Developing Equivalent Forms of a Test of General and Academic Vocabulary

Phil Bennett
Miyazaki International College

Tim Stoeckel
Miyazaki International College

Reference Data:

Bennett, P., & Stoeckel, T. (2013). Developing equivalent forms of a test of general and academic vocabulary. In N. Sonda & A. Krause (Eds.), *JALT2012 Conference Proceedings*. Tokyo: JALT.

This paper outlines the development of a new vocabulary test that assesses written receptive knowledge of the words in the General Service List and the Academic Word List. The test is intended to enable the provision of diagnostic feedback and goal setting over the course of a program of study. To avoid a possible testing effect from repeated assessment, 4 forms of the test were created, each made to the same blueprint. The instrument was field-tested with 334 Japanese university students, and results were analyzed from a Rasch measurement perspective. The vast majority of test items demonstrate good technical quality, test reliability for the 4 forms ranges from .87 to .93, and the 4 test forms have been found to be equivalent for use with Japanese students, within 1 standard error.

本稿では、新たな語彙テストの作成過程の概略を述べる。このテストは、頻出基本単語リスト（GSL）と学術基本単語リスト（AWL）の書面における受容語彙知識を測定するものであり、高等教育および大学教育における学習過程を通して、診断的なフィードバックを与え、目標設定を容易にする目的で作られている。度重なる試験の施行から生じるテスト効果の可能性を回避するため、4形式のテストが作成されており、それぞれは同じ設計書（ブループリント）に基づいている。334名の日本人大学生を対象にこのテストを行い、結果はラッシュモデルで分析した。テスト項目の大多数は性質上正確であり、日本人学生を対象に使用した場合、4形式のテストの信頼性は.87から.93であり、1標準誤差以内であることが判明した。

Vocabulary, once a somewhat neglected aspect of language learning, has now gained a far more prominent position in the field of language acquisition. Several empirical studies have demonstrated high correlations between vocabulary knowledge and performance on tests of the four main language skills (Meara & Buxton, 1987; Milton, Wade, & Hopkins, 2010; Stæhr, 2008). From studies such as these, attempts have been made to estimate the required vocabulary sizes to achieve competence at various language tasks. These estimates show some variation, but the figure of 2,000 words has regularly been put forward as indicative of a "threshold" vocabulary size, without which little can be comprehended (Milton, 2009; Stæhr, 2008).

Vocabulary size is often measured in terms of the number of word families a learner knows. A word family is a headword plus its inflections and closely related derivations. Bauer and Nation (1993) developed a system for determining word family membership based on the criteria of frequency, productivity, predictability, and regularity to grade the affixes used to

produce inflected and derived forms. This system has been employed in the development of several important word lists (Bauman & Culligan, 1995; Coxhead, 2000; Nation, 2006). Studies investigating the number of word families necessary for comprehension of oral interaction beyond a very basic level have proposed figures in the 2,000-3,000 word family range (Milton, 2009; Schmitt, 2010), and learners are likely to require 4,500 word families or more to be able to comprehend a range of written text types and to achieve passing scores on higher level English examinations (Milton & Hopkins, 2006; Nation, 2006; Schmitt, 2010).

If these values are accepted, then language teachers have a benchmark against which to judge learner progress and set appropriate goals. The provision of clear goals that are perceived as important and challenging, yet attainable, is one of the key elements of goal-setting theory as described by Dörnyei (2001). Since most learners of English in either secondary or tertiary institutions follow courses that are at least a year in duration, commitment to learning could be enhanced if regular assessment and individualized vocabulary learning goals were included in language programs.

Word Lists: The Frequency Model and Specialized Needs

Frequency is the standard principle by which vocabulary is organized and sequenced for testing. It is widely recognized that a relatively small number of highly frequent words comprises a very large proportion of typical English texts (Nation, 2001), and the frequency model predicts that the more frequent a word is, the more likely learners are to recognize it (Brown, 2012; Meara, 1992). However, Zipf (as cited in Milton, 2009) has demonstrated that the effects of the model are limited at lower frequency levels. Aizawa's (2006) study of word recognition among Japanese university students found that, beyond the fourth 1,000-word band of English, differences in learners' recognition were no longer statistically significant and were in some cases inconsistent with the predictions of the frequency model. This, coupled with the fact that less frequent words offer progressively lower text coverage, suggests that at some point it would be more beneficial for learners to tailor their vocabulary learning to their individual needs than to study progressively less frequent word bands.

The Academic Word List (AWL; Coxhead, 2000) serves such a purpose for learners in academic settings. The AWL is a list of 570 word families that commonly occur in a range of academic texts. It was compiled as a focused set of lexical items for learners of academic English to study once the words on the General Service List (GSL; West, 1953) have been acquired. The GSL was developed originally to aid the writing of simplified texts for language learners but has also been used to define a minimum vocabulary threshold for comprehension of basic discourse. A frequency-ranked version of the GSL was compiled by Bauman and Culligan (1995). This revised list comprises 2,284 word families and can be divided into two sublists, covering approximately the first and second 1,000 words of English (hereinafter GSL1 and GSL2). While it has been criticized for its age and coverage (Hancioğlu, Neufeld, & Eldridge, 2008), the GSL has been shown to cover around 75% of the words in academic text (Coxhead, 2000) and 80-90% of texts in other genres (Nation, 2001). Taken together, the GSL and AWL provide coverage of around 86% of academic texts (Coxhead, 2000).

Vocabulary Testing Instruments

Two of the more well-known tests of word recognition are the Vocabulary Levels Test (VLT; Nation, 1983; Schmitt, Schmitt, & Clapham, 2001) and the Vocabulary Size Test (VST; Nation & Beglar, 2007). The VLT is primarily intended as a diagnostic

tool, providing feedback on gaps in learners' vocabularies at the 2,000, 3,000, 5,000, and 10,000 word-frequency bands, as well as in a band of words drawn from the AWL. The VST offers a measure of vocabulary size. It contains target items drawn from the first to the 14th thousand-word frequency bands of the British National Corpus. Scores on the VLT and VST are used to estimate the percentage of words known in each tested frequency band and overall vocabulary size, respectively (Beglar, 2010; Nation, 1983). These interpretations, which are derived directly from raw scores, are meaningful to learners and educators and have been used as measures in numerous studies of the relation between vocabulary knowledge and other aspects of second language learning (e.g., Laufer & Ravenhorst-Kalovski, 2010; Stæhr, 2008).

One limitation to both of these instruments is the lack of multiple forms. In their most recent incarnations, only two forms of each instrument have been made available. As a result, repeatedly using either instrument over the course of a program of study to monitor vocabulary growth risks a testing effect.

Equating Tests of Vocabulary Knowledge

When using multiple versions of a test to track vocabulary development, the equivalency of test forms must be established, or the scores need to be transformed to a common scale. However, the primary obstacle for equating L2 vocabulary tests has been meeting the requirement of population invariance, which demands that the equating function be identical for each significant subpopulation (Petersen, 2007). Schmitt et al. (2001) found establishing equivalency of two versions of the VLT to be untenable due to differences in English vocabulary knowledge stemming from learners' various L1 backgrounds.

Purpose

This paper introduces and describes the ongoing development of a new test of vocabulary knowledge. Our objective is to produce an instrument capable of tracking the development of threshold English vocabulary knowledge for Japanese students in academic contexts. To avoid the possibility of a testing effect, four forms of the test were made, each following the same blueprint. The goal was for these forms to be of equivalent difficulty such that raw scores could be used and interpreted interchangeably. By focusing our study on native Japanese speakers, we hoped to eliminate the problems encountered by Schmitt et al. (2001) in equating test forms for speakers from multiple L1 backgrounds.

Such an instrument could serve several valuable purposes. First, it could provide learners with diagnostic feedback on gaps in knowledge of the core vocabulary needed in academic settings. Second, it could help teachers choose texts of appropriate lexical difficulty. Third, it could assist English programs in setting suitable vocabulary learning objectives and determining whether those objectives are being met. Finally, it could provide researchers with a tool for longitudinal studies of vocabulary development where repeated measurement is required.

The following sections will describe the test and its development and report the results of field-testing in terms of item quality, test reliability, and equivalency of test forms.

Instrument Development

Item Development

Test items were designed to assess written receptive knowledge of the GSL1, GSL2, and the AWL. Items were written for 80 target words randomly selected from each of these bands, creating a bank of 240 items.

Test items share many of the same specifications as those in the VST (see Beglar, 2010; Nation & Beglar, 2007). A multiple-choice format was used because of its universal familiarity and because unambiguous results can be quickly obtained. The stem of each item includes the target word in bold typeface followed by a short sentence that uses the word in a natural, nondefining context. This contextualized format has been found to help examinees clarify word meaning (Henning, 1991) and can lead to beneficial washback when compared to discrete point vocabulary measures (Qian, 2008). For the stem of each item, the Corpus of Contemporary American English (http://corpus.byu.edu/coca/) was consulted to confirm that one of the most frequently occurring members of the target word family and its common collocates were used in the example sentence. As in the VST, the stem is followed by answer choices that include the definition of the target word and three distractors.

To avoid construct-irrelevant difficulty (Messick, 1995), test items were written with simplified language. Specifically, items targeting knowledge of the GSL were written with the most frequent 1,000 words of the GSL, and items targeting knowledge of the AWL were written with words from the GSL. A small number of items did not conform to these guidelines, but in each of these cases the words used were among the most frequent 1,000 of either the British National Corpus (accessed at http://www.lextutor.ca/vp/bnc/) or the JACET 8000 list (Aizawa, Ishikawa, & Murata, 2005) (e.g., *conversation*, *rain*), or they were English loanwords in the Japanese language (e.g., *coffee*, *computer*). None of these exceptions was judged to be overly difficult for the target population of examinees.

Though several item features are shared with the VST, a distinct difference is that, for some GSL items (e.g., *metal*, *curve*, *pull*), the four answer choices are in the form of pictures rather than words. It was felt that in cases such as this, pictures would better assess knowledge of the target word than written choices which require less frequently occurring words than the target word itself. This was the approach taken by Nation (2001) in the 1,000-word level version of the VLT.

Expert Review and Piloting

Each test item underwent expert review and was then piloted with learners of English in one Japanese university. The information collected during piloting was utilized to identify items in further need of revision and to estimate item difficulties. It also led to the following two changes in item characteristics. First, in addition to the four choices of word meaning for each test item, a fifth option was added which reads, "I DON'T KNOW THIS WORD" (hereinafter choice E). In addition, the threat of a penalty for wrongly answered items was specified in the test instructions. (Example items are provided in Figures 1 and 2.)

bias: Be careful of **bias** in your writing.

a. grammar mistakes
b. language that is not exact
c. unfair opinions
d. informal language

e. I DON'T KNOW THIS WORD.

Figure 1. Example Text Item

Figure 2. Example Picture Item

Nation (2012) has stated a preference for not using penalties or the *I don't know* option but notes they may be justified when vocabulary tests are used for "proficiency-related decision-making" (p. 13). We introduced these conventions to address the likelihood that scores were being inflated by guessing. Even with explicit directions to skip unknown words, most examinees had far more wrongly answered than skipped items, which suggested that they were guessing. A comparison of data collected before and after these changes revealed a significantly reduced ratio of wrongly answered to skipped items (Bennett & Stoeckel, 2012) and an improvement in Rasch person reliability from .86 to .92. These results are indicative of more accurate estimates of vocabulary knowledge.

Test Form Development

The initial item difficulty estimates obtained during piloting were the basis for distributing the 240 items across four test forms of equal length. Because these estimates came from a small sample, an effort was also made to balance the four forms for parts of speech and for English loanword status in the Japanese language, two variables associated with word difficulty (Daulton, 2008; Milton, 2009). This resulted in test forms A, B, C and D, each of which consists of three 20-item sections to assess knowledge of the GSL1, GSL2, and the AWL. For the purposes of item calibration and test form equating, these test forms were revised by taking some items from their original form and sharing them across the other forms to act as anchors. The end result was four 90-item forms with 30 items at each level.

Field-Testing

The four versions of the instrument were then field-tested and assessed for item quality, test reliability, and test form equivalence under the Rasch measurement model.

Method

A convenience sample of 334 native speakers of Japanese from 21 intact classes at two universities in Japan (university A: n = 205 [137 women, 68 men; TOEIC data unavailable], university B: n = 129 [77 women, 52 men; TOEIC mean = 408.7, SD = 130.5]) participated in this phase of test development. The four 90-item test forms were spiraled in each class section. The data was analyzed with Winsteps software (version 3.72.2). The quality of the links within and between each test form was assessed and found to be satisfactory. Items were then simultaneously calibrated using the Rasch dichotomous model. These item calibrations were used in four separate analyses for converting raw scores to Rasch person measures for each of the test forms.

Results

A preliminary examination of the data revealed satisfactory person fit, item fit, and dimensionality. Item quality was assessed by inspecting point measure correlations and Rasch item fit indices. Four items were flagged as misfitting the Rasch model: *GSL1 include*, *GSL1 offer*, *GSL2 pale*, and *AWL transform*. Inspection of these items revealed ambiguity or grammatical complexity in the wording of the questions. The original item for *AWL transform* is given in Figure 3 as an example.

transform: The **transformation** of the town has had a big effect.

a. terrible damage
b. money that has been spent
c. people arriving from other countries
d. a complete change

e. I DON'T KNOW THIS WORD.

Figure 3. Original Test Item for *AWL transform*

The sentence stem and the four answer choices all contain modified noun phrases, which may have added unnecessarily to item difficulty. Another possibility is that the use of the indefinite article *a* in choice d confused respondents because the definite article *the* is already in the item stem. In addition, all four of the answer choices could constitute examples of transformation. As a consequence, this item was revised as shown in Figure 4. Here, less complex language has been used, and the distractors, while plausible replacements for *transformation* in the sentence stem, are not themselves examples of transformation. The other misfitting items have also been revised and all of these items will be monitored in future test administrations. The remaining 236 items appear to have good technical quality.

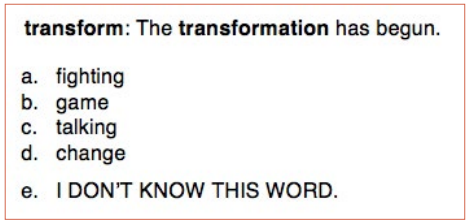

Figure 4. Revised Test Item for *AWL Transform*

Test reliability was assessed by inspecting Rasch person reliability estimates for each test form. Person reliability is an indication of person measure-order reproducibility and is similar conceptually to Cronbach's alpha. Reliability estimates ranged from .92 to .95 for the four 90-item forms and from .87 to .93 with the anchor items removed, indicating that all test versions had acceptable internal consistency (see Table 1).

To assess the relative difficulty of the 60-item forms, Rasch person measures for each possible raw score were compared across the four tests. Partial results are shown in Table 2. At any given raw score, Rasch person measures are within one standard error (*SE*) of each other. However, it is clear that, whereas Forms A and C are nearly identical, Form B is somewhat more difficult (indicated by lower person measures), and Form D somewhat easier. When comparing any person measure from Form B with its closest equivalent on Form D, the difference is about 3 points.

Table 1. Rasch Person Reliability Estimates

Test form	90-item version			60-item version (no anchors)		
	Person reliability	Mean	*SD*	Person reliability	Mean	*SD*
A	.93	57.1	13.6	.89	38.7	8.9
B	.92	60.2	12.8	.87	41.2	8.2
C	.92	61.4	12.5	.88	41.3	8.2
D	.95	57.0	16.7	.93	38.3	10.4

Table 2. Comparison of Raw Scores With Person Measures Across Four Forms

Raw score	Rasch person measure (*SE*)			
	Test form			
	A	B	C	D
37	.82 (.33)	.58 (.34)	.78 (.34)	.90 (.34)
38	.94 (.34)	.69 (.34)	.89 (.34)	1.01 (.34)
39	1.05 (.34)	.81 (.34)	1.01 (.34)	1.13 (.35)
40	1.16 (.34)	.93 (.35)	1.12 (.34)	1.25 (.35)
41	1.28 (.34)	1.05 (.35)	1.24 (.34)	1.37 (.35)
42	1.40 (.35)	1.17 (.35)	1.35 (.34)	1.50 (.36)

Discussion

A primary goal of this project was to develop equivalent test forms so as to avoid the possibility of a testing effect when repeatedly assessing vocabulary growth in a program of study. The instrument displays good item quality and overall reliability, but there are several issues in need of further review.

First, in light of the differences in test form difficulties and our preference for reporting raw scores, a redistribution of items among test forms is necessary to more closely approximate test equivalency. Because Rasch analysis provides difficulty estimates for each item, this is a relatively uncomplicated procedure. However, the stability and precision of item calibrations should first be explored with a larger, more representative sample.

Second, guessing in multiple-choice test formats can inflate estimates of vocabulary knowledge (Milton, 2009; Schmitt, 2010; Stewart & White, 2011). When a vocabulary test is appropriately designed for its intended population, examinees will encounter unknown words, and if these are not accounted for in the scoring rubric, estimates of vocabulary knowledge will be inaccurate. Even with the addition of choice E in our test, some examinees continued to have a rather high proportion of wrongly answered to skipped items (Bennett & Stoeckel, 2012), implying that some correct items were unknown but answered correctly by chance. Because this directly relates to a large body of research on the vocabulary sizes required to accomplish certain tasks in a foreign language (e.g., Laufer & Ravenhorst-Kalovski, 2010; Nation, 2006), further studies are required in this area.

A third issue is the functionality of picture items. The rationale for such items was that some target words could not easily be defined with a limited vocabulary. Though analysis has not flagged any of these items as misfitting the Rasch model, the mental processes involved in answering the two formats are likely to be different, and, as such, there is merit in investigating in greater detail the use of picture items.

A final concern is that items in the GSL1 word band are defined with words from the same frequency band in the current test format. It was expected that the target population would know the vast majority of these words, but results of field-testing demonstrated that this was not the case. Bilingual tests may be a solution to this problem because they would eliminate the difficulty of respondents having to read the answer choices in the L2. However, care must be taken in score interpretation because research has shown that examinees score higher on bilingual tests (Ruegg, 2007).

Although these questions should be addressed, the test in its current format appears to be a useful tool for assessment of threshold vocabulary in Japanese academic contexts. For repeated testing, forms A and C were found to be of approximately equivalent difficulty for this sample, and could be treated as such for low-stakes purposes. This test adds to the instruments currently available to language instructors in that it allows for repeated testing without the risk of a testing effect and enables informed, reliable feedback on each of the word bands that are essential for learners in academic settings. The four test forms are available from either of the authors.

Acknowledgements

The authors wish to express their gratitude to Jeffrey Stewart for his thoughtful feedback on the manuscript.

Bio Data

Phil Bennett is a lecturer at Miyazaki International College. He is interested in all aspects of lexical development, with a current focus on acquisition of metaphorical language. <pbennett@sky.miyazaki-mic.ac.jp>

Tim Stoeckel teaches at Miyazaki International College. His interests include vocabulary teaching and learning and language testing. <tstoecke@sky.miyazaki-mic.ac.jp>

References

Aizawa, K. (2006). Rethinking frequency markers for English-Japanese dictionaries. In M. Murata, K. Minamide, Y. Tono, & S. Ishikawa (Eds.), *English lexicography in Japan* (pp. 108-119). Tokyo: Taishukan-shoten.

Aizawa, K., Ishikawa, S., & Murata, T. (2005). JACET 8000 eitango [JACET 8000 Word List]. Tokyo: Kirihara-shoten.

Bauer, L., & Nation, P. (1993). Word families. *International Journal of Lexicography, 6*, 253-279. doi:10.1093/ijl/6.4.253

Bauman, J., & Culligan, B. (1995). *About the General Service List*. Retrieved from http://jbauman.com/aboutgsl.html

Bennett, P., & Stoeckel, T. (2012). Variations in format and willingness to skip items in a multiple-choice vocabulary test. *Vocabulary Education and Research Bulletin, 1*(2), 2-3.

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing, 27*, 101-118. doi:10.1177/0265532209340194

Brown, D. (2012). The frequency model of vocabulary learning and Japanese learners. *Vocabulary Learning and Instruction, 1*(1), 20-28.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*, 213-238. doi:10.2307/3587951

Daulton, F. E. (2008). *Japan's built-in lexicon of English-based loanwords*. Clevedon, UK: Multilingual Matters.

Dörnyei, Z. (2001). *Teaching and researching motivation*. Harlow, UK: Pearson.

Hancioğlu, N., Neufeld, S., & Eldridge, J. (2008). Through the looking glass and into the land of lexico-grammar. *English for Specific Purposes, 27*, 459-479. doi:10.1016/j.esp.2008.08.001

Henning, G. (1991). *A study of the effects of contextualization and familiarization on responses to the TOEFL vocabulary test items*. Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language, 22*, 15-30. Retrieved from http://nflrc.hawaii.edu/rfl/

Meara, P. (1992). *EFL vocabulary tests* (1st ed.). Swansea, UK: Centre for Applied Language Studies, University College Swansea.

Meara, P., & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing, 4*, 142-154. doi:10.1177/026553228700400202

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749. doi:10.1037//0003-066X.50.9.741

Milton, J. (2009). *Measuring second language vocabulary acquisition*. Bristol, UK: Multilingual Matters.

Milton, J., & Hopkins, N. (2006). Comparing phonological and orthographic vocabulary size: Do vocabulary tests underestimate the knowledge of some learners. *Canadian Modern Language Review, 63*, 127-147. doi:10.1353/cml.2006.0048

Milton, J., Wade, J., & Hopkins, N. (2010). Aural word recognition and oral competence in English as a foreign language. In R. Chacon-Beltran, C. Abello-Contesse, & M. Torreblanca-Lopez (Eds.), *Insights into non-native vocabulary teaching and learning* (pp. 83-98). Bristol, UK: Multilingual Matters.

Nation, I. S. P. (1983). Testing and teaching vocabulary. *Guidelines, 5*(1) 12-25.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press. doi:10.1017/CBO9781139524759

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review, 63*, 59-82. doi:10.3138/cmlr.63.1.59

Nation, I. S. P. (2012, August). *Measuring vocabulary size in an uncommonly taught language*. Paper presented at the International Conference on Language Proficiency Testing in the Less Commonly Taught Languages, Bangkok, Thailand. Retrieved from http://www.sti.chula.ac.th/files/conference%20file/doc/paul%20nation.pdf

Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher, 31*(7), 9-13.

Petersen, N. (2007). Equating: Best practices and challenges to best practices. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales: Statistics for social and behavioral sciences* (pp. 59-72). New York: Springer New York.

Qian, D. D. (2008). From single words to passages: Contextual effects on predictive power of vocabulary measures for assessing reading performance. *Language Assessment Quarterly, 5*, 1-19.

Ruegg, R. (2007). The English vocabulary level of Japanese junior high school students. In K. Bradford Watts, T. Muller, & M. Swanson (Eds.), *JALT2007 Conference Proceedings*, 103-109. Tokyo: JALT.

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Basingstoke, UK: Palgrave Macmillan. doi:10.1057/9780230293977

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing, 18*, 55-88.

Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal 36*, 139-152. doi:10.1080/09571730802389975

Stewart, J., & White, D. A. (2011). Estimating guessing effects on the vocabulary levels test for differing degrees of word knowledge. *TESOL Quarterly*, *45*, 370-380.

West, M. (1953). *A general service list of English words*. London: Longman.

Pre- and Posttest Washback in Paired Oral Classroom Assessments

Nathan T. Ducker
Ritsumeikan Asia Pacific University

Curtis J. Edlin
Ritsumeikan Asia Pacific University

Richard A. Lee
Kurume Institute of Technology

Reference Data:

Ducker, N. T., Edlin, C. J., & Lee, R. A. (2013). Pre- and posttest washback in paired oral classroom assessments. In N. Sonda & A. Krause (Eds.), *JALT2012 Conference Proceedings*. Tokyo: JALT.

Testing is a vital part of the learning process that teachers and curriculum designers can use to motivate students to study, help them monitor their progress, and guide their pre- and posttest learning activities. Successfully implemented testing should therefore have a positive washback effect on students' learning activities in these areas. To gain full benefit from the testing process, once assessments have been carried out and graded, quality feedback should further help students develop good learning habits and focus their efforts on areas that need attention. This paper reports on the review of a speaking program at a private university in Japan in which the teacher-researchers collected data on the washback effect of a cycle of 8 speaking assessments carried out in one semester, in order to improve the speaking program's efficacy in encouraging learner development through the quality and quantity of pre- and posttest learning practice activities.

