Language Testing in Asia

RESEARCH Open Access

A Rasch-based validation of the Vietnamese version of the Listening Vocabulary Levels Test

Hung Tan Ha

Correspondence: hatanhung1991@gmail.com
School of Foreign Languages, University of Economics Ho Chi Minh City, 279 Nguyen Tri Phuong Street, District 10, Ho Chi Minh City, Vietnam

Abstract

The Listening Vocabulary Levels Test (LVLT) created by McLean et al. Language Teaching Research 19:741-760, 2015 filled an important gap in the field of second language assessment by introducing an instrument for the measurement of phonological vocabulary knowledge. However, few attempts have been made to provide further validity evidence for the LVLT and no Vietnamese version of the test has been created to date. The present study describes the development and validation of the Vietnamese version of the LVLT. Data was collected from 311 Vietnamese university students and then analyzed based on the Rasch model using several aspects of Messick's, Educational Measurement, 1989; American Psychologist 50:741–749, 1995 validation framework. Supportive evidence for the test's validity was provided. First, the test items showed very good fit to the Rasch model and presented a sufficient spread of difficulty. Second, the items displayed sound unidimensionality and were locally independent. Finally, the Vietnamese version of the LVLT showed a high degree of generalizability and was found to positively correlate with the IELTS listening test at 0.65.

Keywords: Vocabulary, Rasch model, Validation, Aural knowledge, Listening vocabulary test

Background of the study

Measuring learners' vocabulary has a long history and has been viewed as one of the most important aspects of language education in general and language assessment in particular (Nation, 2013; Schmitt et al., 2020). A sound understanding of the learners' vocabulary knowledge could facilitate various decisions, from the selection of teaching materials, teaching approaches to the extent to which L1 should be used in the classroom. It was generally accepted that the learners should be familiar with at least 95% and preferably 98% of the running words in the text in order to gain adequate comprehension in reading and listening (Laufer, 2013; Nation, 2006; Schmitt et al., 2017; Stæhr, 2009; van Zeeland & Schmitt, 2013). From there, researchers have made a wide range of predictions based on text coverage and vocabulary knowledge. For example, corpus-driven studies have found that a lexical knowledge of the most frequent 3000

word families in Nation's (2006) British National Corpus (BNC) word lists plus proper nouns and marginal words was found to cover 95% of the running words used in daily conversation (Nation, 2006), movies (Webb & Rodgers, 2009b), TV programs (Webb & Rodgers, 2009a), and popular songs (Tegge, 2017).

Although research findings have documented a strong relationship between vocabulary knowledge and reading and listening comprehension (Schmitt et al., 2017; van Zeeland & Schmitt, 2013), most of the data reported in papers could only reflect the participants' orthographic knowledge of vocabulary (Lange & Matthews, 2020). And although research findings showed a strong link between orthographic knowledge of vocabulary and learners' performance in listening comprehension tests (Noreillie et al., 2018; Stæhr, 2008, 2009), evaluating learners' phonological knowledge of vocabulary and predicting their performance in a listening test by measuring their lexical knowledge could be unreliable to some extent (Cheng & Matthews, 2018). As Stæhr (2009), p. 583 pointed out, "Although the results from these studies emphasize that vocabulary knowledge is a determining factor for reading success, such findings simply cannot be transferred to listening; that is, it cannot be assumed that vocabulary knowledge plays an equally significant role and that identical vocabulary size or lexical coverage thresholds will apply to listening comprehension."

In response to such gap in the field, two tests of aural English vocabulary knowledge have been created to date, the AuralLex (A-Lex) (Milton & Hopkins, 2006) and Listening Vocabulary Levels Test (LVLT) (McLean et al., 2015). As a more recent test, the LVLT has been proven to outperform the A-Lex thanks to many strengths. First, each target word is accompanied by a context defining sentence that provides extra information on the word's part of speech and its contextualized meaning, which support examinees in accessing the meaning of the target word (Henning, 1991, cited in McLean et al., 2015), Second, the LVLT inherited the 4-option multiple choices format of the Vocabulary Size Test (VST) (Nation & Beglar, 2007), which allowed the test to examine a deeper depth of vocabulary knowledge compared to the Yes/No format used in the A-Lex. Third, the LVLT measured the first five levels of word frequency in Nation's (2012, cited in McLean et al., 2015) BNC/COCA word list and academic vocabulary from Coxhead's (2000) Academic Word List (McLean et al., 2015). The LVLT also showed positive correlations with parts 1 and 2 of the TOEIC listening subtest (McLean et al., 2015) and the listening component of the General English Proficiency Test (GEPT) (Li, 2019).

Besides being the answer to the dire need for a reliable test of phonological vocabulary knowledge, the LVLT also addressed another burning issue in the field of vocabulary assessment: the trend of developing and using bilingual vocabulary tests. Indeed, bilingual vocabulary tests have received increasing attention since Nguyen and Nation (2011) introduced the first bilingual version of the VST (Nation & Beglar, 2007). To date, five bilingual versions of the VST have been developed in five languages which were Vietnamese (Nguyen & Nation, 2011), Persian (Karami, 2012), Russian (Elgort, 2013), Japanese (Derrah & Rowe, 2015) and Chinese (Zhao & Ji, 2016). Most arguments against monolingual vocabulary tests were related to the interference of construct irrelevant variance such as L2 reading ability and comprehension (Karami, 2012, Karami et al., 2020, Nguyen & Nation, 2011) and such measurement errors were expected to be eliminated in a bilingual test (Karami et al., 2020).

While the development of other bilingual versions of the LVLT seems to be a tempting practice, the assumption that the validity of the revised test could be based on that of the original version and the new test does not require further validity evidence is an "uncritical view of validation" (Schmitt et al., 2020), p. 114. As Schmitt et al. (2020) wrote:

> Current validation theory would view any revised version as a new test, which needs to be validated in its own right. It is no good to assume the validity of a test with new items, and potentially different length and format/modality, based only on the behaviour of the original version. […]
>
> […] We know that speakers of various L1s can have quite different behaviour from one another (Dörnyei & Ryan, 2015), so it is unrealistic to assume that the change of language would not be connected to any other change in examinee behaviour. (p. 114)

To date, only the Japanese version of the LVLT is supported by validation evidence, and no attempt has been made to validate a Vietnamese version of the test. Therefore, a validation study of the Vietnamese version of the LVLT is not only guaranteed but also crucial and essential due to several reasons. First, "Validation is seen as an ongoing process, and so tests can never be 'validated' in a complete and final manner" (Fulcher & Davidson, 2007, cited in Schmitt et al., 2020), p. 113. Unlike the Nation and Beglar's (2007) VST, the validity of the LVLT did not receive the attention it deserves and the test has not been re-validated since its creation in 2015, which could be considered a research gap in the field. Second, vocabulary assessment is an under-researched area in Vietnam, and the lack of measuring instruments could be viewed as one of the major reasons. Considering the limited vocabulary knowledge of Vietnamese English learners, even at the tertiary level (Dang, 2020), using monolingual vocabulary tests for the measurement of vocabulary knowledge of Vietnamese learners of English in elementary, middle or high schools would be viewed as an infeasible practice.

