The Creation and Validation of a Listening Vocabulary Levels Test

Stuart McLean  
*Kansai University, Japan*

Brandon Kramer  
*Momoyama Gakuin University, Japan*

David Beglar  
*Temple University, Japan Campus*

Corresponding author:  
Stuart McLean, Faculty of Foreign Language Studies/Graduate School of Foreign Language Education and Research, Kansai University, 3-3-35, Yamate-cho, Suita-shi, Osaka, Japan 564-8680  
E-mail: stuart93@me.com

This is the final copy which was submitted to *Language Teaching Research*, to be published in early 2015.

Please cite as follows:  
Abstract

An important gap in the field of second language vocabulary assessment concerns the lack of validated tests measuring aural vocabulary knowledge. The primary purpose of this study is to introduce and provide preliminary validity evidence for the Listening Vocabulary Levels Test (LVLT), which has been designed as a diagnostic tool to measure knowledge of the first five 1000-word frequency levels and the Academic Word List (AWL). Quantitative analyses based on the Rasch model utilized several aspects of Messick’s validation framework. The findings indicated that (1) the items showed sufficient spread of difficulty, (2) the majority of the items displayed good fit to the Rasch model, (3) items and persons generally performed as predicted by a priori hypotheses, (4) the LVLT correlated with Parts 1 and 2 of the TOEIC listening test at .54, (5) the items displayed a high degree of unidimensionality, (6) the items showed a strong degree of measurement invariance with disattenuated Pearson correlations of .97 and .98 for person measures estimated with different sets of items, and (7) carelessness and guessing exerted only minor influences on test scores. Follow-up interviews and qualitative analyses indicated that the LVLT measures the intended construct of aural vocabulary knowledge, the format is easily understood, and the test has high face validity. This test fills an important gap in the field of second language vocabulary assessment by providing teachers and researchers with a way to assess aural vocabulary knowledge.

Keywords
item invariance, Rasch model, test validity, unidimensionality, listening vocabulary, vocabulary test
The creation and validation of a listening vocabulary levels test

I Introduction

This paper describes the creation of a Listening Vocabulary Levels Test (LVLT) and the initial efforts to provide validity evidence for its use as a diagnostic instrument designed to measure Japanese learners’ aural vocabulary knowledge of English words from the first five 1000-word frequency levels and the Academic Word List (AWL) (Coxhead, 2000). A quantitative approach as taken by Beglar (2010) when validating the Vocabulary Size Test (Nation & Beglar, 2007) was combined with Schmitt, Schmitt, and Clapham’s (2001) qualitative approach through which examinees were interviewed to determine their knowledge of the target items and impressions of the LVLT. This mixed method test validation approach provides preliminary evidence that the LVLT produces accurate estimates of aural vocabulary knowledge of the targeted word-frequency levels and the AWL.

II Current listening vocabulary tests

Two tests of aural vocabulary knowledge can be found in the second language vocabulary research literature. The first is Fountain and Nation’s (2000) lexically graded dictation test, which was designed as the vocabulary component of a placement examination using five
word-frequency levels based on Thorndike and Lorge’s (1944) frequency list. Although displaying high internal reliability (> .95), it is unclear to what extent the test measures aural vocabulary knowledge, as it has not been thoroughly analyzed and the test format potentially includes confounding variables such as orthographic knowledge.

The second is the AuralLex (A-Lex) test, which was designed to measure aural vocabulary size (Milton & Hopkins, 2006). This digitally administered test uses a yes/no format in which test-takers listen to words selected from the first 5000 lemmatized words from dated corpora along with pseudowords. Although this test can be administered quickly, it can be used with learners at any proficiency level, and it correlates with IELTS listening and speaking subscores at .67 and .71, respectively (Milton et al., 2010), the Lex tests are ‘ideally suited for use in low-stakes environments, where speed and ease of administration is more important than accuracy’ (X_LEX, 2005, p. 4). The A-Lex has two potential weaknesses. First, the target words can be listened to an unlimited number of times, an approach that is unrepresentative of the on-line processing required in most authentic contexts (Buck, 2001). Second, it is unclear to what degree examinees know the meaning of each target word, as distinct from the knowledge of each word’s existence (Read, 2000).

III Gaps in the literature
A listening vocabulary test is needed for three reasons. First, second language learners' orthographic and phonological lexical knowledge can differ significantly (Milton & Hopkins, 2006; AUTHOR & AUTHOR, 2014). Second, on-line phonological processing of lexical forms places greater demands on working memory than off-line orthographic processing. Reasons for this include the greater complexity and large amount of individual variation of speech (Rost, 2011, p. 101). Third, a listening vocabulary test, in contrast to a written vocabulary test, requires knowledge of English phonology, linking, rhythm, and stress patterns.

