The Development of a Phonological Awareness Test for Japanese Learners of English

Paul Joyce

Abstract

Phonological awareness ability has been found to be of importance to L2 proficiency. However, the lack of a validated means of measuring phonological awareness ability for Japanese learners of English has been a gap in the field of L2 phonology. This paper describes the development of a test to address this need. The research instrument employed a minimal pair phonemic discrimination design that focused on six contrasts that are considered difficult for Japanese learners to differentiate. To ensure the trait purity of the test, the influence of word stress knowledge, short-term memory and lexical knowledge were controlled. After piloting, a revised version of the test was administered to 443 Japanese university English majors who ranged from a false beginner to upper intermediate proficiency level. The data was analyzed using both the Classical Testing Theory and Item Response Theory approaches to test evaluation. Since the results showed that all of the test items fitted the Rasch model, the test is considered a valid measure of L2 phonological awareness knowledge for the target population.

Keywords: Classical Testing Theory, Item Response Theory, Japanese learners, phonological awareness, test development

Introduction

Phonological awareness refers to an individual's ability to clearly perceive and produce the sound units of a language. For successful spoken language comprehension, it is widely acknowledged that word recognition must occur accurately (Gillon, 2012; Stone, Silliman, Ehren, & Apel, 2004). In order to enable this to occur, phonological awareness is regarded as one of the most important elements of linguistic competence as it plays an important role in supporting communicative competence (Buck, 2001; Rost, 2002).

A number of language processing models have conceptualized the role of

phonological awareness in first language (L1) listening. While these theories vary in design, it is agreed that to identify words, the listener integrates information from a wide range of sources, and that one of these is phonemic recognition. In the most extreme case, as part of the Cohort model (Marslen-Wilson, 1984), it is proposed that lexical recognition proceeds sequentially, phoneme by phoneme. That is, the acoustic signal is first decoded into representatives of prototypical phonological categories. From this point, through the correct identification of phonemes, candidate words are eliminated, leaving the intended lexical item. There are a number of alternative theories of word recognition, such as the Fuzzy logic model (Massaro, 1994), the TRACE model (McClelland & Elman, 1986), and the Logogen model (Morton, 1969). Although these alternative theories posit a greater role for higher-order units in word recognition, there is a general concurrence on the importance of the phoneme. This emphasis on phonemic awareness is supported by L1 empirical research. Studies in acoustic phonetics suggest that children segment spoken auditory input into discrete phonemes (Gleitman & Wanner, 1982). The ability to discriminate between contrasting sounds has also been found to correlate significantly with measures of children's L1 language comprehension (Marquart & Saxman, 1972). In the case of adults, after collecting over a thousand examples of errors in the perception of speech, it was found that around 27 percent of these were accounted for by segmental phonemic misperceptions (Bond & Garnes, 1980). Thus, for L1 English listening, from both theoretical and empirical perspectives, there is evidence to suggest that the ability to discriminate between the constituent sound contrasts of the language is related to the skill of word recognition.

When considering the role of phonemic awareness in second language (L2) listening, it is important to make a distinction between L1 and L2 linguistic processing. It has been found that when there are meaningful phonetic distinctions in the L2 that are not employed contrastively in the learner's own language, listeners are subject to strong L1 interference (Byrnes, 1984; Major, 2008; Mochizuki, 1981; Mora, 2005; Yamada, Tohkura & Kobayashi, 1997). Thus, unless learners can accurately perceive non-native sound contrasts, phonologically distinct L2 sounds are filtered into existing L1 categories. In this case, since the L2 sounds are stripped of an aspect of their phonological meaning, the learner's listening comprehension ability is impeded.

The importance of phonological awareness for L2 development has been widely reported. Mack (1988) explored the influence of L1 phonological interference on L2

listening comprehension when researching the intelligibility of natural versus computer-generated speech among German native and non-native participants. It was found that over 70 percent of the L2 learners' errors in transcribing natural speech were phonemic in nature. Of these, there were discovered to be a large but unspecified number of errors related to German to English transfer. In one of the few other studies in this area, Pemberton (2003) sought to discover why Cantonese learners were only able to recognize a low proportion of high frequency words from radio news broadcasts. It was observed that when a word contained a phonological mismatch between Cantonese and English, it was transcribed far less accurately when there was not such a mismatch. This result provided further evidence for the influence of learners' L1 phonological background on their L2 listening comprehension. Research findings have also shown that children's ability to accurately perceive and manipulate L2 phonological forms is related to their capacity to read and write in English (Gottardo, 2002; Sparks & Ganschow, 1993; Stanovich, 1988). There have also been studies showing the importance of age (e.g., Flege, Munro, & MacKay, 1995), experience with the foreign language (e.g., Flege, Bohn & Jang, 1997; Levy & Strange, 2008), and type of training (e.g., Giambo & McKinney, 2004; Mora, 2005) on English phonological awareness development.