学習者のアクションは、言語学習の成功の中心となるが、テストとは、教師やカリキュラム設計者が、学生にやる気を起こさせる為に使用する学習過程で重要なもので、学生が自分の進捗状況を知り、テスト前後の学習を高めるのに役立つものである。それ故、より効果的に実施されたテストでは、学生の学習活動にプラスのウォッシュバック効果をもたらすはずである。テストの過程から最大限の利益を得る為に、課題を実施し採点した後、よりよいフィードバックをする事で、学生はさらに良い学習習慣を生み出し、注意を必要とする分野に努力を集中させる事ができる。この論文は、1学期において8つのスピーキング課題を実施し、質の良い練習課題を数多くこなす事によって、テスト前後の学習者の発達を促しスピーキングプログラムの効果を向上させる為に、教師／研究者がデータを収集した日本のある私立大学のスピーキングプログラムの評価をレポートしたものである。

Tests are tools that, amongst other things, help students develop as language learners (Carr, 2011). Students want to do well within their language courses, and thus tests offer extrinsic motivation with regard to grades (Bernard, 2010). Additionally, learners are also intrinsically motivated by their improvement when the language being tested is meaningful to them (Bernard, 2010). Tests also give students a tangible marker to set goals against, which is an important autonomous learning strategy that leads to better language performance (Oxford & Shearin, 1994). Therefore, when course planners and teachers begin to devise tests and assessments for their classes, it is important that the assessment is judged not only on how reliable and valid it is as a summative tool, but also on its potential to positively impact learning (Black & Wiliam, 1998).

This effect on learning is known as test washback (Alderson & Wall, 1993), and is framed in terms of two dimensions: direction, including positive effects (such as motivating the learner to practice) and negative effects (for example, practicing multiple choice questions at the expense of practicing real language use, or discouraging study altogether), and intensity of washback, referring to either strong or weak effects (see Green, 2007, for a discussion).

The majority of washback research has focused on high-stakes testing while little research has been done on classroom-based testing. One of the reasons that classroom-based assessment may be receiving little attention is the belief that "high-stakes tests have more power to modify teacher and learner behavior whereas low-stakes tests, such as classroom-based assessments, are not central to decision-making and therefore have fewer consequences," as reported by Munoz and Alvarez (2010, p. 2). However, the need for classroom-based research has been called for by several researchers, (Munoz & Alvarez, 2010; Watanabe, 2005; Xie & Andrews, 2012). Watanabe (2005) argued that more research in this area is needed in order to answer questions such as how to motivate students through tests and to find out what sort of feedback is most useful for students. Furthermore, the majority of the research conducted on washback has dealt with teachers' responses to tests rather than learners' reactions with regard to test preparation and follow-up (Xie & Andrews, 2012). Therefore, this study's goal was to further understand the washback effect of classroom-based testing on students' learning actions.

The few studies of classroom tests that do exist show that students' thorough understanding of the expectations and goals of tests plays a large part in determining whether a positive washback effect is produced or not. Munoz and Alvarez (2010) reported that students' awareness of assessment goals led to them focusing their efforts on better performance on speaking tests. Similarly, Green (2007) found that students' understanding of test requirements might be a greater mediator of learning attainment than course content. Additionally, Xie and Andrews (2012) (citing Struyven, Dochy, & Janssens, 2005) suggested that students choose the appropriate learning strategies to match their perceptions of what a test entails. Therefore, it was expected that students' learning actions would be mainly focused on successful test completion rather than personal language learning goals.

While for some, washback is limited to pretest influence (Peirce, 1992; Berry, 1994), for others, washback has a broader meaning, extending to effects on students taking an exam, feedback received, and subsequent decisions (Bailey, 1999; Brown & Hudson, 2002). Given this wider description of washback, feedback has an important moderating effect on the positive or negative washback of a test. For example, Cameron and Pierce (1994) and Kluger and DeNisi (1996) reported that positively voiced feedback (to encourage students), with no focus on the objective goals of a task, had a negative effect on students' attitudes toward study and subsequent assessment performance (as cited in Black & Wiliam, 1998, p. 14). The importance of good feedback on students' successful studying cannot be underestimated, as Hattie (1999) pointed out: "The most powerful single moderator that enhances achievement is feedback" (p. 9). In order to be valuable in terms of positive washback, feedback needs to be diagnostic, detailed, relevant, and useful (Shohamy, 1992; Black & Wiliam, 1998; Munoz & Alvarez, 2010; Munoz, Casals, Gaviria, & Palacio, 2004). Hattie and Timperley (2007) and Black and Wiliam (1998) both further explained that the most effective kinds of feedback involve students both receiving feedback on a performed task and being able to identify how to improve their performance.

In this paper, we report on both the pretest and posttest washback effects of a cycle of speaking assessments conducted eight

times in a semester in a mandatory, intermediate-level, general English course at an international university in Japan. The testing procedure was designed with the intention of maximizing students' speaking opportunities and promoting confidence in their oral abilities. Studying the washback effects of the testing process can help course designers and teachers to understand if a course is well designed in terms of promoting students' proactive, pretest, out-of-class study, and studying the posttest washback effects of the testing cycle can further inform designers and teachers if the test feedback is fulfilling the important educational role of helping students improve their performance.

Course Description and Data Collection

The study was carried out on an intermediate-level, multi-skill, mandatory general English course in an international university in southern Japan with 3,208 domestic (Japanese) students and 2,526 international students from 83 different countries. The majority of students in the course had completed elementary and preintermediate level English classes, while a small proportion of students matriculated directly into the intermediate course on attainment of a paper-based TOEFL score in the 460-479 range. While the majority of students were Japanese, a small number of Korean and Chinese students (fluent in Japanese) studied English alongside their Japanese counterparts and their responses are also included in the data.

The speaking component of the course consisted of eight individual speaking tests developed using task-based role-play activities created from chapter themes and conversation topics contained in the required textbook for the course (Tanka & Most, 2007). The tests emphasized communication strategies. In particular, the main communication strategies were

- initiating conversations,
- introducing topics,
- maintaining conversations,
- overcoming communication breakdown, and
- giving reasons and support.

Students completed the task-based role-plays in pairs, while the teacher assessed task completion and oral proficiency. While two students were completing the assessment, the remaining students carried out other work and waited for their turn. In order to reduce the anxiety associated with testing and to encourage students to feel relaxed during the assessments, the grade for an individual assessment was only 1.5% of the total grade for the course. After the first role-play conversation, all subsequent assessments were designed to recycle and repeat previously covered skills and language using a new topic or context. In this way, each assessment aimed to challenge students to practice previously learned material and reinforce the use of good communication strategies.

The eight assignments were delivered in a cycle of three phases: a task introduction lesson, a task practice lesson, and a task assessment and self-review. In the introduction phase, teachers provided students with a set of worksheets detailing (a) the assessment task and a checklist of the communication strategies upon which teachers' assessments would be based; (b) key vocabulary and language forms useful for satisfactory completion of the tasks; and (c) example conversations (audio file and scripts) and questions designed to raise the students' awareness of the language used by the speakers. For the practice phase, teachers and students were provided with further practice activities and time for students to practice the task with a partner and receive teacher feedback about their general performance. In the assessment phase, students completed a role-play with student partners while the teacher assessed the students' completion of the task using a checklist. Following the assessment, students completed a self-review sheet and teachers gave students feedback related to both their completion of the

task and other areas of their speaking proficiency. Postassessment, students were encouraged to use their feedback to improve areas of speaking proficiency as directed by their teachers; however, no additional class time was set aside for this work. Given the large number of sections (15-20 per semester), it was difficult to determine if the content was consistently delivered in the manner described above.

Testing took place in a very limited time (teachers managed up to a dozen pair interviews in a 95-minute class), and given the complicated nature of the construct of oral proficiency (see Brown, 2003, for a discussion), standardization of grading was difficult. Therefore, to keep the grading uncomplicated and standardized across a large number of sections, the students' assessment scores were calculated based on completion of the task only. Additional feedback was provided on students' oral proficiency, and teachers were encouraged to select one or two areas about which to give students advice on how to improve (see Appendix A). Given the large numbers of sections, teachers, and students involved, it was difficult to determine what feedback was given and how students used it at the time.

Completing three phases of an assessment eight times in one semester was both time and labor intensive for teachers and students alike. However, in a previous study of student activity on the international campus, it was found that despite the setting, students failed to take full advantage of the opportunities to practice English with international students (see Lee, Browne, & Kusumoto, 2011). Therefore, the course designers sought to develop an approach to teaching speaking that would give students as much opportunity as possible to practice speaking in English and further provide students with both the skills to communicate in English on campus and the motivation to practice speaking autonomously. Subsequently, in order to judge the success of the course, we were keen to find out if the testing process promoted students' proactive learning and to what extent. Additionally, as a testing cycle finished, we wanted to know if students were then able to use teacher feedback to further develop their language practice. If students were not proactively practicing outside of class time and not using their feedback to further improve their proficiency, the designers believed that aspects of the assessment process would need to be redesigned. With this in mind, the following research questions were asked:

- Did the testing process promote students' proactive learning and if so, to what extent?
- As each testing cycle finished, were students then able to use teacher feedback to further develop their language practice?

Data was collected in three stages. A bilingual Japanese/English pilot survey was delivered to two classes, totaling 42 respondents. A follow-up structured interview was carried out with 24 random members of the two classes to check the pilot survey. The questions in the semi-structured interview were initially asked in English and supplemented with Japanese by the interviewer when necessary. The results were recorded on paper, but not digitally. A total of 203 students out of the 327 enrolled in the course responded to the final survey (students who participated in the pilot survey were excluded). The survey was voluntary and anonymous. There was no compulsion for students to take the survey as the course had already been completed. In addition, a bilingual disclaimer explaining the purpose of the survey was included on the first page of the survey.

Results

Unless specifically indicated, responses from the preliminary survey and the interviews reflect the results of the final survey. Key results of the follow-up interview are in Appendix B, and the final survey questions and results are in Appendix C. Once the survey results were collected, the results were analysed for evidence of positive and negative washback.

To answer the first research question (if the testing process promoted students' proactive learning and to what extent), we analysed the responses to Survey Questions 3 and 4 (see Appendix C, Q3 & Q4), which asked about the frequency and duration of student practice. Out of 145 respondents who answered that they practiced, 63% said that they practiced three or more times per testing cycle. Most commonly, students practiced for more than 30 minutes per practice. More specifically, 27.5% of the practices were between 20 and 30 minutes, 35.2% were between 30 minutes and an hour, and 12.4% of students said that their practices were longer than an hour. In terms of their practice foci (see Appendix C, Q10), the majority of respondents (134) reported that practice was aimed at *completion of task*, while the remaining criteria received a nearly evenly distributed numbers of responses: conversational management activities (90), fluency and pronunciation issues (86), accuracy (82), and using the correct vocabulary (84). That students focused on task completion was underscored by the kinds of activities they reported completing in preparation for the test (see Appendix C, Q6). The most popular practice activities were: practicing with a partner from class (84), memorizing key vocabulary (70), planning exactly what to say (68), and writing out a script (67). Further practices with peer advisors, students from other classes, or international students comprised a total of 59 responses.

To further address the first research question, we analysed responses to Survey Question 5 concerning students' motivation for practicing for the tests (see Appendix C, Q5). The majority of students reported that their main motivation was to improve their speaking ability (67.9%), while only 23.6% practiced in order to improve their grade. In fact, the results revealed that only 4% of the students did not practice due to the low weighting of the individual tests.

Additional responses relevant to the first research question were revealed in students' responses regarding the value of the testing process (see Appendix C, Q15) in that they found the tests to be good for helping them self-monitor their improvement (62%), a good opportunity to converse in English (44.5%), and beneficial in pushing students to study (25.5%). These results slightly contrasted with the interim oral interviews (see Appendix B) in which students indicated that the tests were mainly beneficial in pushing students to study (11 responses), in contrast to student indications that they provided a good opportunity to converse in English (6 responses). Students' positive perceptions of the testing cycle were further highlighted by their high levels of satisfaction with the speaking programme's outcomes (see Appendix C, Q14). In short, students believed that they were improving their oral abilities as they took the tests, which was an integral part of the washback effect of these tests.

To address the second research question (whether students were able to use teacher feedback to further develop their language practice), Survey Questions 8, 11, 12, and 13 were analyzed (see Appendix C). While there may be some variation in how teachers gave feedback, all students were supposed to receive the same grading form from their teacher. Thus, it was important to know if this form could be effectively used by students to review their tests. Multiple items on the grading form were unclear to students. The majority of students were able to discern the meaning of the task's requirements (see Appendix C, Q8). For example, 89.8% said they understood *introduce the topic*, 73.9% said they understood *maintain the conversation*, and 72.6% said they understood *give opinions and support*. However, the linguistic skills pertaining to language proficiency were not well understood. For example, *enunciation* had a positive response of only 39.2%, *syntax* a positive response rate of 46.1%, and *accuracy* a positive response rate of 50%.

In answer to concerns over students' ability to understand the feedback form, 94.5% of students reported that they were able to understand their teacher's written feedback (see Appendix C,

Q11). Yet, the survey results indicate that most students either did nothing with their test results (34.7%), or passively remembered their weak points (46.5%) for the next test (see Appendix C, Q12). Only one student each reported practicing weak points arising from the test results or taking the results to discuss them with a peer advisor; 3.5% reported discussing their results with their classmates; and 6.5% reported discussing their results with their teacher. As for the reasons why students did not review, no single answer clearly stood out as a reason (see Appendix C, Q13). The one result we expected to see more of was *it won't improve my grade*—yet only four students reported this. Conversely, nearly one-fourth of the respondents to this question wanted to review but either did not have sufficient time (21), or did not know how to use their teacher's feedback (25).

Discussion

The testing approach was successful in motivating students to proactively study for the test. Typically, students practiced three times per test for an average of 45 minutes. With eight tests per semester, this results in a typical student completing 18 hours of additional speaking practice—clear evidence of positive washback from the testing cycle. Students usually practiced in at least one of four ways: conversation practice, memorizing vocabulary, writing out a script, or making a list of key points to cover in the test. All of these items focused on the graded portion of the test and revealed that students intended to complete the task and improve their test scores. As no score was given for proficiency items, such as fluency or accuracy, students did not focus on improving their overall oral proficiency. These results correspond with reports that students' learning activities are strongly influenced by perceptions of test requirements (see Green, 2007; Munoz & Alvarez, 2010; Xie & Andrews, 2012). The results allowed us to see that the course achieved two of its goals by getting students to further practice speaking outside of the classroom and to develop autonomous study habits.

The presurvey interview responses indicated that the tests were a strong motivating factor in making students study and subsequent data collection further supported this. One concern was grade weighting. Considering that it has been argued that low-stakes tests such as classroom-based assessment are not central to decision-making and therefore have few consequences (Munoz and Alvarez, 2010), we were concerned that the points distribution of 1.5% of the students' overall grade would have a negative washback effect on students' motivation to study. However, with only four students responding that the low grade weighting stopped them from studying, the results directly contradicted that particular long-held belief about washback. Additionally, the majority of students reported that their main motivation to study for the tests was to improve their ability to communicate orally in English, rather than to get a good grade. Additional results showing the students' satisfaction with the testing process in relation to communicative ability also support the idea that students perceived the tests as useful in improving their English communication skills.

In contrast to the positive pretest washback of the tests, the posttest effects were mostly negative. An important consideration relating to students' posttest activities was the effect of test design on feedback. Students indicated that they could understand their teacher's feedback; however, many students indicated problems understanding the proficiency section of the grading form. For example, less than 40% of students understood *enunciation*, while 89.8% understood *introduce the topic*. Student responses indicated that teachers were either not taking the time to clarify these words with students or not discussing their impact on students' oral proficiency, which may have been due to the washback effect extending to teachers' actions and their placing more emphasis on the section directly related to

students' scores. This reiterates the need for course designers and teachers to consider how to ensure that feedback given to students is sufficiently diagnostic, detailed, and relevant, as well as understood by the participants, in order to facilitate better use of feedback, as argued by Shohamy (1992), Black and Wiliam (1998), Munoz et al. (2004), and Munoz and Alvarez (2010).

The most important finding in terms of posttest washback was that the majority of students did not actively use their teachers' feedback. There could be several reasons for this. Some students reported time constraints—perhaps because there was only a short interval between testing cycles (less than 2 weeks), so there was no time to work on using feedback before the next testing cycle started. Some students cited no additional grades, and a number of students simply had more productive (in their opinion) things to do. Some students reported not knowing how to review; though it was unclear whether this was due to a lack of study skills or a lack of understanding of the technical terms on the grading form. This evidence highlights how important it is that teachers allocate time to help students understand and learn how to use feedback. Finally, attitudes toward tests may impact students' review behaviours. Many Japanese students will have seen previous tests (such as entrance exams) as a barometer of achievement and may not be inclined to see tests as diagnostic tools that carry the requirement of further related study by the student. Whatever the reasons for students' nonuse of feedback, this study highlights that it is the responsibility of course designers and teachers to find ways to actively engage students in well-directed, feedback-driven, postassessment study as suggested by Shohamy (1992), Black and Wiliam (1998), and Hattie and Timperley (2007).

Limitations of This Study

Despite a large number of participants, this study should be considered exploratory as it highlights important areas for further research in order to fully understand the washback effect of classroom-based tests. The first issue concerns the quality of the data relating to students' practices. We do not know if students had a dual focus when preparing for the tests. For example, when writing out a script, did students aim for higher accuracy or, as they were memorizing vocabulary, did they repeat the vocabulary item many times in a sentence to develop their fluency? This aspect of the quality of a student's practice needs to be further understood to help teachers better advise their students and to help course planners understand the effects the test design has on students' behaviours and subsequent language learning. Additionally, the data indicate that the testing cycle had a motivational effect, encouraging students to study further. Unfortunately, the data are limited; further study in this area would help inform teachers and course designers as to how to better adjust their courses to encourage students to work on their speaking autonomously. Similarly, despite data which indicate that students felt the tests were beneficial in improving their speaking ability, we do not have any data regarding the ways in which this subsequently impacted the washback effect of the assessment cycle. Further investigations into the motivational effects of the testing process would also reveal if students are, after course completion, motivated and able to continue practicing during vacation periods.

Posttest, this study indicates the importance of teacher-assisted focus on feedback. However, we collected no data on the types of feedback that teachers were giving. We need further information on the quality, focus, type, and quantity of feedback in order to discern whether improvements in this area should be focused on course-wide procedures, teacher-centered instruction, or student motivation. Another consideration in terms of postassessment washback and our research design is that many students did use the results of their tests to monitor their progress through test scores. This needs deeper investigation. *Remember my weak points for next time* (see Appendix C, Question

12) can cover a wide range of activities, such as subconsciously readjusting the focus of practice for the following test, actually putting emphasis on checking sentences when writing scripts, or simply thinking, "I hope I get a better grade than last time." It is difficult for anyone to articulate the mental processes that he or she goes through when describing a learning activity, so future research could employ think-aloud-protocols for data collection to further elucidate the processes that students go through when they prepare and review for tests.

Conclusion

In contrast to other studies on washback, we examined evidence of the pre- and posttest washback effect of an oral classroom assessment cycle by looking at students' learning actions rather than the effects on teaching. This study has shown that classroom-based oral assessments do have some positive washback effects on learners' actions before taking a test, as highlighted by students' further study. The study also indicated that there were negative washback effects, as indicated by the limited range of activities that students undertook. Furthermore, the study suggested that there were posttest washback effects of classroom-based assessments, in that students did not pay attention to feedback and subsequent remedial study. As such, we hope this study provides course planners and teachers with information useful to setting realistic program goals centred on the learner and judging a course's effectiveness in terms of achieving those goals.

Additionally, given that the evidence pointed to negative washback after a test had been taken, this study highlighted the need for a clear focus on feedback during the assessment cycle. Here again, it is unclear whether teacher action had a mediating role or whether the test procedures and grading led directly to nonuse of feedback. Finally, in order to utilize the potential of classroom-based assessment, not just as a summative tool, but as a practical way of improving classroom-based language learning, this study highlighted the need for further investigations that consider (a) the actions of learners, such as strategies that students use to manage their practice for tests and monitor progress; (b) data on the motivational processes surrounding tests and students' perceptions of the testing process; and (c) the strategies that students use once they have received their feedback.

Bio Data

Nathan Ducker is an English language instructor at Ritsumeikan Asia Pacific University, Japan. His research interests include autonomy, speaking skills and the curriculum, and willingness to communicate. He can be contacted at <ducknath@apu.ac.jp>

Richard A. Lee is an English lecturer at the Kurume Institute of Technology, Japan. His research interests are related to speaking skills, curriculum design, pragmatics, study abroad, and the impact of language learning context. He can be contacted at <leera1@mac.com>

Curtis Edlin is an English language instructor at Ritsumeikan Asia Pacific University, Japan. His research interests are largely related to speaking, including prosody, spoken grammars, and the effects of cultural legacy on discourse in speaking. He can be contacted at <edlinc85@apu.ac.jp>

References

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics, 14*, 115-129.

Bailey, K. M. (1999). Washback in language testing. *TOEFL monograph series*. Princeton, NJ: Educational Testing Service.

Bernard, J. (2010). *Motivation in foreign language learning: The relationship between classroom activities, motivation, and outcomes in a university language-learning environment* (Honors theses, Paper 74). Dietrich College, Pittsburgh, PA. Retrieved from http://repository.cmu.edu/hsshonors/74

Berry, V. (1994). Current assessment issues and practices in Hong Kong: A review. In D. Nunan, R. Berry, & V. Berry (Eds.). *Bringing about change in language education* (pp. 31-34). Hong Kong: The University of Hong Kong, Department of Curriculum Studies.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*, *5*, 7-74.

Brown, J. D. (2003). Promoting fluency in EFL classrooms. *Proceedings of the 2nd Annual JALT Pan-SIG Conference*. Retrieved from http://jalt.org/pansig/2003/HTML/Brown.htm

Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing*. Cambridge, UK: Cambridge University Press.

Carr, N. T. (2011). *Designing and analysing tests*. Oxford: Oxford University Press.

Cameron, J., & Pierce, D. P. (1994). Reinforcement, reward, and intrinsic motivation: A meta analysis. *Review of Educational Research, 64*, 363-423.

Green, A. (2007). *IELTS washback in context*. Cambridge: Cambridge University Press.

Hattie, J. (1999). *Influences on student learning*. Inaugural lecture: Professor of education, University of Auckland. Available from http://www.education.auckland.ac.nz/webdav/site/education/shared/hattie/docs/influences-on-student-learning.pdf

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research 77*, 81-112.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*, 254-284.

Lee, R., Browne, K., & Kusumoto, Y. (2011). Measuring EFL learners' environment: English contact and use outside the classroom at a Japanese international university. *Polyglossia*, *20*, 15-25.

Munoz, A., & Alvarez, M. (2010). Washback of an oral assessment system in the EFL classroom. *Language Testing*, *27*, 33-49.

Munoz, A., Casals, S., Gaviria, S., & Palacio, M. (2004). Guidelines for oral assessment. *Cuadernos de Investigacion*, Doc. 23.

Oxford, R., & Shearin, J. (1994). Language learning motivation: Expanding the theoretical framework. *The Modern Language Journal, 78*, 12-28. Retrieved from http://www.jstor.org/stable/329249

Peirce, B. N. (1992). Demystifying the TOEFL reading test. *TESOL Quarterly, 26*, 665-689.

Shohamy, E. (1992). Beyond proficiency testing: A diagnostic feedback testing model for assessing foreign language learning. *The Modern Language Journal, 76*, 513-521.

Struyven, K., Dochy, F., & Janssens, S. (2005). Students' perceptions about evaluation and assessment in higher education: A review. *Assessment & Evaluation in Higher Education, 30*, 325-341.

Tanka, J., & Most, P. (2007). *Interactions 1: Listening and speaking* (Silver Ed.). New York: McGraw-Hill.

Xie, Q., & Andrews, S. (2012). Do test design and uses influence test preparation? Testing a model of washback with structural equation modeling. *Language Testing, 30*, 49-70.

