Akase *Language Testing in Asia* (2022) 12:5
https://doi.org/10.1186/s40468-022-00155-8

Language Testing in Asia

RESEARCH Open Access

Longitudinal measurement of growth in vocabulary size using Rasch-based test equating

Masaki Akase

Correspondence: m_akase@nagano-nct.ac.jp
National Institute of Technology, Nagano College, 716 Tokuma, Nagano-shi, Nagano-ken 381-8550, Japan

Abstract

The purpose of this study is to equate and further validate three forms of the vocabulary size test (VST) created by Aizawa and Mochizuki (2010). These three forms, VST 1, 2, and 3, were administered to a cohort of 189 high school students ranging in age from 16 to 18 in April of their 1st, 2nd, and 3rd year of high school. Although these alternate forms were designed to be of equal difficulty, formal equating of the three forms was never carried out. In order to verify whether gains in test scores were due to growth in vocabulary size or differences in difficulty among the three forms, a fourth form comprised of items selected from VST 1–3 was created and administered in December of year 3. The four test forms were then equated using Rasch analysis, placing persons and items on a single, uniform logit scale. The results indicated that (1) the three original forms of the VST all showed good fit to the Rasch model, (2) differences in test difficulty among the original three forms were minor, and (3) the four VST forms, linked via a single Rasch analysis, can be appropriately used for estimating gains in students' VS across their high school career. In addition, follow-up analyses indicated considerable overlap in item difficulty among word frequency bands, suggesting that word frequency was not the sole indicator of difficulty of vocabulary items, and that learner progress in vocabulary learning tended to be uniform and parallel across all frequency bands. Overall, the study illustrates a method for creating a valid and reliable measure of growth in VS over an extended period of time and provides insight into the relationships among word frequency, word difficulty, and progress in vocabulary learning.

Keywords: Vocabulary size, Vocabulary test, Longitudinal measurement, Rasch analysis

Introduction

Researchers and teachers agree that learning vocabulary is an essential part of second language (L2) acquisition. The importance of learning vocabulary is revealed in three ways. Firstly, vocabulary is the primary building block of understanding for communication. As Wilkins (1972) puts it, "without grammar, very little can be conveyed; without vocabulary nothing can be conveyed" (pp. 111–112). Secondly, sufficient vocabulary helps L2 learners expand the four skills. Regarding this, Nation (1994) notes, "a rich vocabulary makes the skills of listening, speaking, reading, and writing

easier to perform" (p. viii). Finally, more vocabulary leads learners to learn more, which means that in all stages of learning, vocabulary is central to learning content and is therefore an essential component of almost every aspect of our lives (Webb & Nation, 2017). Accordingly, increasing vocabulary size (VS) must be a key goal for enhancing communication in a foreign language.

Literature review

Vocabulary size and depth

In order to assess various facets of vocabulary knowledge (Henriksen; 1999; Nation, 2001, 2020; Read; 2000), several types of vocabulary test have been developed thus far (e.g., Meara & Jones, 1990; Nation, 1990; Read, 2000; Schmitt, 2000; Nation & Beglar, 2007; Wesche & Paribakht, 1996). Two well-established descriptive dimensions of vocabulary knowledge are those of size/breadth (hereafter size) and depth/quality (hereafter depth) (Anderson & Freebody, 1981). While size refers to *how many* words are known, depth is concerned with *how well* those words are known (Schmitt, 2000). Thus, vocabulary knowledge can be assessed from the perspective of both size and depth. However, measuring both types of vocabulary knowledge can be difficult and impractical in classroom settings where time is limited.

Although the size-depth distinction has been widely taken up (e.g., Read, 2004), size has been regarded as a more useful measure of L2 vocabulary knowledge partly because of its simple polling method in which multiple target words are measured at a single time. It also has been presumed to be the primary aspect of vocabulary knowledge due to its importance in the form-meaning link for vocabulary use (e.g., Meara, 2002; Laufer et al., 2004; Schmitt, 2010), and generally, researchers have paid more attention to the significance of measuring learners' VS.