As discussed, there is reason to believe that L2 phonological awareness has an important role in L2 language learning. However, since the differences in the participants' first and second languages affect the findings of phonological awareness research (Geva & Siegel, 2000), the results from such studies vary depending on the two languages that are being compared. When considering the large phonological differences between Japanese (L1) and English (L2), it becomes clear that research involving other language pairs, especially those between two alphabetic languages, may not be readily generalisable to Japanese learners of English. The research that has been conducted on the phonological awareness knowledge of Japanese EFL learners has employed a range of different research instruments. Unfortunately, the tests used in such studies have frequently not been provided (e.g., Miyawaki, Strange, Verbrugge, Liberman, Jenkins & Fujimura, 1975; Yamada & Tohkura, 1992), which constrains follow-up studies. In addition, basic psychometric data about the tests used in such research, such as their reliability, has often not been supplied (e.g., Flege, Takagi & Mann, 1996; MacKain, Best & Strange, 1981), which means that it is unclear how much random error is contained in the scores. Furthermore, often the tests used have focused

on single phonological contrasts between the languages, primarily, /r/ and /l/ (e.g., Mochizuki, 1981; Takagi & Mann, 1995). Such measures are suitable for the narrowly focused purpose for which they were intended. However, to address wider research questions relating to the relationship between L2 phonological awareness and L2 listening or reading, or to provide a more holistic assessment of a Japanese EFL learner's phonology awareness, a more broadly based test would be of value.

This paper will seek to help address this need by introducing some of the issues concerned with testing phonological awareness and describing the process of developing a discrete-point phonological awareness test. With this purpose in mind, there will be a discussion of the test development methodology used and an account of the results from the two stages of test piloting.

Methodology

The principles of test development

Test design. There are a large number of test formats for assessing L2 phonological awareness. However, it is most commonly operationalised through a minimal pair (AX) phonemic discrimination test (e.g., Dreiier & Larkins, 1972; Mochizuki, 1981; Mora, 2005). Minimal pairs are words that differ by one phoneme only. When engaging in AX tasks, participants hear sets of word pairs that contain the target phonemic contrasts. For instance: plod...prod

After hearing each word pair, the test takers are required to indicate whether they have heard the same word repeated or encountered two different lexical items. To ensure the test is valid, typically 25 percent of the pairs are distracters that consist of the same word uttered twice (e.g., Wepman, 1975; Yamada, Tohkura, & Kobayashi, 1997; Mora, 2005). These items do not form part of the scored section of the test.

The AX task approach to evaluating phonological knowledge was adopted as it minimizes the influence of unrelated variables that could contaminate the test scores. Most notably, by not presenting the assessed words within sentences, contextual clues were not introduced that could assist in phonemic perception. To further improve the trait purity of the test, a number of further measures were taken. Firstly, to avoid conflating the target construct with word stress knowledge, it was decided that the target lexical items would be limited to monosyllabic words. When the test was recorded, approximately the same pitch and intensity was placed on each of the lexical

stimuli. Secondly, as short-term memory capacity has been found to influence phonemic recognition (Mochizuki, 1981), it was important that the selected task did not place any strain on the learners' short-term memory (STM). To alleviate the effect of memory, the lexical items containing the assessed phonemic contrast were presented with only a short time lapse between them. Lastly, since it has been observed that word familiarity significantly correlates with phoneme identification (Yamada, Tohkura, & Kobayashi, 1997; Mora, 2005), it was necessary to control for lexical knowledge. Therefore, the lexical pairs largely consisted of very low frequency words that were considered to be largely unknown to the participants. The likelihood of selecting such vocabulary was considered high since experienced language teachers have been shown to be capable of predicting with which words students are unfamiliar (Brutten, 1981).

As discussed, the AX phonemic discrimination test format held a number of advantages. However, since it is binary in nature, to achieve sufficiently high reliability, a large number of items were required. Therefore, both to determine the number and content of the test items, piloting was essential.