Further, the second language vocabulary assessment literature makes it clear that there is a need for a test that measures aural English vocabulary knowledge. This gap in the field has been noted by multiple second language vocabulary researchers. For instance, Milton (2009) stated, 'strangely, it is an aspect of knowledge that has attracted very little systematic interest from researchers' (pp. 92-93), while Stæhr (2009), in an attempt to investigate the relationship between lexical knowledge and listening comprehension, lamented that 'no suitable test of phonological vocabulary size existed' (p. 597).

IV The Listening Vocabulary Levels Test

The primary purpose of this study is to introduce and provide preliminary validity evidence for a new aural vocabulary test, the LVLT. This test is intended as a diagnostic and
achievement instrument that measures knowledge of English lexis from the first five 1000-word frequency levels and the AWL (Coxhead, 2000) for either pedagogical or research purposes. The 1000 to 5000 word-frequency levels are measured with 24 items per level, while the AWL is measured with 30 items. The 150-item test, if given as a whole, can be completed in 30 minutes; however, specific test sections can be administered depending on the needs of the researcher or teacher.

1 Test format

The LVLT uses the same multiple-choice format as the Vocabulary Size Test (VST). This format provides multiple benefits: (a) The difficulty of each distractor can be manipulated; (b) marking is efficient, reliable, and can be conducted electronically; (c) item analyses can be conducted easily; (d) it supports item independence, a necessary criterion for Rasch analysis; and (e) participants can demonstrate their knowledge of the meaning of each item. Each item has four options, from which examinees must choose the Japanese response with the closest meaning to the target word. The four options are provided in the examinees’ first language in order to isolate the construct of aural vocabulary knowledge from other constructs such as L2 reading ability. An example item is shown below (the translations in parentheses were not visible to examinees):
[Examinees hear: 'School: This is a big school.]

1.
   a. 銀行 (bank)
   b. 海の動物 (sea animal)
   c. 学校 (school)
   d. 家 (house)

As shown above, examinees listen to a single reading of the target word followed by the word in a non-defining context sentence. Each item is read only once, as (a) this is consistent with the demands made on a listener’s aural vocabulary knowledge in most authentic L2 listening situations (Buck, 2001, pp. 170-171), (b) it is consistent with the format used by major institutional listening test formats such as the TOEFL, and (c) reading each question multiple times would extend the length of the test, reducing the number of items that can be administered. The defining context sentence clarifies the target word's part of speech, helps the examinees perceive the word as an item of language use, and provides associational assistance in accessing the meaning of the target word (Henning, 1991). There is a 5-second pause between the reading of each item, a length of time that was found in pilot
testing to provide examinees with sufficient time to process the aural input while maintaining efficiency. A 15-second pause is provided between test levels to allow examinees to turn test pages and prepare for the next section.

The test instructions (see Appendix A for an English translation), which were piloted and revised multiple times, were written in the examinees’ native language (Japanese) to promote rapid and accurate comprehension. To reduce the effects of guessing, the instructors verbally instructed the participants in English, 'If you do not know a word at all, please leave it blank. However, if you think there is a chance that you may know the word, please try to answer.' After reading the instructions, the examinees completed two recorded practice questions before beginning the test.

The audio portion of the test was recorded by a male native speaker of General American English because American English is the form primarily taught in the Japanese educational system. The recording was conducted in a sound-proof music studio using high quality recording equipment (Tascam DR-40 Linear PCM Recorder), and later cleaned and edited using the free open-source digital audio editing program Audacity <audacity.sourceforge.net>.

2 The source of target vocabulary
The LVLT items are from Nation's (2012) British National Corpus (BNC)/Corpus of Contemporary American English (COCA) word lists. The first and second 1000-word family lists of the BNC/COCA were derived from a specially designed 10 million token corpus that includes 6 million tokens from spoken British and American English. The corpus was created to provide a set of high frequency word lists suitable for teaching and course design (Nation, 2012). The lists for the third 1000-word family and above were created using BNC/COCA rankings after removing the word families from the first 2000 words of the BNC/COCA. The BNC/COCA word lists provide a strong basis for the LVLT, as they are representative of both British and North American varieties of English and are partly based on a spoken corpus. As Webb and Sasao (2013) stated, 'the new BNC/COCA lists should be representative of current English, and provide a far better indication of the vocabulary being used by native speakers today than the lists used for the creation of the earlier versions of the VLT' (p. 267).

The word family unit was used for the LVLT, as it was the basis for the creation of the twenty-five 1000-word BNC/COCA frequency lists (available at <victoria.ac.nz/lals/about/staff/paul-nation>). The word family unit is a reasonable choice when measuring vocabulary knowledge, frequency, and coverage because even low proficiency learners have some control of word building devices and they can perceive both
a formal and semantic relationship between regularly affixed members of a word family (Nation & Beglar, 2007). There is also evidence that the word family is a psychologically real unit (Nagy et al., 1989; Bertram, Baayen, & Schreuder, 2000; Bertram, Laine, & Virkkala, 2000).