The development and validation of the Vietnamese version of the LVLT not only provide validity evidence for the original LVLT in another context, but also can fill an important gap in vocabulary research in Vietnam. Moreover, the LVLT is arguably one of the only two vocabulary tests known in the field that assess the vocabulary knowledge of the 1000-, 2000-, 3000-, 4000-, 5000-word levels in the BNC/COCA word list plus an academic word level from the AWL, which means that the test allows scholars to capture vocabulary development from a very early stage of language learning as well as the acquisition of academic vocabulary of learners studying in academic contexts. Researchers can also use the tests for longitudinal studies that investigate vocabulary development of Vietnamese learners studying both inside and outside of Vietnam, which is also a very under-researched area.

Research questions

In their validation study, McLean et al.'s (2015) utilized the Rasch model based on four aspects of Messick's (1995) validation framework to provide validity evidence for the LVLT and found that:

1. The test items showed sufficient spread of difficulty and displayed a good fit to the Rasch model.
2. The test distinguished learners of different levels of language proficiency and performed in accordance with a hypothesized order of difficulty.
3. The LVLT correlated positively with another test of listening proficiency at .54.
4. Test items presented a high degree of unidimensionality.
5. The test items showed a strong degree of measurement invariance with different sets of items.

Following their lead, the present study also used the Rasch model to provide validation evidence for the Vietnamese LVLT based on several aspects of Messick's (1995) validation framework. Besides, additional analyses were also carried out to provide necessary validity evidence concerning Rasch items and persons reliability and separation statistics as well as local independence as suggested by Aryadoust et al. (2021).

In general, the present validation study was guided by the following research questions:

1. Do the test items show a sufficient spread of difficulty and display a good fit to the Rasch model?
2. Does the test distinguish learners of different levels of language proficiency and perform in accordance with a hypothesized order of difficulty?
3. Does the test positively correlate with another test of English listening proficiency at nearly .60?
4. Do test items display a strong degree of unidimensionality and local independence?
5. Do the test items show a strong degree of measurement invariance with different sets of items?

Methodology

Participants

The participants in this study included 311 Vietnamese EFL learners (96 males and 215 females), all of whom were second-year students of various academic majors except the English language at a highly ranked university in Vietnam. Convenience sampling was applied. The participants were the students in 8 business English classes which the researcher was the lecturer-in-charge. The participants' ages ranged from 20 to 23. All the participants were native speakers of Vietnamese, and none had lived in a country where English is the official language. In addition to having completed at least 9 years of formal English education from elementary to high school, the participants shared similar educational backgrounds. At the time of data collection, the students who took part in this study were attending the Business English Level 4 courses. As a prerequisite for attending this course level, they had already passed the 1st, 2nd, and 3rd levels of business English courses, the participants' IELTS scores suggest an average English language proficiency of A2-B1.

Instruments

The Listening Vocabulary Levels Test

The primary assessment instrument was a translated version of the *Listening Vocabulary Levels Test* (McLean et al., 2015), a 150-item multiple-choice test which was first

designed to measure Japanese learners' aural vocabulary knowledge of the first five-word frequency levels (1000, 2000, 3000, 4000, 5000) from Nation's (2012, cited in McLean et al., 2015) BNC/COCA word lists and an academic vocabulary level from the AWL (Coxhead, 2000). The 150-item test consisted of 24 items per level for the first five 1000-word frequency levels (1000–5000) and 30 items for the AWL (McLean et al., 2015).

The general format of the LVLT included two parts: the audio recording and the answer sheet. The audio portion of the test had the total running time of 28:30 min, with approximately 4:30 min for each of the five-word frequency levels and 5:51 min for the AWL; therefore, the whole 150-item test could be administered and completed within a 30-min time frame (McLean et al., 2015). It was recorded in a sound-proof music audio and was read by a male native speaker of General American English since American English has been widely taught in Japanese schools (McLean et al., 2015). The answer sheet utilized the same multiple-choice, four-option format as the Vocabulary Size Test (VST) (Nation & Beglar, 2007). The test takers were expected to listen to a single reading of the target word followed by a defining context sentence which provides extra information on the word's part of speech and associational assistance for the comprehension of the word's meaning (Henning, 1991, cited in McLean et al., 2015) and then select the target word written in their first language. The four options of each item were given in the learners' first language in order to "isolate the construct of aural vocabulary knowledge from other constructs such as L2 reading ability" (McLean et al., 2015), p. 7. There was a 5-s pause between the reading of each item so that learners could have sufficient time to process the aura input and might still maintain efficiency, a 15-s pause was given between test levels for the preparation for the next section (McLean et al., 2015). Each item was read only once. An example item of the LVLT is illustrated below (McLean et al., 2015):

1. *[time, they have a lot of time.]* (This is what the learners hear and, therefore, is invisible on the answer sheet)
   a. お 金
   b. 食べ 物
   c. 時間
   d. 友 だち

The Vietnamese version of the Listening Vocabulary Levels Test

The primary assessment instrument in this study was a Vietnamese version of the *Listening Vocabulary Levels Test*. The Japanese version of the LVLT was first translated by professional translators who were native speakers of Vietnamese, all the translators involved in this study were fluent in Japanese and had obtained N1 level, the highest level in the Japanese-Language Proficiency Test (JLPT). The translation was then carefully reviewed by the researcher himself and the translators, the translation was revised multiple times. The English version of the test provided on https://brandonkramer.net/resources/ was utilized for the comparison and revision of the target words and distractors. The Vietnamese equivalents were contextualized based on both the Japanese words in the options and the context defining sentences read in the recording. Due to

linguistic ambiguity, one English/Japanese word could have several Vietnamese meanings in the same context. For example, the word "stone" could be translated into "viên đá" (a small stone) and "t ng đá" (a big stone), while using "đá" alone could lead to even more serious ambiguity. In order to tackle this problem, the most relevant equivalents were listed with a "/" between them. An example of such an item is shown below:

2. *[stone, she sat on a stone]* (This is what the learners hear and, therefore, is invisible on the answer sheet)

a. viên đá/ t ng đá
b. cái ghế
c. t m th m
d. cành cây

The final translation was then given to two Vietnamese teachers of English for review. The teachers listened to the recording and answered the test items correctly without any misunderstanding or confusion, suggesting an appropriate translation of the LVLT.

The IELTS listening test

The present study employed the International English Language Testing System (IELTS), a standardized and globally accepted English test widely used for assessing English language proficiency of the test takers in a great variety of contexts such as education, employment, and immigration as an instrument for the measurement of participants' English listening proficiency. The IELTS was jointly developed by the British Council, The University of Cambridge Local Examination Syndicate (UCLES), and IDP Education Australia (Pearson, 2019; Quaid, 2018). There were four parts in the IELTS listening test: parts 1 and 2 included a conversation and a prompted monologue with transactional purposes and parts 3 and 4 consisted of a discussion dialogue and a monologue in academic contexts (Alavi et al., 2018; Phakiti, 2016). Cronbach's alpha reliability coefficient for the IELTS listening test was .805, which was high and strongly confirmed sound internal consistency.

Data collection

The Vietnamese version of the LVLT was administered in the first week of the course and an IELTS listening test was given to 234 out of 311 participants in the following week. All the participants were well informed of the significance and purposes of the study as well as the confidentiality, anonymity, and security of the collected data. All the students took part in the study voluntarily and were well aware that they could withdraw from the study at any time. The participants were also instructed to try their best to answer every question and to leave an item blank in case the word was completely unfamiliar to them. The tests were administered through speakers and all participants confirmed that they could hear the test items clearly. At no time did the

researcher and the students encountered any technical problems and difficulties hearing the recordings. The tests were administered in approximately 30 min and all the students were given the same amount of time.