Watanabe, Y. (Interviewee), & Newfields, T. (2005). Insights in language testing [Interview transcript]. *JALT Testing & Evaluation SIG Newsletter*. Retrieved from http://jalt.org/test/wat_new.htm

Appendix A

Sample Grading Rubric

Completion of the task: Satisfies the requirements of the test item						
Greet partner (10)	☐					
Introduce topic (10)	☐					
Recommendation 1 (20)	☐					
Recommendation 2 (20)	☐					
Recommendation 3 (20)	☐					
Ask and answer clarification questions (20)	☐					
Score	0-49	50-59	60-69	70-79	80-89	90-100

Hesitations and Halts in Speech, Mispronunciations and Enunciation, Connected Speech	
	<= Weaker ═══ Stronger =>
hesitations, halts	
mispronunciations, enunciation	
connected speech	

Word Order and Agreement, Grammar and Tense, Ability to Comprehend and Negotiate the Dialogue.	
	<= Weaker ═══ Stronger =>
Syntax (word order and agreement)	
Accuracy (grammar and verb tense)	

Body Language, Eye Contact, Voice Projection, Ability to Introduce and Maintain Conversation, Overcome Communication Breakdown	
	<= Weaker ═══ Stronger =>
body language, eye contact, voice projection	
introducing new topics	
maintaining the conversation	
Overcoming communication breakdown	

Range and use of vocabulary is content appropriate, English only (i.e., no unneccesary use of first language)	
	<= Weaker ═══ Stronger =>
vocabulary range	

Appendix B

Selected Responses from the Structured Interview Questions

Do you think that these tests are useful for improving your English?

	Response count
Yes, the content is useful	4
Yes, the tests make me study / practice	11
Yes, I can understand how to improve my skills	1
Yes, they are a good chance for us to really speak English	6

We have a lot of tests on this course - do you think this number is too many, just right, not enough?

	Response count
Too many	8
Just right	13
We should do more	3
No response	2

Appendix C

Finalized Survey and Results (irrelevant data tables omitted)

3. How many times do you usually practice as homework for a speaking test? 宿題として、大体何度スピーキングテストの練習をしますか?

Five times 5回	Four times 4回	Three times 3回	Two times 2回	One time 1回	Never 全くしない
15.9%	10.3%	37.2%	28.2%	9%	2.8%

Note. $n = 145$.

4. If you practice, how much time do you spend practicing? 練習にどれくらいの時間を費やしますか

More than 1 hour 1時間以上	More than 30 minutes 30分以上	20 - 30 minutes 20－30分	10 - 20 minutes 10－20分	0 - 10 minutes 10分以下
12.4%	35.2%	27.5%	21.4%	3.4%

Note. $n = 145$.

5. I prepared carefully because 私は熱心に準備しました、何故なら・・・

I worried about getting a high score for my GPA GPA 高得点を取り、GPAを上げたいから。	I always prepare carefully for tests 私は常にテスト勉強を熱心にするから。	I wanted to improve my ability to speak スピーキング能力を向上させたいから。
23.6%	8.6%	67.9%

6. If you do practice at home, which of these activities do you do to practice? 下記のどのような方法で練習しますか?

	Response count
Listening to the audio files on blackboard ブラックボードの音声ファイルを使う。	21
Practicing with a partner from class 同じクラスのパートナーと一緒に練習する。	84
Practicing with a partner from another class 他のクラスのパートナーと一緒に練習する	22
Practice with an international student 国際学生と一緒に練習する。	26
Practice with a PA* from SALC* SALCのPAと一緒に練習する。	11
Write out a script 台本を書き出す。	67
Memorize key vocabulary 重要な単語を記憶する。	70
Practice key phrases 成句を練習する。	29
Practice the key grammar 重要な文法を練習する。	34
Use shadowing シャドーイング(聞いた英語をすぐに追いかけて声に出す学習法)する。	21
Plan exactly what to say 何を言うか全て決めておく。	68
Other その他(詳しく書いて下さい	2

* PA - Peer Advisor (formerly called Teaching Assistant)

**SALC - Self Access Learning Center

7. Which of these is true for you:
下記のどれが当てはまりますか

I didn't prepare because
私はあまり準備をしませんでした、何故なら・・・

	Response count
The grade was only 1.5 % of my total grade テストの評価は全体の1.5％でしかないから。	4
The tests didn't motivate me やる気が出る課題がないから。	5
I had other more important things to do 他に優先すべきものがあるから。	22
I didn't know how to prepare 準備の仕方が分からないから。	9
I had prepared enough in class 授業中に与えられる時間だけで十分だから。	13

8. Which of these words from the speaking test form do you understand? スコアシートに記されている、どの項目を理解していますか?***

	Yes, I understand 完全に理解している項目	I am not sure どちらともいえない	No, I don't understand 全く理解していない項目
Introduce the topic	89.8%	10.2%	0.0%
Use transitions to signal questions	57.4%	36.1%	6.5%
Maintain the conversation	73.9%	22.5%	3.6%
Give opinions and support	72.6%	22.6%	4.7%
Close the conversation	76.6%	19.6%	3.7%
Hesitations	54.6%	31.6%	13.8%
Halts	39.8%	37.5%	22.7%
Mispronunciations	56.1%	29.2%	14.6%
Enunciation	39.2%	35.7%	25.1%
Connected speech	69.2%	23.8%	7.0%
Syntax	46.1%	35.4%	18.5%
Accuracy	50.0%	31.6%	18.4%
Body language	87.2%	11.7%	1.1%
Voice projection	79.0%	17.0%	4.0%
Introducing new topics	82.3%	15.4%	2.3%
Maintain the conversation	84.1%	13.5%	2.4%
Overcoming communication breakdown	62.4%	28.8%	8.8%
Vocabulary range	78.0%	19.7%	2.3%

***A Japanese translation of these items was not included at this stage in order to determine if students understood the English only grading rubric and English only teacher explanations

10. If you practice, what do you focus on? (You can choose more than one). 練習する際、何に注意しますか?(複数選択可)

	Response count
I didn't practice 私は準備をしませんでした。	15
Completion of task / Completing the conversation (introduction, opinions, questions, maintain the conversation, closing) 課題にそって会話を発展させること。(前置き、意見、質問、会話を保つ、結句)	134
Hesitations, halts, pronunciation, enunciation, connected speech ためらい、口ごもり、発音あやまり、発声	86
Word order, subject verb agreement, grammar, tense 語順、呼応、文法、時制	82
Body language 身振り、手振り	49
Introducing, maintaining conversations, overcoming communication breakdown 説明、会話を保つ、途切れた会話からの立ち直り、	90
Vocabulary 語彙	84
I don't choose one thing, I just try to complete the conversation 会話をとぎれさせない練習はするが課題内容チェックはしません。	11
Other その他(詳しく書いて下さい)	5

Comprehension of feedback

11. Which is true for you? 下記のどれが当てはまりますか?

I understand my teacher's written feedback 先生が書いたフィードバックの内容を理解出来ます。	I don't understand my teacher's written feedback 先生が書いたフィードバックの内容を理解出来ません。	I can't read my teacher's written feedback 先生が書いたフィードバックが読めません。
94.5%	3.0%	2.5%

12. What do you usually do with your feedback form? フィードバックの内容を見てあなたはどうしますか?

I look once to check my score, but I don't review 自分の得点は確認しますが復習はしません。	34.7%
I do nothing 特に何もしません。	7.9%
I look and remember my weak points for next time フィードバックを満見て次回の為に自分の弱点を覚えておきます。	46.5%
I review my teacher's written feedback and discuss with my teacher 先生が書いたフィードバックを元に復習し、先生に助言を求めます。	6.4%
I review my teacher's written feedback and discuss with a SALC PA 先生が書いたフィードバックを元に復習し、SALCのPAに助言を求めます。	0.5%
I review my teacher's written feedback and discuss with my classmates 先生が書いたフィードバックを元に復習し、クラスメイトに助言を求めます。	3.5%
I look at the form, then I practice my weak points carefully フィードバックを見て自分の弱点をよく練習します。	0.5%

If you practice after the test please explain how
テストの後に練習した事があれば、どのように練習したか教えて下さい。

13. If you don't review, please can you explain why:
テストの後フィードバックを参考にしない理由は次のうちどれですか?

	Response count
The teacher didn't tell me to review 復習する様に言われてないから。	5
I don't have enough time 時間がないから。	21
It won't improve my grade 成績に関係がないから。	4
I don't know how to review 復習の仕方が分からないから。	25
I am not interested 興味が無いから。	7
I had other more important things to do 他にやるべき事があるから。	17
I had more fun things to do 他に楽しめる事があるから。	3
Something else? その他。出来るだけ詳細に説明して下さい。	2

14. Considering the speaking test, which of these things do you think you have specifically improved this semester? スピーキングテストを考慮した上で、下記のどの項目 が特に上達したか教えて下さい。

	I have definitely improved this 間違いなく上達した項目は	I have maybe improved this 上達したかもしれない項目は	I have not improved this 上達していない項目は
Choosing correct vocabulary in conversations 会話の中で正しい単語を使う能力	31.4% (61)	**59.3% (115)**	9.3% (18)
Using correct grammar in conversations 会話の中で正しい文法を使う能力	32.1% (63)	**54.1% (106)**	13.8% (27)
Speaking smoothly – (fluently) スムーズに話す能力	**52.0% (102)**	39.3% (77)	8.7% (17)
Speaking quickly – (fluently) 早く話す能力	40.4% (78)	**46.1% (89)**	13.5% (26)
Correct pronunciation 正確な発音能力	29.2% (56)	**54.7% (105)**	16.1% (31)
Correct intonation 正確なイントネーション	30.1% (58)	**48.7% (94)**	21.2% (41)
Confidence in speaking 自信を持って話す能力	**52.8% (102)**	37.3% (72)	9.8% (19)
Speaking skills (such as starting a conversation with a stranger or explaining again if your partner doesn't understand) 会話能力(例―他人と会話を始められる)(英語で自分から話しかけるスキル)	**49.5% (96)**	42.8% (83)	7.7% (15)
Speaking on more complicated topics than before 以前よりも難しい話題について話す能力	42.3% (83)	**43.4% (85)**	14.3% (28)
Speaking with international students better than before 以前よりも国際学生と上手く話す能力	**45.1% (88)**	43.1% (84)	11.8% (23)

	I have definitely improved this 間違いなく上達した項目は	I have may-be improved this 上達したかもしれない項目は	I have not improved this 上達していない項目は
Talking about a wider variety of subjects than before 以前よりも広域の話題について話す能力	38.7% (75)	**47.9% (93)**	13.4% (26)

15. Do you think that speaking tests are a good way to improve your English? (You can choose more than one). 英語の能力を高める為にスピーキングテストは役に立つと思いますか？(複数選択可)

Yes, I can check my improvement はい。英語力の上達が確認出来ます。	62% (124)
Yes, they make me study はい。テストが勉強する動機になります。	25.5% (51)
Yes, I have a chance to speak English 英語を話す機会が持てます。	44.5% (89)
No. Can you explain?	1.5% (3)

16. We have had 8 speaking tests this semester. Do you think this number is . . . 今回のセメスターで 8 回のスピーキングテストを行いました。この回数についてあなたはどう思いますか？

too many 一多すぎると思います。	23.3% (47)
just enough 一ちょうど良いと思います。	65.8% (133)
not enough 一十分ではないと思います	5.9% (12)
I have no opinion 一特に意見がありません。	5.0 % (10)

Test Taker Attitudes to Response Time Length in Speaking Tests

Kristen Sullivan
Shimonoseki City University

Reference Data:
Sullivan, K. (2013). Test taker attitudes to response time length in speaking tests. In N. Sonda & A. Krause (Eds.), *JALT2012 Conference Proceedings*. Tokyo: JALT.

This paper forms part of a larger study concerned with the effect of response time length on responses to TOEFL iBT independent speaking tasks. Test takers are currently given 45 seconds to complete their responses to independent speaking questions. However, given the nature of spontaneous interactive speech, I question whether 45 seconds is indeed enough time for test takers to fully develop their response and demonstrate their best ability. In this study, 36 university students responded to 3 independent speaking test tasks that were allocated different response time lengths (45 seconds, 90 seconds, and 135 seconds). Participants also completed 2 sets of surveys designed to question their attitudes toward the tasks, their performance on the tasks, and their preferences regarding response time length. In this paper I look specifically at the survey results and report on test taker attitudes to TOEFL iBT independent speaking task response time lengths.

本論文は、TOEFL iBTのIndependent Speaking (IS)問題において解答時間が解答にどのような影響があるかを調べる研究調査の一部を成している。現在、IS問題の解答時間が45秒で設定されている。しかし、即興発話の性質を考慮すると、受験者が実力を出すのに45秒が果たして十分だと言えるかどうか。よって、本論文はTOEFLのIS問題の解答時間の妥当性を調査対象とする。本調査では、36人の大学生が3つのIS問題に対して解答した。それぞれの問題に異なる解答時間（45秒、90秒、135秒）が割り当てられた。また、学生は、テスト問題に対する考え方や解答の出来具合に対する反応、解答時間に対する希望を尋ねる2種類のアンケートにも答えた。本論文はこの2種類のアンケートの結果を報告し、受験者のIS問題の解答時間に対する考え方を調べる。

THE PURPOSE of this paper is to report on the initial findings of a study that is investigating the influence of response time length on responses made to TOEFL iBT (Test of English as a Foreign Language Internet-Based Test) independent speaking tasks. The larger study is investigating the effect of three different response times on test taker scores, the discourse of the responses, and test taker attitudes toward the tasks and task conditions, with the aim to identify if and how response time length influences performance. This paper includes an initial descriptive analysis of the results of two sets of surveys that investigated test taker responses to their performance on speaking tasks taken under different response time conditions, and their overall attitudes toward response time length for speaking tests.

The TOEFL iBT Test

The TOEFL iBT test is a high stakes, gate-keeping test of academic English proficiency. For students at Japanese universities, TOEFL iBT scores are commonly used as a language requirement benchmark for participation in study abroad programs at English-language universities. The Liberal Democratic Party also recently proposed that the TOEFL test be used for admission to public universities (Yoshida, 2013). The TOEFL iBT test is a highly researched exam (c.f., Chapelle, Enright, & Jamieson, 2008) that was introduced in 2005 as the successor of the TOEFL CBT (Computer-Based Test) and TOEFL PBT (Paper-Based Test). As its name suggests, the TOEFL iBT test is conducted completely by computer, with test items and responses transferred via the Internet for administration and scoring. The test is made up of four sections (reading, listening, speaking, and writing) and the speaking section, which is the focus of this paper, consists of two main types of test items: independent tasks (which require the test taker to respond to a prompt that asks them to give and explain their opinion) and integrated tasks (which require test takers to summarize written and spoken texts).

The TOEFL iBT test is the first version of the TOEFL test to include a speaking section. In the speaking section, test instructions and questions are presented to the test taker via the test station computer (both on screen and via audio recordings) and test-taker spoken responses are automatically recorded by the test station. The TOEFL iBT speaking section can thus be categorized as a semi-direct test of spoken English.

TOEFL iBT Speaking Tasks: A Background to Response Time Lengths

In this study I focused on the independent speaking tasks, in particular the free-choice speaking tasks. In these tasks, test takers are given 15 seconds to prepare and 45 seconds to give their response to a prompt that appears on their test station computer screen. The preparation and response times are shown to test takers via a countdown clock that also appears on their computer screen.

The concern I wanted to address is whether the 45-second response time limit is enough to allow test takers to demonstrate their best ability, especially in consideration of the nature of spontaneous interactive speech which is characterized by hesitations, self-repair, and "constraints of breath and spoken language processing" (Hughes, 2002, p. 77). The issue of response time length is important if we are concerned with creating test tasks which "bias for best" possible performance (Swain, 1984, pp. 195-196). It must also be considered from the viewpoint of achieving task authenticity and maintaining a positive impact on how teachers and learners practice speaking in the classroom. It is also a potential factor that can influence test taker perceptions of task difficulty and authenticity—that is, their perceptions of the face validity of the test—which can affect how the test taker responds to test items and whether they accept their test results (c.f., Wigglesworth & Elder, 2010).

While response time length does appear to be an important test condition for speaking tests, a review of the literature reveals that there is a dearth of research on the topic. Rather, the effect of planning time length has been the predominate focus of research to date. This can most probably be explained by the fact that interview tests (i.e., direct speaking tests) are more commonly in use for speaking tests rather than semi-direct tests such as the TOEFL iBT test.

While not a main focus of research, the issue of the response time length of TOEFL speaking tasks was investigated by the ETS (Educational Testing Service) in the development stages of the TOEFL iBT test. An early theoretical study concluded that the collection of larger samples of the test taker's oral proficiency would be desirable and that test takers "should have the op-

portunity to produce a total of at least 10 to 15 minutes' worth of speech for assessment" (Butler, Eignor, Jones, McNamara, & Suomi, 2000, p. 13). How the 10 to 15 minutes should be divided and allocated among the various test items in the speaking section, however, was not touched upon.

In a prototype study looking at integrated speaking test items for the TOEFL iBT test, Enright, Bridgeman, Eignor, Lee, and Powers (2008) sought to investigate whether test scores were affected by different response lengths. They found that while mean test scores were slightly higher for test takers who had 120 seconds (versus 60 seconds) and 90 seconds (versus 150 seconds) of response time, this difference was not statistically significant. This finding was used to justify the prioritizing of factors such as "domain coverage, expert opinion, availability of text materials, and cost of development" over issues of "task characteristics and administration conditions" in iBT speaking test development (Enright et al., 2008, p. 128). Indeed, the shortest response time length of 60 seconds is now in use for the integrated speaking tasks in the TOEFL iBT test.

These kinds of decisions are important for test creators who also need to consider the human and monetary resources involved in the implementation of their test items (c.f., Bachman and Palmer, 1996). The testing of speaking ability is indeed resource intensive. However, this balancing act needs to be conducted in a way that does not negatively affect the test taker's performance on and experience of the test. In regard to the TOEFL iBT speaking section this is an issue that deserves deeper consideration.

Study Aims and Design

To further investigate these issues, I aimed in this study to replicate the Enright et al. (2008) study by conducting a similar experiment, this time looking at the effect of response time length on responses to free-choice independent speaking test items. The lack of past research gave few clues to guide the choice of response time lengths to be used in this study. However, it made sense to compare the current time of 45 seconds with response times of double (90 seconds) and triple (135 seconds) this amount of time. In addition, coming from an understanding that it is important to also consider test takers' perceptions of the test and test conditions in test development and validation endeavors (Wigglesworth & Elder, 2010) participants also responded to a series of surveys questioning their attitudes toward the speaking tasks.

The participants in this study were 36 undergraduate and study abroad students studying within the economics faculty of a small, public university in rural Japan. The participants were recruited through the use of on-campus advertisements and announcements, and participation was voluntary. The difficulty in finding participants within a nonlanguage department meant that all respondents regardless of year level, proficiency level, and past experience with the TOEFL test were invited to participate in the study. There were 16 male and 20 female participants. By year level, four were 1st-year students, five were 2nd-year students, 14 were 3rd-year students, nine were 4th-year students, and four were exchange students. There were 29 native Japanese speakers, one bilingual Japanese-Spanish speaker, five native Chinese speakers, and one Turkish speaker.

Participants were required to respond to three speaking tasks taken under three different response time conditions (45 seconds, 90 seconds, and 135 seconds). Response time length, speaking task questions, and the order in which these were presented to participants were organized into different combinations to take into account any effects of item topic and task order. The experiment conditions were designed to replicate those of the actual TOEFL iBT test as much as possible, and the test item prompts were taken from official TOEFL iBT practice

tests (Educational Testing Service, 2007; Educational Testing Service, 2009).

After each speaking task participants were asked to respond to a survey questioning their attitudes toward the task and their performance on the task. Furthermore, after completion of all three tasks participants responded to another survey about response time preferences. This paper offers a descriptive analysis of responses to these surveys to gain an initial understanding of test taker beliefs about response time lengths for speaking tasks. The paper will firstly introduce the findings of the most imperative questions asked in both surveys, before discussing these findings and their implications for the speaking section of the TOEFL iBT test and response time lengths in speaking tests in general.

Results

Immediate Posttask Survey Results

The following questions were asked in each immediate post-task survey. This section will descriptively analyze test-taker responses to these questions, comparing their responses by response time length.

1. Did you have enough time to answer the question?
2. Could you complete your response?
3. Was it difficult to respond to the question?
4. Were you satisfied with your response?
5. Could you demonstrate your ability?
6. If the response time were longer would you be able to give a better response?

Adequacy of Response Time Length to Answer Questions and Complete Responses

Being able to complete one's response is obviously important not just in regard to the test taker's eventual score, but also for how the test taker feels about the test item itself. As Table 1 shows, more than half of the participants felt that they had enough time to answer the question under each response time condition. However, the 90-second response time seems to have been most adequate, with 72% of respondents saying they had enough time under this condition. In contrast, 42% of students said they felt they did not have enough time under the 45-second condition, and 31% of students responded that they had too much time under the 135-second condition.

Table 1. Question 1 Results (Did you have enough time to answer the question?)

Response	45 seconds	90 seconds	135 seconds
Too much time	1 (3%)	5 (14%)	11 (31%)
Enough time	20 (56%)	26 (72%)	23 (64%)
Not enough time	15 (42%)	5 (14%)	2 (6%)

In response to question 2 that asked test takers if they could complete their response, slightly more than half of participants felt they could complete their response under the longer response time conditions. In contrast, slightly more than half of test takers felt they could not complete their response under the shorter 45-second condition (see Table 2).

Table 2. Question 2 Results (Could you complete your response?)

Response	45 seconds	90 seconds	135 seconds
Yes	15 (42%)	21 (58%)	19 (53%)
No	21 (58%)	15 (42%)	17 (47%)

Perceptions of Item Difficulty

In regard to test taker perceptions of item difficulty, as Table 3 shows, slightly more participants felt that the items with longer response time conditions were difficult to respond to. However, it is interesting to note that more than 50% of respondents felt that it was difficult to respond to each question, regardless of response time length.

Table 3. Question 3 Results (Was it difficult to respond to the question?)

Response	45 seconds	90 seconds	135 seconds
Yes	20 (56%)	22 (61%)	24 (67%)
No	16 (44%)	14 (39%)	12 (33%)

Satisfaction with Response and Performance

As Table 4 shows, the majority of students were dissatisfied with their response regardless of response time length. However, it is noteworthy that 83% of students were dissatisfied with their response made under the 90-second condition. This is interesting given that students were more likely to respond that they had enough time to answer the question and complete their response under this time condition, which suggests that satisfaction with response may not be related to perceptions of having sufficient time to answer the test question.

Table 4. Question 4 Results (Were you satisfied with your response?)

Response	45 seconds	90 seconds	135 seconds
Yes	13 (36%)	6 (17%)	9 (25%)
No	23 (64%)	30 (83%)	27 (75%)

For this question participants were also asked to explain why they were satisfied or dissatisfied with their response. Participant comments were coded by the researcher and the results are summarized below. For each response time condition, satisfaction with the content of the response (i.e., being able to give enough reasons, being able to give a complete response, and being able to say what they wanted to say) was given as a reason. Only for the 45-second condition was using the response time efficiently (i.e., using up all of the response time and completing the response within the set response time) offered as a reason.

As for the reasons why participants were dissatisfied with their responses, for all response time conditions not being able to complete the response and not being able to give enough information were offered. Only for the 45-second condition was not being able to respond at all (i.e., stopping halfway and having no ideas to talk about) given as a reason, and only for the 90-second condition was having time left over or finishing the response too quickly mentioned.

Here it is interesting to see that test takers nominated response content rather than grammar, vocabulary, or pronunciation as a key reason for their satisfaction or dissatisfaction with

their response. The connection between task completion and satisfaction is also of interest. It is also noteworthy that students were concerned about how they used their response time, and that there was a perception that finishing early is not good. This is all important because reasons for test taker satisfaction or dissatisfaction are presumably connected to how test takers perceive their response will be scored, and these perceptions can affect how they approach making their responses, and in larger terms, how they actually prepare for the test.

For question 5, which asked test takers if they felt they could demonstrate their ability through the task, the response patterns were quite consistent over all response time conditions. Regardless of response time length, approximately two-thirds of test takers felt that they could not demonstrate their ability through the test task (see Table 5).

Table 5. Question 5 Results (Could you demonstrate your ability?)