Test content. As discussed in the Introduction, when there are meaningful phonological distinctions in the L2 that are absent from the listener's native language, there has been found to be strong L1 interference (Mochizuki, 1981; Byrnes, 1984; Yamada, Tohkura & Kobayashi, 1997; Mora, 2005). In the case of Japanese, L1 phonological transfer impedes the mapping of a number of phonetic categories that exist in English. For the purposes of this study, there was a focus on six contrasts deemed particularly difficult for Japanese learners to differentiate (Kenworthy, 1987). These contrasts are summarized in Table 1 below:

Table 1.

Phonemic Contrasts Present in English but Absent from Japanese

Phonemic Contrasts	Example
/r/–/l/	rink – link
/h/–/f/	hall - fall
/s/–/θ/	sink - think
/b/–/v/	berry - very
/z/–/s/ (in a final position)	peers - pierce
/ʌ/–/ɔ:/	hush - harsh

Statistical Analyses. In constructing and validating the test of phonological awareness, two measurement models were used: Classical Testing Theory (CTT) and Item Response Theory (IRT).

Classical Testing Theory. CTT analysis primarily addresses the difficulty, discriminability, and reliability of both individual test items and complete tests. The difficulty of a test question, otherwise known as its Item Facility (IF), relates to the proportion of test takers who correctly complete the item. It has been found that the greater the spread of IF values for a particular test, the lesser the test score dispersion (Ebel, 1979). Therefore, in accordance with a widely recommended guideline (see Henning, 1987; Tuckman, 1972), the number of items with an IF value of between .33 and .67 was closely monitored.

Item discrimination (ID) concerns how well a test item differentiates between the stronger and weaker examinees. And, a point-biserial correlation of .25 or above is widely regarded as acceptable (Henning, 1987). Test reliability simply refers to how consistently a scale measures a target construct. For the purposes of educational research, a coefficient in excess of .70 is commonly cited as acceptable (e.g., Nunnally, 1978; Kline, 1999). However, it was hoped that internal consistency would reach .80. CTT provides a useful basis to evaluate the psychometric characteristics of a test. Nevertheless, this approach to testing is sample-dependent. In other words, it is not possible to compare individuals across different tests and items across different groups of test takers. A statistical model that overcomes this limitation is IRT.

Item Response Theory. There are a number of different IRT models. For the purposes of this research, the one-parameter or Rasch model was used and the analysis was undertaken using Quest (Adams & Khoo, 1993). An important aspect of the one-parameter IRT models is that they are sample independent. Furthermore, both an item's difficulty and a person's ability are placed onto the same continuum. The scale usually has a mean of 0.0 and a standard deviation (SD) of 1.0. Consequently, the vast majority of difficulty estimates typically range between +3.0 and -3.0. A value of +3.0 corresponds to a person with a high ability or an item that is very difficult. Conversely, a figure of -3.0 pertains to a person with a low ability level or a very easy item. To most effectively discriminate between the test takers, the item difficulty values should mirror the person ability estimates. As it was expected that most test takers would be of mid-range ability,

there was a corresponding requirement for test items of comparable difficulty. Thus, both from the standpoint of the CTT and IRT approaches, developing a large proportion of test items of mid-range difficulty was considered important. However, to match person ability estimates along the entire scale, it was also important that there were test items throughout the difficulty scale.

IRT evaluates how well the observed data fits the statistical model. An infit mean-square fit of less than 1.0 indicates that the data has a better than expected correspondence to the model. However, extreme over-fit values suggest the presence of redundant items. Conversely, an infit figure of greater than 1.0 signifies that the data has a worse than expected match to the model. An extreme under-fit value is indicative of an unusual or inappropriate response pattern, and suggests that the item is mis-performing for the target population. As mentioned, test questions with extreme infit values are considered problematic. Therefore, in accordance with recommended practice (McNamara, 1996), the items with an infit mean square of less than .75 or greater than 1.3 were excluded from the later research instruments.

As previously mentioned, an important advantage of IRT over classical analysis is that it allows items to be compared across different tests. This comparison is enabled through common anchor items that are administered as part of each test form. For example, the anchor item parameter values from a first test can be used to calibrate the anchor items in secondary tests. The difference in the difficulty values of the two sets of anchor items is used to calibrate the statistical values of the remaining items in the secondary test.

General Procedure

As discussed in the Test Design section, L2 phonological awareness was operationalised through an AX auditory discrimination task. Each of the three research instruments contained 80 questions. However, only the 60 pairs that included different words were scored. One point was awarded for each correct answer. The instructions were presented in both aural and written form, and an example item was provided. To forewarn the participants of the onset of the assessed material, the question number immediately preceded each word pair. The students indicated whether each particular word pair was identical or different by shading the appropriate bubble on a mark card. The listening material was produced and delivered through high quality audio

equipment. The tests took around twelve minutes to administer.

