and Frawley, W. (1985) Oral proficiency testing: a critical analysis. *The Modern Language Journal* 69, 337-345.

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.

McNamara, T. (1996) *Measuring Second Language Performance*. London: Longman.

Nakane, I. (2012) Silence. In C. B. Paulston, S. F. Kiesling and E. S.Rangel (eds) *The Handbook of Intercultural Discourse and Communication* (pp. 158–179). London: Blackwell Publishing Ltd.

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

Rubin, D. L. (1992) Nonlanguage factors affecting undergraduates' judgments of non-native 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. 199.

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 [TM] Speaking Section and What Kind of Training Helps?* TOEFL iBT [TM] Research Report. RR-09-31. Princeton, NJ: Educational Testing Service.