Response	45 seconds	90 seconds	135 seconds
Yes	13 (36%)	14 (39%)	14 (39%)
No	23 (64%)	22 (61%)	22 (61%)

Potential to Improve Performance with Longer Response Time

Contrary to the very similar response patterns that we saw for question 5, for question 6, which asked participants if they could improve their performance with more response time, we see a split in responses between the 45-second condition and the 90-second and 135-second conditions. As Table 6 shows, after completing tasks under the longer response time lengths, approximately two-thirds of test takers felt that they would not be able to do better with more time. In contrast, after making their response under the 45-second condition approximately two-thirds of test takers felt that they could improve with more time. This suggests that while test takers felt that 45 seconds was not enough time, they thought that a response time of 90 to 135 seconds was sufficient.

Table 6. Question 6 Results (If the response time were longer would you be able to give a better response?)

Response	45 seconds	90 seconds	135 seconds
Yes	23 (64%)	11 (31%)	12 (33%)
No	13 (36%)	23 (64%)	22 (61%)
Not sure	0 (0%)	2 (6%)	2 (6%)

For this question too, participants were asked to explain why they believed their response would or would not improve with more time. A common reason given for all time conditions was that they could not think of any further ideas to talk about. For the 90- and 135-second conditions a limitation in English ability was given as a reason. In addition, several participants responded that if anything they would prefer more preparation time.

For test takers who answered that their response would improve with more time, for the responses made under the 45- and 90-second conditions, being able to add more information was given as the primary reason, which parallels the importance given to response content raised in question 4 about response satisfaction. Interestingly, for both the shortest condition (45 seconds) and the longest condition (135 seconds) test takers

said they could improve their response with more time because they would be able to prepare or think more about their answer. Only for responses made under the 45-second condition was being more relaxed given as a reason, which suggests that for certain test takers the shorter time condition caused some level of anxiety.

Overall Attitudes to Response Time Lengths

After completion of all three tasks, students were asked to respond to a final questionnaire that surveyed their overall attitudes to the speaking tasks, particularly the response lengths. Here student responses to three pertinent questions asked in the survey will be examined:

1. Under which response time could you best demonstrate your ability?
2. For this kind of speaking test, how much response time do you need to best demonstrate your ability?
3. What do you think about having response time limits in speaking tests like this one?

Best Response Time Condition to Demonstrate Ability

Data in Table 5 shows that regardless of response time condition, participants were generally pessimistic about their performance with approximately two-thirds of test takers responding that they could not demonstrate their ability through the tasks. Thus, it is interesting that after completing all tasks the majority of test takers felt more positively about their performance under the longer time conditions with 56% responding that they could best demonstrate their ability under the 90-second condition and 28% under the 135-second condition (see Table 7).

Table 7. Response Time Length That Allowed Test Taker to Best Demonstrate Ability

45 seconds	90 seconds	135 seconds
6 (17%)	20 (56%)	10 (28%)

Test takers were also asked to explain their response to this question. For the 17% of students who nominated the 45-second time condition, the reason given was that they did not have that much to say anyway and would thus find it difficult to speak for longer. For the 90-second condition, feeling that they had enough time to complete their response, give reasons, and say what they wanted to say, without having too much time left over, were common reasons given by test takers. In addition to this, and in common with responses given by students who nominated the 135-second condition, were the ideas of having enough time to think about their response and not feeling pressured. This repeats the connection between response time length and anxiety levels first raised in question 6 in the immediate posttask survey. A common theme among these responses is the importance of having the appropriate amount of time to complete one's response. It seems that for students with not much to say this meant a shorter response time, and for students with more to say this meant a longer response time.

Preferred Response Time Length

This question asked test takers to nominate their own preferred response time length and to explain this preference. Test taker responses are summarized in Table 8.

Table 8. Preferred Response Time Length

Less than 60 seconds	60 seconds	90 seconds	Around 120 to 150 seconds
1 (4%)	7 (26%)	8 (30%)	11 (41%)

Note. For purposes of simplification, six responses are not included in Table 8: one participant responded 45 or 90 seconds; two participants responded 60 to 90 seconds; one participant responded 45 or 135 seconds depending on the topic; one participant responded more than 90 seconds; one participant responded 5 minutes. In addition, three participants did not respond with a particular time length; two of these participants responded that they would prefer more pretask preparation time. Thus for this question $n = 27$.

As Table 8 shows, 96% of test takers felt that 60 seconds or more is required to best demonstrate their ability under this type of task. The majority of test takers thus did not believe that the current response time length of 45 seconds was enough to do their best. The main reason given for this preference was that the longer response time lengths would enable online planning. Other reasons were that the longer response times would allow them not to feel pressured, to give more detail, to give a logical response, and to deal with problems with delivery.

Test Taker Opinions About Time Limits in Speaking Tests

Given the trend toward test taker preferences for longer response time lengths, one may assume that test takers would prefer speaking test tasks with no specific time limits. However, perhaps surprisingly, the majority of test takers responded that they do not mind the presence of time limits (see Table 9). Indeed, many replied that having a time limit gives an indication of examiner or scorer expectations, and thus provides them with a goal for the task. Another reason was that a time limit forces the test taker to get to the point, again having a goal-setting function. Interestingly, several students also referred to issues of test practicality and fairness in their responses, which shows that test takers are very sophisticated in their understanding of the behind-the-scenes issues that go along with high stakes assessment.

Table 9. Test Taker Opinions about Time Limits in Speaking Tests

Time limits are okay	I would prefer no time limit	I don't care
25 (69%)	10 (28%)	1 (3%)

For the 28% of test takers who responded that they would prefer not to have a response time limit, the most common reason was that it is a distraction and puts pressure on the test taker, which is an issue that has already been touched upon above. Another reason was that the necessary amount of response time depends on the question itself, suggesting that test taker reactions to item content are also a factor that should be considered.

Discussion

These initial descriptive results suggest that most test takers are not satisfied with the current TOEFL iBT independent speaking response time limit of 45 seconds and believe that they can perform more strongly under longer response time conditions. It seems that for most test takers 45 seconds is not enough to make a complete response, especially in terms of depth of content. It also seems that for some, the shorter time length can result in enhanced feelings of anxiety, which could affect response qual-

ity in various ways. Concern about finishing early and having time left over is also an indication that many test takers believe that doing so will have a negative effect on their score; indeed, many noted that having a time limit gives an important indication of the type of response that is being asked of them. This is an important reminder of the need for test developers to clearly state expectations and criteria. If they do not, students will be forced to come up with their own conclusions, which may negatively impact how they go about responding to the task and actually preparing for the test.

Future Research

The results discussed in this paper are purely descriptive and must still be subjected to further statistical analysis. Moreover, it is important to remember that these results indicate only how students reacted to the item characteristics; analysis of the effect of response time on scores and actual performance is still to be conducted and will be reported in a future paper.

In their study on integrated speaking tasks, Enright et al. (2008) concluded that a lack of a statistically significant difference in scores for responses made under different time conditions allows test creators to prioritize practicality concerns when deciding test conditions. However, given the high stakes nature of the TOEFL iBT test and the fact that test taker attitudes toward a test may affect not only how they perform on the test day but also how they prepare for the exam, greater consideration should be given to the impact that shorter response time lengths may have.

The inclusion of a speaking section in the latest version of the TOEFL test is a great step forward. Now that the test is in use, more research is necessary to ensure that the TOEFL iBT test provides a positive experience for test takers, both on and in the lead up to the test day.

Bio Data

Kristen Sullivan is a lecturer at Shimonoseki City University and cowriter of *Impact Conversation 1 & 2*. She is interested in the teaching, learning, and assessment of speaking, as well as interactions between language learner identity and language use. <kris@shimonoseki-cu.ac.jp>

References

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Butler, F., Eignor, D., Jones, S., McNamara, T., & Suomi, B. (2000). *TOEFL 2000 speaking framework: A working paper*. TOEFL Monograph Series, Report No. 20. Princeton, NJ: Educational Testing Service.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language*. New York: Routledge.

Educational Testing Service. (2007). *The official guide to the new TOEFL iBT*. New York: McGraw-Hill.

Educational Testing Service. (2009). *The official guide to the TOEFL test* (3rd ed.). New York: McGraw-Hill.

Enright, M. K., Bridgeman, B., Eignor, D., Lee, Y., & Powers, D. E. (2008). Prototyping new assessment tasks. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 97-144). New York: Routledge.

Hughes, R. (2002). *Teaching and researching speaking*. London: Pearson Education.

Swain, M. (1984). Large-scale communicative language testing: A case study. In S. J. Savignon (Ed.), *Initiatives in communicative language teaching: A book of readings* (pp. 185-201). Reading, MA: Addison-Wesley.

Wigglesworth, G., & Elder, C. (2010). An investigation of the effectiveness and validity of planning time in speaking test tasks. *Language Assessment Quarterly, 7*, 1-24.

Yoshida, R. (2013, April 8). LDP panel binds TOEFL to degrees. *The Japan Times*. Retrieved from http://www.japantimes.co.jp

Testing Interactional English Conversation Skills in a University Speaking Exam

William Collins
Nagasaki University

Reference data:

Collins, W., (2013). Testing interactional English conversation skills in a university speaking exam In N. Sonda & A. Krause (Eds.), *JALT2012 Conference Proceedings*. Tokyo: JALT.

In this study I looked at ways to help learners develop conversation skills such as turn-taking, backchanneling, using clarifying cues to repair communication breakdown, and making comments. The study was conducted over 2 semesters and concerned the ability of students to use the skills in audio recordings under varying degrees of preplanning limitation. Students in both semesters took a speaking exam, then transcribed the recordings and self-evaluated their use of the different skills. The students in the second semester periodically did 2 reflective-listening exercises. The results suggest that the exercises had some positive impact on improving the students' use of the conversation skills.

本論は、順番交替, 相槌、発言や理解に関わる問題の修復等の会話技能を学生に身に着けさせる方法に関する研究を報告する。1年間、2つのクラスの学生が会話を定期的に録音し、会話形式の期末試験では、ペアで4つの会話を録音した。それぞれの会話がだんだん、事前練習が制限されることで、本当の会話状況でどれだけその技能を使えたかを研究した。2学期目の学生グループが学期中反省的なリスニング学習をし、それが学生の試験結果にどのような影響を与えたかという点も本研究で報告する。

One of the most difficult challenges faced by EFL students is learning how to manage conversational interaction (Kramsch, 1986; Young, 2008). Pragmatic skills such as getting, holding, and keeping a turn, backchanneling, using clarifying cues to repair comprehension breakdown (Barraja-Rohan, 2011), and giving feedback comments (Mori, 2002) are difficult for students because they must use them under the time pressure of a conversation. Learners have a strong desire to improve their conversation skills in their L2. Given that the language classroom is the default setting in which EFL students will do this, teachers must find ways to ensure that the learning activities approximate as closely as possible the conditions of actual conversation if students are to become accustomed to real-time conversation.

This paper reports the results of a 1-year study concerning testing improvements in conversational interaction skills, conducted in a 1st-year university English communication class. The key components of the study were (a) a set of active-listening comment and clarifying strategies for increasing the participation of the listener in the conversation; (b) regular pair-recording and, in the second semester, a set of post-recording reflective-listening activities for raising students' awareness of the strategies; and (c) a speaking exam in which pair-recording was done under varying degrees of preplanning limitation including controlling whether

students had previously recorded on the given speaking topic or listened to their partner's story. The study was conducted in Nagasaki University's English Communication courses for 1st-year non-English majors. Prior to the introduction of pair-recording, exams were written that focused on vocabulary and written cloze exercises based on conversation dialogues. In student surveys I conducted, students expressed a desire for more in-class speaking practice and also said they did not know how to tell if their speaking skills were improving. Against this background, I sought a way students could increase in-class speaking time and a tool for measuring improvements in conversations.

Literature Review

Cook (1989) stressed the continuum in spoken discourse between more "one-way" speech, and discourse which has a high degree of *reciprocity*, which he defined as discourse in which "the receiver can influence the development of what is being said." (p. 60). Cook argued that the distinction between written and spoken discourse is a matter of degree and can be placed on a cline defined by several criteria: planned—unplanned, socially structured—less socially structured, aided by writing—unaided by writing, and less reciprocal (one-way)—more reciprocal (two-way). Cook argued further that one of the hardest aspects of teaching conversation is the gaining, holding, and yielding of turns.

Sinclair and Coulthard (1975) first proposed the Birmingham Model for analyzing classroom interaction. In this framework, classroom discourse was divided into five ranks: lesson, transaction, exchange, move, and act. The three types of exchanges were eliciting, informing, and directing and there were three parts or moves to an exchange: initiation (opening move), response (answering move) and follow-up (acknowledging move). Finally, moves were further subdivided into acts, the smallest unit of spoken discourse. In the following exchange from Brazil (1995) each teacher and student turn is called a *move*, the units marked within slash marks are called *acts*, each three-turn unit (question/answer/follow-up) is called an *exchange* and the continuation of this sequence about pharaohs and pyramids until the teacher moves on to another segment of the lesson is called a *transaction*.

T: They were pharaohs. / Erm do you know anything about them? / They were great for building something you make in math. /

S: Pyramids

T: Pyramids yes. / Why did they call them pyramids? / Paul. / (Brazil, 1995, p. 17)

Francis and Hunston (1992) modified this framework so that it was "flexible and adaptable enough to cope with a wide variety of discourse situations [including] casual conversations between friends." (p. 123). The ranks in Francis and Hunston's framework were: interaction, transaction, exchange, move, and act. This framework was applied in the current study to student interactions in the form of recorded conversations in which one student told a story about a personal experience and the other listened and responded using backchanneling, rejoinder, and comment acts. Collectively, these three types of listener-response acts will be termed active-listening strategies in the study (see Figure 1).

Rost (2002) argued that "collaborative listening, in which learners interact with each other, is established as a vital means of language development" (p. 143). Rost identified comment strategies including (a) responding—providing a personal, relevant response to information or ideas presented, and (b) inferencing—drawing inferences based on incomplete information.

Barraja-Rohan (2011) stressed the importance for ESL students of learning interactional skills such as the turn-taking system,

self-repair, and displaying common understandings. She argued that explicit instruction and practice of conversation analysis techniques and having students analyze transcriptions of their own talk can aid them in becoming more aware of such interactional skills. Displaying common understandings was also stressed in Mori (2002), who discussed the adjacency-pair highly typical of talk-in-interaction in which the listener acknowledges the comment or answer the speaker has given by repeating all or part of the speaker's words or by producing explicit assessments such as "That's amazing." In Mori, students did not produce assessments as much as they could have but rather fell into a question-answer-question interview pattern. Hyvärinen (2008) stressed the particular relevance of evaluation feedback by the listener in narratives such as the storytelling conversations in the current study. The importance of repair in maintaining sequential development of talk-in-interaction in the face of comprehension breakdown was emphasized by Hutchby and Wooffitt (1998) who found examples of both self-initiated repair and other-initiated repair in their data.

Collins and Ruhl (2008) explored the impact of pair-recording and active-listening on students' enjoyment of and confidence in their English conversations. Students in the study reported that pair-recording and active-listening helped them enjoy English more and improved their conversations. Washburn and Christianson (1996) argued that pair-taping helped students achieve higher fluency and listening comprehension.

Research Questions

1. Given regular practice using active-listening comment and clarifying strategies in rehearsed conversations throughout the semester, how well would students be able to use them in an unrehearsed conversation?
2. How would reducing the degree of planning allowed for the storytelling conversations affect the turn-taking dynamics between speaker and listener, particularly the speaker's ability to recognize and respond to listener clarifying cues, such as word and sentence repetition, and the listener's ability to get a turn?
3. What impact would reflective listening activities, such as writing comments while listening back to recorded conversations and editing and redoing conversations, have on students' subsequent ability to use the strategies?

Method

Participants

The study was conducted at Nagasaki University in Japan with 70 medical students over 1 year in a 1st-year English Communication class. Thirty students participated in the first-semester group and 35 in the second. The students' proficiency levels ranged from high intermediate to advanced, based on the results of a university-administered G-TELP (General Test of English Proficiency), an English proficiency test similar to the TOEIC that tests reading and listening skills. The first-semester group had a mean score of 227.6, equivalent to 525 TOEIC, while the second-semester group had a mean score of 233.4, equivalent to 539 TOEIC (Ogasawara, in press). The two student populations were of comparable age, ethnic, and linguistic background.

Methodology

The study employed the Francis and Hunston (1992) adaptation of the Birmingham Model of spoken discourse analysis to measure degree of listener-participation and speaker-response in recorded storytelling conversations. In both semesters of the study, students regularly recorded their conversations. In the second semester of the study, two reflective listening activities

were introduced to assess their impact on improving students' awareness of and ability to use the active-listening and clarifying strategies. This was evaluated by comparing the student exam results of the class in the first semester that did not do the reflective listening activities (the control group) with the second-semester group that did (the study group).

Speaking Exam Format and Evaluation Criteria

For the speaking exam students recorded eight conversations, four as speaker and four as active-listener. In each conversation one student told a story about a personal experience while the partner listened and responded with active listening. For the first recording as speaker, the students were permitted to choose a partner they had already recorded with and retold the same story. For the other three recordings as speaker, I assigned a new topic and a partner with whom the speaker had not previously recorded. Students were not permitted to use any notes during recording. A sample of the story topics is shown in Appendix A. Grading was according to the following criteria:

As speaker (storyteller)

1. Was the student able to tell her story smoothly without too many pauses?
2. Was the content of the story enough (i.e., did it include information like when and where the story took place and who was in the story, as well as an evaluation, such as whether it was a good or bad experience, what the student learned from it or how it changed or affected the student)? Was it long enough (at least 2 minutes)?
3. Did the speaker notice when her partner signaled she didn't understand? Was she able to make her partner understand?
4. When her partner made a comment or asked a question, did the speaker pause to acknowledge it before continuing with her story?

As listener (active-listener)

Was the listener able to

1. use not just the backchanneling strategies, but also the comment strategies (see active-listening strategies below);
2. use the advanced active-listening strategies; and
3. signal when she didn't understand her partner and get clarification?

The focus of the current study was on the exam results concerning listener evaluation criteria and speaker evaluation criteria (3) and (4). Since the objective was to examine whether the students could use the strategies in unrehearsed conversations as well as in rehearsed conversations only the raw data reported in Table 1 will be discussed.

Active-Listening Strategies: Exam's Focus on Storyteller-Listener Interaction

The speaking exam in the study aimed to measure improvements in students' ability to use the active-listening strategies in recorded conversations over the course of the semester. In this section the target listening strategies, that were the focus of the exam, and the two reflective-listening activities used with the study's second-semester group are introduced.

Active-Listening Strategies

The strategies were taught and practiced in class and consisted of "basic" and "advanced" strategies (Collins and Ruhl, 2008). The basic strategies included *backchanneling* ("Oh yeah?" / "Oh really?" / "Uh-huh"), *comments or rejoinders* ("That's + adjec-

tive" / "Wow!" / "No Way!" / "Oh no!"), and *clarifying cues* (repeating an unfamiliar word or phrase). The advanced strategies consisted of personalizing, speculating, and generalizing, examples of which are shown in Figure 1.

Interlocutor's Comment: **I recently went camping with some friends.**
Personalizing
[Oh, I ---- too.]
Oh, I enjoy camping too. / Oh, I recently went camping too
[Oh, I ---- but ----]
Oh, I like camping, but I haven't gone recently
[Oh really? (In my case) ----]
Oh really? I have never been camping. / I want to go camping!
Generalizing
[(doing----)] is (adjective), isn't it?
Going camping is fun, isn't it?
[(Noun)] is (adjective), isn't it?
Camping is fun, isn't it?
It's (adjective) [(to do---- / doing----)] isn't it?
It's nice to go camping, isn't it?
[I think a lot of people ----]
I think a lot of people go camping in spring.
[I've heard that ----]
I've heard that camping is very nice this time of year.
Speculating
[I guess ----]
I guess that was (a lot of) fun. / I guess you had a good time.
[I bet ----]
I bet that was (a lot of) fun. / I bet you enjoyed that.
[---- must have been ---- / ([done]----)]
That must have been fun. / You must have had a good time.
Question
How ----
How is it going? / So how do you like it? / How was it?
What ----
What was that like? / What happened next?
Repeat sentence (to show surprise, strong feeling).
Only one person came? / You forgot your wallet?

Figure 1. Advanced Active-Listening Strategies

Regular Story-Conversation Pair-Recording

In addition to the recordings made for the speaking exam, a portion of each regular class throughout the semester was set aside for pair-recording of story-conversations. In the second semester of the study, students periodically transcribed these; after each transcription, they used the transcripts as the basis for two reflective-listening exercises.

Reflective Listening Exercise 1: Student Comment Protocols

In the first exercise, while listening back to their recorded conversations and reading their transcriptions, the listeners identified points in the conversation where they had wanted to get a turn but had been unable to and wrote comments in their

L1 as to the reason why. Appendix B shows one student's comment protocol.

Reflective Listening Exercise 2: Editing and Redoing Conversations

This activity was conducted in two stages. First, after the listener had finished the comment protocol, she wrote down ideas for comments she planned to make in the second recording. Then the students recorded the conversation again. The listener was not permitted to refer to notes during the recording. The transcript for this student's second recording is shown in Appendix C.

After the second recording, the speaker who told the story listened back to the conversation, transcribed it, and noted spots in the conversation where she had failed to respond to listener conversational cues. The aim of this activity was to further improve the reciprocity of the interaction, so that not only did the listener respond to the speakers' talk, but the speaker also managed to pause in her storytelling to respond to the listener's comments. After completing this review, the students recorded the conversation for a third and final time. Appendix D shows the student-transcription of the final conversation.

Speaking Exam: Preplanning Limitations and Blend of Self- and Teacher-Evaluation

The evaluation for the speaking exam was done in two stages. First, the students transcribed their conversations and, using Francis and Hunston's (1992) rank-scale, identified types, number, and mean-length of listener acts (see Speaking Exam section below). Then I checked the data and assigned a number score of between 60 and 100 based on how well each student met the evaluation criteria enumerated above (see Speaking Exam Format and Evaluation Criteria). Students recorded eight storytelling conversations in two 90-minute classes at the end of the semester for the speaking exam. For two of the recordings (one as speaker, one as listener) they were allowed to choose the topic and partner and practice the conversation before the exam, while the other three partners and topics were assigned by me at the start of the exam. The amount of planning and rehearsing that was permitted students before recording was incrementally reduced in the four conversation recordings. The topics and partners were chosen to ensure that they met the following conditions:

Recording 1: Speaker and listener had previously recorded together on same topic. (Retelling of same story with same partner.)

Recording 2: Both speaker and listener had recorded on the topic but with different partners.

Recording 3: The speaker had recorded with a different partner on that topic but the listener had not recorded on this topic before.

Recording 4: Students were assigned a partner with whom they had not recorded any conversations during the semester and given a random story topic.

Measuring Listener Participation: Student-Compiled Data

To measure improvements in listener participation and speaker response to listener cues, the students compiled data based on their conversation transcripts. I subsequently checked the data. (The results needed some adjustments in the data originally reported by the students and will be discussed later.) The data measuring listener participation included *ratio of listener back-channeling acts to comment acts, ratio of basic comment to advanced*

comment acts and *mean word-length of comment acts*. In addition, as a way to measure effectiveness of listeners' use of clarifying cues, students identified points in each conversation where they had not understood something their partner had said. This was then used to calculate the *ratio of the number of times students noted inability to understand the speaker to the number of successful clarifications*. Finally as a way to measure reciprocity, the *ratio of acknowledged to unacknowledged comment moves* was calculated.

Findings: Speaking Exam Results

The students transcribed only the four conversations they participated in as listener. The first three ratios were based on student counts of their moves, classification of them as back-channeling or comment, and advanced or basic comments. The listener also noted parts in each of the four conversations where they had not understood what their partner said to calculate comprehension breakdown to successful clarification ratio. Finally, the listener identified instances in each conversation of acknowledged and unacknowledged comment cues.

Teacher Verification and Adjustment of Student-Reported Data

The following error-distribution in counting and classifying acts were found. Thirty-two percent of students in the first semester and 27 percent in the second semester confused moves with acts, resulting in undercounts of the total number of listener-acts. Twelve percent of first-semester and 9 percent of second-semester students incorrectly classified comment acts as backchanneling, resulting in undercounts of comment acts in both semesters. Ten percent of first-semester and 8 percent of second-semester students incorrectly identified basic comment acts as advanced, resulting in over-counts of advanced acts in both semesters.

I collected the students' transcriptions and the results of each of the five measurements and determined the mean for each of the measurements. The results for the two semesters of the study are shown in Tables 1 and 2.