Participants

The data was gathered in Japan at a university that specializes in foreign languages. The participants were all native Japanese L1 speakers, who were enrolled as full-time English language major undergraduates. In terms of proficiency, the learners were ranged from a false beginner to an upper intermediate level. In terms of performance on the paper and pencil TOEFL, the participants' scores ranged from approximately 357 to 513 (see Bonk, 2001), which converts to scores of between 70 and 180 on the TOEFL Computer-Based Test. As the selection of the participants was determined by the cooperation of their EFL teachers, a convenience sample was used.

Results

Test Administration One

Procedure. The primary purpose of the first test administration was to pilot a sufficiently large number of items to generate a sizeable bank of psychometrically high quality items for the second test administration. Although this suggests the use of a long test, such an approach would risk test fatigue and a subsequent decrease in the reliability of the item data. Therefore, three test versions were produced and the items from these tests were placed upon the same scale through the deployment of 19 common anchor items. A total of 132 students participated in the pilot study.

Results and discussion. Perhaps the most notable aspect of the results related to the difficulty of the three tests. Considering the binary nature of the research instrument, the mean average scores were fairly low. Specifically, Form One had a grand mean of 39.52 (65.86%), Form Two 35.77 (59.61%), and Form Three 34.16 (56.93%).

Table 2.

Descriptive Statistics for Test Administration One

	n	*k*	*M*	*SD*	*min.*	*max.*	*rel.* (α)
Form One	50	60	39.52	5.29	30	48	.61
Form Two	39	60	35.77	6.97	23	50	.78
Form Three	43	60	34.16	5.87	24	49	.68

Given the difficulty of the test, it is unsurprising that a relatively large proportion of the items fell within the target IF range of .33 to .66. On Form One of the test, 43.33% of items met this criterion, 48.33% for Form Two, and 41.66% for Form Three. Nevertheless, owing to the binary nature of the test, when the three sets of results were placed onto a common logit scale, the person ability estimates were in excess of the item difficulty estimates. As shown in Table 3, Forms One, Two and Three yielded person estimates of between .15 and .87, while the item estimates were between .00 and -.25.

Table 3.

Inferential Statistics for Test Administration One

	Person Estimates			Item Estimates			
	M	*SD*	*Rel.*	*M*	*SD*	*Rel.*	Misfitting Items (< .75, > 1.3)
Form One	.87	.52	.62	.00	1.44	1.00	0
Form Two	.42	.69	.79	-.18	1.38	1.07	3
Form Three	.15	.57	.70	-.25	1.35	1.07	5

Partially as a consequence of the test difficulty, the instruments yielded promising reliability values. With alpha figures of .61 (Form One), .78 (Form Two), and .68 (Form Three), none of the measures attained the target .8 internal consistency target. Nevertheless, since the test score consistency was reasonably high, a large number of items fulfilled the .25 ID goal. In the case of Form One, there were 21 (35%) such items, for Form Two 29 (48%), and for Form Three, 26 (43%). Yet, since there were 19 anchor items that were contained in all three test versions, there were actually only 58 different test items that met the ID criterion. As the most effective of these test items would be combined for the second test administration, it was expected that the test discriminability would rise further. Lastly, three items from Form Two, and five items from Form Three recorded extreme Rasch fit values. In all of these cases, the items underfitted the model. Owing to their poor correspondence with the IRT model, these test items were omitted from the second test administration.

Test Administration Two

Procedure. A revised version of the phonological awareness test was used to

determine whether the items were functioning as anticipated. In total, there were 443 learners who participated in the second pilot study. The test instructions and a transcription of the test materials are available in Appendix A and Appendix B.

Results and discussion. The descriptive results showed that the internal consistency of the test scores was sufficiently high (Cronbach's alpha = .80). The improvement was mainly due to the large proportion (61.66%) of test items that met or exceeded the .25 ID target. The improvement in the reliability of the test over those used in Test Administration One was due to the selection of the most psychometrically robust items from the three pilot tests, and the increased sample size.

Table 3.

Descriptive Statistics for Test Administration Two

k	*M*	*SD*	*min.*	*max.*	*rel.* (α)
60	39.26 (65%)	7.48	20.00	59.00	.80

The mean average score was a relatively high 39.26 (65.43%). Nevertheless, since the research instrument was comprised of binary items, the results showed that many of the participants were yet to develop an awareness of the full range of target contrasts. Overall, the test scores ranged between 20 and 57. While the minimum score points to the difficulty that some of the participants found with the test material, the highest score reveals that there were learners who were able to accurately distinguish between the phonemic contrasts used in the test. As was the case for Test Administration One, a large proportion of the test items (55.00%) fell within the target IF range.