Data analysis

Data were scored dichotomously, put into an Excel spreadsheet, and then exported to WINSTEPS 4.8.0 (Linacre, 2021) and SPSS. A Rasch analysis for dichotomous items was then carried out. The Rasch model had a great number of strengths; it facilitates the detection of measurement flaws like item misfitting, multidimensionality, and local dependence (Aryadoust et al., 2021; Müller, 2020). Wright stressed that the special feature of the Rasch model was "it allows for separating parameters of objects and agents, that is of children and test items [….] the Rasch item analysis model is the only model which retains parameter separability. From Rasch's point of view this separability is a sine qua non for objective measurement" (Lord & Wright, 2010), p. 1289. In addition, Pearson product-moment correlations, a Z-test, and several sets of one-way ANOVA, Dunnett's T3, and Tukey's post hoc tests were also conducted for data analysis.

Results

This section reports and discusses the validity of the Vietnamese version of the LVLT from the five aspects of construct validity described by Messick (1995): content, substantive, structural, generalizability, and external.

Content aspect of construct validity

The content aspect of construct validity determines "the boundaries of the construct domain to be assessed" (Messick, 1995), p. 745. This facet consists of three components: content relevance, representativeness, and technical quality. First, the content relevance addresses "the relationship between the test items and the construct being measured (receptive knowledge of the form-meaning relationships of words)" (Webb et al., 2017), which has already been discussed at length in McLean et al. (2015). The test was carefully designed to measure vocabulary knowledge of English words from the first five-word frequency levels and the AWL "through a retrofit and redesign of previous VST items" (McLean et al., 2015). The test items were divided into sections in accordance with the frequency of occurrence on the BNC/COCA word lists. These principles suggest that the LVLT could be representative of the construct domain.

Representativeness

The first method for evaluating representativeness is examining strata (H) and separation (G) statistics, both indices refer to the number of statistically different levels of item difficulty and person ability in the data (Linacre, 2021). G and H can be derived using the formulas:

G = True standard deviation/Average measurement error

$H = (4 \times G + 1)/3$

Concerning the relationship between G and H, Wright and Masters (2002) wrote:

> G itself is a more conservative "Separation Index" than H. For instance, suppose that the "true" standard deviation of a sample is the same as the average measurement error. Then G=1, and the test reliability is 0.5, warning us that we don't know whether observed differences within the sample are real differences or merely measurement error. H is (4+ 1)/3, i.e., roughly 2. This indicates that the opposite ends of the "true" distribution are measurably different, implying that, if the observed measures are sufficiently far apart, they probably reflect real differences. (p. 888)

Item strata and separation statistics should be greater than 2 for a healthy test (Linacre, 2021). Low strata and separation values (< 2) may mean that the test fails to differentiate 2 levels of item difficulty. Table 1 gives information on the item and person separation and reliability. The Vietnamese version of LVLT showed separation statistics of 4.61 and 7.01 for person and item respectively. In other words, the test was able to differentiate 7 levels of item difficulty, and more than 4 levels of person ability were differentiated by measurement among the test takers. The Vietnamese LVLT also showed an item strata statistic of 9.68, confirming that the test has more than two statistically distinct difficulty levels. Reliability statistics, which indicate the reproducibility of the item measures if the items were given to another group from the same population, or the reproducibility of person measures if they were tested again (Bond & Fox, 2015), were also high. The Vietnamese version of LVLT had 96% and 98% of confidence about the measure of persons and items correspondingly. All of these could be taken as supportive evidence for the test's representativeness.

Another way for examining representativeness is to check whether (1) the test consists of a sufficient number of items, (2) the empirical item hierarchy shows sufficient spread, and (3) whether there are gaps in the item difficulty hierarchy. All of these aspects were clarified in Fig. 1, which illustrates the linear relationship between 311 examinees and 150 test items. Each "#" and "." indicates 3 and 1–2 test takers, respectively. More able persons were toward the top of the figure and less able persons were toward the bottom of the Wright map, the same went for more difficult items and easier items, in the order given.

Test items were labeled according to their frequency level and the item number on the test form. For example, item 4000-89 belonged to the fourth 1000-word frequency level and was the test item number 89. Items from the Academic Word List were labeled AWL. Figure 1 shows that there were items represented throughout the difficulty

Table 1 Separation and reliability statistics

	Total	Count	Measure	Realse	IMNSQ	ZSTD	OMNSQ	ZSTD
Person	311 Input		311 Measured		Infit		Outfit	
Mean	93.1	150.0	.88	.22	1.00	– .1	1.02	.0
P. SD	23.0	.1	1.07	.03	.18	1.7	.55	1.5
Real RMSE	.23	True SD	1.04	Separation	4.61	Person reliability		.96
Item	150 Input		150 Measured		Infit		Outfit	
Mean	192.9	311.0	– .04	.19	1.00	.0	1.02	.0
P. SD	77.7	.2	1.80	.17	.11	1.8	.31	1.8
Real RMSE	.25	True SD	1.78	Separation	7.01	Item reliability		.98