According to Meara (1996), VS is more important for L2 learners, especially for those with a small lexicon. Read (2000, p.115) acknowledges that "despite the fact that the size tests may seem superficial, they can give a more representative picture of the overall state of the learner's vocabulary than an in-depth probe of a limited number of words". Furthermore, having a larger VS means that there is greater potential for learners to understand the language that is encountered (Webb & Paribakht, 2015). Consequently, the current study focuses on growth in VS, as opposed to development of depth of vocabulary knowledge, as the more practical and possibly more important dimension of growth for high school students studying EFL. Growth in VS can be conveniently defined as the amount of increase in words recognized, as measured by vocabulary size tests of equivalent difficulty.

Measuring vocabulary size

Several researchers in Japan have used Nation's (1990, 2001) word frequency approach to make vocabulary size tests better suited for Japanese EFL learners. For example, Aizawa (1998) created his own vocabulary size test (VST) and added four major improvements. Firstly, he increased the number of items in each level band, aiming to improve the reliability and predictive validity for estimating the VS of each frequency band. Secondly, he wrote the definitions of the items in Japanese, eliminating incorrect answers due to not understanding the definition while knowing the word. Thirdly, he

selected words based on word-item (lemma[1]) counting instead of word family[2] counting, which many vocabulary specialists believe is better tuned to learner knowledge of morphology compared to word family. Specifically, many EFL Japanese learners may be unfamiliar with the various word forms, such as noun, adjective, adverb, etc., that is assumed when basing VS estimates on word families (see Kremmel, 2016; McLean, 2018). Finally, he based frequency calculations on a locally developed corpus, namely the Hokkaido University English Vocabulary List (HUEVL; Sonoda, 1996). The HUEVL is based on approximately 9 million words from Time Magazine plus a total of 2.7 million words from the US Department of Energy Corpus from 1989 to 1993. Thus, the HUEVL represents English as used for current topics and science. It is organized into five level bands from junior-high to college-advanced level. Soon after Aizawa created his test, Mochizuki (1998) revised the HUEVL and created a revised version of the test that reflected higher familiarity with loan words, and grouped words into seven word-frequency levels: 1000- to 7000-word levels. Mochizuki's (1998) test included 30 items in each level band, in which test takers identify, from six alternatives, the two English words that match the two Japanese definitions. An example item from the Mochizuki (1998) test is shown in Table 1.

The 2000- to 7000-word levels of Mochizuki's (1998) test were further analyzed by Kasahara (2006) using the FACETS (Linacre, 2005) Rasch measurement software package. He replaced 21 items which were deemed too easy or which did not match a hierarchical Rasch model with new items selected from the Japan Association of College English Teachers (JACET) List of 8000 Basic Words (JACET, 2003). He argued that his revisions maintained the high reliability of the test and improved its validity by removing items that did not fit a unidimensional scale structure, as indicated by FACETS' fit statistics. Later, Aizawa and Mochizuki (2010) devised three forms of the VST with 26 words in each level band. They also provided a formula for estimating the total VS of junior and senior high school learners, which has been widely used in Japan (e.g., Yashima, 2002; Kosuge, 2003; Katagiri, 2009; Akase & Uenishi, 2015). Using these VSTs, estimated VS was calculated separately for each level band by multiplying the percent of items answered correctly by 1000, with the total VS then estimated by summing the results for the seven level bands.

As mentioned above, the word frequency approach has generally been used to construct VSTs. However, aside from frequency, a wide range of other factors have been acknowledged to make words relatively easy or difficult to learn, including phonotactic regularity, structural or morphological complexity, word class, imageability of concept, and word meaningfulness (see Laufer, 1997; de Groot & van Hell, 2005; de Groot, 2006). These intralexical factors are also influenced by the regularity of the language: the more regular the lexis of a language, and the more that a word or phrase conforms to the language's norms, the easier it is to learn (Schmitt & Schmitt, 2020). These multiple sources of word difficulty might call into question whether the unidimensional Rasch model is appropriate for estimating difficulty. However, distinctions between psychological dimensionality and measurement dimensionality have been well discussed in

[1]The term lemma is more restricted and includes only the base word and its inflections. Lemmas are made up of a headword (e.g., "add") and its inflections ("adds," 'adding," "added").
[2]Word families are usually held to include the base word (e.g., "add"), all of its inflections ("adds," "adding," "added") and its common derivatives ("addition," "additions," "additional," "additionally," etc.).