The LVLT covers the first five 1,000-word frequency levels of the BNC/COCA, which provides adequate coverage for numerous listening genres. Nation (2006) reported that the first five 1000-word levels of the BNC provides approximately 96-97% coverage of conversations. Moreover, Nation has stated that knowledge of the first five BNC/COCA 1,000-word bands provides 98% of coverage of talk-back, interviews, and conversations (I.S.P. Nation, personal communication, August 5, 2014), a figure that generally results in adequate comprehension of aural texts (van Zeeland & Schmitt, 2012). Furthermore, Webb and Sasao (2013) stated that “mastery of the 5000 word level may be challenging for all but advanced learners, so assessing knowledge at the five most frequent levels may represent the greatest range in vocabulary learning for the majority of L2 learners” (p. 266).

3 Test creation

The LVLT items from the first five 1,000-word frequency levels were created through a retrofit and redesign of previous VST items. This approach was adopted because the
functioning of many of these items had been previously validated with similar participants and in a similar EFL context (Beglar, 2010). There were not enough items available on the Japanese translation of the Vocabulary Size Test <victoria.ac.nz/lals/about/staff/publications/paul-nation/Vocab_Size_Test_Japanese.pdf> to meet the goal of piloting at least 25 items per level (30 for the AWL), however, so item specifications were reverse engineered from descriptions of previous tests (e.g., Nation & Beglar, 2007) and specification-driven test assembly was implemented as recommended in Fulcher and Davidson (2007).

The retrofitting of items for the 1000- to 5000-word levels was accomplished by adapting items from three versions of the Vocabulary Size Test. Two versions were downloaded from <victoria.ac.nz/lals/about/staff/paul-nation> and the third version was obtained through personal correspondence with I.S.P. Nation. The items were then re-assigned to their appropriate BNC/COCA level. For instance, *period* and *basis* were moved from the first to the second 1000-word level. Moreover, items such as *nil*, present in the second 1000-word frequency level of the Vocabulary Size Test, and not present in the first five 1000-word levels of the BNC/COCA corpus, were removed from the LVLT. The context sentence for each item was then presented to volunteers in early pilot testing with pseudoword replacements for
each target word to ensure the test was not conflating the construct of L2 contextual inferencing with aural vocabulary knowledge.

Finally, the key and distractors for each item were directly translated to the core Japanese meanings for all answer choices. For example, the key for the target word shoe was written as 'the thing you wear on your foot' in version B of the Vocabulary Size Test. Rather than directly translate 'the thing you wear on your foot' into Japanese, 靴 (kutsu), the Japanese word for shoe, was used. These direct translations were thought to be less cognitively demanding for examinees while they search for the best match to the target word in the limited time available. When the translation of the target word was an English loanword in Japanese (e.g., test exists in Japanese as テスト [tesuto]), the direct translation was replaced with an alternative Japanese word or phrase, as examinees would otherwise be able to select the correct answer through phonological matching alone. This solution was preferable to removing the item because loanwords are a legitimate part of the Japanese language and thus deserve to be represented in the LVLT.

The AWL was included on the LVLT for three reasons: (a) given the large number of academic English programs throughout the world, it is important that test users can assess knowledge of AWL vocabulary; (b) while the AWL covers an average of 10% of tokens in academic texts (Coxhead, 2000), Dang and Webb (2014) found that the AWL 4.41%
coverage of academic spoken English; and (c) we wanted to provide maximum flexibility to
users of the LVLT by allowing them to easily assess knowledge of academic lexis.

AWL items were created using the item specifications listed in Appendix B. The
AWL is divided into 10 sublists according to word frequency, with 60 words in each of the
first nine sublists and 30 in the tenth (Coxhead, 2000). To ensure an even distribution of
items, three target words were chosen from each of the first nine sublists and two from the
tenth using a random number generator. The final item was chosen at random from the entire
AWL. The actual target word for each test item was the headword of the AWL word family
(as listed in Coxhead, 2000). After each target word was chosen, distractor choices were
randomly selected from the same sublist as the target word until the desired part of speech
was obtained. If a suitable distractor could not be found in the same sublist, the same process
was repeated one sublist lower (i.e., higher frequency sublist).

Both qualitative and quantitative piloting took place to ensure that each distractor was
a plausible choice. After selecting the target items, a generic sentence that provided context
but no clues to the target word’s meaning was written for each item, consulting concordancer output from the BNC/CBOCA
corpus <www.lextutor.ca/conc/eng/> for authentic examples when the target word had more
than one possible meaning or use. If a sentence fit only some of the distractors, the non-
conforming distractor was discarded and a new one was chosen randomly until one that fit the necessary criteria was identified. Lastly, each example sentence was checked to ensure that only words from the first and second 1000-word frequency levels were included. The items were repeatedly piloted with a small group of native Japanese speakers until no significant problems remained.