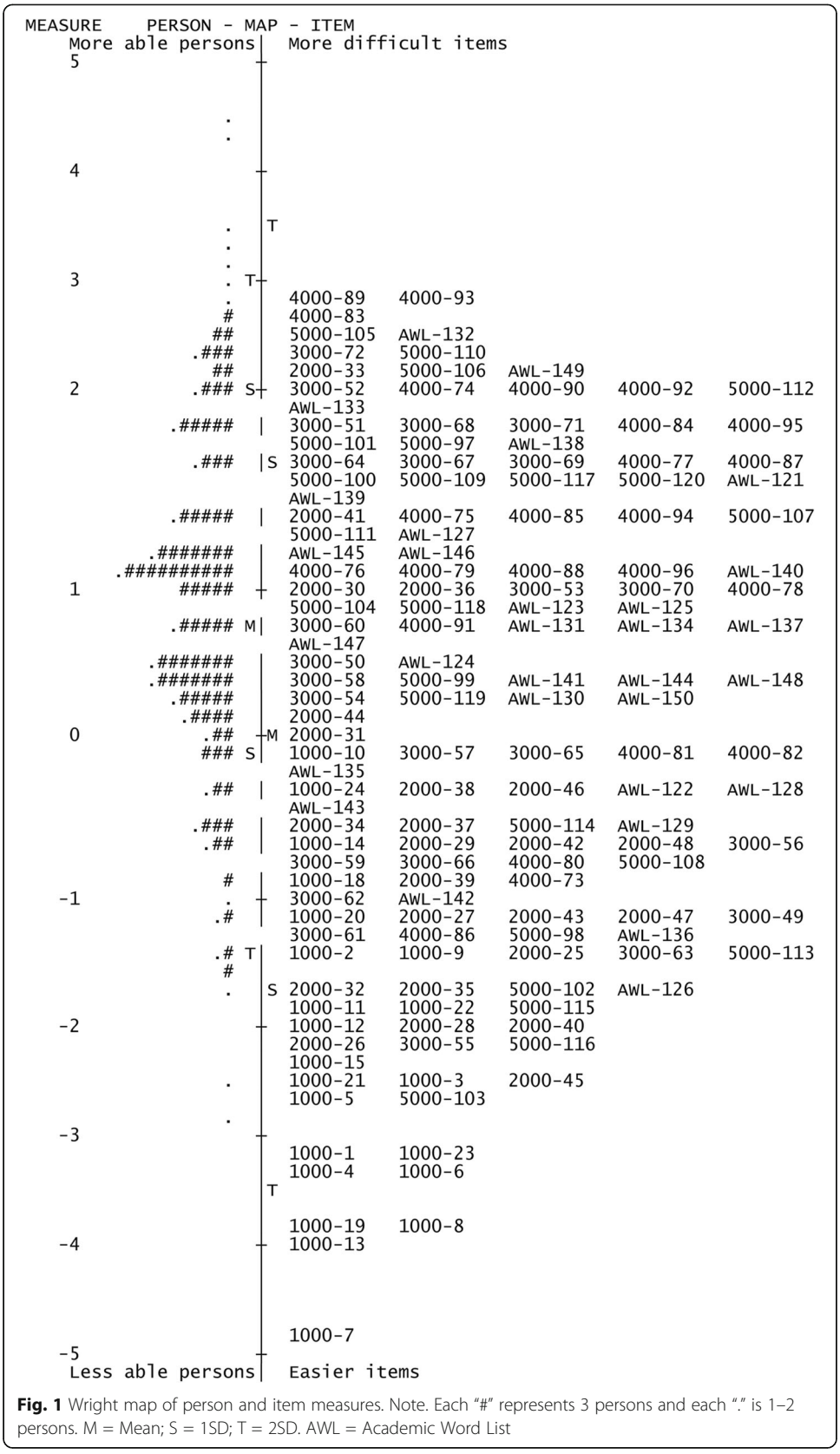

Fig. 1 Wright map of person and item measures. Note. Each "#" represents 3 persons and each "." is 1–2 persons. M = Mean; S = 1SD; T = 2SD. AWL = Academic Word List

hierarchy and that no significant gaps were present in the item difficulty hierarchy, indicating a strong degree of representativeness (RQ1).

Technical quality

Technical quality could be evaluated by inspecting how well the empirical data fit the Rasch model (Smith Jr., 2004), using the Rasch Infit and Outfit mean-square (MNSQ) statistic. A cutoff point for determining item fit must be decided first, and each researcher prefers a different threshold for infit and outfit statistics, as Aryadoust et al. (2021) commented, "There is no universal agreement on fit statistics in Rasch measurement" (p. 6). Still, a rule of thumb was given for the present study based on the suggestions made by Wright and Linacre (1994), Smith et al. (1998), Linacre (2003, 2010, 2017), Smith (2005), Wilson (2005), and Bond and Fox (2015). It has been generally agreed that Mnsq metrics of 05–1.5 indicated a good fit to the Rasch model and could be considered productive for measurement. Researchers have also suggested that while Mnsq indices of 1.5–2 could be considered unproductive to the test, those values might not necessarily degrade the test's results. Mnsq values of greater than 2, however, were perceived as a signal of unexpected observations that might present severe underfit to the Rasch model and could distort or degrade the test's result (Linacre, 2017). However, not every Mnsq index of higher than 2 should be deemed significantly underfitting, the significance of underfit must be confirmed by the standardized z score (ZSTD). Only items with both Mnsq and Zstd values greater than 2 could be considered significantly underfitting (Aviad-Levitzky et al., 2019). Items with Mnsq statistics lower than 0.5 were perceived as too predictable and thus might overfit the Rasch model. An inspection of item fit statistics spotted no overfit.

Table 2 presents a list of test items with Mnsq values over 1.5. Out of the ten items with Mnsq metrics over 1.5, only two items had Mnsq indices greater than 2, and only one of them had the Zstd values of slightly over 2. However, ZSTD indices were believed to be "most useful when datasets consist of < 250, beyond which they can become inflated" (Aryadoust et al., 2021, p. 27). The fact that the present study collected data from 311 students might be considered the factor contributing to the inflated Zstd values. A qualitative inspection of the most misfitting response strings pointed out that

Table 2 Summary of possible misfitting items

Item	Infit		Outfit	
	MNSQ	ZSTD	MNSQ	ZSTD
1000-1	1.13	.50	2.70	2.50
1000-16	1.02	.35	2.58	1.86
5000-117	1.49	7.79	1.93	7.50
AWL-142	.96	− .36	1.71	3.40
4000-89	1.11	1.11	1.60	2.69
1000-19	1.06	.27	1.55	.97
AWL-138	1.23	3.67	1.55	4.39
3000-69	1.32	5.41	1.52	4.74
1000-5	1.01	.12	1.51	1.21
1000-7	1.04	.28	1.50	.84

the underfit was caused by only four persons (approx. 1.28%), suggesting no major flaws in these items. More importantly, the two out of 150 items mentioned only represented a small proportion of 1.33% misfit rate, indicating a very good fit to the Rasch model (RQ1).

Another method of inspecting technical quality was examining local independence. One indication of the possible violation of local independence is overfitted, which was not spotted in the analysis of fit statistics. Another way of investigating was analyzing the standardized residual correlations. The Rasch model required that dependence should not exist between test items (Bond & Fox, 2015). Wendy Yen (1984, 1993) suggested a Q3 statistic (also known as Q3 coefficient) which was used to detect dependency between pairs of items and persons. Some researchers believed that a Q3 efficient exceeding 0.30 could be a sign of a violation of local independence (Chen & Thissen, 1997; Christensen et al., 2017; Liu & Maydeu-Olivares, 2013). However, Dr. John Michael Linacre argued, "local dependence would be a large positive correlation. Highly locally dependent items (Corr. > +.7) [….] share more than half their "random" variance, suggesting that only one of the two items is needed for measurement" (Linacre, 2021), p. 426. Hence, "Correlations need to be around 0.7 before we are really concerned about dependency" (Linacre, 2021), p. 427. In other words, a correlation of 0.7 between two variables indicates a shared variance of 0.7 × 0.7 = 0.49 = ~ 0.5 of each item's variance. Therefore, the correlation of 0.7 should be taken as the threshold value between two variables measuring effectively the same thing (Linacre, 2021). The results of an analysis of the standardized residual correlations showed that two item pairs had the residual correlations of larger than 0.4, which were items 1000-22 and 2000-40 (correlated at .46) and items 1000-3 and 1000-4 (correlated at .53). Even for the greatest correlation of 0.53, the two items only shared 0.53 × 0.53 = 28% of the variance in their residuals in common, which means that 72% of their residual variances differed. This could be taken as supportive evidence that the Vietnamese version of LVLT is acceptable in terms of local independence (RQ4).

Substantive aspect of construct validity

The substantive aspect of construct validity could be evaluated by examining whether the empirical item hierarchy was presented as expected by theoretical hypothesis and whether the pattern of responses was consistent with that item hierarchy (Smith Jr., 2004). The hypothesis for item hierarchy was that words at higher levels of frequency would be easier than those at lower frequency levels (Beglar, 2010). Therefore, the hypothesized order of item difficulty was 5000 > 4000 > 3000 > 2000 > 1000. The words in the AWL was not given a hypothesized priority due to the fact that they come from different frequency levels. A one-way ANOVA was conducted to investigate whether the mean score statistically dropped from one frequency level to the next. Both Welch and Brown-Forsythe statistics were significant (p = 0.000). The ANOVA was significant, F (4,155) = 386.610, p = .000. Tukey's and Dunnett's T3 post hoc tests indicated that all comparisons were significant except between the 3000 and 5000 levels. Figure 2 displays the mean item difficulties and their 95% confidence intervals for the five frequency levels. The figure generally supported the given hypothesis regarding item difficulty.