Table 1 Mochizuki's (1998) VST format (1000-word level)

1. 小麦粉を焼いた菓子 (2)			2. 集まり，会 (4)		
(1) birthday	(2) cookie	(3) fork	(4) party	(5) star	(6) sweater

the literature (Sick, 2010; McNamara, 1996 p. 270–271), and employing a Rasch approach for VST analysis is well established (e.g., Beglar, 2010; McLean, Kramer, & Beglar, 2015).

Although it is acknowledged that many factors affect word difficulty, VSTs have generally adopted word frequency as the sole indicator of difficulty for practical reasons (Nation & Anthony, 2016). This is due in part to the development of corpus linguistics, which has made it easy to categorize words from various sources and genres into frequency bands. Nevertheless, the extent to which frequency determines difficulty in a VST is an unresolved issue. Kasahara (2006), for example, found that while mean differences in difficulty between the frequency bands were orderly and progressive, there was considerable overlap in the range of difficulty of the items within the six frequency bands that he analyzed. Ha (2021) on the other hand, found that frequency bands, with the exception of the 5000-word level, were not only progressive in difficulty but showed very little overlap.

A related question is whether growth in VS is characterized by the progressive mastery of the words comprising each frequency band, in order, or whether learners make more or less equal progress in each frequency band as a course of instruction progresses. To my knowledge, no studies have directly addressed this issue.

To gain a better understanding of growth in vocabulary, this study focused on measuring longitudinal growth in VS of Japanese EFL high school students. Measuring longitudinal growth presents two problems. One is a possible practice effect. If learners take the same test several times, motivated learners might be led to looking up or learning words on the test, which can lead to overestimating their VS on future tests. Aizawa and Mochizuki's (2010) constructed three forms of their VST with the goal of avoiding this practice effect. However, these three forms were assumed to be of equal difficulty without any formal equating.

The other issue with the measurement of growth over time is a possible testing effect. The three VST forms might vary somewhat in difficulty. When attempting to measure growth, it is not clear whether higher scores on a testing occasion are due to learners' growth, slight differences in test difficulty, or some combination of both. In this study, equating these three forms using Rasch analysis is offered as a solution to both problems.

Applying Rasch measurement to vocabulary size tests

According to Sick (2008a, 2008b), Rasch measurement is used to assess the quality of tests and questionnaires. It constructs true interval-scale measures from raw scores. It also plays an important role in construct validation. In addition, most importantly for the purpose of this research, Rasch analysis can be used to equate tests by creating a common scale based on Rasch logits.

Figure 1 is a simplified example of using Rasch measurement theory to link two tests with common items. Test 1 is easier than test 2. If a group of people take both tests, it is expected that they will get a lower percent score on test 2, the more difficult test. To be able to compare their scores, it is necessary to place the test scores on a common scale. Rasch analysis enables the analyst to place all test items on a single logit scale, based on the difficulty of a subset of common items. Item difficulty is estimated based on the percentage of test takers who answered each item correctly. The item scores are then converted to Rasch logits, which represent a probability that a test taker of a certain ability will be able to answer an item correctly. Easier items have lower logits, such as − 2 and − 1, and more difficult items have higher logits such as 3, 4. The zero logit is set to the average difficulty of item 4 on both tests. Items 4, 5, 6, which appear on both tests, are common items that can be used as linking items.

A Rasch analysis constructs an interval level scale which is invariant between test occasions. No matter which items are selected, persons with greater ability will tend to do better and regardless of a persons' ability, there will be a greater chance of success on an easy item than a more difficult one. In addition to producing interval-level measures for both person ability and item difficulty that can be employed in further statistical analyses, a Rasch analysis can be used to evaluate test targeting using a Wright map (Engelhard Jr. and Wang, 2021), a figure which shows the simultaneous locations of persons and items along a common Rasch logit scale.

Rasch measurement may also provide an alternative method of estimating total VS. The dichotomous Rasch model (Rasch, 1980) is an estimation of the probability that a test taker of an estimated ability will answer an item of an estimated difficulty correctly. The Rasch dichotomous model can be expressed using the following equation,

$$P(X_{ni} = 1) = \frac{e^{(\theta n - \delta i)}}{1 + e^{(\theta n - \delta i)}}$$

in which θ represents the ability level of the test taker, δ represents the difficulty of an item on the same scale, and e is the natural log (2.171). Assuming the data fit the