V Methodology

1 Participants

The participants in this study were 214 Japanese university students. 186 participants (56 female and 130 male) were low proficiency first- and second-year students studying in various departments in a private university. TOEIC scores were unavailable for these students, but students with similar English proficiency from a different department had a mean TOEIC score of 280. Fourteen participants were first-year female university students with a mean TOEIC score of 521 ($SD = 108$; Range 340-615) of various academic majors attending a different private university. Finally, fourteen (9 female and 5 male) participants were third- and fourth-year students at a public university majoring in English, with a mean TOEIC score of 826 ($SD = 73$; Range 730-915).

2 Instruments
Listening vocabulary levels test. The LVLT was initially made up of 29, 26, 30, 27, and 25 items at the first, second, third, fourth, and fifth 1000-word frequency levels, respectively, and 30 items from the AWL (k = 167). The items from each 1000-word frequency level and the AWL were divided into two halves, due to classroom time constraints. The recorded tests for the first and second halves were 15.7 and 16.3 minutes, respectively.

Parts 1 and 2 of the TOEIC listening section. Parts 1 and 2 of the TOEIC listening section were administered to obtain an estimate of the participants’ English listening proficiency. In Part 1 (k = 10) examinees choose the best description of a printed picture from a series of four audio prompts. In Part 2 (k = 30) examinees hear a question or statement and select the best response from a series of three audio prompts. The Rasch item reliability estimate was .97.

3 Procedures
Participants completed the first half of the LVLT during the first week of class. In the following week, the second half of the LVLT was administered followed by the TOEIC subsections. The test administered through speakers, sound checks confirmed that all participants clearly heard the instructions and practice items, and at no time did the participants mention any difficulties hearing the recordings.
4 Analyses

Test data were input into an Excel spreadsheet and then exported to WINSTEPS 3.80.1 (Linacre, 2013) and calibrated using the Rasch dichotomous model (Rasch, 1960). The Rasch model has numerous strengths, such as providing equal interval item and person measures, item and person fit indices, detailed information concerning item and person dimensionality, as well as item and person reliability estimates.¹

A preliminary Rasch analysis indicated that each level of the test had at least 24 items that performed well. Therefore, the best-performing 24 items were retained for each 1000-word band. All 30 AWL items also performed well and were therefore retained. This resulted in a 150-item test.

VI Results

In this section we report the results of Rasch analyses used to investigate four aspects of validity described by Messick (1995) as well as carelessness and guessing. We also examine interpretability as defined by the Medical Outcome Trust Scientific Advisory Committee (1995) and report on a qualitative investigation of test functioning with 22 participants.
Content aspect of construct validity

The content aspect of construct validity consists of three parts: content relevance, representativeness, and technical quality (Messick, 1995). First, content relevance, which concerns the relevance of the test items to the targeted construct, was described in subsection 2, The source of target vocabulary, and subsection 3, Test creation.

Representativeness. Representativeness concerns the degree to which the test is sensitive to variations in the measured construct (Borsboom et al., 2004). Representativeness can be operationalized in terms of whether, (a) a sufficient number of items is included on the test, (b) the empirical item hierarchy shows sufficient spread, and (c) gaps exist in the empirical item hierarchy. Each of these aspects is addressed in Figure 1, which shows the linear relationship between the 214 test-takers and 150 items. On the far left side is the Rasch logit scale. Persons are indicated by ‘#’ (representing three test-takers) and ‘.’ (representing one or two test-takers). More able persons (i.e., higher-scoring persons) are toward the top of the figure and less able persons are toward the bottom. The items are labelled according to their word-frequency level and the item number on the test form; thus, the most difficult item, item 5000-126, was in the fifth 1000-word frequency level and was test item 126. Items from the
Academic Word List are labelled AWL. More difficult items (i.e., items with fewer correct responses) are toward the top of the figure and easier items are toward the bottom.

As shown in Figure 1, 150 items is sufficient to measure the participants’ aural vocabulary knowledge, as the items ranged in difficulty from -4.79 logits (item 5000-127; spider) to 3.34 logits (item 5000-126; slaughter), while 96.7% of the person ability estimates were between -1.00 and +3.00 logits. No floor or ceiling effects were present, and because of the large number of items on the test form, the person ability estimates were precise, as indicated by the mean SE of .21 logits (SD = .02). The spread of the item calibrations was also assessed through the Rasch item strata statistic, which indicates the number of statistically distinct strata of item difficulty estimates that are separated by at least three errors of measurement. The strata statistic of 9.09 indicated that the items were separated into nine statistically distinct levels of difficulty. Finally, no significant gaps were present in the item hierarchy.
**Figure 1.** Wright map of person measures and item calibrations. Each ‘#’ is 3 persons and each ‘.’ is 1-2 persons. M = Mean; S = 1 SD; T = 2 SD. AWL = Academic Word List.
Technical quality. Technical quality of the items was first determined by inspecting the items’ fit to the Rasch model using the Rasch Infit mean-square (Infit MNSQ) statistic. A value of ±2 standard deviations was used to determine the Infit MNSQ criterion (McNamara, 1996). As the standard deviation was .08, this resulted in a strict Infit MNSQ criterion of .84-1.16. Overfit, which indicates a possible violation of the assumption of local independence, was indicated by Infit MNSQ statistics under .84. Item 3000-69 (solution) slightly overfit the model with an Infit MNSQ statistic of .83. Underfit is viewed as a more serious problem than overfit because it indicates unexpected responses to an item by persons with ability estimates near the item’s difficulty estimate. Underfit was indicated by Infit MNSQ statistics over 1.16. Four of the 150 items slightly underfit the model: items 5000-121 (trench; Infit MNSQ = 1.26), 4000-101 (evacuated; Infit MNSQ = 1.25), 5000-132 (crook; Infit MNSQ = 1.23), and 5000-134 (warfare; Infit MNSQ = 1.19). An inspection of the most misfitting response strings indicated that the underfit was caused by seven or fewer persons (3% or less) out of the 214 test-takers in each case, indicating no serious flaws in these items. Furthermore, the five items listed above represent only a 3.33% misfit rate (i.e., 5 out of 150 items); thus, the LVLT items showed very good fit to the Rasch model overall.