First-Semester Findings

The results for the first semester of the study showed sharply different levels of listener participation between the first recording and the remaining three. The first conversation was recorded with a high degree of planning; the two partners had already recorded on this same topic previously. Having already worked through negotiation of meaning and clarification in the initial recordings, the listeners reported no instances of communication breakdown in their re-recording on the speaking exam. As the degree of preplanning was reduced in each successive recording, the mean number of listener acts (obtained by adding together the backchanneling and comment acts) decreased steadily from 27 to 22 to 18 to 15, and the listener used fewer comment and more backchanneling acts. The mean ratio of backchanneling to comment acts and of basic to advanced comment acts also shifted towards the less active end of the continuum. The mean word-length of listener comment acts also decreased (the latter uptick resulting from fewer comment moves). The incidence of listener comprehension-breakdown in the second through fourth recordings averaged between 3 and 4, and in each the listener was unable to use clarifying cues to negotiate meaning with the speaker. Finally, in the ratio of acknowledged to unacknowledged listener comment cues, the number of unacknowledged cues averaged between 3 and 4, while the number of acknowledged cues declined from 2 to between 1 and 0.

Table 1. Mean Results for Five Measurements of Four Recordings (First-Semester Exam)

Type of measurement	First	Second	Third	Fourth
Listener backchanneling to comment acts ratio	15:12	12:10	12:6	10:5
Basic to advanced comment acts ratio	10:2	8:2	5:1	4:1
Comment act mean word-length	3.5	2.8	2.6	2.7
Comprehension breakdown to successful clarification ratio	0:0	3:0	3:0	4:0
Acknowledged to unacknowledged comment ratio	2:4	1:3	0:4	1:4

Second-Semester Findings

There was greater listener-participation across the five measurements in the second-semester group of students who had regularly done the two reflective-listening exercises. As the degree of preplanning was reduced in each successive recording, the number of listener acts decreased slightly from 35 to 31 to 29 to 28, a much smaller decrease than in the first-semester group. The decline in the mean number of comment acts, from 18 to 14, was smaller in the second-semester group than the first-semester group's 13 to 5. The mean number of advanced active-listening comments the listeners were able to use decreased over the four recordings, but the mean number of advanced comments was higher in each recording compared to the first-semester group. Even with no preplanning (fourth exam), students were still able to produce a mean of 3 advanced comments, compared to just one comment for the first-semester group. As with the first-semester group, there was a decline in the mean word-length of listener comment acts, but the mean for all four recordings was higher than in the first-semester group. The mean number of successful clarifications during instances of reported communication breakdown was also higher, averaging between 2 and 3 in the second group compared with 0 in the first. Finally, in the ratio of acknowledged to unacknowledged listener comment cues, the number of acknowledged cues averaged between 4 and 3 in the second-semester group compared with between 2 and 0 in the first-semester group.

Table 2. Mean Results for Five Measurements of Four Recordings (Second-Semester Exam)

Type of measurement	First	Second	Third	Fourth
Listener backchanneling to comment acts ratio	17:18	16:15	15:14	14:14
Basic to advanced comment acts ratio	12:6	10:5	10:4	11:3
Comment act mean word length	4.1	3.9	3.6	3.5
Comprehension breakdown to successful clarification ratio	0:0	3:2	3:2	3:3
Acknowledged to unacknowledged comment ratio	4:1	3:3	3:2	4:1

Discussion

The findings obtained in the study provided evidence to help answer each of three research questions raised at the outset of the paper. The first question concerned ability to use the comment and clarifying strategies in unrehearsed conversation. Comparison of the first-semester control groups' exam results for the first recording (rehearsed) with the other three recordings (unrehearsed) revealed a sharp decline in all of the measures of strategy use, suggesting that explicit instruction and regular recording alone were not sufficient to enable students to use the strategies in unrehearsed conversations. The second question dealt with the impact of reduced planning on the turn-taking dynamic. In the first-semester exams, the decline in both the absolute number of comments and the ratio of basic to advanced comments as well as the rise in both unsuccessful clarification cues and unacknowledged comment cues suggested that reduced planning sharply diminished the contribution of the listener to the conversation. The final research question looked at the impact of regular reflective listening activities on students' ability to use the strategies. The higher degree of listener participation across all four measurements in the second-semester students' recordings suggested that those students' regular performance of reflective-listening activities such as writing comments, editing conversations, and redoing conversations had a positive impact on students' ability to use the target strategies.

Limitations

There are a number of limitations to the current study. The decision to base the findings on student-compiled data raises questions concerning the reliability of the findings. While the different categories were carefully defined and repeatedly discussed and practiced with students, individual perceptions of what constituted a countable move, or which type of move it was, were subject to some degree of imprecision and error. While the author checked all of the students' data, there is a need to establish reliability estimates for the data obtained. The criteria for determining what was an unacknowledged cue was necessarily subjective as it was based on each individual listener's sense of which of their own cues called for some kind of acknowledgement from the speaker and which were more naturally passed over by the speaker without sacrificing the reciprocity that made for satisfying interaction. Finally, the relatively small student population and their high proficiency-level makes it difficult to infer a general efficacy of reflective-listening exercises in improving listener-participation.

Conclusion

The findings in the current study suggest that the study group's ability to use the active-listening strategies in unrehearsed conversations was improved by the reflective-listening activities and the student analysis of their own transcribed conversations using the rank scale of spoken discourse in the Francis and Hunston (1992) framework. The wide disparity between the first rehearsed recording and the other three unrehearsed ones in the control group's ability to use the various conversation skills was not seen in the study group's exam results. In the three recordings where the students had not recorded together, the frequency of listener response acts, ratio of comment to backchanneling acts, and ratio of advanced to basic comment acts was consistent with the first recording. Listeners also had a similar rate of success in using clarifying cues to repair comprehension breakdown and receiving speaker acknowledgement to their comment-cues.

Student feedback suggested that the reflective-listening exercises were helpful in raising the awareness of both speakers and listeners in ways to maintain speaker-listener reciprocity. Listen-

ers commented that transcribing, editing, and redoing their conversations helped them understand the advanced comment cues and how to use them and also helped them more assertively signal comprehension breakdown. The speakers similarly commented that the editing exercise helped them notice and respond to the signals better.

Future research should test the findings with a larger student group. Further, students' perceptions of how to distinguish between cues that need speaker acknowledgement and those that can be passed should be explored in greater detail. Finally, larger samples of transcribed student conversations might prove fruitful in developing a comprehensive corpus of EFL learner language so that patterns in speaker and listener interaction could be studied in greater depth.

References

Barraja-Rohan, A. M. (2011). Using conversation analysis in the second language classroom to teach interactional competence. *Language Teaching Research, 15*, 479-507.

Brazil, D. (1995). *Classroom and spoken discourse*. Birmingham, UK: University of Birmingham Press.

Collins, W., & Ruhl, D. M. (2008). Speaking and listening skills through storytelling, talking journals, and active listening. In K. Bradford Watts, T. Muller, & M. Swanson (Eds.), *JALT2007 Conference Proceedings*. Tokyo: JALT.

Cook, G. (1989). *Discourse*. Oxford: Oxford University Press.

Francis, G., and Hunston, S. (1992) Analyzing everyday conversation. In M. Coulthard (Ed.), *Advances in spoken discourse analysis*. London: Routledge.

Hyvärinen, M. (2008). Analyzing narratives and story-telling. In P. Alasuutari, L. Bickman, & Brannen, J. (Eds.), *The Sage handbook of social research methods*. London. Sage.

Hutchby, I., and Wooffitt, R. (1998). Conversation analysis: Principles, practices and applications. Oxford: Polity Press.

Kramsch, C. (1986). From language proficiency to interactional competence. *The Modern Language Journal, 70*, 366-72.

Mori, J. (2002). Task design, plan, and development of talk-in-interaction: An analysis of a small group activity in a Japanese language classroom. *Applied Linguistics 23*, 323-347.

Ogasawara, S. (in press). G-TELP Kokusai eiken ni yoru TOEIC sukoa no yosoku —G-TELP reberu 3 ni yoru yosokusiki no sakusei to kousatu [An inquiry into developing a formula for estimating TOEIC scores based on G-TELP scores]. In Yamasaki Tosihiko sensei go tainin kinen ronbun syu [A collection of articles commemorating the retirement of Professor Toshihiko Yamaoka]. Tokyo: Kairyudo.

Rost, M. (2002). Teaching and researching listening. Harlow, UK: Pearson Education.

Sinclair, J. M., & Coulthard, M. (1975). *Towards an analysis of discourse*. Oxford: Oxford University Press.

Washburn, N., & Christianson, K. (1996). Teaching conversation strategies through pair-taping. *Internet TESL Journal*. Retrieved from http://iteslj.org/Techniques/Christianson-PairTaping.html

Young, R. F. (2008). *Language and interaction: An advanced resource book*. London: Routledge.

Appendix A

Speaking Exam Story Topics Sampling

1. Tell me about a time you had a big change in your life,
2. A memory or experience that meant a lot to you
3. A scary experience you had
4. A time you pushed yourself to do something you didn't think you could do
5. A time you thought "I know I shouldn't do this but . . ."

Appendix B

Sample Student Comment Protocol

Note. The points where the listener didn't understand a word were left blank by the student. For convenience these parts are inserted into the listener's transcription based on the speaker's transcription. Comments made by the listener are shaded. They were originally in Japanese and have been translated.

A: I remember a great day I had

B: Uh-huh?

A: It's a precious memory for me.

B: Oh, that's good!

A: A few years ago I visited my grandfather with my family during summer vacation.

B: ----------[My partner seemed to be concentrating and I didn't want to disturb her so I hesitated.]

A: The day was special because it was his birthday.

B: Birthday? Ohh, that's nice.

A: So I thought I wanted to do something special for him.

B: Oh I see.

A: My grandfather is a farmer.

B: Oh farmer? [I couldn't think of a comment. My partner seemed to be concentrating and didn't notice.]

A: So I helped him with his farming.

B: Oh that's nice!

A: His farm was very large and he was growing a variety of vegetables and fruits.

B: Ohhh. Uh-huh. [I didn't understand well what my partner said, but thought I should say something.]

A: He told me to cut watermelons with him.

B: Oh yeah? [I couldn't think of a comment.]

A: He seemed so happy because he wanted to work in the field with me.

B: Uh-huh.

A: He knocked watermelons before he cut.

B: Not? [I didn't understand a word.]

A: I thought it was strange and asked him why he knocked them.

B: Uh-huh.

A: He answered, "When I knock them and heard good sound, they are fit to eat."

B: Oh, I see. [I couldn't understand well, but I had some idea about the general meaning.]

A: Yes. So I thought farming is interesting.

B: Yes, I think so too.

A: Then I cut them and we brought them to the market.

B: Market? Uh-huh.

A: When we got to the market I was surprised it was very lively.

B: Lively? [I wasn't sure about word, but I had some idea.]

A: He was spoken to many people and introduced me to them gladly.

B: Oh, nice.

A: I was also happy.

B: Yeah.

A: The watermelons were displayed near the front.

B: Oh that's good.

A: Right away customer bought one of them.

B: Oh really? That's great!

A: When I saw the watermelons bought, I was very happy.
B: Yes, that's happy!
A: He was happy too.
B: Yeah? [I wanted to say more but couldn't think what to say.]
A: Now I live in Nagasaki, so I rarely see him.
B: Oh yeah? That's too bad.
A: But I hear he is fine.
B: Oh that's good.
A: I want him to come to Nagasaki and I will show him various places.
B: That sounds nice!

Appendix C

Redo 1 (Conversation Recorded After Listener's Edit)

Note. Additions by the listener are shaded.

A: I remember a great day I had
B: Uh-huh?
A: It's a precious memory for me.
B: Oh, please tell me!
A: A few years ago I visited my grandfather with my family during summer vacation.
B: Oh you did? That's nice.
A: The day was special because it was his birthday.
B: Birthday? Ohh, that's nice.
A: So I thought I wanted to do something special for him.
B: Oh I see.
A: My grandfather is a farmer.
B: Oh farmer? My grandfather is a farmer too. [Personalizing]
A: So I helped him with his farming.
B: Oh that's nice! I guess he was very happy. [Speculating]
A: His farm was very large and he was growing a variety of vegetables and fruits.
B: Oh he was? Wow!
A: He told me to cut watermelons with him.
B: Oh yeah? So how did that go?
A: He seemed so happy because he wanted to work in the field with me.
B: Uh-huh.
A: He knocked watermelons before he cut.
B: He not?
A: Yes.
B: What's not?
A: Knocked! [gesture]:
B: Ohh, I see. Knocked?
A: Yes. I thought it was strange and asked him why he knocked them.
B: Uh-huh.
A: He answered, "When I knock them and heard good sound, they are fit to eat."
B: Oh, I see. They are fit to eat.
A: So I thought farming is interesting.
B: Yes, I think so too.
A: Then I cut them and we brought them to the market.
B: Market? Uh-huh.

A: When we got to the market I was surprised it was very lively.
B: Lively? What's lively?
A: Um, many customers.
B: Oh, I see. Lively?
A: Yes. He was spoken to many people and introduced me to them gladly.
B: I bet he was proud of you. [Speculating]
A: I was also happy.
B: Yeah.
A: The watermelons were displayed near the front.
B: Oh they were? That's good.
A: Right away customer bought one of them
B: Oh really? That's great!
A: When I saw the watermelons bought, I was very happy.
B: Yes, I bet you were! [Speculating]
A: He was happy too.
B: Yes, I guess he was. [Speculating]
A: Now I live in Nagasaki, so I rarely see him.
B: Oh yeah? That's too bad.
A: But I hear he is fine.
B: Oh that's good.
A: I want him to come to Nagasaki and I will show him various places.
B: That sounds nice!

Appendix D

Redo 2 (Conversation Recorded After Speaker's Edit)

Note. The places where the speaker responded to listener cues are shaded. Listener moves are noted in boldface: the number of the move, [the number of words for each move], and (the type of move). BC = backchanneling.

A: I remember a great day I had
B: Uh-huh? **1 [1] (BC)**
A: It's a precious memory for me.
B: Oh, please tell me! **(meta-move)**
A: A few years ago I visited my grandfather with my family during summer vacation.
B: Oh you did? **2 [3] (BC)** That's nice. **3 [2] (comment)**
A: The day was special because it was his birthday.
B: Birthday? **4 [1] (BC)** Ohh, that's nice. **5 [3] (comment)**
A: So I thought I wanted to do something special for him.
B: Oh I see. **6 [3] (BC)**
A: My grandfather is a farmer.
B: Oh farmer? **7 [2] (BC)** My grandfather is a farmer too. **8 [6] (comment)**
A: Ohh really?
B: Yeah! **9 [1] [1] (BC)**
A: So I helped him with his farming.
B: Oh that's nice! **10 [3] (comment)** I guess he was very happy. **11 [6] (comment)**
A: Yes, he was. His farm was very large and he was growing a variety of vegetables and fruits.
B: Oh he was? **12 [3] (BC)** Wow! **13 [1] (comment)**

A: He told me to cut watermelons with him.
B: Oh yeah? **14 [1] (BC)** So how did that go? **15 [5] (question)**
A: Good. He seemed so happy because he wanted to work in the field with me.
B: Uh-huh. **16 [1] (BC)**
A: He knocked watermelons before he cut.
B: He knocked? **17 [2] (clarifying)**
A: Yes. I thought it was strange and asked him why he knocked them.
B: Uh-huh. **18 [1] (BC)**
A: He answered, "When I knock them and heard good sound, they are fit to eat."
B: Oh, I see. **19 [3] (BC)** I've heard it's important to check fruit before eating. [Generalizing] **20 [9] (comment)**
A: Yes. So I thought farming is interesting.
B: Yes, I think so too. **21[5] (comment)**
A: Then I cut them and we brought them to the market.
B: Market? **22 [1] (BC)** Uh-huh. **23 [1] (BC)**
A: When we got to the market I was surprised it was very lively.
B: Yes, sometimes markets are very lively. [Generalizing] **24 [6] (comment)**
A: He was spoken to many people and introduced me to them gladly.
B: I bet he was proud of you. [Speculating] **25 [7] (comment)**
A: I was also happy.
B: Yeah. **26 [1] (BC)**
A: The watermelons were displayed near the front.
B: Oh they were? **27 [3] (BC)** That's good. **28 [2] (comment)**
A: Right away customer bought one of them
B: Oh really? **29 [2] (BC)** That's great! **30 [2] (comment)**
A: When I saw the watermelons bought, I was very happy.
B: Yes, I bet you were! [Speculating] **31 [5] (comment)**
A: He was happy too.
B: Yes, I guess he was. [Speculating] **32 [5] (comment)**
A: Now I live in Nagasaki, so I rarely see him.
B: Oh yeah? **33 [2] (BC)** That's too bad. **34 [3] (comment)**
A: But I hear he is fine.
B: Oh that's good. **35 [3] (comment)**
A: I want him to come to Nagasaki and I will show him various places.
B: That sounds nice! **36 [3] (comment)**

University Student Knowledge of Loanwords Versus Non-loanwords

Raymond Stubbe
Shintaro Hoke
Kyushu Sangyo University
Chris O'Sullivan
Kitakyushu University

Reference Data:

Stubbe, R., Hoke, S., & O'Sullivan, C. (2013). University student knowledge of loanwords versus nonloanwords. In N. Sonda & A. Krause (Eds.), *JALT2012 Conference Proceedings*. Tokyo: JALT.

Students (N = 408) from 3 Japanese universities took 2 vocabulary tests of their receptive and productive knowledge of English loanwords versus nonloanwords. Six loanwords (LWs) and 6 nonloanwords (NLWs) from each of the 8 JACET8000 levels were tested in a passive recognition yes-no test followed by an passive recall translation (English to Japanese) test of the same 96 items. Overall, students showed knowledge of 57% more LWs than NLWs on the yes-no test, but knew 195% more on the translation test. The differences between LW and NLW results decreased as English ability levels increased. LWs were better known than NLWs at every frequency level on the translation test and recognized more often on all but 2 of the higher frequency levels on the yes-no test. These results have implications for vocabulary teachers and testers, in terms of the differences in the learning difficulty of LWs versus NLWs, as well as the risks these differences pose for vocabulary assessment.

日本国内の3大学に通う大学生（408人）を対象に、英語における借用語と非借用語の受容的・生産的理解に関する2つの語彙テストを実施した。JACET8000における8つの頻度レベルそれぞれから借用語6語と非借用語6語の計96語を選び、受動型のYes / No認識テストと、英語を日本語に直す能動型の生産的翻訳テストを実施した。その結果、被験者はYes / Noテストにおいて借用語の得点が非借用語よりも57%高く、翻訳テストでは195%も高かった。また、英語能力レベルが低い大学の被験者ほど借用語と非　借用語の得点差が大きかった。テスト別に見ても、翻訳テストでは、被験者は全ての頻度レベルで非借用語よりも借用語を多く理解し、Yes / Noテストでも、被験者は2つの高頻度レベルを除き、非借用語よりも借用語を多く認識した。これらの結果は語彙を教える教師やテスト作成者への示唆となろう。

Loanwords (LW) can be thought of as "lexical items in two or more languages which are identified by speakers as related by their form," regardless of meaning (Uchida, 2001, p. 9). When learners pair L1 and L2 words as LWs, "they make a connection between L2 stimuli and L1 representations stored in the mental lexicon" (Uchida, 2001, p. 9). Encouraging students to notice and use English LWs common in their native language is "a very effective vocabulary expansion strategy" (Nation, 2003, p. 2).

The Japanese language contains thousands of English LWs, "many of which are well-established and in universal use" (Kay, 1995). It has been estimated that about half of the most common 3,000 words of English have some borrowed form in Japanese (Daulton, 1998). Of a random selection of words contained in the Reading Sections (parts four and five) of two official TOEIC® Bridge Practice Tests (Ashmore et al., 2007), it has been found that 53% of them were English LWs in Japanese. Stubbe (2010) suggested that LW recognition was significantly

better than nonloanword (NLW) recognition, especially among lower level students. The lower ability students in that study knew 44% of all words, but 143% more LWs than NLWs, while the higher level students knew 72% of all words but only 76% more LWs. Additionally, the low students knew only 60% of the LWs whereas the high students knew 85% of them. It is "possible that the low-level students have more difficulty in recognizing LWs which they already know in L1" (Stubbe, 2010, p. 718) than their high-level counterparts.

In the pilot to this present study (Stubbe & Yokomitsu, 2012), it was found that Japanese university students' receptive knowledge of a random selection of 60 English LWs across all levels of the JACET List of 8,000 Basic Words (JACET, 2003) (hereinafter J8000) was on average almost twice their passive knowledge of an equal number of NLWs from the same frequency levels, as measured by a yes-no checklist vocabulary test (means of 78.8% for LWs versus 40% for NLWs). That investigation also found that those same students' recall knowledge as measured by an English to Japanese (L2 to L1) translation test of the same LWs was on average three times greater than their productive knowledge of the same NLWs (46% for LWs and 13.2% for NLWs). Thus it was concluded that LW status strongly influenced student lexical knowledge across all levels of the J8000 (Stubbe & Yokomitsu, 2012).

The nearly 50% drop in item means reported in Stubbe and Yokomitsu (2012) (49.6% versus 25.0% of the full 120 items for yes-no and translation tests respectively) may suggest that students were simply overestimating their lexical knowledge on the yes-no test. However, Waring and Takaki (2003) reported a nearly 70% decrease in mean scores between a similar recognition checklist test (15.3 of 25 items) and an L2 to L1 translation test (4.6 of the same 25 items). It is possible that students taking the yes-no test in Stubbe and Yokomitsu (2012) as well as the recognition checklist test in Waring and Takaki (2003) were signaling items which they thought they recognized and believed they knew a meaning of, whereas the translation test results of both studies showed that their translations were often lacking or faulty. In other words there appears to be a considerable gap between thinking one knows a word and actually being able to produce a correct translation for that word. In the pilot study only 45.6% of the 120 items were attempted on the translation test, with 45.3% of those being incorrect (Stubbe and Yokomitsu, 2012). Waring and Takaki (2003) also conducted a multiple-choice test of the same 25 pseudowords, and reported a mean of 10.6 (42.4%). Discussing Waring and Takaki (2003), Nation and Webb (2011, p. 282) wrote:

> Thus only a small number of words were learned well (per the results of the translation test), but quite a large number were learned at least partially. If only the translation test had been given, the amount of vocabulary learning from the reading would have been greatly underestimated.

Similarly, it is possible that the students involved in the Stubbe and Yokomitsu (2012) study signaled knowledge of words which they had partial knowledge of, and this could account for a portion of the gap between those yes-no and translation test scores.

Method

In preparation for this research project a pilot study was undertaken to evaluate the words to be tested as well as the testing instruments to be employed. In this pilot, four LWs and four NLWs were randomly selected from the top half and the bottom half of each of the eight J8000 word frequency levels; for a total of 64 items for each group (Stubbe & Yokomitsu, 2012). To improve the separation between adjacent J8000 levels for the

present study it was decided to sample words only from the bottom half of each level (e.g., words 501-1000 for the 1K level). Thus half of the items used in the pilot study were eliminated from the item pool for this study. Rasch analysis using Winsteps (Linacre, 2011) was performed on the remaining 64 items for both tests to determine which words, if any, had poor model fit statistics. In total, 20 words were found to not perform well on either the yes-no or translation test or both, and were also excluded from this study's item pool. Hence, only 44 of the words from the pilot were included in this study. It was also decided to decrease the number of words tested from 128 to 96 to lessen the burden of marking the expected 400 plus translation tests. To complete the desired item pool of 96 words (6 LWs and 6 NLWs from each J8000 level), 52 additional words (25 LWs and 27 NLWs) were randomly selected as required from the eight levels of the J8000. In creating this item pool, consideration was not given to word class (nouns, verbs, etc.) primarily because LWs are usually found in Japanese as nouns (Daulton, 2008), and restricting this study to a comparison of LW and NLW nouns was deemed too restrictive and cumbersome. As it turned out, 44 of the 48 LWs were nouns, compared to the 28 NLW nouns.