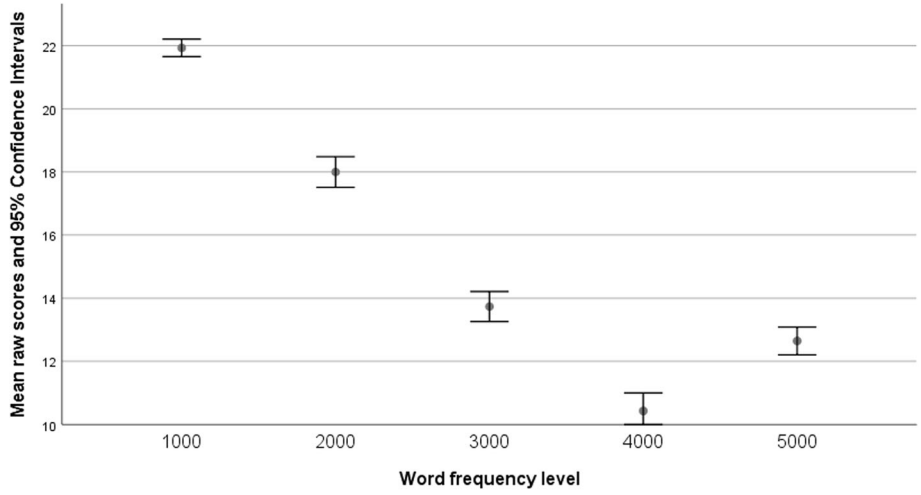

Fig. 2 Mean item difficulties and 95% confidence intervals

Data concerning the 4000 and 5000 frequency levels, however, did not conform to the proposed hypothesis, which could be explained in certain ways. First, this study was conducted in an English as a foreign language (EFL) context (Vietnam), where learners' exposure to English input was limited. Second, the fourth and fifth levels of word frequency are mid-frequency levels and the lack of L2 input in the EFL context "may reduce the effects of lexical frequency for less frequent words. For example, there may be sufficient lexical input within the classroom and course books to differentiate knowledge of the highest frequency words [….]. However, the same may not always hold true of slightly less frequent words [….], because words at the 4000 level may not always be encountered much more often than those at the 5000 word level in the EFL context" (Webb et al., 2017), pp. 47–48.

Since vocabulary knowledge is a strong predictor of language proficiency, scores on the LVLT were hypothesized to reflect learners' English listening proficiency. To warrant this claim, the IELTS listening test scores of 234 students were examined. It was also hypothesized that IELTS band scores greater than 6.0, which were indicated by answering correctly more than 23 out of 40 items in the IELTS listening test, would suggest high language proficiency. IELTS band scores of 4.5, 5.0, and 5.5, which were indicated by scores from 13 to 22, were supposed to suggest intermediate proficiency. Scores from 12/40 and below, which reflected IELTS band scores of 4.0 and lower, were assumed to be an indication of low proficiency.

The participants were then divided into high proficiency (n = 40), intermediate proficiency (n = 116), and low proficiency (n = 78) groups. First, a one-way ANOVA, a Dunnett's T3 and a Tukey post hoc tests were run to see if there were significant differences between the three groups' listening proficiency. Both Welch and Brown-Forsythe statistics were significant (p = 0.000). The ANOVA was significant, F (2,231) = 530.249, p = .000. Tukey's and Dunnett's T3 post hoc tests showed that the students' performance between all groups differed significantly. After the significant difference between the three groups' listening proficiency was confirmed, another set of one-way ANOVA, Dunnett's T3, and Tukey's post hoc

tests were conducted to determine if the phonological vocabulary knowledge of the three groups differed significantly. The hypothesis was that greater aural knowledge of vocabulary would result in greater listening proficiency. All the necessary assumptions were checked and met. The ANOVA was significant, F (2,231) = 64.719, p = .000. Tukey's and Dunnett's T3 post hoc tests indicated that all pair-wise comparisons were statistically significant. Results from the analyses confirmed that higher aural vocabulary knowledge would lead to higher listening proficiency. These may be taken as supportive evidence for the substantive aspect of the test's construct validity (RQ2).

Structural aspect of construct validity

The structural aspect of construct validity could be evaluated by examining the unidimensionality (the degree to which the test measures only one underlying latent trait). The most commonly used method in language assessment to investigate unidimensionality was principal component analysis of residuals (PCAR). The principal component analysis (PCA) of standardized residuals was carried out to test whether the Vietnamese version of the LVLT measured a single construct, given that both the analyses of the VST (Beglar, 2010) and the LVLT (McLean et al., 2015) resulted in very strong unidimensionality. Table 3 shows the standardized residual variance of the test, measured in eigenvalue units. The total amount of raw variance explained by Rasch measurement was 38.3% of the variance in the residuals (eigenvalue = 92.3), which was well consistent with the data reported in McLean et al. (2015). The observed variance explained by the measure was identical to the expected variance in the model and the unexplained variance in the first contrast was only 4.97, accounting for 2.1% of the variance, much smaller than the variance explained by the items, which all together suggested a perfect fit to the Rasch model.

However, the eigenvalue was larger than 2.0, and therefore, further investigation was demanded. Table 4 gives data about the correlation of the item clusters. It is clear that the lowest disattenuated Pearson correlations of the item clusters in PCA contrasts were about 0.75. This means that the items in those clusters shared 0.75 × 0.75 = more than 56% of the variance in their residuals in common, indicating that they measured the same thing and that the clusters represented strands rather than dimensions

Table 3 Standardized residual variance in eigenvalue units

	Eigenvalue	Observed	Expected
Total raw variance in observations	241.3286	100.0%	100.0%
Raw variance explained by measures	92.3286	38.3%	38.3%
Raw variance explained by persons	36.2003	15.0%	15.0%
Raw variance explained by items	56.1283	23.3%	23.3%
Raw unexplained variance (total)	149.0000	61.7%	61.7%
Unexplained variance in 1st contrast	4.9741	2.1%	3.3%
Unexplained variance in 2nd contrast	3.3754	1.4%	2.3%
Unexplained variance in 3rd contrast	3.1087	1.3%	2.1%
Unexplained variance in 4th contrast	2.9187	1.2%	2.0%
Unexplained variance in 5th contrast	2.6583	1.1%	1.8%

Table 4 Approximate relationships between PERSON measures

PCA contrast	ITEM cluster	Pearson correlation	Disattenuated Pearson correlation
1	1–3	0.6105	0.7629
1	1–2	0.7404	0.8557
1	2–3	0.8679	1.0000
2	1–3	0.6901	0.7868
2	1–2	0.8607	0.9587
2	2–3	0.7915	0.9205
3	1–3	0.6241	0.7520
3	1–2	0.8252	0.9483
3	2–3	0.8239	0.9281
4	1–3	0.6883	0.8103
4	1–2	0.8411	0.9461
4	2–3	0.8450	0.9561
5	1–3	0.6924	0.8214
5	1–2	0.8127	0.9415
5	2–3	0.8754	0.9743

(Linacre, 2021). Taken together, the Vietnamese version of the LVLT was most likely to measure the unidimensional construct, that was, aural vocabulary knowledge (RQ4).