Table 4.

Inferential Statistics for Test Administration Two

Person Estimates			Item Estimates			
M	*SD*	*Rel.*	*M*	*SD*	*Rel.*	Misfitting Items (< .75, > 1.3)
.90	.81	.83	.00	1.05	.98	0

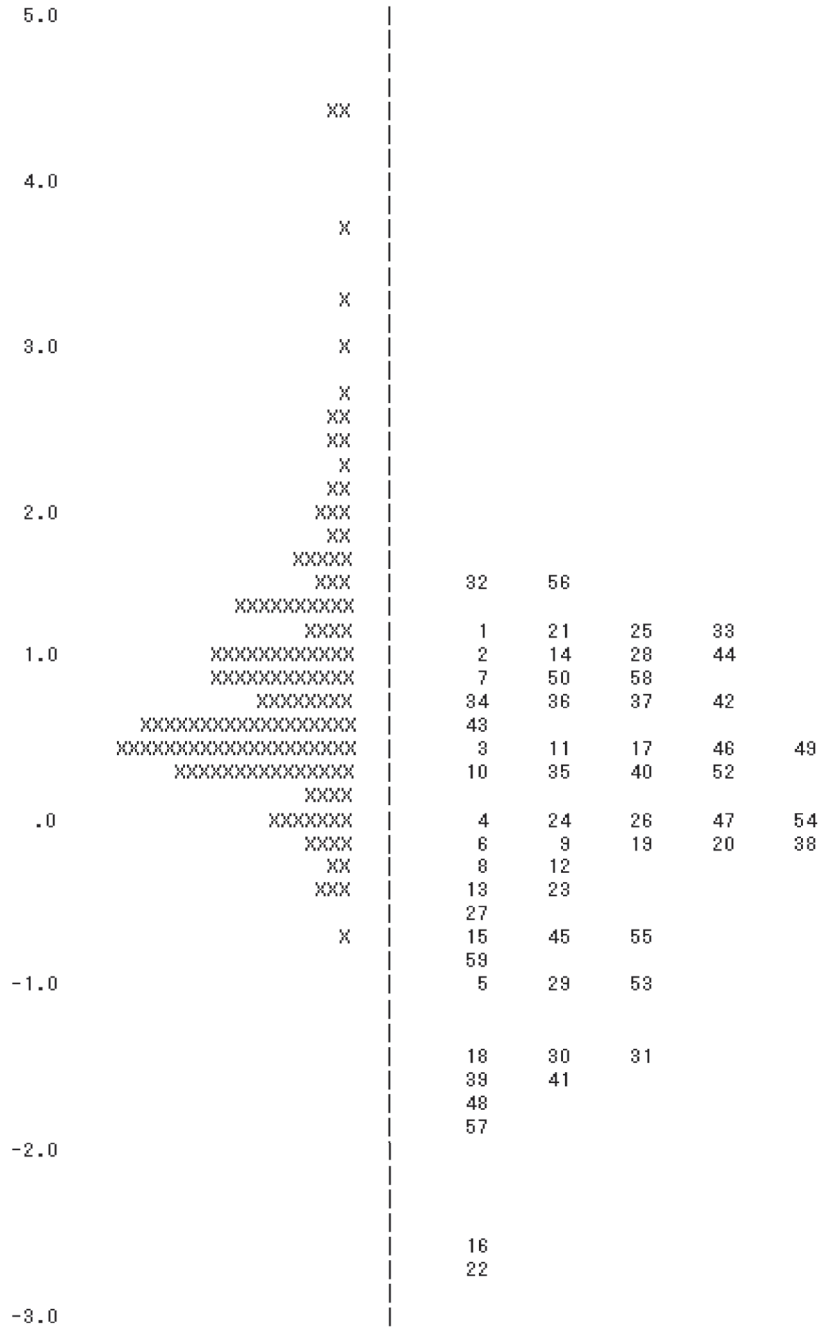
The IRT results were found to be consistent with the descriptive findings. That is, given that the item difficulty estimates fell between -2.98 and 1.48 logits with a mean of

.00 (SD = 1.05), there were some items at the easier end of the difficulty spectrum, but few really challenging ones. In contrast, the case ability estimates ranged between -.80 and 3.5 with a mean of .90 (SD = .81). Thus, while there were no participants of a really low ability level, there were some who performed very well. Lastly, the item infit values for the items all fell within the acceptable range.