These 96 items (see Appendix) were used to create two vocabulary tests, the first being a receptive yes-no vocabulary test. The second test was a passive recall test of the same 96 items from English into the students' L1. This latter test was given in part to ensure students knew a proper translation of the English words as opposed to a usage found only in Japanese (for example *trump*, which means *playing cards* in Japanese). Students were given the option not to participate in this research. The following waiver appeared on the top of both test forms, in English and Japanese:

> This is not a test. This is an optional level check. This form will help teachers better understand and improve the vocabulary program. By completing this form you agree to participate in this research. If you do not wish to participate please turn the form face down and do not mark it. Your information will be held confidentially and your responses will not be used to identify you. Your class grade will not be affected by filling in this form or not.
>
> この用紙はテストではありません。任意のレベルチェックです。レベルチェックはボキャブラリー研究の理解と向上に役立ちます。この用紙を記入することにより、この研究に参加することに同意することを意味します。参加を希望しない場合、記入せずに用紙を裏返してください。個人情報は厳守され、回答は個人の特定には利用されません。この用紙の記入の有無により、成績に影響はありません。

To maximize pairings of the yes-no and translation tests, participants were given the yes-no test at the beginning of one class in July or August, 2012, and received the translation test toward the end of that same class. Yes-no test forms were then marked by running them through an optical scanner and the resulting data was converted into an Excel file for analysis. The translation test forms were hand-marked by three markers: one of the authors and two 3rd-year students. To check interrater reliability, 30 translation test forms were copied three times and marked by each marker in addition to the other forms they marked. These 30 forms were then culled from the data pool for separate analysis. Interrater agreement between these three markers on these 30 forms was 92% on the correct and incorrect responces (test questions left blank by the participants were excluded from this analysis). The 30 test forms (10 from each rater, selected randomly) were then replaced in the data pool.

Participants

Students from 21 classes in three Japanese universities ($N = 408$), with TOEIC scores ranging from about 200 through 450, participated in this study.

Results and Discussion

Similar to the pilot study, the yes-no test mean was nearly double that of the translation test (48.3 and 25.7 of the 96 words, respectively). Standard deviations (*SD*) were 17.6 and 11.6, respectively, with scores ranging from 5-87 on the yes-no test and 0-56 on the translation test. Test reliabilities (Cronbach's alpha) were high at .96 and .92, respectively.

Table 1. Descriptive Statistics for Yes-No and Translation Tests (N = 408, k = 96)

Test	Mean	*SD*	Range	Low–High	Reliability
Yes/no	48.3 (50.35%)	17.6	82	5–87	.96
Translation	25.7 (28.60%)	11.6	53	3–56	.92

Note. k = number of words tested

Table 2 breaks the test results down by university, which are listed from the highest English ability level through the lowest level (U1-U3). Both test means had a direct relationship with proficiency level. Additionally, the amount of variance or the standard deviation (*SD*) as well as the differences between yes-no means to the translation means both had an inverse relationship with proficiency level, similar to the findings of Stubbe (2012). As ability level increased so too did test scores, while variance as well as the gap between recall and recognition knowledge decreased. This decrease in the gap between recall ability and recognition ability as proficiency levels increased was also found in Hu and Nation (2000), who observed that students comprehending 90% of the words in a text had a smaller recall versus recognition knowledge gap than students at an 80% comprehension level. Differences between the yes-no test means for the three universities were all statistically significant, as were the differences between the translation test means. A one-way ANOVAs confirmed that the differences between the university means were significant (F (2, 405) = 185.7 and 85.4, p < .0001, for the yes-no and translation tests, respectively). Post hoc analysis (Turkey HSD) revealed that the differences between all university pairings (U1 and U2; U1 and U3; and U2 and U3) were statistically significantly (alpha was set at p = .0167, using a partial Bonferroni adjustment for three comparisons).

Table 2. Descriptive Statistics by University for Yes-No (YN) and Translation Tests (k = 96)

University	***n***	**YN Means**	**Tr. Means**	**YN *SD*s**	**Tr. *SD*s**	**YN *M* / Tr. *M***
1	159	59.5 (62.0%)	35.2 (36.7%)	10.1	6.9	1.69
2	53	49.5 (51.6%)	25.9 (27.0%)	12.9	8.0	1.91
3	196	38.9 (40.5%)	17.9 (18.6%)	18.1	9.5	2.17
Overall	408	48.3 (50.3%)	25.7 (28.6%)	17.6	11.6	1.88

Note. k = number of words tested

Table 3 breaks down yes-no and translation test results by loanword status (48 LWs and 48 NLWs). On the yes-no test the LWs had 57.4% more reports than the NLWs. On the translation test, however, the LWs were known practically three times often more than the NLWs. Also, similar to the pilot to this study (Stubbe & Yokomitsu, 2012), the yes-no NLWs mean is almost the same as the translation LWs mean. In fact, post hoc analysis

(Turkey HSD) revealed that only the difference between this pairing (yes-no NLWs and translation LWs) was not statistically significant. This may suggest that the relative difficulty of recognizing NLWs versus LWs on a yes-no test is comparable to the increase in difficulty when moving from passively recognizing LWs on a yes-no test to translating LWs into Japanese on a translation test.

Table 3. Descriptive Statistics by J8000 Level for Yes-No (YN) and Translation Tests ($N = 408$, $k = 48$)

Test	Mean (%)	*SD*	Range	Low–High
YN LWs	29.6 (61.7%)	9.5	43	3–46
YN NLWs	18.8 (39.2%)	8.8	41	0–41
Tr LWs	19.2 (40.0%)	7.6	32	3–35
Tr NLWs	6.5 (13.5%)	4.8	26	0–23

Note. k = number of words tested

Figure 1 breaks down the data presented in Table 3 by J8000 frequency level (1K through 8K). With the exception of the jumps at 6K and 8K, the translation NLW results best follow the pattern predicted by word frequency level (Milton, 2009). It can be noticed that the 8K jump was common to both LWs and NLWs on both tests, and replicates findings observed in Aizawa (2006) as well as Stubbe and Yokomitsu (2012). Also contrary to frequency level expectations, LWs on both tests jumped considerably from 3K to 4K, possibly because one of the 4K LWs *helicopter* had a high score relative to the other words at those two frequency levels.

A comparison of LW results at the 7K level with NLW results at higher levels reveals that on the yes-no test 7K LW scores were higher than 4K NLW scores. On the translation test 7K LW scores were higher than 3K NLW scores. LW scores at the 8K level surpassed NLW scores at the 3K level on both tests. It is possible that loanword status may be as important as or even more important than frequency level when considering the learning difficulty of new vocabulary for Japanese learners.

This trend of LWs scores exceeding NLWs was not universal however. At the 1K level the difference between yes-no LWs and NLWs was slight and actually reversed at the 3K level. A closer look at the yes-no LWs and NLWs results by university (Table 4) revealed that at the 1K level, U1 had a slightly higher NLW mean (5.75 versus 5.90, LWs and NLWs respectively), but at the 3K level the difference was substantial (3.88 versus 4.44). On the other hand, for the mid-level university (U2), the LW and NLW means on the yes-no test were 5.21 and 5.58 respectively at the 1K level, while at the 3K level they were nearly even at 5.15 and 5.17, respectively. It appears that the LWs exceeding NLWs trend was significantly reversed for U2 at the 1K level and at the 3K level for U1. This could be due to these higher level students having a progressively better grasp of NLWs at the 1K and 3K levels. Why this reversal did not appear at the 2K level warrants further investigation. With the lowest level university (U3), LW means exceeded NLW means at all eight J8000 levels on both tests, possibly reflecting their general lack of knowledge of NLWs as suggested in Stubbe (2010).

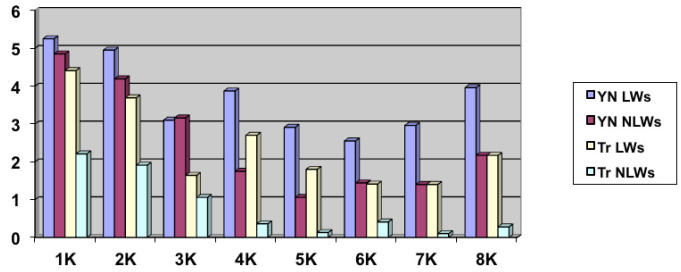

Figure 1. LW and NLW Results for Yes-No and Translation Tests

Note: The Y axis represents the mean score (maximum of 6), and the X axis represents the eight frequency levels of the J8000.

Table 4. Yes-No LW and NLW Test Results by University and J8000 Frequency Level

J8000 level	U1 LWs	U1 NLWs	U2 LWs	U2 NLWs	U3 LWs	U3 NLWs
1K	5.75	5.90	5.21	5.58	4.85	3.80
2K	5.74	5.38	5.21	4.42	4.23	3.16
3K	3.88	4.44	3.15	3.17	2.45	2.12
4K	4.64	2.58	3.74	1.60	3.29	1.13
5K	3.61	1.33	2.96	0.89	2.33	0.89
6K	3.58	1.52	2.62	1.72	1.68	1.30
7K	3.42	1.39	3.04	0.92	2.57	0.82
8K	4.55	1.81	4.23	1.02	3.42	0.89
Overall	4.40	3.04	3.77	2.42	3.10	1.76

Note. *k* = six words per level

Table 5. Translation LW and NLW Test Results by University and J8000 Frequency Levels

J8000 level	U1 LWs	U1 NLWs	U2 LWs	U2 NLWs	U3 LWs	U3 NLWs
1K	5.23	3.58	4.26	2.58	3.79	1.01
2K	4.47	2.85	3.68	1.85	3.06	1.16
3K	2.36	1.75	1.64	1.02	1.06	0.52
4K	3.35	0.72	2.58	0.26	2.19	0.09
5K	2.76	0.27	1.83	0.09	1.02	0.02
6K	2.16	0.65	1.51	0.51	0.78	0.19
7K	1.63	0.16	1.42	0.09	1.22	0.06
8K	2.69	0.54	2.40	0.21	1.68	0.08
Overall	3.08	1.32	2.42	0.83	1.85	0.39

Note. *k* = six words per level

Item Analysis

An item (or word) analysis also revealed the strength of the students' ability to recognize and translate LWs over NLWs. Only five words on the translation test scored zero: *captive, casualty, cripple, exacerbate,* and *relentless* (from the J8000 frequency levels: 7, 4, 7, 7, and 8, respectively). All of these are NLWs. Meanwhile, the top scoring words were all LWs: *park, cup, drama, corner,* and *helicopter* (from the J8000 frequency levels: 1, 1, 2, 1, and 4; with scores of 370, 369, 345, 335 and 330 of the total 408 participants, respectively). These results may explain the 4K LW and 8K jumps mentioned above and displayed in Figure 1.

A high-low item analysis, in which the 96 words were sorted according to translation score then split into two groups of 48, was also performed. Results revealed that 77% of the words in the high group were LWs, with 23% being NLWs. Naturally these percentages were reversed for the low group. Both of these

item analyses support the predominance of LWs over NLWs in students' second language lexicons.

Finally, the 28 nouns found in the 48 NLWs were compared to the 20 non-noun NLWs. Perhaps surprisingly, the non-nouns were better known (had higher mean scores) on both tests, before and after accounting for differences in J8000 level. Hence, it appears that not considering *word class* during item selection may not have unfairly biased the results reported above.

Conclusion

This study was an investigation into the recognition and recall of English loanwords in Japanese versus NLWs across all levels of the J8000 frequency listing. At the lower frequency levels (beyond 3K) on the yes-no test and at all levels on the translation test, LW knowledge was significantly greater than NLW knowledge. However, at the 1K and 3K levels on the yes-no test, the higher level university students recognized more NLWs than LWs. This result may suggest that although Japanese university students' know and recognize more LWs than NLWs, the difference diminishes at the higher word frequency levels with higher level students, whose overall vocabulary sizes are larger.

This study does suffer from a number of limitations. Although translation tests do check for student knowledge of a word's basic meaning, they do not guarantee the students can use the word appropriately. Some qualitative research, such as interviewing some of the students, could have provided a means of checking for such appropriate usage ability. As well, possible reasons behind unexpected results such as *helicopter* could be uncovered. The 92% interrater reliability amongst the three translation test markers was also a little weak. A *Facet Analysis* (Linacre, 2012) is needed to show which items were most adversely affected on the translation test. These items then could be deleted from the analysis to determine whether the results and conclusions remain valid. The selection of only six LWs and six NLWs from each J8000 could be considered too small to capture a truly representative sampling, and thus allowed for the skewing of the results. The LW jump at the 4K level, for example, was likely due to the influence of the single word *helicopter*. Sampling a greater number of words from fewer J8000 levels could help alleviate this weakness.

Despite these weaknesses, these results may have implications for both vocabulary teachers and testers. Even at the lower word frequency levels (4K through 8K) LW status does seem to have a strong influence on which words students are familiar with. Knowing which words are LWs out of a list of vocabulary to be taught or used in a classroom could help teachers better assist students in their lexical development. The LWs in a list could be reviewed first, focusing on potential variances with native-English usages, before teaching the likely more difficult NLWs. For vocabulary testers (who often rely on word frequency lists like the J8000), knowing which items are LWs while developing a test should help to better predict item performance. Not knowing which items in a test are LWs could lead to some startling results.

Acknowledgements

The authors would like to thank the staff of the Language Education and Testing Center (LERC) of Kyushu Sangyo University for their assistance in marking the translation tests.

Bio Data

Raymond Stubbe holds an MA in Applied Linguistics and TESOL from the University of Leicester, England; and a BEd from the University of Victoria, Canada. He has been a lecturer at Kyushu Sangyo University since 2009. His research interests include vocabulary acquisition and testing. <raymondstubbe@

gmail.com>

Shintaro Hoke received his MEd from Fukuoka University of Education in 2002; and his BEd from Fukuoka University of Education in 2000. He has been teaching English at Kyushu Sangyo University since 2005. His research interests include developmental education for university students, grammar acquisition, and grammar teaching methodology. <hoke@ip.kyusan-u.ac.jp>

Chris O'Sullivan has been teaching at Kitakyushu University for 10 years. He is currently interested in trilingualism.

References

Aizawa, K. (2006). Rethinking frequency markers for English-Japanese dictionaries. In M. Murata, K. Minamide, Y. Tono & S. Ishikawa (Eds.), *English Lexicography in Japan* (pp. 108-119). Tokyo: Taishukan-Shoten.

Ashmore, E., Carter, E., Duke, T., Hauck, M., Locke, M., & Shearin, R. (2007). *TOEIC Bridge kousiki gaido & mondaisyu* [TOEIC Bridge official guide & braindumps]. Princeton, NJ: Educational Testing Service.

Daulton, F. (1998). Japanese loanword cognates and the acquisition of English vocabulary. *The Language Teacher 22*(1). Retrieved from http://jalt-publications.org/tlt/issues/1998-01_22.1

Daulton, F. (2008). *Japan's built-in lexicon of English-based loanwords.* Clevedon, UK: Multilingual Matters.

Hu, H.-C., M., & Nation, I. S. P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, *13*(1), 403-430. Retrieved from http://nflrc.hawaii.edu/rfl/PastIssues/originalissues.html

JACET Basic Word Revision Committee. (2003). *JACET list of 8000 basic words.* Tokyo: Japan Association of College English Teachers.

Kay, G. (1995). English loanwords in Japanese. *World Englishes, (14)*, 67-76.

Linacre, J. M. (2011). *Winsteps* [Computer program]. Beaverton, OR: Winsteps.com. Retrieved from http://www.winsteps.com//index.htm

Linacre, J. M. (2012). *Facets computer program for many-facet Rasch measurement* (Version 3.70.0). Beaverton, OR: Winsteps.com. Retrieved from http://www.winsteps.com//index.htm

Milton, J. (2009). *Measuring second language vocabulary acquisition*. Bristol, UK: Multilingual Matters.

Nation, I. S. P. (2003). The role of the first language in foreign language learning. *Asian EFL Journal 5*(2). Available from http://asian-efl-journal.com/journal-2003/

Nation, I. S. P., & Webb, S. (2011). *Researching and analyzing vocabulary*. Boston, MA: Heinle.

Stubbe, R. (2010). Exploring the lexical challenge of the TOEIC® *Bridge*. In A. M. Stoke (Ed.), *JALT2009 Conference Proceedings* (pp. 710-712). Tokyo: JALT.

Stubbe, R. (2012). Do pseudoword false alarm rates and overestimation rates in yes/no vocabulary tests change with Japanese university students' English ability levels? *Language Testing. 29*, 471-488. Available from http://dx.doi.org/10.1177/0265532211433033

Stubbe, R., & Yokomitsu, H. (2012). English loanwords in Japanese and the JACET 8000. *Vocabulary Education and Research Bulletin, 1*(1), 10-11. Retrieved from http://jaltvocab.weebly.com/publications.html

Uchida, E. (2001). *The use of cognate inferencing strategies by Japanese learners of English* (Unpublished doctoral dissertation). Essex University, UK.

Waring, R., & Takaki, M. (2003). At what rate do learners learn and retain new vocabulary from reading a graded reader? *Reading in a Foreign Language, 15*(2), 130-163. Retrieved from http://nflrc.hawaii.edu/rfl/October2003/waring/waring.pdf

Useful Information Teachers and Administrators Should Know About the TOEIC

Brian D. Bresnihan
University of Hyogo

Reference Data:
Bresnihan, B. D. (2013). Useful information teachers and administrators should know about the TOEIC. In N. Sonda & A. Krause (Eds.), *JALT2012 Conference Proceedings*. Tokyo: JALT.

In recent years, a total of over half a million TOEIC Institutional Program (IP) Tests and TOEIC Bridge IP Tests are administered to college and university students in Japan yearly. This paper covers various issues concerning the use of the TOEIC test in schools, such as using the TOEIC, a norm-referenced test, for criterion-referenced test purposes, which results in misuses of the scores, as does using the scores without reference to their standard errors of difference. It also explains why it is unreasonable for schools to expect that most of their students' TOEIC scores will increase, even after a semester or year of studying, and points out that using the TOEIC test tends to hinder the teaching, practicing, and learning of certain English language abilities, namely those that are not specifically addressed on the TOEIC test.

近年、総計５０万以上のTOEICインスティテューショナルプログラム(IP)テストとTOEICブリッジIPテストが、日本国内の大学生に対して執り行われている。この論文では、ノームレファレンステストであるTOEICを、クライテリオンレファレンステストの目的に使用することや、スタンダードエラーズオブディファランスを参照せずにスコアを使用することにより起こる、スコアの間違った解釈等、様々な問題点を取り上げている。また、何故、半年や1年の学習後でさえ、ほとんどの学生のスコアが上がることを学校が期待するのが不合理なのか説明し、TOEICの使用が、ある種の（TOEICで特に取り上げていない）英語能力の指導や練習や学習の妨げになっていることを指摘している。

Before the mid 1990s, the TOEIC test, produced by the Educational Testing Service (ETS), was practically unknown outside Japan's business community, for which it was created. Now it is the most sat for test of its kind in the world. In Japan, students of all ages are now sitting for it, or for the TOEIC Bridge test, which is a version for those with lower English language abilities. Its use has become particularly pervasive in Japan's colleges and universities, increasing every year. Over 440,000 TOEIC (IIBC, n.d.b, p. 8) and over 100,000 TOEIC Bridge (IIBC, n.d.a, p. 7) tests were administered in 2011, the last date for which statistics are available.

The standard TOEIC, the concern of this paper, tests listening and reading. Speaking and writing are not tested. (The TOEIC Speaking Test and the TOEIC Writing Test are totally separate tests.) Test takers receive a Listening score, a Reading score, and a Total score, which is simply the first two scores added together. There are two types of administrations: Secure Program (SP) and Institutional Program (IP). People who take the TOEIC test at their place of work or study are taking an IP test. Those who take the test at an official testing site are taking an SP test.

Certain issues arise when schools use the TOEIC test and its scores. In this paper, the following will be addressed:

- The TOEIC test was not created to test students' English abilities, how much students learned, or how well students performed in a class, but, according to Wilson, one of the leading TOEIC researchers for ETS, to measure workers' "English-language proficiency in the international work environment" (1993, p. 2).
- Even if TOEIC scores are used only to compare students for placement, the standard errors of difference must be used in order to make decisions fairly and correctly.
- TOEIC scores are not as precise as they appear to be, and they are not able to measure English abilities as exactly as some may think.
- It is unlikely that most school programs provide enough classroom hours of English language study for most of their students to be able to increase their TOEIC scores without extensive additional study on their own.
- Requiring the use of the TOEIC test or TOEIC scores tends to inhibit the use of other materials and practices which are beneficial and necessary for students to attain full, well-rounded acquisition of all English language abilities and result in overall competence.

Students as TOEIC Test Takers: The Issues of Validity and Reliability

"The TOEIC test is designed for use by organizations working in an international market where English is the primary language of communication" (CGI, 2000, p. 2). "It measures the everyday English skills of people working in an international environment. The scores indicate how well people can communicate in English with others in business, commerce, and industry" (ETS, 2012, p. 2) "in the global workplace. The test does not require specialized knowledge or vocabulary; it measures only the kind of English used in everyday work activities" (ETS, 2007, p. 2). Hardly any Japanese students have worked in the international business world or have had opportunities to use English in such situations. This lack of background knowledge and experience will cause at least some students to have difficulties understanding the contexts and contents of at least some test items. Examples of these are corporate development, investments, marketing, labor relations, plant management, board meetings, and various technical areas (ETS, 2007, 2012). These shortcomings will result in lower scores because of deficiencies other than those related to English abilities, that is, the students' lack of comprehension of the ways, interactions, contents, and circumstances of the international business world. It will also weaken the TOEIC's validity as a test of these students' English abilities. "A test is valid if it measures what it says it measures" (Kubiszyn & Borich, 1987, p. 278), and it is less valid the more other content or issues affect the test results. This lack of validity is what led to the discrediting, in the eyes of most scientists, if not the general public, of the results of IQ testing: The test results were influenced by factors other than intelligence, which itself is a very unquantifiable capacity (Gould, 1996; Poundstone, 2003, pp. 23-42). Modern standardized testing of such things as learning potential and language proficiency developed out of IQ testing (Poundstone, 2003, pp. 35-36).

Furthermore, as most students in Japan's colleges, universities, and high schools are admitted based on tests that usually include English tests, the students on any one campus, or in any one department, have a much narrower range of English proficiency than does the general population. This suggests that the great majority of their TOEIC scores will also fall into much more restricted ranges, resulting in weaker reliability (Stratton Ray, personal communication, 1 Dec 2008). ETS itself warns that, "If you have a sample of candidates who are very similar to

each other, the reliability of the test within that specific homogeneous group will be quite low. If there is no (or very little) variation among candidates' test scores then, by definition, there can be no accurate estimate of reliability" (CGI, 1998, p. IV.3). "A test is reliable . . . if it consistently yields the same, or nearly the same, ranks over repeated administrations during which we would not expect the trait being measured to have changed" (Kubiszyn & Borich, 1987, p. 291), and it is less reliable the more the rankings vary, which will be the case if a group of test takers have very similar abilities.

Without strong validity and strong reliability, TOEIC scores have little meaning.

Two Types of Tests: Criterion-Referenced and Norm-Referenced

Tests can be separated into two basic types, each of which provides different information about the test takers. Brown (1995) explained the various differences between them. Most tests created for educational purposes, possibly outside of placement, are *criterion-referenced tests*, which try to determine what and how much of certain information or skills a student knows or can perform. If the test is a posttest, the hope is that a great majority of the students will demonstrate mastery of the materials and skills they have studied by scoring highly on the test. For pretests, there is no desire for mastery to be demonstrated. Instead, the purpose is to discover what students already know or can do and what they still need to learn. With this information, the teachers can decide what to teach. The makers of criterion-referenced tests know the details of the individual test takers' abilities well and they create test items that measure precise details of what the test takers will be expected to know and be able to do. In the case of posttests, the students know in detail what the test will cover and are expected to study those specific materials and practice the skills to prepare for the test.

The other type is *norm-referenced tests*, which attempt to measure overall proficiency. If these are also standardized tests, which are administered to large numbers of people in many locations at the same time, as they usually are, the test makers know little if anything about the test takers. The test items of norm-referenced tests must cover a wide range of materials and abilities. Makers of such a test hope that there is no small, identifiable set of materials or precise skills that test takers could study to help them to achieve higher scores. Otherwise, the test would not measure overall proficiency. Of course, the test results cannot provide precise details about what individual test takers know or can do. The test makers also hope that, when all of the scores from one administration are gathered, they demonstrate a normal distribution, that is, that few scores are very high or very low and that most scores fall in a range around the middle of the scale.