Generalizability aspect of construct validity

The generalizability aspect of construct validity addresses "the extent to which score properties and interpretations generalize to and across population groups, settings, and tasks" (Messick, 1995, p. 745). This aspect of construct validity can be investigated by examining the degree to which item difficulty and person ability statistics are consistent across measurement contexts without measurement error (Smith Jr., 2004; Wolfe & Smith Jr., 2007). The test items at each frequency level including the AWL were randomly divided to create two 75-item versions of the test.

Rasch item reliability, separation, and strata statistics for the 150-item version, the first and the second 75-item versions were 98% (separation = 7.01, strata = 9.68), 99% (separation = 8.59, strata = 11.78) and 97% (separation = 6.23, strata = 8.64) respectively. Rasch person reliability and separation statistics of the 150-item, the first and the second 75-item test forms were .96 (4.61), .92 (3.37) and .91 (3.15), correspondingly. These together indicated that the three versions of the Vietnamese LVLT produced similar person ability estimates and were free of measurement errors.

Pearson product-moment correlations were computed between the scores of 150-item test form and two 75-item versions of the test to determine the relationship between the three sets of test items. Table 5 displays the results of this analysis. It can be

Table 5 Pearson correlation among the 150-item test form and two 75-item test forms

Test	1st 75-item test	2nd 75-item test
2nd 75-item test	.911**	–
150-item test	.979**	.976**

Note. "**" indicates that correlation is significant at 0.01 level (2-tailed) (N= 311)

observed that the Pearson correlation coefficients of the three sets were all above .90, the level at which multicollinearity occurs. The high correlations between the two randomly selected sets of items and the original test strongly confirmed item invariance. These could be considered to be positive evidence for the test's generalizability (RQ5).

External aspect of construct validity

The external aspect of construct validity refers to "the extent to which the test's relationships with other tests and nontest behaviors reflect the expected high, low, and interactive relations implied in the theory of the construct being assessed" (Messick, 1989), p. 45. In order to examine the relationship between the LVLT and other tests measuring the related construct, an IELTS listening test was given to 234 out of 311 participants. It was hypothesized that the LVLT and the IELTS listening test scores would be positively correlated as the IELTS listening test assesses a wide variety of aural language skills and abilities, including phonological knowledge of vocabulary. It was also hypothesized that the correlations between the IELTS-LVLT would be lower than the within-LVLT correlations (the correlations between scores from different test items of the LVLT), because all the test items in the LVLT was created to measure only one construct, aural vocabulary knowledge. In order to measure within-LVLT correlations, the correlations between students' scores on the LVLT and on each vocabulary level were examined. The correlations between participants' scores on the IELTS listening test and each word level in the LVLT were also measured. Then, a Z-test was performed based on Meng et al.'s (1992) method to test if there were statistically significant differences between two groups of correlation coefficients (within-LVLT and IELTS-LVLT). The results are presented in Table 6.

A positive, strong correlation of .652 was found between the LVLT and the IELTS listening test scores. Moreover, it was also found that the IELTS listening test scores strongly correlated with the scores on each level of the LVLT (r = .455, .593, .571, .582, .472, .648). Additionally, the Z-test showed that the within-LVLT correlations were significantly higher than the IELTS-LVLT correlations. All of these generally confirmed the proposed hypotheses and could be taken as supportive evidence for the external aspect of the Vietnamese version of the LVLT (RQ3).

Discussion

Adopting the Rasch's (1960) dichotomous model based on Messick's (1989, 1995) framework of validation, the present study aimed at providing validity evidence for both

Table 6 Difference between within-LVLT and IELTS-LVLT correlations

	LVLT	IELTS Listening Test	Z	*p*
1000 level	.696**	.455**	5.70	.000
2000 level	.894**	.593**	10.31	.000
3000 level	.885**	.571**	10.30	.000
4000 level	.903**	.582**	11.16	.000
5000 level	.829**	.472**	9.78	.000
AWL level	.929**	.648**	11.55	.000

Note."**" indicates that correlation is significant at 0.01 level (2-tailed) (N= 234)
AWL Academic Word List, *LVLT* Listening Vocabulary Levels Test

the Vietnamese LVLT and its original version. As suggested in Wright and Stone's (1999, cited in Aryadoust et al., 2021) comprehensive framework, validity evidence of a test should be reflected in (1) metrics of psychometric validity which include unidimensionality, local independence, and fit statistics, and (2) metrics of reliability consisting of reliability and separation values for items and persons.

In general, the test displayed strong values of person and item reliability (Table 1), which is an indication of the stability of the scoring system. Separation and strata statistics were also higher than 2 for persons and items. This, together with the Wright map of persons and items measures (Fig. 1), strongly suggests that the test presented a sufficient spread of difficulty and were sensitive enough to distinguish test takers of different levels (Linacre, 2021).

The test items' fit values were examined using more lenient criteria than those applied in McLean et al. (2015). However, this does not mean that the test items in the Vietnamese LVLT were intentionally given a free pass. In fact, McLean et al. (2015) utilized McNamara's (1996, cited in McLean et al., 2015) criterion for determining only the items' infit Mnsq, and they did not report or provide arguments for the outfit Mnsq and the Zstd values of the test items. Therefore, it could be said that the present study provided a broader view regarding the items' fit statistics. Although some items were indeed noisy, especially item 1000-1, in general, the test items presented very good fit to the Rasch model with less than 2% of misfit rate.

The test items' unidimensionality and local independence were also carefully investigated by the analysis of standardized residual correlations and principal component analysis of residuals. Principal component analysis and standardized residual correlations analysis are the most suitable methods for examining unidimensionality and local independence compared to other methods that use fit metrics and reliability coefficients (Aryadoust et al., 2021; Linacre, 2021). The items in the Vietnamese LVLT were proven to have really strong unidimensionality and were free of local dependence.

The generalizability and external aspects of the test were also carefully examined. The test items presented a very strong degree of measurement invariance with Pearson correlations of greater than .90 between randomly divided sets of items and very high item and person reliability statistics (>.90) for all sets of items. The Vietnamese LVLT and the IELTS listening test were strongly correlated at .652. Different vocabulary levels of the test were also found to positively correlate with the IELTS listening test at 0.455–0.593. The correlation was especially high between the academic word level and the IELTS listening test (.648), signaling a strong relationship between academic vocabulary knowledge and academic listening proficiency.

The Vietnamese LVLT also shows a really good degree of practicality in terms of administration, scoring, and score interpretation. The test can be easily administered in a standard, quiet classroom with pens or pencils, papers, a basic computer, and good speakers. Little or zero training is required for the administration of the test and neither is it needed for grading. The test could be reliably completed in approximately 35–40 min including instructions and other administrative tasks. Tests scores could be interpreted by using a stringent cut-off point for vocabulary level mastery suggested by McLean and Kramer (2015) and McLean et al. (2015) or by using vocabulary scores as instructed by Ha (Ha, H. T.: *Exploring the relationships between various dimensions of receptive vocabulary knowledge and L2*

listening and reading comprehension, n.d.). The test has good potential to be delivered in both paper- and computer-based, online formats. Scores on the LVLT were proven to have strong correlations with tests of English listening proficiency such as the TOIEC listening test (McLean et al., 2015), GEPT listening subtest (Li, 2019). Moreover, Ha's (Ha, H. T.: *Exploring the relationships between various dimensions of receptive vocabulary knowledge and L2 listening and reading comprehension*, n.d.) comprehensive study on the relationship between receptive vocabulary knowledge and receptive language skills did illustrate a linear, strong relationship between students' scores on the Vietnamese LVLT and the IELTS listening and academic reading tests. The study suggested that the LVLT could be used either in combination with other tests of English proficiency or in isolation and can still be a very powerful predictor of learners' success in academic listening and reading comprehension (Ha, H. T.: *Exploring the relationships between various dimensions of receptive vocabulary knowledge and L2 listening and reading comprehension*, n.d.).