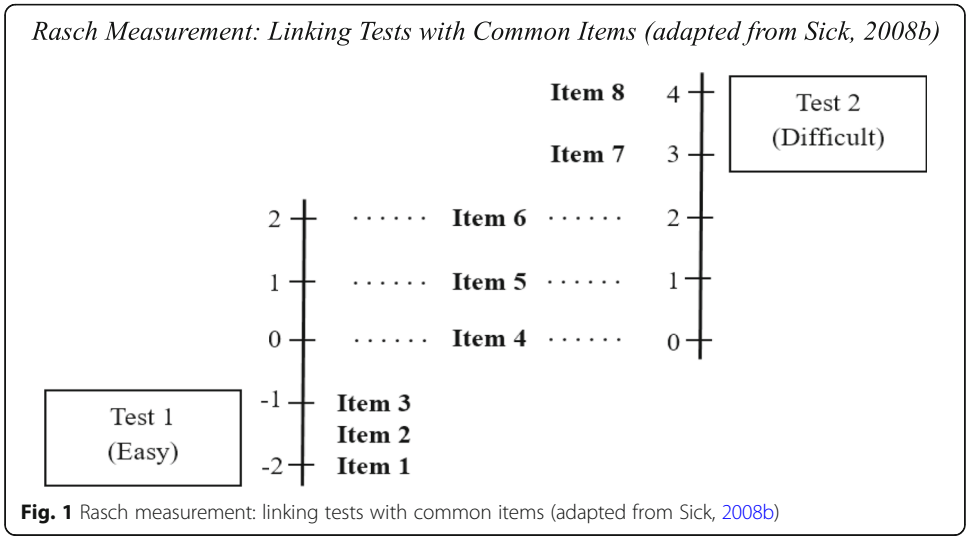

Fig. 1 Rasch measurement: linking tests with common items (adapted from Sick, 2008b)

Rasch model, the formula can be used to determine the probability that student n will answer item i correctly ($X = 1$).

The above equation can also be used to estimate the number of items within a set that a student will answer correctly. For example, by calculating the mean item difficulty of a set of items sampled from a frequency band, it is possible to estimate the number of items in that set that will be answered correctly by a student of known ability. It is then possible to infer the total number of words in that frequency band that is known by the student, based on the mean item difficulty of the sample (see Gibson and Stewart, 2014, for a more detailed explanation).

An important difference from the formula proposed by Aizawa and Mochizuki (2010), which will hereafter be referred to as the raw score method, is that the Rasch method incorporates information from all test items, not only the items in that level band. In the VS estimation using the Rasch method, the estimated number of words known in a frequency band is equal to the mean probability of success multiplied by 1000. The total VS is the sum of estimated words known for each level.

This study attempts to equate and validate the three forms of a VST created by Aizawa and Mochizuki (2010) and to determine the validity of using these three test forms to measure longitudinal growth in VS during three years of high school study. The larger purpose of the study is to lay the groundwork for investigating factors that affect or influence growth in VS across an extended period, such as a course of EFL study during secondary school. However, in the author's opinion, that goal cannot be separated from the need to methodically link the instruments used to measure VS. Above all, it is necessary to account for and eliminate both a testing effect and practice effect. Without these steps, measurement of growth in VS will be ambiguous in that it will not be clear whether observed changes in longitudinal measures are due to growth in VS, differences in difficulty of the VSTs, or a combination of both. Thus, the first three research questions form a logical sequence.

RQ1: To what extent do the three forms of Aizawa and Mochizuki's VST (2010) fit the Rasch measurement model?

RQ2: If the three forms of the VST fit the Rasch model individually, can a fourth form (VST 4) be created to validly link them to a common scale?

RQ3: To what extent can the four forms then be used to unambiguously measure growth in VS during a three-year high school career?

Two additional research questions address characteristics of word difficulty and growth in VS with potential pedagogical implications:

RQ4: To what degree do the words within frequency bands vary in difficulty, as estimated using a common scale of measurement?

RQ5: To what degree is growth in VS influenced by word frequency? Specifically, is progress characterized by the progressive mastery of each frequency band, or do learners make parallel and equivalent progress in multiple frequency bands?

Method

Participants

Originally, 204 Japanese EFL high school students majoring in science and engineering at a National Institute of Technology (NIT) were recruited for this study. The study

began in April 2018. However, during the 3-year high school duration of the study, 15 students were unable to participate in all parts of the study, dropping the pool to 189 participants (36 female, 153 male). Participants were informed that participation was voluntary and that no data obtained from the study would not affect their school grades.