2 The substantive aspect of construct validity
The substantive aspect of construct validity concerns the rationale for the item difficulty and the person ability hierarchies shown in Figure 1. Our hypothesis for the item hierarchy was that higher frequency words would be easier than lower frequency words (e.g., see Beglar, 2010; Milton, 2009, Chapter 5, for support for this hypothesis); thus, the hypothesized order of difficulty was 1000 < 2000 < 3000 < 4000 < 5000. The words in the AWL come from various word-frequency levels, so we did not make an a priori hypothesis regarding its difficulty. A one-way ANOVA was conducted to investigate this hypothesis. The independent variable was word-frequency level and the dependent variable was the mean raw scores for each of the five word-frequency levels. The assumptions for the analysis were met except that the variances among the groups were unequal, Levene statistic \((4,1065) = 16.71, p < .001\); thus, the Welch and Brown Forsythe tests were utilized. As both tests were statistically significant, only the results of the Welch test are reported. The ANOVA was significant, \(F(4,525.95) = 513.83, p < .001\), partial \(\eta^2 = .56\), and Dunnett’s T3 post-hoc test indicated that all pairwise comparisons were significant at \(p < .05\) except for the 4000 and 5000 word-frequency levels, which did not differ significantly. These results generally supported the a priori hypothesis concerning item difficulty. The means and 95% confidence intervals for each of the five word-frequency levels are displayed in Figure 2.
Persons were hypothesized to differ in their aural vocabulary sizes based on their listening proficiency as measured by Parts 1 and 2 of the TOEIC test. The participants were first divided into three groups based on their TOEIC listening scores: High proficiency ($n = 71$), Intermediate proficiency ($n = 71$), and Low proficiency ($n = 72$). A one-
way ANOVA was conducted to determine if the three groups listening proficiency differed significantly. The independent variable was TOEIC group (three proficiency levels) and the dependent variable was the Rasch person ability estimates for the TOEIC listening section. The assumptions for the analysis were met except that the variances among the groups were unequal, Levene statistic (2,211) = 23.78, $p < .001$; thus, the Welch and Brown Forsythe tests were utilized. As both tests were statistically significant, only the results of the Welch test are reported. The ANOVA was significant, $F(2, 103.50) = 113.44, p < .001$, partial $\eta^2 = .61$, and Dunnett’s T3 post-hoc test indicated that all pairwise comparisons were significant at $p < .001$. Having established that the three groups differed significantly in terms of listening proficiency, it was hypothesized that the groups would also differ in terms of their Rasch person ability estimates on the LVLT. The hypothesis that High proficiency > Intermediate proficiency > Low proficiency was investigated by conducting another one-way ANOVA. The independent variable was the same three TOEIC proficiency groups and the dependent variable was the Rasch LVLT person ability estimates. The assumptions of the analysis were checked and met. The ANOVA was significant, $F(2, 211) = 25.14, p < .001$, partial $\eta^2 = .19$, and Bonferroni post hoc tests indicated that all pairwise comparisons were significant at $p < .001$; thus, the hypothesis that higher listening proficiency would be accompanied by a greater listening vocabulary size was supported.
Based on previous investigations of the relationship between lexical knowledge and listening comprehension (e.g., Alderson, 2005, cited in Schmitt, 2010, p. 5), we also hypothesized that the Rasch person ability estimates for the LVLT and the TOEIC listening section results would correlate at approximately .60. This hypothesis concerns the strength of the ordering of persons on the LVLT and the measure of listening proficiency. The Pearson correlation of .54 ($p < .001$) indicated that the LVLT and the TOEIC scores had a moderately strong relationship that was slightly below the hypothesized magnitude.

Finally, as recommended by Davies et al. (1999), the test was administered to three native English speakers who had passed the highest level of a Japanese proficiency test and three native Japanese speakers who had completed an English-language undergraduate or graduate program abroad. The native English speakers’ scores ranged from 146 to 150, and the native Japanese speakers’ scores ranged from 143 to 145 (maximum score = 150). Incorrect responses were primarily from the 5000-word or AWL sections. These results confirm that very proficient examinees can score highly on the LVLT.