Conclusion

This study provides evidence supporting the validity of the Vietnamese version of the LVLT, which can be taken as validity evidence for the LVLT, an aural vocabulary test that measures knowledge of English words from the first five-word frequency levels from Nation's (2017) BNC/COCA word lists and the Academic Word List (Coxhead, 2000). I believe that the Vietnamese LVLT could be of great value and help to Vietnamese teachers and researchers as it offers an instrument for the measurement of learners' phonological knowledge of vocabulary which can serve as a part of a needs analysis to inform the predictions and decisions teaching, testing, and designing language courses and programs.

The LVLT inherits the 4-option multiple-choice format of the VST, which has been warned to potentially foster the strategic examinee guessing effect, which could result in overestimation of vocabulary size as much as 26% (Gyllstad et al., 2015; Schmitt et al., 2020). McLean et al. (2015) had to carry out in-depth qualitative investigations into the effect to make sure that it did not have overwhelming influences on test scores. However, due to certain reasons, such investigations were not conducted in the present study, which should be considered to be a major limitation.

As McLean et al. (2015) suggested, future research should aim to create different versions of the LVLT in other languages and the tests' functioning requires further quantitative and qualitative investigation. Vietnamese researchers are urged to provide further validity evidence for the test and to use the Vietnamese LVLT in combination with its written form to examine the relationship between phonological and orthographic knowledge of vocabulary. Future research on the Vietnamese LVLT should also pay special attention to the mentioned strategic guessing effect.

Abbreviations

BNC: British National Corpus; COCA: Corpus of Contemporary American English; LVLT: Listening Vocabulary Levels Test; NVLT: New Vocabulary Levels Test; VST: Vocabulary Size Test; IELTS: International English Language Testing System; AWL: Academic Word List; ANOVA: Analysis of variance; PCAR: Principal component analysis of residuals; PCA: Principal component analysis; ESL: English as a Second Language; EFL: English as a Foreign Language; MNSQ: Mean square; ZSTD: Z standard; TOEIC: Test of English for International Communication; GEPT: General English Proficiency Test; UCLES: University of Cambridge Local Examination Syndicate; JLPT: Japanese-Language Proficiency Test

Acknowledgements

Not applicable.

Author's contributions
The author confirms sole responsibility for the following: study conception and design, data collection, analysis and interpretation of results, and manuscript preparation. The author read and approved the final manuscript.

Funding
The author received no financial support for the research, authorship, and/or publication of this article.

Availability of data and materials
The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Competing interests
The author declares that he has no competing interests.

Received: 14 April 2021 Accepted: 25 June 2021
Published online: 09 August 2021

References

Alavi, S. M., Kaivanpanah, S., & Masjedlou, A. P. (2018). Validity of the listening module of international English language testing system: multiple sources of evidence. *Language Testing in Asia*, *8*(8). https://doi.org/10.1186/s40468-018-0057-4.

Aryadoust, V., Ng, L. Y., & Sayama, H. (2021). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing*, *38*(1), 6–40. https://doi.org/10.1177/0265532220927487.

Aviad-Levitzky, T., Laufer, B., & Goldstein, Z. (2019). The new Computer Adaptive Test of Size and Strength (CATSS): Development and validation. *Language Assessment Quarterly*, *16*(3), 345–368. https://doi.org/10.1080/15434303.2019.1649409.

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing*, *27*(1), 101–118. https://doi.org/10.1177/0265532209340194.

Bond, T., & Fox, C. (2015). *Applying the Rasch model*, (3rd ed.,). Lawrence Erlbaum Associates. https://doi.org/10.4324/9781315814698.

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using Item Response Theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265–289. https://doi.org/10.3102/10769986022003265.

Cheng, J., & Matthews, J. (2018). The relationship between three measures of L2 vocabulary knowledge and L2 listening and reading. *Language Testing*, *35*(1), 3–25. https://doi.org/10.1177/0265532216676851.

Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical values for Yen's Q3: Identification of local dependence in the Rasch model using residual correlations. *Applied Psychological Measurement*, *41*(3), 178–194. https://doi.org/10.1177/0146621616677520.

Coxhead, A. (2000). A new Academic Word List. *TESOL Quarterly*, *34*(2), 213–238. https://doi.org/10.2307/3587951.

Dang, T. N. Y. (2020). Vietnamese non-English majored EFL university students' receptive knowledge of the most frequent English words. *VNU Journal of Foreign Studies*, *36*(3), 1–11. https://doi.org/10.25073/2525-2445/vnufs.4553.

Derrah, R., & Rowe, D. E. (2015). Validating the Japanese bilingual version of the Vocabulary Size Test. *International Journal of Languages, Literature and Linguistics*, *1*(2), 131–135.

Dörnyei, Z., & Ryan, S. (2015). *The psychology of the language learner revisited*. New York: Routledge.

Elgort, I. (2013). Effects of L1 definitions and cognate status of test items on the vocabulary size test. *Language Testing*, *30*(2), 253–272.

Gyllstad, H., Vilkaitė, L., & Schmitt, N. (2015). Assessing vocabulary size through multiple-choice formats: issues with guessing and sampling rates. *ITL International Journal of Applied Linguistics*, *166*, 276–303.

Karami, H. (2012). The development and validation of a bilingual version of the Vocabulary Size Test. *RELC Journal*, *43*(1), 53–67. https://doi.org/10.1177/0033688212439359.

Karami, H., Nejad, M. K., Nourzadeh, S., & Shirazi, M. A. (2020). Validation of a bilingual version of the vocabulary size test: comparison with the monolingual version. *International Journal of Bilingual Education and Bilingualism*, 23(4), 368–380. https://doi.org/10.1080/13670050.2017.1391744.

Karami, H., Nejad, M. K., Nourzadeh, S., & Shirazi, M. A. (2020). Validation of a bilingual version of the vocabulary size test: comparison with the monolingual version.

Laufer, B. (2013). Lexical thresholds for reading comprehension: What they are and how they can be used for teaching purposes. *TESOL Quarterly*, *47*(4), 867–872. https://doi.org/10.1002/tesq.140.

Li, C. H. (2019). Using a Listening Vocabulary Levels Test to explore the effect of vocabulary knowledge on GEPT listening comprehension performance. *Language Assessment Quarterly*, *16*(3), 328–344. https://doi.org/10.1080/15434303.2019.1648474.

Linacre, J. M. (2003). Rasch power analysis: size vs. significance: standarized chi-square fit statistic. *Rasch Measurement Transactions*, *17*(1), 918 Retrieved from https://www.rasch.org/rmt/rmt171n.htm.

Linacre, J. M. (2010). When to stop removing items and persons in Rasch misfit analysis? *Rasch Measurement Transactions*, *23*(4), 1241 Retrieved from http://www.rasch.org/rmt/rmt234g.htm.