The TOEIC test is a norm-referenced standardized test. Therefore, it cannot provide details about exactly what a test taker has learned in a class or what a test taker knows or does not know, can or cannot do. Instead, it gives information about how an individual's English proficiency compares with others who took the same test.

Interpreting TOEIC Scores

The possible TOEIC Total score range is from 10 to 990, and the possible Listening score and Reading score ranges are from 5 to 495. If an individual takes the TOEIC test twice (at times A and B), one cannot just subtract the A scores from the B scores to determine if any increases in the scores indicate true increases or by implication demonstrate probable increases in English language abilities. Instead, the standard error of difference must be used with each score to find out if the differences in the scores, when subtracted, are wide enough to confidently state that the test taker's B scores are truly higher than the A scores.

The same method needs to be used to determine if the scores of any two test takers are the same or different (PsyAsia, 2013a; CGI, 1998, p. IV.6-IV.7). Tests "are always associated with some degree of error. . . . An *obtained* score has a *true* score component (actual level of ability, skill, knowledge) and an *error* component (which may act to lower or raise the obtained score)" (Kubiszyn & Borich, 1987, p. 304). If the difference in two scores is less than the error or confidence band created using the standard error of difference, then neither score can truthfully be said to demonstrate higher ability. Scientists, mathematicians, and testing researchers are aware of this, and so created the practice of using error or confidence bands to make the measurements they gather more precise. Most people, however, do not know about these statistical procedures or the importance and necessity of using them.

Unfortunately, ETS does not publish the standard error of difference for the TOEIC Total score, which is the score many administrators, teachers, and students are concerned about. However, ETS does say that the standard error of difference for both the Listening score and the Reading score is about +/-35 points. This allows for a comparison of scores with 68% confidence. To be 95% confident in one's decisions, two standard errors of difference, or +/-69 points, need to be used (CGI, 1998, p. IV.6-IV.7). "Why bother with the 95-percent level . . . . If you are going to make important decisions about a student, a conservative approach appears warranted. . . . If you are concerned about the effects of a 'wrong' decision (that is, saying a real difference in achievement exists when it is really due to chance), take the conservative approach" (Kubiszyn & Borich, 1987, p. 321).

Table 1 presents the TOEIC IP Test scores for 10 of 25 freshman university students who were studying together in three English language classes, a convenience sample. Each class met once a week for 90 minutes, 15 weeks per semester. The students were placed in the classes in the Japanese equivalent of alphabetical order. They took the TOEIC test on campus twice, with 6 months between the two administrations, which included a summer break of 2 ½ months. These scores are typical examples of the approximately 1,200 students' scores from which they were selected, though there are also many students' scores with less variety and some with more. They were specifically chosen to demonstrate the points that will be made. The last three columns on the right give the changes in scores, that is, the amount each student's scores were higher or lower on the second test than the first. Taking a few minutes to look over and consider these scores may make following the discussion somewhat easier.

Table 1. Student TOEIC Scores on Two Administrations

	Total Score		Listening Score		Reading Score		Change in Score		
Student	Test 1	Test 2	Test 1	Test 2	Test 1	Test 2	T2-T1	L2-L1	R2-R1
A	405	515	240	255	165	260	110	15	95
B	480	390	240	220	240	170	-90	-20	-70
C	395	495	240	270	155	225	100	30	70
D	500	585	280	350	220	235	85	70	15
E	530	490	295	315	235	175	-40	20	-60
F	460	450	270	230	190	220	-10	-40	30
G	445	480	210	270	235	210	35	60	-25
H	440	470	290	250	150	220	30	-40	70
I	635	625	305	270	330	355	-10	-35	25
J	385	465	205	225	180	240	80	20	60

In comparing scores for individual students in Table 1, there are only four changes in scores that demonstrate an increase

in Listening or Reading score with 95% confidence, that is, of more than 69 points: student D's change in Listening score and students A, C, and H's changes in Reading scores. Student B had the second highest score in Test 1, but the change in Reading score of more than 69 points indicates a truly lower score on Test 2 than on Test 1. No true increase or decrease in the other 15 Listening and Reading scores can be claimed with 95% confidence, including those of student I, who has the highest Total score on both tests. One thing to note is that students A and C, two of the three students whose Listening scores demonstrate true increases, had two of the three lowest Total scores on the first administration.

As for comparing scores between students, only the Listening scores of students G and J on the first administration are more than 69 points lower than any of the others' Listening scores, and so can be claimed to be truly lower with 95% confidence. All of the rest are statistically the same. On the second administration, all but one of the other students' Listening scores can be said to be truly lower than student D's, the highest, with 95% confidence. These eight scores are the same statistically as each other, and five of them are statistically the same as the second highest Listening score, student E's. Similar comparisons can be made for the Reading scores.

These kinds of comparisons and considerations ought to be carried out whenever TOEIC scores are being used, in order to understand what the scores are indicating and to use them fairly when judging and comparing test takers' English abilities. No matter how TOEIC scores are used, if standard errors of difference are not considered, it will lead to unfair and unwise decisions and practices, such as giving a higher score to one student than another based on scores that are within two standard errors of difference of each other.

TOEIC Scores as Precise Measurements of English Language Abilities

Some schools use TOEIC scores or gains in TOEIC scores as at least a partial measure for grading students in classes. Some schools use achievement of a certain TOEIC score as a criterion for awarding certificates of completion for courses. The latter usage may be justified, if it is in combination with other criteria, and if the administrators believe it is legitimate to use the TOEIC test to measure their students' English language abilities. Yet it may lead to misuse, for example, if students are denied being awarded a certificate in areas or fields not related to international business just because they did not achieve a high enough TOEIC score. The former usage, however, is definitely not intended by ETS and is not supported by norm-referenced tests, and so is erroneous. ETS publishes information on how to interpret and use TOEIC scores (CGI, 1998; ETS, 2007, 2012), yet it seems that this information may not be well known or understood by many administrators and teachers.

In the present age, people want numbers and measurements to support claims and ideas in all fields. This is considered as providing scientific proof. TOEIC scores seem to provide this proof. The scores appear to be precise measurements of test takers' English language abilities. However, Cameron (1963, p. 13) stated, "Not everything that can be counted counts, and not everything that counts can be counted," a claim seemingly so innovative, radical, and yet correct that it is often accredited incorrectly to Albert Einstein. This is the case with language abilities, which have no physical aspects, though language itself is manifested physically when we write or speak. Language abilities are aspects of our thinking, our will, and our feelings. They are part of our inner being and inner self, not part of our physical bodies, even though we use our physical bodies to make use of them. They cannot be assigned meaningful, precise, numerical scores, just as IQ cannot. Therefore, when a school

uses something like TOEIC scores, it needs to do so with great caution and with careful attention to the fairness and truthfulness of the usage.

Classroom Study and Increases in TOEIC Scores

Many students, teachers, and administrators would like to know how much time it takes to improve foreign language abilities enough to be demonstrated in increased test scores. In an attempt to answer this question, Saegusa (1985) generated multiple correlations and regression equations using pairs of TOEIC scores from workers who had been studying English in courses arranged by their companies. He then used these and standard errors of measurement to determine how much classroom study time would be needed to expect most learners to improve their TOEIC scores by certain amounts. He concluded that, "less than 80 hours of (English language) instruction is not very effective. In such classes, a majority will make little or no progress. If effectiveness is given top priority, at least more than 100 hours of instruction, and ideally 200 hours of instruction, as a unit should be recommended" (p. 174). He also determined that approximately 400 classroom hours of English language instruction would be needed for most students to raise their TOEIC Total scores from 450 to 600 or from 600 to 730 (p. 181).

As most Japanese college and university English language classes meet for just 90 minutes once a week for 15 weeks per semester, students would need to attend a minimum of five such classes, with nine being preferable, in order for most of them to be expected to raise their TOEIC scores. Such a schedule is likely to be found only in programs in which students major in English. With this information, and the information concerning standard errors of difference, the scores in Table 1 seem much less unusual. It is almost expected that few students would demonstrate improvement by achieving truly higher scores, as they did not spend enough time studying, unless they also studied English extensively outside of their classes. The seemingly large variability of many of the individual students' scores would also be expected, as "jumping around is in the nature of TOEIC scores" (Childs, 1995, p. 73), due, at least partially, to the sizes of the standard errors of difference.

In addition, Saegusa (1985, p. 167) explained that, generally, the company classes consisted of about 10 people per class, met for 2 hours two or three times a week (for a total of 50 to 200 hours during a period of 3 to 6 months), and were taught by native English speakers. Attendance was 80%. The English language study requirements at most colleges and universities in Japan do not meet most of these conditions. Therefore, it is possible that the estimates of classroom English study time needed for most students to raise their TOEIC scores by the amounts suggested by Saegusa would prove to be too low. Also, because Saegusa used standard errors of measurement, which are used for determining the range in which a test taker's true score falls based on a single obtained test score (PsyAsia, 2013b), where he should have used standard errors of difference, his estimates of the number of classroom hours of English language study needed for most students to achieve the gain scores he spoke of are probably about 30% too low (Bresnihan, 2010, p. 213-214).

Teaching for the TOEIC

ETS's initial head TOEIC researcher, Woodford, wrote, "The way in which we test can inform the manner in which we teach" (1982, p. 2). It can also distort the way we teach. It is not unusual for English language classes at Japanese colleges and universities to use TOEIC-like materials, which are the focus of a great many textbooks, and to have students do drill work with them. When asked about the practice of teaching the TOEIC test in an interview, an ETS representative seemed confused by the question and finally replied, "The student needs to be motivated to learn English and NOT simply to pass the test. . . . TOEIC is a test not

a language, so *teaching TOEIC* is not really an option. The best thing to do is to teach English focusing on proficiency rather than rules or vocabulary" (Wood, 2010, p. 44). If a class focuses on TOEIC-like drills, the students will not be encouraged to study or practice English in other ways or with other materials. On paper, these classes may appear systematic, efficient, and rigorous. In reality, though, such styles of teaching are stifling and ignore a great many other ways of learning and acquiring language and other aspects of language usage that students need to learn, not to mention materials to use. For example, there seems to be no reason to be able to read or understand novels or poems or to learn how to speak or write clearly and accurately, because these things are not on the TOEIC test. Yet, such abilities will surely benefit anyone who is interested in using English.

ETS has produced charts (for example, CGI, 2000; ETS, n.d.) that give expected speaking and writing abilities related to TOEIC Listening and Reading scores, although at the same time explaining that these are general claims and cannot be used as definitive statements about any particular person's abilities. The charts are based on research published by ETS (for example, CGI, 1998; Liao, Qu, & Morgan, 2010; Wilson, 1989, 1993; and Woodford, 1982), even though Liao, Qu, and Morgan concluded their study of the standard TOEIC test and the TOEIC Speaking and Writing tests by stating, "The results . . . confirm that there are four separate language skills measured by the TOEIC tests. It is natural that different language skills are correlated with each other to a certain degree; however, each test measures distinct aspects of English language proficiency that cannot be adequately assessed by the other tests. Examinees should take all of the TOEIC tests in order to gain a full understanding of the complete spectrum of their language proficiency skills" (p. 13.11). Hirai (2013), in his comparisons of TOEIC Total scores with direct tests of both speaking and writing, also found ETS's claims based on the standard TOEIC scores to overestimate abilities in the productive skills, stating that "Japanese people's business speaking/writing skills . . . are substantially lower than the levels the general public might expect of them from their TOEIC [Total] scores" (p. 124). These findings suggest the possibility that these other abilities may not be fostered as much as some claim by practicing only listening and reading.

Conclusion

Unfortunately, despite, or perhaps because of, its widespread usage, the TOEIC test is often used in ways that testing experts, even the makers of the TOEIC test, do not support. The TOEIC test can only measure general listening, reading, and overall English language proficiency, and only for those who are familiar with the settings, circumstances, and basic content of the test items. Although ETS has now added students as target users of the TOEIC test in its promotion materials, the contexts and contents of the test are still aimed at people who use English in international business situations. As a norm-referenced test, the TOEIC cannot determine what a student has learned in a class, what class materials a student knows and does not know, or what functions a student can or cannot perform well. Even if the scores are being used only to find out how students compare with each other, the standard errors of difference must be used along with the scores. In addition, the English abilities of Japanese college and university students on a given campus or in a given department are more similar to each other than desirable for TOEIC scores to be strongly reliable measures of English language ability.

Even if all of the above problems with TOEIC score usage were rectified, it is unreasonable and unfair for most Japanese colleges and universities to expect their students' TOEIC scores to increase during a semester or even a year because they do not offer nearly enough classroom hours of English language study for this to happen. It is depressing and demotivating for students and teachers when it appears that most students'

scores do not increase and many go down, due to not taking the standard errors of difference into consideration.

In any case, measurements of English proficiency are only estimates. TOEIC scores fool us into thinking otherwise and distract us from engaging in more beneficial practices and setting our sights on more useful goals. Using TOEIC scores for evaluative purposes, or even just requiring the TOEIC test to be taught or taken, has a very restrictive effect on what and how teachers teach and what and how students study and learn. Choice and motivation become connected with and distorted by the idea of increasing TOEIC scores rather than improving English language ability.

Quite opposite to what administrators might hope, using a test like the TOEIC in place of classroom-based tests "minimizes the possibilities that their program will look good" (Brown, 1995, p. 18). It also minimizes the possibilities that teachers and the students will look good. The wellsprings of teaching and learning are self-motivation and freedom. The limiting and conforming tendencies of using the TOEIC test in schools work against these impulses.

Bio Data

Brian Bresnihan teaches EFL at University of Hyogo. Before this, he did the same plus administrative and supervisory work in Temple University Japan's intensive English language program in Tokyo and at a small school that no longer exists in Hiroshima. Between those two positions, he spent 4 years studying at Teachers College, Columbia University (including 2 years coordinating the in-house ESL program) and teaching ESL part-time in a few programs in New York City. <brian@econ.u-hyogo.ac.jp>

References

Bresnihan, B. D. (2010). *Possible reliability problems affecting use of TOEIC IP Test scores*. Kobe: Institute for Policy Analysis and Social Innovation, University of Hyogo.

Brown, J. D. (1995). Differences between norm-referenced and criterion-referenced tests. In J. D. Brown & S. O. Yamashita (Eds.), *Language testing in Japan* (pp. 12-19). Tokyo: JALT.

Cameron, W. B. (1963). *Informal sociology: A casual introduction to sociological thinking*. New York: Random House.

CGI (The Chauncey Group International). (1998). *TOEIC technical manual*. Princeton, NJ: Author.

CGI (The Chauncey Group International). (2000). *TOEIC can-do guide: Linking TOEIC scores to activities performed using English*. Princeton, NJ: Author. Retrieved from http://www.ets.org/Media/Research/pdf/TOEIC_CAN_DO.pdf

Childs, M. (1995). Good and bad uses of TOEIC by Japanese companies. In J. D. Brown & S. O. Yamashita (Eds.), *Language testing in Japan* (pp. 12-19). Tokyo: JALT.

ETS (Educational Testing Service). (2007). *TOEIC user guide: Listening & reading*. Princeton, NJ: Author. Retrieved from http://www.ets.org/Media/Tests/Test_of_English_for_International_Communication/TOEIC_User_Gd.pdf

ETS (Educational Testing Service). (2012). *TOEIC examinee handbook: Listening & reading*. Princeton, NJ: Author. Retrieved from http://www.ets.org/Media/Tests/TOEIC/pdf/TOEIC_LR_examinee_handbook.pdf

ETS (Educational Testing Service). (n.d.). *TOEIC can-do levels table*. Princeton, NJ: Author. Retrieved from http://www.amideast.org/sites/default/files/otherfiles/hq/advising%20testing/toeic20can-do-20levels20table1.pdf

Gould, S. J. (1996). *The mismeasure of man* (Rev. ed.). New York, NY: W. W. Norton.

Hirai, M. (2013). Correlations between BULATS Speaking/Writing and TOEIC scores. In R. Chartrand, S. Crofts, & G. Brooks (Eds.), *Literacy: SIGnals of emergence. Proceedings of the 11th Annual JALT Pan-SIG Conference* (pp. 118-125). Hiroshima: Hiroshima University. Retrieved from http://www.pansig.org/2013/JALTPanSIG2013/Proceedings/The2012Pan-SIGProceedings.pdf

IIBC (The Institute for International Business Communication). (n.d.a). *TOEIC Bridge data & analysis 2011*. Tokyo: Author. Retrieved from http://www.toeic.or.jp/toeic_en/pdf/data/TOEIC_Bridge_DAA2011.pdf

IIBC (The Institute for International Business Communication). (n.d.b). *TOEIC Test data & analysis 2011*. Tokyo: Author. Retrieved from http://www.toeic.or.jp/toeic_en/pdf/data/TOEIC_DAA2011.pdf

Kubiszyn, T., & Borich, G. (1987). *Educational testing and measurement* (2nd ed.). Glenview, IL: Scott, Foresman.

Liao, C. W., Qu, Y., & Morgan, R. (2010). *The relationships of test scores measured by the TOEIC Listening and Reading test and TOEIC Speaking and Writing tests*. Princeton, NJ: ETS. Retrieved from http://www.ets.org/Media/Research/pdf/TC-10-13.pdf

Poundstone, W. (2003). *How would you move Mount Fujii? Microsoft's cult of the puzzle: How the world's smartest companies select the most creative thinkers*. New York: Little, Brown and Company.

PsyAsia (PsyAsia International Pte.). (2013a). *Standard error of difference*. Singapore: Author. Retrieved from http://www.psyasia.com/support/Knowledgebase/Article/View/28/0/standard-error-of-difference

PsyAsia (PsyAsia International Pte.). (2013b). *Standard error of measurement*. Singapore: Author. Retrieved from http://www.psyasia.com/support/Knowledgebase/Article/View/27/0/standard-error-of-measurement

Saegusa, Y. (1985). Prediction of English proficiency progress. *Musashino English and American Literature, 18*, 165-185. Tokyo: Musashino Women's University.

Wilson, K. (1989). *TOEIC research report, No. 1: Enhancing the interpretation of a norm-referenced second-language test through criterion referencing: A research assessment of experience in the TOEIC testing context*. Princeton, NJ: ETS. Retrieved from http://www.ets.org/Media/Research/pdf/RR-89-39.pdf

Wilson, K. (1993). *TOEIC research summaries, No. 1: Relating TOEIC scores to oral proficiency interview ratings*. Princeton, NJ: ETS. Retrieved from http://www.ets.org/Media/Research/pdf/TOEIC-RS-01.pdf

Wood, J. (2010). TOEIC materials and preparation questions: Interview with an ETS representative. *The Language Teacher, 34*(6), 41-45. Tokyo: JALT. Accessed from http://jalt-publications.org/tlt/articles/109-readers-forum-toeic-materials-and-preparation-questions-interview-ets-representativ

Woodford, P. (1982). *TOEIC research summaries: An introduction to TOEIC: The initial validity study*. Princeton, NJ: ETS. Retrieved from http://www1.ets.org/Media/Research/pdf/TOEIC-RS-00.pdf

Vocabulary: What Should We Test?

Paul Sevigny
Ritsumeikan Asia Pacific University
Kris Ramonda
Kwansei Gakuin University

Reference Data:
Sevigny, P., & Ramonda, K. (2013). Vocabulary: What should we test? In N. Sonda & A. Krause (Eds.), *JALT2012 Conference Proceedings*. Tokyo: JALT.

Diagnostic Yes/No tests are a recommended and much researched assessment tool (Read, 2007; Nation, 2008), yet there is little research into how to apply them to address the mismatch between pre-existing course vocabulary lists from commercial textbooks for a particular level and learners' actual vocabulary knowledge. This study looked at a vocabulary battery of 240 words adopted with a textbook for a pre-intermediate level English course at a Japanese university. During the 1st week of instruction, a Yes/No test including nonwords (pseudo-words) was administered in three forms with 85 items each. Approximately 100 students took each form. On the average, test takers claimed they knew 75% of the items on the list. A low false alarm rate supports Shillaw's (1996) findings that the use of nonwords could be lessened significantly in the Japanese context.

Yes/No形式の語彙診断テストは、学力診断用に奨励されかつ研究がされているツールである（Read, 2007; Nation, 2008）が、特定のレベルおよび学習者用に使用される市販教科書と、既に存在している語彙リストとの齟齬を分類する方法に関する研究は、ほとんど存在しない。当研究では、日本の大学における準中級英語科目で使用されている教科書から抽出した240語の語彙群について報告するものである。講義開始から1週間の間に、無意味語を含む85語ずつ3種類のYes/No形式の診断テストが行われた。各テストをおよそ100名ずつの学生が受けた。平均して、受験者はリストの75%の単語を熟知語であると判断した。無意味語を「知っている」と回答した割合が低かったことから、無意味語の使用は日本では大幅に減らすことができるというShillaw（1996）の結果を肯定する結果となった。

The initial impetus for this study came from a top-down English program evaluation and the development of curricular research teams. Teacher researchers were concerned that preexisting vocabulary achievement tests were being used for grading purposes without determining the learners' knowledge of the items being tested at the beginning of the semester. The problem was the need for content validation with respect to vocabulary related goals and testing. Assessing learners' knowledge with a Yes/No test at the beginning of the semester was the first step in evaluating the common vocabulary list and achievement tests in a program with large reading and vocabulary classes at a Japanese university.

Literature Review

Yes/No tests traditionally present a word without context and ask the participant to indicate whether the word is known or not. In this report *checklist* will be used interchangeably with

Yes/No test, as the difference only refers to whether the learner circles, checks, or clicks to indicate whether an item is known or not. Nation (2008) and Read (2007) recommended the use of Yes/No vocabulary tests (in addition to the use of Vocabulary Levels Tests [VLTs]) for placement purposes. While the VLT provides general placement information with regard to vocabulary frequency bands, Read suggested that Yes/No tests can assess vocabulary size relative to specific lists, and thus may be used by programs to develop their own assessment tools. Additionally, because of its simplicity, the Yes/No format is "informative and cost effective" (Read, 2007, p. 113).

Shillaw (2009) presented a careful overview of Yes/No tests and the many efforts to establish the reliability and validity of the tests. Zimmerman, Broder, Shaughnessy, and Underwood (1977) measured the vocabulary size of native English speakers and reported validity of a word recognition test based on correlation with verbal scores on the Scholastic Aptitude Test. Anderson and Freebody (1983) studied the vocabulary size of 5th grade native English speakers and reported that Yes/No test results were more reliable than multiple-choice tests for measuring vocabulary size. This led to the first research into Yes/No tests for determining the size of L2 learners' vocabulary. The rationale was that the simplicity and efficiency of the Yes/No format would allow for sampling the large number of items necessary for estimating vocabulary size. Meara and Buxton (1987) compared a Yes/No test with a multiple-choice test in predicting nonnative speakers' grades on the First Certificate in English examination and reported that only the Yes/No scores had a significant correlation with grades. By the turn of the century, these tests had become established measures of vocabulary size for L2 learners (Nation, 2001; Read, 2000). In support of Yes/No tests, Cameron (2002) stated, "Eventually, after sufficient contextualized encounters, a word will be recognized when it is met in context or in isolation . . . it does not seem unreasonable to test to see how much vocabulary can be recognized without extended linguistic or textual contexts" (p. 151).

More recently, Mochida and Harrington (2006) completed an in-depth review of the Yes/No test as a tool for testing receptive vocabulary. They reviewed a number of studies that attempted to correlate results from Yes/No tests with other test forms—multiple-choice and translation tests. They stated, "The results show that the Yes/No test is a reliable measure of the kind of vocabulary knowledge measured by the VLT and, presumably, similar multiple-choice tests" (p. 91). Finally, they concluded that the Yes/No test has compelling practical advantages warranting further attention from L2 testers and teachers, such as incorporating their use in word recognition tasks.

The use of nonwords started as an attempt to validate test takers' judgments. The nonwords follow the phonetic rules of English and provide a window into determining whether students are honestly stating their familiarity or unfamiliarity with vocabulary items. Nonwords are typically created by changing one or two letters in real words; for example, *foggy* becomes *wuggy*. A second method of creating nonwords, called pseudo-derivation, uses unconventional base + affix combinations, for example, *adjustation* (Shillaw, 2009). A false alarm is an instance when a learner reports knowing a nonword. Read (2007) recommended using nonwords to correct the total score for each learner on a Yes/No test by simply taking the number of Yes/No responses to real words minus the number of Yes responses to nonwords and finding the resulting vocabulary size.