3 The structural aspect of construct validity

The structural aspect of construct validity concerns the hypothesized dimensionality of the construct being measured (Messick, 1989). We hypothesized that the LVLT measures a single primary construct, aural vocabulary knowledge, given that an analysis of
the Vocabulary Size Test (Beglar, 2010) yielded strongly unidimensional results. This hypothesis was tested using the Rasch PCA of item residuals. The Rasch model extracts the first major dimension in the data, and if the data are fundamentally unidimensional, no systematic relationships should be present in the residual variance (i.e., the residuals should be random). In these data, the Rasch model accounted for 38.0% of the variance (eigenvalue = 92.0), a figure that matched what would be expected if the data fit the Rasch model perfectly. Unidimensionality, however, is not dependent on the amount of variance accounted for by the primary latent dimension, but rather by the size and interpretation of residual contrasts. The first residual contrast accounted for 2.2% of the variance (eigenvalue = 5.4). While this percentage of residual variance is extremely low and strongly suggests unidimensionality, the eigenvalue was above the chance level of 2.0, so the residual loadings of the individual items were inspected. Six items had a residual loading above .40 (1,000-20, 1,000-22, 1,000-25, 1,000-19, 2,000-42, and 5,000-127). All six were extremely easy items with difficulty estimates ranging from -2.22 to -4.79. One item (AWL-140) had a residual loading below -.40 and a difficulty estimate of 2.48. Thus, because six easy items formed one set of related residuals and one difficult item formed the second set, no meaningful second dimension appeared in the data (i.e., item difficulty is not a meaningful dimension). In addition, the number and strength of the residual loadings did not constitute a
factor using Stevens’ (2002, p. 395) criteria. Together these results suggest that the LVLT measures one construct, presumably aural vocabulary knowledge.

4 The generalizability aspect of construct validity

Generalizability concerns the principle of invariance, which was described by Rasch (1960):
'The comparison between two stimuli should be independent of which particular individuals were instrumental for the comparison; and it should also be independent of which other stimuli within the considered class were or might also have been compared' (p. 332). One meaning of this statement is that if test-takers are assessed with different sets of items from the LVLT, the resulting person ability estimates should be similar, provided that the items fit the Rasch model and display sufficient variance in terms of item difficulty. Generalizability was tested by creating two test forms. The items at each frequency level and the AWL items were randomly divided to create two 75-item test forms, and Rasch person ability estimates were obtained. If the items measure the same construct, then the disattenuated Pearson correlation of the two sets should be above .90 (i.e., the level at which multicolinearity exists). In this case, the Pearson correlation was .87 and the disattenuated correlation was 1.00. Thus, the two randomly chosen sets of items produced strongly similar person ability estimates that confirmed item invariance.
Generalizability was also investigated by determining the degree to which three versions of the LVLT were free of measurement error. Rasch item reliability estimates and item separation statistics were calculated for the full 150-item test and the two 75-item forms described above. The Rasch item reliability estimates (Rasch item separation statistics) for the three tests were .98 (6.57), .98 (7.20), and .97 (5.71), respectively.

5 Carelessness and guessing
Carelessness is defined as test-takers incorrectly answering items they have a high probability of knowing, while guessing is defined as test-takers correctly answering items they have a low probability of knowing. Both carelessness and guessing can be empirically investigated with the Rasch model. Carelessness was investigated using two criteria. First, items that were 1 or more logits below the test-taker’s person ability estimate (probability of a correct response = .73), yet were answered incorrectly, were identified (hereafter CUTHI 1), then the process was repeated using a difference of 2 or more logits (probability of a correct response = .88) (hereafter CUTHI 2). Rasch person ability estimates were produced using the original 150-item data set, the data produced using CUTHI 1, and those produced using CUTHI 2. Pearson correlation coefficients were calculated to determine the strength of the relationships among the three variables. The correlation coefficients between the original data and CUTHI 1 and CUTHI 2 were .94.
and .99, respectively. These results indicated that carelessness had a negligible effect on the measurement of aural vocabulary knowledge for these test-takers.

Guessing was also investigated using two criteria. First, items that were 1 or more logits above the test-taker’s person ability estimate (probability of a correct response = .27) yet were answered correctly, were identified (hereafter CUTLO 1), then the process was repeated using a difference of 2 or more logits (probability of a correct response = .12) (hereafter CUTLO 2). Rasch person ability estimates were produced using the original 150-item data set, the data produced using CUTLO 1, and those produced using CUTLO 2. Pearson correlation coefficients were calculated. The correlation coefficients between the original data and CUTLO 1 and CUTLO 2 were .98 and .99, respectively. These results indicate that guessing also had a negligible effect on the measurement of aural vocabulary knowledge for these test-takers.