Linacre, J. M. (2017). Teaching Rasch measurement. *Rasch Measurement Transactions*, *31*(2), 1630–1631 Retrieved from https://www.rasch.org/rmt/rmt312.pdf.

Linacre, J. M. (2021). A User's Guide to WINSTEPS® MINISTEP Rash-model computer programs. Program Manual 4.8.0. Available at: https://www.winsteps.com/winman/copyright.htm (accessed 22 February 2021).

Liu, Y., & Maydeu-Olivares, A. (2013). Local dependence diagnostics in IRT modeling of binary data. *Educational and Psychological Measurement*, *73*(2), 254–274. https://doi.org/10.1177/0013164412453841.

Lord, F., & Wright, B. D. (2010). Fred Lord and Ben Wright discuss Rasch and IRT Models. *Rasch Measurement Transactions*, *24*(3), 1289–1290.

Lange, K., & Matthews, J. (2020). Exploring the relationships between L2 vocabulary knowledge, lexical segmentation, and L2 listening comprehension. *Studies in Second Language Learning and Teaching.*, *10*(4), 723–749. https://doi.org/10.14746/ssllt.2020.10.4.4.

McLean, S., & Kramer, B. (2015). The creation of a New Vocabulary Levels Test. *Shiken*, *19*(1), 1–11.
McLean, S., Kramer, B., & Beglar, D. (2015). The creation and validation of a listening vocabulary levels test. *Language Teaching Research*, *19*(6), 741–760. https://doi.org/10.1177/1362168814567889.
Meng, X. L., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin*, *111*(1), 172–175.
Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement*, (3rd ed., pp. 13–103). Macmillan.
Messick, S. (1995). Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741.
Milton, J., & Hopkins, N. (2006). Comparing phonological and orthographic vocabulary size: do vocabulary tests underestimate the knowledge of some learners? *The Canadian Modern Language Review*, *63*(1), 127–147. https://doi.org/10.3138/cmlr.63.1.127.
Müller, M. (2020). Item fit statistics for Rasch analysis: can we trust them? *Journal of Statistical Distributions and Applications*, *7*(1), 1–12. https://doi.org/10.1186/s40488-020-00108-7.
Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review*, *63*(1), 59–82. https://doi.org/10.1353/cml.2006.0049.
Nation, I. S. P. (2013). *Learning vocabulary in another language*, (Second ed.,). Cambridge: Cambridge University Press.
Nation, I. S. P. (2017). The BNC/COCA Level 6 word family lists (Version 1.0.0) [Data file]. Available at: https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-lists (accessed 22 February 2021).
Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, *31*, 9–13.
Nguyen, L. T. C., & Nation, I. S. P. (2011). A bilingual vocabulary size test of English for Vietnamese learners. *RELC Journal*, *42*(1), 86–99. https://doi.org/10.1177/0033688210390264.
Noreillie, A. S., Kestemont, B., Heylen, K., Desmet, P., & Peters, E. (2018). Vocabulary knowledge and listening comprehension at an intermediate level in English and French as foreign languages: an approximate replication study of Stæhr (2009). *ITL - International Journal of Applied Linguistics*, *169*(1), 212–231. https://doi.org/10.1075/itl.00013.nor.
Pearson, W. S. (2019). 'Remark or retake'? A study of candidate performance in IELTS and perceptions towards test failure. *Language Testing in Asia*, *9*(17). https://doi.org/10.1186/s40468-019-0093-8.
Phakiti, A. (2016). Test-takers' performance appraisals, appraisal calibration, state-trait strategy use, and state-trait IELTS listening difficulty in a simulated IELTS Listening test. *IELTS Research Reports Series*, *6*, 1–140.
Quaid, E. D. (2018). Reviewing the IELTS speaking test in East Asia: theoretical and practice-based insights. *Language Testing in Asia*, *8*(2). https://doi.org/10.1186/s40468-018-0056-5.
Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danmarks Paedogogiske Institute.
Schmitt, N., Cobb, T., Horst, M., & Schmitt, D. (2017). How much vocabulary is needed to use English? Replication of van Zeeland & Schmitt (2012), Nation (2006) and Cobb (2007). *Language Teaching*, *50*(2), 212–226.
Schmitt, N., Nation, P., & Kremmel, B. (2020). Moving the field of vocabulary assessment forward: the need for more rigorous test development and validation. *Language Teaching*, *53*(1), 109–120. https://doi.org/10.1017/S0261444819000326.
Smith, E. (2005). Effect of item redundancy on Rasch item and person estimates. *Journal of Applied Measurement*, *6*(2), 147–163.
Smith Jr., E. V. (2004). Evidence for the reliability of measures and validity of measure interpretation: a Rasch measurement perspective. In E. V. Smith Jr., & R. M. Smith (Eds.), *Introduction to Rasch measurement: Theory, models and applications*, (pp. 93–122). JAM Press.
Smith, R. M., Schumacker, R. E., & Bush, M. J. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Mearsurement*, *2*, 66–78.
Stæhr, L. (2008). Vocabulary size and the skills of listening, reading and writing. *The Language Learning Journal*, *36*(2), 139–152.
Stæhr, L. (2009). Vocabulary size and the skills of listening, reading and writing. *The Language Learning Journal*, *36*(2), 139–152. https://doi.org/10.1080/09571730802389975.
Tegge, F. (2017). The lexical coverage of popular songs in English language teaching. *System*, *67*, 87–98.
van Zeeland, H., & Schmitt, N. (2013). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics*, *34*(4), 457e479. https://doi.org/10.1093/applin/ams074.
Webb, S., & Rodgers, M. P. H. (2009a). The vocabulary demands of television programs. *Language Learning*, *59*(2), 335–366. https://doi.org/10.1111/j.1467-9922.2009.00509.x.
Webb, S., & Rodgers, M. P. H. (2009b). The lexical coverage of movies. *Applied Linguistics*, *30*(3), 407–427. https://doi.org/10.1093/applin/amp010.
Webb, S., Sasao, Y., & Balance, O. (2017). The updated Vocabulary Levels Test. *ITL - International Journal of Applied Linguistics*, *168*(1), 33–69. https://doi.org/10.1075/itl.168.1.02web.
Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Routledge.
Wolfe, E. W., & Smith Jr., E. V. (2007). Instrument development tools and activities for measure validation using Rasch models: Part 2 – Validation activities. *Journal of Applied Measurement*, *8*(2), 204–234.
Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(3), 370 Retrieved from https://www.rasch.org/rmt/rmt83b.htm.
Wright, B. D., & Masters, G. N. (2002). Number of Person or Item Strata (4G+ 1)/3. *Rasch Measurement Transactions*, *16*(3), 888 Retrieved from https://www.rasch.org/rmt/rmt163f.htm.
Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three parameter logistic model. *Applied Psychological Measurement*, *8*(2), 125–145 10.1177/014662168400800201.
Yen, W. M. (1993). Scaling performance assessments: strategies for managing local item dependence. *Journal of Educational Measurement*, *30*(3), 187–213. https://doi.org/10.1111/j.1745-3984.1993.tb00423.x.
Zhao, P., & Ji, X. (2016). Validation of the Mandarin version of the Vocabulary Size Test. *RELC Journal*. https://doi.org/10.1177/0033688216639761.

Publisher's Note