Calculating average false alarm rates for populations allows for identifying populations that are generally overconfident. In a study done in Japan (Barrow, Nakanishi, & Ishino, 1999), participants reported knowing an average of 1.26 nonwords on tests with 15 nonwords, yielding a false alarm rate of 8.4%. Milton (2009) provided averages from studies outside Japan with much higher false alarm rates. Shillaw (1996) found Japanese learners

to be very conservative, almost never falsely claiming knowledge of nonwords. Stubbe, Stewart, and Pritchard (2010) and Stubbe (2012) reported false alarm rates for low-intermediate Japanese university students of 4-5%.

Perhaps the most compelling of studies attempting to investigate the reliability and validity of Yes/No tests with the use of nonwords was Shillaw's (1996) use of Rasch scaling techniques to examine three of Meara's (1992) Yes/No tests that were each comprised of 40 real words and 20 nonwords. These tests were administered to seven classes of Japanese university students. In a rather complex Rasch analysis comparing the results of two tests taken by the same groups of students, Shillaw reported higher correlations when comparing scores of real words only versus the scores which included all words (real words and nonwords). Shillaw pointed out the marginal value of nonwords in contributing to test variance; he concluded that on these assessments and for these learners, the presence of nonwords had little effect on their test performance.

Context of the Study

The context for the current study was the pre-intermediate English level of a large EFL program at a Japanese university. The courses were divided by skills into a two-credit reading and vocabulary course and a four-credit listening, speaking, writing, and grammar course. This research was situated in the pre-intermediate English reading and vocabulary course. At the time of this study, learners were placed using the paper-based TOEFL test without a diagnostic VLT. The course utilized a commercial textbook and aimed to increase learners' receptive vocabulary knowledge for readings in the textbook.

The English Program followed a 5-year curriculum cycle, implementing an all-new curriculum every 5 years. Prior to the beginning of the new curriculum cycle in 2011, textbooks were adopted as a base for each course. In the pre-intermediate level, *Interactions Access: Reading* (Hartmann & Mentel, 2007) was selected. The first seven chapters and corresponding 240 items in chapter word banks were adopted as the base curriculum and common course vocabulary list. The items were taught and tested in the first two semesters of the curriculum cycle. Various computer-based vocabulary activities were constructed based upon the new common list for individual student practice. Regular, summative vocabulary assessments, accounting for 30% of learners' grades, consisted of multiple-choice and matching items and tested receptive, form-meaning connections (the ability to recognize a word and recall its meaning).

After two semesters in the new curriculum cycle, students' average score on vocabulary quizzes and tests was above 95%. Although high scores are encouraging to all stakeholders, teachers began to voice concerns that the high grades might be a result of learners already knowing the vocabulary. The current study began with the desire to ensure that learners have a worthwhile learning experience.

Research Issues

The goal of this research was to determine how much of the established course vocabulary list learners believed they already knew at the beginning of the course. For the purpose of this study, the extent of that knowledge was considered for the group as a whole and not for individuals. That is, the measure of interest was the percentage of learners who reported they knew a particular item. A further goal was to determine the corpus-based frequency of each item.

The following research questions were investigated:

1. How much of the common course vocabulary list is already known by most of the learners?

2. How does the learners' familiarity with each word (item facility) relate to established corpus-based frequency-band data?

Method

Participants

The participants in this research were 300 university learners of pre-intermediate English (TOEFL scores 400-437). Of these learners, 89% were Japanese and the remaining 11% were learners from China and Korea who were already fluent in Japanese. All participants were 1st-year university students, most having placed directly into pre-intermediate English, but some continuing from a previous semester in elementary English.

Procedure

Frequency Information

The first step was to record item frequencies for all 240 words (see Appendix, column *f*) from Web Vocabprofile (Cobb, 2006), an adaptation of Heatley and Nation's (1994) Range. The profiler provides lemmatized word frequency information: K1, K2, AWL, and Off List; that is, the same frequency is assigned to all members of one lemma, or headword. For example, the word *problematic* is a member of the word family *problem*, which falls in the K1, or most frequent one-thousand word families. Following are item frequency categories:

- K1—word from first thousand most frequent word families,
- K2—word from second thousand most frequent word families,
- AWL—word from Academic Word List, and
- Off List—word not included in K1, K2, and AWL.

Yes/No Test

The 240 items of the common course list were organized by frequency, alphabetized, and then divided into three groups to make three test forms. The intention was to create three sets of words that presented variety as to frequency, spelling, meaning, and word length, rather than having alphabetically ordered segments or chapter themes grouped together. Similar-sounding items were intentionally separated when possible. The rationale for having three separate test forms was out of a concern that test taker fatigue could impact participant responses. Nation (2008) suggested using from 50 to 100 items in such an assessment. In the current study, 80 items were used in each test. The same five nonwords were added to control for overconfidence, yielding a total of 85 words on each Yes/No test. The downloadable application Wuggy 0.2.0b3, available from the Center for Reading Research at Ghent University, was used to create the following nonwords of similar length: *wuggy*, *ecution*, *pregime*, *mengel* and *runster*. The word lists from the three tests (A, B, and C) are in the Appendix.

After their level was determined via the school placement test, learners were randomly assigned to classes. Learners for this study came from six teachers' pre-intermediate English classes (two teachers' groups completed each test form). There were three large classes (60+ students) and three slightly smaller classes (40+ students). Large and small classes were paired to form three groups of approximately equal size, each of which received one test: Test A, Test B, or Test C. The three tests were administered on the second day of instruction during the spring semester of 2012. Each test was administered to two classes. See Figure 1 for instructions and examples.

Figure 1. Instructions and Examples From the Yes/No Vocabulary Tests

Instructions:
This is a vocabulary test. Please indicate whether you know the word or not, "I know this word" or "I don't know." By "knowing" a word, we mean that you are able to recognize its basic meaning.

これは語彙知識を判断するテストです。それぞれの単語を知っているかどうか、該当する選択肢を選んでください。単語を「知っている」ということは、その単語の基本的な意味が分かるということです。

Examples

1. travel
   a) I know this word.
   b) I don't know.

2. wuggy
   a) I know this word.
   b) I don't know.

Scoring

The test was administered using a content management system called Blackboard 6.2. The students indicated their choice by clicking a radio button. Blackboard 6.2 yields score reports for individual learners, group averages, and individual test items, including the percentage of learners answering each item correctly. Three scoring procedures were used in this study:

1. *average test score:* the average percent of the real words that the learners claimed to know,
2. *item facility:* the percentage of learners who claimed to know that item, and
3. *false alarm rate:* the total number of false alarms made by all participants divided by the total number of nonwords presented on the three forms.

Results

Blackboard 6.2 does not automatically save learners' answers during a test. Those who do not use the "save" or "save all" function receive a zero. If a test taker does not save any answers, it is evident as all item responses appear as "not answered." There were 34 participants who did not save their answers and thus were removed from the study, resulting in a sample size of 300.

Average Test Scores

Average test scores are reported in Table 1. The total number of words known minus the number of nonwords (incorrectly selected) yields vocabulary size (Read, 2007). The results showed that almost all learners correctly rejected all five nonwords. The adjusted average vocabulary size using Read's formula would yield about 59 out of 80. In other words, on the second day of class, learners claimed to know about 75% of all words on the common course list.

Table 1. Average Test Scores (N = 300)

Test	Average number known words (k = 80)	Average number non-words correctly rejected (k = 5)
A (n = 111)	55 (69%)	4.7 (94%)
B (n = 101)	61 (76%)	4.8 (97%)
C (n = 88)	64 (80%)	4.8 (96%)
Total	60 (75%)	4.8 (96%)

Note. k = number of words on test

Item Facility Results

The percentage of learners reporting to know each test item is reported in the Appendix. In classical testing theory, this measure is known as the item facility and is calculated by taking the number of students who reported to know a word, divided by the total number of responses. They are listed in descending order in percentage form, with the items that 100% of all learners reported knowing at the top. The frequency band from Web Vocabprofile for each item is reported to the right of each item facility result.

Discussion

False Alarm Results

The false alarm average of 4.8% reveals that the learners in this study were reasonably conservative when deciding whether an item was known or not. As we added just 5 nonwords to each test containing 80 real words, we presented learners with 94% real words on the test. Meara (1992) presented 40 real words out of a total of 60 words, or a proportion of 67% real words. Due to time constraints and concern about test-taker fatigue, we opted for a smaller pool of nonwords. More importantly, in the light of the commonly low false alarm rates reported by other researchers in Japan (Shillaw, 1996; Stubbe, 2012; Stubbe et al., 2010), decreasing the proportion of nonwords in the Japanese context seemed reasonable.

Item Frequencies

While there are clear visual correlations between the K1 words and the most familiar items at the top of the Appendix, any obvious pattern obscures with decreasing item facility. Applying more sophisticated statistical analyses to the results in the Appendix could provide more nuanced understanding of what makes some items easier than others to acquire. For example, item facility scores give an indication of learners' familiarity with collocations, for which frequency data is not readily available. In some instances where there is disparity between the item facility and item frequency, there may be a need to reclassify item frequencies. For example, *computer* is listed as AWL, but now is likely to be in the most frequent one thousand words. *Email* and *online* are OL, but the frequency of these items has also increased dramatically. Meara (2010), in the preface to the second edition of his *EFL Vocabulary Tests*, cited the effect of digital communications on word frequency:

> *Text*, for example, was a relatively infrequent word in 1992, largely confined to a couple of very specific genres. Thanks to mobile phones, *text* must be one of the most frequent words to occur in everyday spoken English in 2010. (p. 3)

Knowing a Word Receptively

One limitation of the Yes/No test for receptive knowledge, as it was used in the current study, might be the provision of the decontextualized, written word alone. There is more to knowing a word receptively than just seeing it—for example, knowing a word by hearing it. Another possible way to construct this test could have been to include both the written word and an audio recording of the same word. Learners could have worn headphones and clicked a button to play the audio pronunciation as they looked at the written form of the word. Providing both audio and visual channels could impact the results and might serve to identify words that have already been partially learnt receptively. In the current study, the primary interest was in measuring participant knowledge of the form-meaning of the target items, since the course vocabulary tests were designed in a similar fashion. For these reasons, only the written form-meaning was presented.

A second way to alter the form of a Yes/No test would be to provide context around the given vocabulary item, that is, to provide the word in a clear sentence and have the learner indicate whether the underlined word is known or unknown. The added dimension of sentence level context in Yes/No tests was investigated by Shillaw (2009). He compared the results of two Yes/No tests, one with context provided (with instructions that encouraged test takers to use it) and one with no context for the lexical items. The results showed a statistically significant and higher affirmation rate for the Yes/No test with context provided. If sentence level context can trigger the learner to recognize an item, then the item may already be partially acquired receptively.

The decontextualized Yes/No test given in this study provides less receptive context than either of the two alternatives above. This lack of added context presents the least likely reading scenario for learners in a natural context; thus, it might be logical to infer that the affirmation rate on the Yes/No test in this study, with the addition of aural or sentence context, would increase if listening or context were added. When encountering words in authentic contexts, a number of other linguistic features such as aural, visual, and syntactic cues could aid in accessing partially learnt receptive lexical items. On one hand, there is the possibility that with added context the learner might not actually know the word, but could infer the meaning from its lexical environment. Furthermore, a context-rich environment could place an additional burden on working memory, especially if other words were unknown to the learner, or presented in cognitively difficult-to-process grammatical structures. This could potentially distract the learner from the target item. Moreover, adding these other dimensions also takes more time, both in creating and taking the test, which would diminish some of the simplicity and efficiency that makes the Yes/No test so valuable. Nonetheless, studying how added context in Yes/No tests affects the reliability of the results deserves more attention.

Conclusion

When curriculum or learner populations (or both) are in flux, it is inevitable for teachers and administrators at some point to ask the question posed in the title of this paper—Vocabulary: What should we test?

Language programs should include tests of vocabulary levels. However, program administrators need to test lexical knowledge that corresponds to both the needs of the learners and the levels of courses in which learners will be placed. Thus, they need something different than a norm-referenced test like the VLT. Diagnostic Yes/No tests, as employed in this study, could provide administrators with valuable item facility data for creating custom placement tests. Diagnostic Yes/No tests can also provide important data about the level of mastery a cohort of learners has in relation to a common course list so that teachers can move specific vocabulary items from receptive to productive-mode tasks and assessments. The results of Yes/No tests can also move a course that has adopted a specific textbook towards the process of *adapting* the use of that textbook's vocabulary lists.

The mismatch between predetermined vocabulary lists in commercial textbooks for a particular level and students' actual vocabulary knowledge in corresponding levels can lead to inefficient allocation of teacher and time resources. This study provides one possible solution to address this mismatch. The Yes/No test is quick and easy to administer and allows for agility and flexibility in tailoring vocabulary items to a specific student population. However, this type of assessment is not without limitations. There can be reliability issues due to overconfidence or misinterpretation of what it means to know a word on the part of the test taker. This can be controlled to some extent by including nonwords in the assessment and by providing clear examples of what knowing a word means in the test instructions.

Bio Data

Paul Sevigny is a Lecturer at the Center for Language Education, Ritsumeikan Asia Pacific University in Beppu, Japan. His areas of research interest are extensive reading, vocabulary, and the diffusion of best practices.

Kris Ramonda is an Associate Lecturer of English at Kwansei Gakuin University in Japan. His research interests include vocabulary acquisition, extensive reading, and metaphor in language.

References

Anderson, R. C., & Freebody, P. (1983). Reading comprehension and the assessment and acquisition of word knowledge. In Hutson, B. A. (Ed.), *Advances in reading/language research*, 2 (pp. 231-256). Greenwich, CT: JAI Press.

Barrow, J., Nakanishi Y., & Ishino, H. (1999). Assessing Japanese college students' vocabulary knowledge with a self-checking familiarity survey. *System*, *27*, 223-247. doi:10.1016/S0346-251X(99)00018-4

Cameron, L. (2002). Measuring vocabulary size in English as an additional language. *Language Teaching Research*, *6*, 145-173. doi:10.1191/1362168802lr103oa

Cobb, T. (2006). *Web Vocabprofile* [Computer program]. Accessed from http://www.lextutor.ca/vp/

Hartmann, P., & Mentel, J. (2007). *Interactions access: Reading*. Singapore: McGraw-Hill.

Heatley, A., & Nation, P. (1994). *Range* [Computer program]. Victoria University of Wellington, New Zealand. Available from http://www.vuw.ac.nz/lals/

Meara, P. (1992). *EFL vocabulary level tests*. Swansea, UK: University of Wales, Swansea.

Meara, P. (2010). *EFL vocabulary level tests* (2nd ed.). Swansea, UK: University of Wales, Swansea.

Meara, P., & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing*, *4*, 142-154. doi:10.1177/026553228700400202

Milton, J. (2009). *Measuring second language vocabulary acquisition*. Bristol, UK: Multilingual Matters.

Mochida, A., & Harrington, M. (2006). The Yes/No test as a measure of receptive vocabulary knowledge. *Language Testing*, *23*, 73-98.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nation, I. S. P. (2008). *Teaching vocabulary: Strategies and techniques*. Boston: Heinle Cengage Learning.

Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.

Read, J. (2007). Second language vocabulary assessment: Current practices and new directions. *International Journal of English Studies*, *7*(2), 105-125.

Shillaw, J. (1996). *The application of Rasch modelling to yes/no vocabulary tests*. Unpublished manuscript, University of Tsukuba, Ibaraki, Japan. Retrieved from http://www.lognostics.co.uk/vlibrary/

Shillaw, J. (2009) Putting yes/no tests in context. In T. Fitzpatrick & A. Barfield (Eds.), *Lexical processing in second language learners: Papers and perspectives in honour of Paul Meara* (pp. 13-24). Bristol, UK: Multilingual Matters.

Stubbe, R., Stewart, J., & Pritchard, T. (2010). Examining the effects of pseudowords in yes/no vocabulary tests for low level learners. *Kyushu Sangyo University Language Education and Research Center Journal*, *5*, 5-23.

Stubbe, R. (2012). Do pseudoword false alarm rates and overestimation rates in yes/no vocabulary tests change with Japanese university students' English ability levels? *Language Testing*, Advance online publication. doi:10.1177/0265532211433033

Zimmerman, J., Broder, P. K., Shaughnessy, J. J., & Underwood, B. J. (1977). A recognition test of vocabulary using signal-detection measures and some correlates of word and nonword recognition. *Intelligence 1*(1), 5-13.

Appendix

Pre-Intermediate Course Vocabulary Yes/No Test Results

Results: Yes/No % and Item Frequencies

Test A $n = 111$	K%	*f*	Test B $n = 101$	K%	*f*	Test C $n = 88$	K%	*f*
easy	100.0	K1	address (n)	100.0	K1	alone	100.0	K1
move (vb)	100.0	K1	building (n)	100.0	K1	child (children)	100.0	K1
people	100.0	K1	carry	100.0	K1	cities (city)	100.0	K1
too (adv)	100.0	K1	different	100.0	K1	country (countries)	100.0	K1
computer	99.1	AWL	in	100.0	K1	live (vb)	100.0	K1
fun	98.2	K2	mean (vb)	100.0	K1	power (n)	100.0	K1
however	98.2	K1	money	100.0	K1	problem	100.0	K1
life	98.2	K1	team	100.0	AWL	travel (vb)	100.0	K1
small	98.2	K1	women	100.0	K1	busy (adj)	98.9	K2
wonderful	98.2	K1	work (vb)	100.0	K1	large	98.9	K1
active	97.3	K1	choose	99.0	K1	monster	98.9	OL
store (n)	97.3	K1	drive (vb)	99.0	K1	put (vb)	98.9	K1
married	96.5	K1	energy	99.0	AWL	information	98.9	K2
teach (taught)	96.5	K1	from	99.0	K1	aunt	97.8	K2
introduce	96.4	K1	second (adj)	99.0	K1	grow	97.8	K1
uncle	95.6	K2	famous	98.0	K1	plant (vb)	97.8	K1
across (adv)	94.7	K1	generation	98.0	AWL	population	97.7	K1
afraid	94.7	K2	on	98.0	K1	revolution	97.7	AWL
bag (n)	94.7	K2	volunteer (n)	98.0	AWL	volleyball	97.7	OL
similar	94.7	AWL	public (adj)	96.0	K1	email (vb)	96.6	OL
customer	93.8	K2	take care of	96.0	K1	in front of	96.6	K1
search (vb)	93.8	K2	neighbor	94.1	K1	outside (adj)	96.6	K1
draw (drew)	92.0	K1	position (n)	94.1	K1	sell (vb)	96.6	K1
gym	91.2	OL	daily	93.1	K1	stage (n)	96.6	K1
research (n)	91.2	AWL	percent	93.1	AWL	feelings	95.5	K1
close (adj)	90.3	K1	huge	91.1	OL	online	94.3	OL
quit	89.3	OL	private (adj)	91.1	K1	scientist	94.3	K1
corner (n)	88.5	K2	street children	91.1	K1	symbol	93.3	AWL
volume	87.6	AWL	terrible	91.1	K2	wonder (vb)	93.3	K1

Test A $n = 111$	K%	*f*	Test B $n = 101$	K%	*f*	Test C $n = 88$	K%	*f*
rent (n)	86.7	K2	transport (n)	91.1	AWL	double (vb)	93.2	K2
environment	85.0	AWL	average (n)	90.1	K1	homeless (adj)	93.2	K1
wedding	82.3	OL	either	90.0	K1	neighborhood	92.0	K1
familiar	81.4	K1	lonely	88.1	K2	dirty (adj)	91.1	K2
mall	77.9	OL	product	88.1	K1	drugs	91.0	OL
fix (vb)	76.8	K1	journalist	87.1	OL	entertainment	90.9	K2
unimportant	76.1	K1	repair (vb)	87.1	K2	according to (prep)	88.6	K1
childhood	74.3	K1	vision	87.1	AWL	full time	88.6	
available	73.5	AWL	habit	86.1	K2	realize	88.6	K1
olive tree	72.6		single-parent family	86.1		category	87.6	AWL
traditional family	70.8		unfair	86.1	K1	contain	87.5	K1
cousin	69.9	K2	benefit (n)	85.1	AWL	deliver	87.5	K2
retire	69.9	K2	equal (adj)	85.0	K1	publish	87.5	AWL
improvement	69.0	K2	release (vb)	84.2	AWL	traditional	87.5	AWL
influence (n)	69.0	K1	purpose (n)	83.2	K1	awake	85.2	K2
suggestion	69.0	K1	apologize	82.2	K2	prepare	85.1	K1
calculate	68.1	K2	financial (adj)	82.2	AWL	crowded (adj)	84.3	K1
demonstrate	68.1	AWL	charity (n)	81.2	OL	conversation	83.9	K2
psychologist	66.4	AWL	inform	81.2	K2	version	83.9	AWL
emotions	65.5	OL	focus on	79.2	AWL	basics	83.1	OL
virtual shopping mall	65.5		academic (adj)	76.2	AWL	teenager	83.0	OL
contrast (n)	64.6	AWL	crime	74.5	K2	site (n)	82.8	AWL
tend to	62.8	K2	uninteresting	73.3	OL	responsibility	81.8	K2
branch (n)	61.9	K1	AIDS (n)	73.0	AWL	occur	81.6	AWL
behavior	61.1	K2	orders (n)	72.3	K1	marriage	79.5	K1
logic	61.1	AWL	tough	72.3	K2	gather	77.3	K1
gender	60.2	AWL	barbecue (n)	71.3	OL	replace	77.0	K2
desires (n)	59.3	K1	great-grandparents	69.3		application (app)	73.9	K1
nuclear families	58.4		survey (n)	67.3	AWL	wealth	72.7	K1
predict	58.0	AWL	garage (n)	66.3	K2	unfamiliar	69.3	K1
portable	57.5	OL	argue	65.3	K2	relatives	69.0	K1
homelessness	56.6	OL	mammal	65.3	OL	complicated (adj)	68.2	K2
reward (n)	52.2	K2	politics	65.3	K1	megacity	65.5	OL

Test A $n = 111$	K%	*f*	Test B $n = 101$	K%	*f*	Test C $n = 88$	K%	*f*
symbolize	50.4	AWL	cost of living	63.4		make sense	64.8	
point out	46.0	K1	satisfaction	61.4	K2	profit (n)	63.2	K1
anxious	45.1	K2	emotional	56.4	OL	take responsibility	62.8	
divorce (n)	44.2	OL	competition	55.4	K2	struggle (vb)	59.8	K1
evidence (n)	44.2	AWL	resident	52.5	AWL	illegal (adj)	57.5	AWL
feminist	44.2	OL	socialize	49.5	OL	adulthood	57.0	AWL
appropriate (adj)	40.2	AWL	annual (adj)	41.6	AWL	adulthood	56.8	AWL
mixture	37.2	K2	conflict (n)	39.6	AWL	emotionally	56.8	OL
purchases (n)	37.2	AWL	donate	37.6	OL	household	53.5	OL
embarrassing (adj)	36.8	OL	radical (adj)	31.7	AWL	ethnic group	52.3	AWL
prevention	34.8	K1	extended family	30.7		first-born (n)	51.1	
hierarchy	27.7	AWL	prediction	27.7	AWL	anti- (prefix)	44.9	OL
alternate (vb)	27.4	AWL	conventional	25.7	AWL	format (n)	42.0	AWL
optimistic	22.1	OL	oak tree	21.8		hardship	41.9	K1
problematic	22.1	K1	home improvement products	20.8		eye scan	36.0	
gourmet (adj)	14.2	OL	reunion	20.8	OL	density	23.3	OL
runster	8.0	NW	hormone	19.8	OL	generosity	14.9	K2
ecution	8.0	NW	life expectancy	11.9		brag (vb)	14.0	OL
mengel	7.1	NW	ecutian	5.0	NW	runster	10.2	NW
Freud	6.2	OL	runster	5.0	NW	ecutian	4.6	NW
wuggy	6.2	NW	mengel	3.0	NW	pregime	3.5	NW
pregime	6.2	NW	pregime	2.0	NW	wuggy	2.2	NW
census	5.3	OL	wuggy	0.0	NW	mengel	0.0	NW

K% = % of learners reporting the item known (item facility in percent)

f = item frequency for that lemma (from Web Vocabprofile, Cobb, 2006)

K1 = first thousand words

K2 = second thousand words

AWL = Academic Word List

OL = off list, not in K1, K2, or AWL

NW = nonword

Blank = frequency not obtained