The IRT results were found to be consistent with the descriptive findings. The IRT findings are displayed the form of a Wright Map (see Figure 1). The left side of the map shows the candidates with each X representing three test taker, while the right side of the chart displays the test items. The ability of the candidates and difficulty of the items are presented vertically. The candidates at the top of the map scored highest, and the higher the item on the scale, the greater its difficulty. The Item Infit Mean Square values (.92 to 1.06) for all of the questions fell within the acceptable range.

As shown by the map, the items were on average found to be less difficult than the persons were able. In terms of the items, the difficulty estimates fell between -2.98 and 1.48 logits with a mean of .00 (SD = 1.05) and it is noticeable that although there were some items at the easier end of the difficulty spectrum, there were few really challenging questions. On the other hand, while there were no participants of a really low ability level, there were some who performed very well. Aside from the ability of the test to discriminate between the higher ability candidates, an area for improvement relates to the gaps between the item cluster containing questions 32 and 1, and 10 and 4. The development of items to fill these gaps is an area for further research.

Figure 1. Wright Map for Test Administration Two



Each X represents three participants

Conclusion

This paper has summarized the development and initial validation of an L2 phonological awareness test. As has been discussed, the first phase of test construction focused upon the design and content of the measure. This was followed by two stages of piloting and the analysis of the results through both Classical Testing Theory and Item Response Theory to ensure that the test scores derived from the research instrument were reliable. It is hoped that the test that was produced will provide a basis for those wishing to explore the field of L2 phonology further.

In terms of the results themselves, it was notable how difficult the participants found the task. As previously discussed, the research instrument consisted of a series of binary-choice items that required participants to differentiate between two monosyllabic words. Since the participants were majoring in English language and had at least seven years of English language education, they might have been expected to score more highly. However, despite the binary nature of the research instrument, only 65 percent of the participants' responses were found to be accurate. Therefore, consistent with previous studies (e.g., Flege, Takagi & Mann, 1996; Mochizuki, 1981; Yamada, Tohkura and Kobayashi, 1997), the study has confirmed that even Japanese English language majors have great difficulty distinguishing naturally produced phonemes that are not used contrastively in their own language. This result may reflect the emphasis in Japanese EFL education upon reading and grammar over the development of L2 listening skills.

On the other hand, as displayed by the Wright Map, while the instrument was capable of discriminating effectively between the majority of test takers, it would benefit from the introduction of more challenging items to help distinguish between the candidates who achieved higher ability estimates. However, given the discrete nature of the test material, it is difficult to increase the difficulty of the items without sacrificing the naturalness of the spoken language used and thereby endangering the validity of the test. Nevertheless, there remains the option of adjusting the selected test methodology. Since the AX task is binary in nature, the participants had a fifty percent chance of guessing the correct answer. To reduce the influence of guessing, it is possible to add additional answer choices to the task (see Harris, 1969; Lado, 1961). Although such tests have been used far less frequently than the AX methodology, they have been shown to increase item difficulty. For instance, when Mochizuki (1981) asked candidates

to identify the odd word out from three answer choices, accuracy was found to decrease from 85% to 58%. It should be noted that the researcher ascribed the difference in performance to additional memory burden. However, there is room to explore this issue further. In addition, there is also scope to investigate the background of those students who performed very well on the test. Variables of interest include the proficiency of the students, time spent abroad, and age of acquisition.

The results from this study need to be viewed in light of its limitations. The research was conducted with a relatively homogenous group of participants; eighteen to twenty-two year old Japanese university students, who ranged from a false beginner to upper intermediate proficiency level. The uniformity of the sample limits the range of participants with which the test can reliably be used. Thus, before drawing any conclusions about the appropriateness of the test for a more diverse population, the materials should be carefully piloted.

References

Adams, R. J., & Khoo, S. T. (1993). Quest: the interactive test analysis system (Version 2.1) [computer program]. Camberwell, Australia: Australian Council for Educational Research.

Bond, Z. S., & Garnes, S. (1980). Misperceptions in fluent speech. In R. A. Cole (Ed.), *Perception and production of fluent speech*. (pp. 115-132). Hillsdale, NJ: Lawrence Erlbaum Associates.

Brutten, S. R. (1981). An analysis of student and teacher indications of vocabulary difficulty. *RELC Journal*, *12*(1), 66-71.

Buck, G. (2001). *Assessing Listening*. Cambridge: Cambridge University Press.