6 Interpretability

Another important issue concerns test interpretability, which is the degree to which qualitative meaning can be assigned to the quantitative measures produced by a test instrument (Medical Outcomes Trust Scientific Advisory Committee, 1995). Interpreting LVLT results will vary, as test users can select the levels that are appropriate for their program or research purposes.
First, the LVLT can be used to diagnose learners’ knowledge in the beginning of a course of study and estimate achievement throughout the course of study (i.e., formative assessment) and upon completion of a course (i.e., summative achievement). For instance, in courses designed for lower proficiency learners, instructors might set a course goal that learners acquire knowledge of the 2000 high-frequency words of English, while in an academic English program, the goal might be to acquire AWL lexis. In both cases, these words are so important that any gaps detected with the LVLT should be addressed.

A second use of the LVLT is to assess learners’ readiness for a particular course of study. Instructors could first estimate the aural vocabulary load of course materials. Given that research has shown that 98% coverage is sufficient for comprehending informal narratives (van Zeeland & Schmitt, 2012) and academic speech (Stæhr, 2009), the LVLT can be used to estimate learners’ knowledge of lexis at particular word-frequency levels to determine if they have the necessary lexical knowledge to comprehend course materials. For instance, learners would need to correctly answer at least 45-46 of the 48 items that make up the 1000 and 2000 word-frequency levels on the LVLT to comprehend aural texts made up of the first 2000 high frequency words of English.

7 Qualitative investigations into the LVLT
In line with Schmitt's (1999) argument that construct validity evidence can be provided by directly establishing that examinees’ lexical knowledge matches their responses on an instrument, interviews were conducted with 22 volunteer examinees (mean Rasch ability estimate on the LVLT = 1.97, SD = .83) on the same day or the day following the administration of the second half of the LVLT. The participants' knowledge of the first eight items of each 1000-word frequency band and the first 10 AWL items \(k = 50\) was investigated by conducting one-on-one interviews. None of the interviewees reported checking the meaning of the LVLT items after taking the test. They also reported that they thought that the LVLT was a reasonable measure of aural vocabulary, and that they had no difficulty understanding the test directions and format.

The interviewer first read the item stem and the participant was asked to provide an L1 (Japanese) translation. When the participant was unable to do this they were shown the four answer choices and asked to select one while explaining their rationale. Two interviewers assessed the interviewees' responses by assigning a know or don't know rating for each item. Inter-rater reliability, as estimated by Cohen's kappa during piloting and the main study, was .96 \(p < .001\). Second, each interviewee answered two practice test items and then two new items, one from the first 1000-word frequency level (name) and one
from the fifth 1000-word frequency level (shudder), under time constraints similar to those of LVLT using a think-aloud protocol.

The interviewees' know and don’t know ratings as judged by the interviewers were compared with the interviewees’ responses on the LVLT, and it was found that the LVLT measured the interviewees’ aural lexical knowledge accurately (Phi = .81, p < .001) (see Table 1 for a summary of the results). The interviewees knew the target word in the interview but did not answer correctly on the LVLT for 2.1% of the items. This result is consistent with the Rasch results concerning carelessness. Further probing during the interviews indicated that for these items, one or more distractors were usually too semantically similar to the target word, causing confusion.

Table 1. Comparison of Interview Results with the LVLT Results

<table>
<thead>
<tr>
<th>LVLT response</th>
<th>Correct</th>
<th>Incorrect</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knew</td>
<td>916 (83.3%)</td>
<td>23 (2.1%)</td>
<td>939 (85.4%)</td>
</tr>
<tr>
<td>Did not know</td>
<td>28 (2.5%)</td>
<td>133 (12.1%)</td>
<td>161 (14.6%)</td>
</tr>
<tr>
<td>Total</td>
<td>944 (85.8%)</td>
<td>156 (14.2%)</td>
<td>1100</td>
</tr>
</tbody>
</table>
Three lines of evidence indicate that the interviewees rarely guessed blindly. First, in only 2.5% of the cases interviewees correctly answered items on the LVLT yet were unable to demonstrate this knowledge during the interview. This small figure is consistent with the results of the Rasch analysis. Second, the interviewees reported blindly guessing on an average of only 8.3% of the items that they had no knowledge of on the 150-item LVLT. For instance, a student who had no knowledge of 20 items of the 150-item LVLT would have guessed on only one or two items. Most interviewees confirmed that they tried to follow the instructions to not answer items of which they had no knowledge. Third, the think-aloud protocol indicated that interviewees had enough time to select an answer based on knowledge of the target word, but not enough time to deduce the answer by excluding distractors. This might suggest that in such a situation interviewees would guess blindly. However, considering the high level of agreement between the test responses and interview responses (Phi = .81, p < .001), it is clear that the 22 interviewees' very rarely guessed blindly and only skipped unknown items. It should be remembered that these interviewees were disproportionately of higher ability, so this result should not be generalized to lower proficiency examinees.