Byrnes, H. (1984). The role of listening comprehension: A theoretical base. *Foreign Language Annals*, *17*, 317-329.

Dreiier, B., & Larkins, J. (1972). Non-semantic auditory discrimination: Foundation for second language learning. *Modern Language Journal*, *56*, 227-230.

Ebel, R. L. (1979). *Essentials of educational measurement* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Flege, J. E., Bohn, O. -S., & Jang, S. (1997). Effects of experience on non-native speakers'

production and perception of English vowels. *Journal of Phonetics, 25*, 437-470.

Flege, J. E., Munro, M. J., & MacKay, I. R. A. (1995). The effect of age of second language learning on the production of English consonants. *Speech Communication, 16*, 1-26.

Flege, J. E., Takagi, N., & Mann, V. (1996). Lexical familiarity and English language experience affect Japanese adults' perception of /r/ and /l/. *Journal of the Acoustical Society of America, 99*(2), 1161–1173.

Geva, E., & Siegel, L. S. (2000). Orthographic and cognitive factors in the concurrent development of basic reading skills in two languages. *Reading and Writing: An Interdisciplinary Journal, 12*, 1-30.

Giambo, D. A., & McKinney, J. D. (2004). The effects of phonological awareness intervention on oral English proficiency of Spanish-speaking kindergarten children. *TESOL Quarterly, 38*(1), 95-117.

Gillon, G. T. (2012). *Phonological Awareness: From Research to Practice*. New York: Guildford Press.

Gleitman, L. R. & Wanner, E. (1982). Language Acquisition: The state of the art. In E. Wanner & L. R. Gleitman (Eds.) (pp. 3-48). *Language acquisition: The state of the art*. Cambridge: Cambridge University Press.

Gottardo, A. (2002). The relationship between language and reading skills in bi-lingual Spanish-English speakers. *Topics in Language Disorders, 22*, 46-70.

Harris, D. P. (1969). *Testing English as a Second Language*. New York: McGraw-Hill.

Henning, G. (1987). *A guide to language testing: development, evaluation, research*. Cambridge, MA: Newbury House.

Kenworthy, J. (1987). *Teaching English pronunciation*. London: Longman.

Kline, P. (1999). Handbook of Psychological Testing (2nd ed.). Routledge: London.

Lado, R. (1961). *Language testing: The construction and use of foreign language tests*. London: Longman.

Levy E. S., & Strange, W. (2008). Perception of French vowels by American English adults with and without French language experience. *Journal of Phonetics*, 36, 141-157.

Mack, M. (1988). Sentence processing by non-native speaker of English: evidence from the perception of natural and computer-generated anomalous L2 sentences. *Journal of Neurolinguistics, 3*, 293-316.

MacKain, K. S., Best, C. T. & Strange, W. (1981). Categorical perception of English /r/ and /l/ by Japanese bilinguals. *Applied Psycholinguistics, 2*, 368-390.

Major, R. C. (2008). Transfer in second language phonology. In J. G. Hansen Edwards & M. L. Zampini (Eds.) (pp. 63-94). *Phonology and second language acquisition*. Amsterdam: John Benjamins

Marquart, T. P. & Saxman, J. H. (1972). Language comprehension and auditory discrimination in articulation deficient kindergarten children. *Journal of Speech and Hearing Research, 15*, 382-389.

Marslen-Wilson, W. D. (1984). Function and process in spoken word recognition. In H. Bouma & D. Bouwhis (Eds.), *Attention and performance X: Control of language processes*. (pp. 125-150). Hillsdale, NJ: Erlbaum.

Massaro, D. (1994). Psychological aspects of speech perception. *Handbook of psycholinguistics*. New York: Academic Press.

McClelland, J. & Elman, J. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18*, 1-86.

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

Miyawaki, K., Strange, W., Verbrugge, R., Liberman, A. M., Jenkins, J. J., & Fujimura, O. (1975). An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English. *Perception & Psychophysics, 18*, 331-340.

Mochizuki, M. (1981). The identification of /r/ and /l/ in natural and synthesized speech. *Journal of Phonetics, 9*, 283-303.

Mora, J. C. (2005). Lexical knowledge effects on the discrimination of non-native phonemic contrasts in words and non-words by Catalan/Spanish bilingual learners of English. *Proceedings of the ISCA Workshop on Plasticity in Speech Perception*. (pp. 43-46). UCL, London, UK.

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review, 76*, 165-178.

Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw-Hill.

Pemberton, R. (2003). *Spoken word recognition and L2 listening performance; An investigation of the ability of Hong Kong learners to recognise the most frequent words of English when listening to news broadcasts*. (Unpublished doctoral thesis). University of Wales, Swansea.

Rost, M. (2002). *Teaching and Researching Listening*. Applied Linguistics in Action

Series. Pearson Education Press.

Sparks, R., & Ganschow, L. (1993). Searching for the cognitive focus of foreign language learning difficulties: Linking first and second language learning. *Modern Language Journal, 77*, 289-302.

Stanovich, K. (1988). The right and wrong places to look for the cognitive focus of reading disability. *Annals of Dyslexia, 38*, 154-177.

Stone, C. A., Silliman, E. R., Ehren, B. J., & Apel, K. (2004). *Handbook of language & literacy: Development and disorders*. New York: Guilford Press.

Takagi, N., & Mann, V. (1995). The limits of extended naturalistic exposure on the perceptual mastery of English /r/ and /l/ by adult Japanese learners of English. *Applied Psycholinguistics, 16*, 379-405.

Tuckman, B. W. (1972). *Conducting Educational Research* (2nd ed.). NY: Harcourt Brace Jovanovich.

Wepman, J. M. (1975). Auditory perception and imperception. In W. M. Cruickshank & D. P. Hallahan (Eds.), *Perceptual and learning disabilities in children: Research and theory*, Vol. 2, 259-293. Syracuse: Syracuse University Press.

Yamada, R. A. & Tohkura, Y. (1992). The Effects of Experimental Variables on the Perception of American English /r/ and /l/ by Japanese Listeners. *Perception and Psychophysics, 52*(4), 376–392.

Yamada, R. A. Tohkura, Y. & Kobayashi, N. (1997). Effect of word familiarity on non-native phoneme perception: Identification of English /r/, /l/ and /w/ by native speakers of Japanese. In A. James, & J. Leather, (Eds.), *Second Language Speech. Structure and Process*. (pp. 103-117). Berlin: Mouton de Gruyter.

Appendix

Appendix A: Test Directions

You are going to hear lots of word pairs. Sometimes the two words will be the same and sometimes they will be different. If you think the word pairs are the same, mark "*a*" on your mark sheet. If you think they are different words mark "*b*". Here is an example:

The two words were different. To answer this question correctly, you needed to choose "*b*". For the rest of the questions, mark "a" if you think the words are the same, and "b" if you think they are different.

Appendix B: Test Administration Two

1.	verb	verve	**28.**	curb	curve	**55.**	sluice	sleuth
2.	gloat	groat	**29.**	plod	prod	**56.**	dace	dace
3.	biz	viz	**30.**	bland	bland	**57.**	lump	rump
4.	limb	rim	**31.**	blink	brink	**58.**	hies	Hythe
5.	fawned	horned	**32.**	lank	lank	**59.**	clucks	crux
6.	dose	dose	**33.**	marl	mull	**60.**	bowed	vowed
7.	lens	Rennes	**34.**	bail	veil	**61.**	ply	pry
8.	veld	veld	**35.**	sues	thews	**62.**	sways	swathe
9.	clipped	crypt	**36.**	darns	darns	**63.**	boos	booth
10.	lance	larns	**37.**	luge	rouge	**64.**	loam	roam
11.	clause	crores	**38.**	sluice	slews	**65.**	barn	bun
12.	sous	sous	**39.**	fang	fang	**66.**	foist	hoist
13.	bine	vine	**40.**	seam	seam	**67.**	clave	crave
14.	luxe	rucks	**41.**	bowel	vowel	**68.**	voles	voles
15.	bib	bib	**42.**	loon	rune	**69.**	bile	vile
16.	latch	ratch	**43.**	Lab	lav	**70.**	laud	roared
17.	baize	bathe	**44.**	ties	tithe	**71.**	becks	vex
18.	ob	ob	**45.**	scythe	scythe	**72.**	gland	gland
19.	rends	rends	**46.**	Bros	broth	**73.**	furls	hurls
20.	leers	rears	**47.**	Fran	flan	**74.**	harsh	hush
21.	clack	clack	**48.**	reft	reft	**75.**	glaze	grays
22.	vac	vac	**49.**	blanch	branch	**76.**	soar	thaw
23.	clique	creak	**50.**	sari	Surrey	**77.**	bide	vied
24.	sane	thane	**51.**	lieu	rue	**78.**	baas	baas
25.	bows	vows	**52.**	gibe	jive	**79.**	feud	hewed
26.	blessed	blessed	**53.**	plank	prank	**80.**	lacy	racy
27.	douse	dhows	**54.**	beard	veered			