Furthermore, mismatches between word knowledge as determined by the interviews and test responses were partly dependent on the learners' aural vocabulary size. More able
learners were slightly more likely to have missed test items for target words they knew, while less able learners were more likely to have correctly guessed an item when they could not demonstrate knowledge of the target word. This finding supports the argument for a conservative cut-off score, particularly for lower proficiency examinees, recommended as only missing zero or one item for each level. This recommendation is supported by the finding that knowledge of 98% of the vocabulary comprising aural texts is necessary for good comprehension (van Zeeland & Schmitt, 2012) and the first 2000 words of English play a critical role in all aural texts.

VII Conclusion
This study provides preliminary validity evidence for the LVLT, an aural vocabulary test that assesses knowledge of the first five 1000-word frequency levels and AWL vocabulary for diagnostic and achievement purposes. The quantitative results showed that items showed sufficient spread of difficulty, the majority of the items displayed good fit to the Rasch model, hypotheses concerning item difficulty and person ability were largely supported, the LVLT correlated fairly strongly with a shortened version of the TOEIC listening test, LVLT items formed a fundamentally unidimensional construct, multiple versions of the test were highly reliable, and carelessness and guessing were minimal. The qualitative
results indicated that the LVLT has high face validity, the format is easily understood by examinees, and it is an accurate measure of examinees’ knowledge of aural vocabulary.

In future work with the LVLT, the test should be expanded to include lower frequency word levels to accommodate higher proficiency learners. Versions of the LVLT also need to be created in other languages and their functioning investigated both quantitatively and qualitatively, as this study provides initial validation for the test format used in this study. Creating written forms of the LVLT designed to measure written receptive vocabulary would allow teachers to directly compare learners’ aural and written vocabularies.

The LVLT fills an important gap in the field of second language vocabulary assessment by providing a comprehensive measure of aural vocabulary knowledge. The test form and audio file are freely available and can be downloaded from <www.lvlt.info>.

Notes
1. Rasch reliability estimates are more conservative than Cronbach alpha estimates but are preferable because they are based on data that conform stochastically to the Rasch model. As such, the Rasch reliability estimate is an indication of the repeatability of the linear measures constructed by the Rasch model.
2. This definition of carelessness could also include unexpected gaps in test-takers’ lexical knowledge as found in AUTHOR, AUTHOR, and AUTHOR (2014), while this definition of guessing could also include unexpected specialized knowledge.

Acknowledgements

The authors would like to thank Dr. Rob Waring of Notre Dame Seishin University and Professor Trevor Holster of Fukuoka Women’s University for their helpful comments and advice on an earlier draft of this paper. We are also grateful to the anonymous Language Teaching Research reviewers for their constructive and insightful critiques and comments that have improved this work. Any remaining faults are the sole responsibility of the authors.

References


Rost, M. (2011). Teaching and researching listening. Harlow: Pearson Education.


Appendix A: Translation of LVLT instructions

This is a vocabulary test. Each English word will be read together with an example sentence. Select the Japanese word from the choices (a~d) that has the closest meaning to the English word being read.

Each question will be read only once.

Example Problem
1. a. 食べた (ate)  
b. 待った (waited)  
c. 見た (saw)  
d. 寝た (slept)

The correct answer is b.

If you do not know a word at all, please leave it blank. However, if you think there is a chance that you may know the word, please try to answer.

Let’s practice some problems.

Practice Problem 1
a. 強い (strong)  
b. 幸せな (happy)  
c. 食べすぎる (eats too much)  
d. 親切な (kind)  

Practice Problem 2
a. ～について話します (talk about)  
b. ～を運ぶ (carry)  
c. ～に名前を書く (write your name on)  
d. ～を振る (shake)

Because this is a listening test, please do not speak until it is finished. Let’s begin.
Appendix B: Specifications for New Items

Example Item: School: This is a big <school>.

a. 銀行 [bank]
b. 海の動物 [sea animal]
c. 学校 [school]
d. 家 [house]

**Overall**
- The target word in isolation and the context sentence should be read only once
- There should be 5 seconds between questions
- There should be 15 seconds between test sections
- The answer key should be randomly generated
- Avoid gender-biased language and have balanced gender representation throughout the test

**Target words**
- Written in citation form
- From frequency list based on established corpus
- Random sampling of words from each word-frequency level

**Context sentence**
- Context sentences in the first two 1000-word levels should be written using vocabulary within the first 1000-word level
- Context sentences in the third 1000-word level and above should be written using vocabulary within the first two 1000-word levels
- In cases where the part of speech is ambiguous, the most common form should be used based on frequency data
- The accompanying sentence should be as contextualized as possible without giving hints to the meaning of the target word.

**Distractors**
- The Japanese translations should be the most direct translations possible, unless it is a *katakana* loan word or a very difficult Japanese word, in which case a simple explanation should be provided in Japanese.
• Distractors should be of similar word frequency and difficulty level as the target word
• To as great a degree as possible, all distractors should be equally plausible in the context sentence