Before entering NIT, all participants had studied English for at least 3 years in Japanese junior high school. They had also taken Foreign Language Activities in grades five and six in elementary school, where they developed a foundation of communication abilities through English. During the three-year enrollment in NIT, their general English proficiency could be estimated as ranging from grade pre-2 (CEFR A2) to grade 2 (CEFR B1) of the STEP Eiken test, which is the most widely used standardized test for assessing test takes' English proficiency in Japan. Basically, most students felt the importance of studying English and a few highly motivated students achieved Eiken Grade Pre-1 (CEFR B2). In April, 2021, at the end of their 3rd year, all students took the Test of English for International Communication (TOEIC), a commercial test which is widely used by Japanese companies when recruiting employees in Japan. The average TOEIC score of the participants was around 400.

Materials and procedure

The three forms of Aizawa and Mochizuki's VST (2010) were employed in this study. Because of time constraints and test difficulty, six level bands representing 1000- to 6000-word levels were used. Higher level bands were not included because it was assumed that few of the participants would have encountered such low-frequency words, and with a multiple-choice format such as the Aizawa and Mochizuki VSTs, widespread guessing of unknown words could both overestimate VS and lower reliability (Stewart, 2014). Each level has 26 items, which means the total number of items was 156 for each form. The three VST forms (VST 1, VST 2, and VST 3) were administered to the participants in April from 2018 (first-year) to 2020 (third-year), respectively.

As mentioned previously, no formal equating of test difficulty of the three VST forms has ever been carried out. Consequently, a longitudinal study would conflate learner increases in VS with differences in test difficulty. Although Rasch analysis is frequently employed for equating test forms, the problem was that these forms did not have any common items that could be used to link them. In order to link the three forms of a VST, it was necessary to make a fourth version, VST 4, which consisted of words selected from each level of each test form. Specifically, for each vocabulary frequency band, 42 items were selected from VST 1 and 8 items from both VST 2 and VST 3, making a total of 26 items for each vocabulary band, equivalent in length to VST 1, 2, and 3.

Prior to selecting items for VST 4, separate Rasch analyses were carried out on VST 1–3 using Winsteps version 4.7 (Linacre, 2019) in order to determine that the three forms fit the Rasch model individually and were thus suitable for linking. Items were selected for VST 4 based on having very good fit to the Rasch model and to represent the full range of difficulty found in each frequency band. However, items that were answered correctly by all or all but a few participants were avoided because items with extreme scores cannot be used as linking items. Consequently, it was expected that VST 4 would be slightly more difficult, on average, than VST 1–3. VST 4 was administered to the same participants in December 2020.

Analyses

Following the individual analyses of VST 1–3 and the administration of VST 4, data from the four forms were analyzed using the stacking method (Wright, 1996; Bond, Yan, & Heene, 2021, p. 203). The stacking method combines data from all four administrations, initially treating persons tested at different times as different persons, in order to estimate all item difficulties concurrently. The common items comprising VST 4 provide a basis for linking the four forms. This technique places all forms of the test on a common logit scale, allowing unambiguous measurement of mean item difficulty of the three original test forms, as well as of person ability at different times. In all, the stacking analysis combined data from a total of 468 items (156 items times 3) plus 756 students (189 students times 4). Figure 2 provides a graphic illustration of the procedure.

Following the stacking analysis, mean item difficulties of VST 1 to VST 4 were compared using a one-way repeated measures analysis of variance (ANOVA) in order to assess whether the original forms were of equivalent difficulty. Following this, the total VS for each student at each administration was estimated using the Rasch estimation method. Mean VS at each administration was then plotted and a one-way repeated measures ANOVA used to assess whether growth in mean VS was statistically significant and substantially meaningful.

Finally, the range of difficulty found within each vocabulary frequency band was plotted and compared in order to ascertain the degree to which word frequency alone affects the likelihood of students knowing the word. The analysis of frequency bands was further supplemented by plotting growth in VS by frequency band in order to examine whether VS within frequency bands tends to grow at differential or parallel rates.

Results

Individual Rasch analyses of VST 1–3

The results of the individual Rasch analyses for VST 1–3 are presented in Table 2. The individual item fit statistics were explored to examine the technical quality of the three

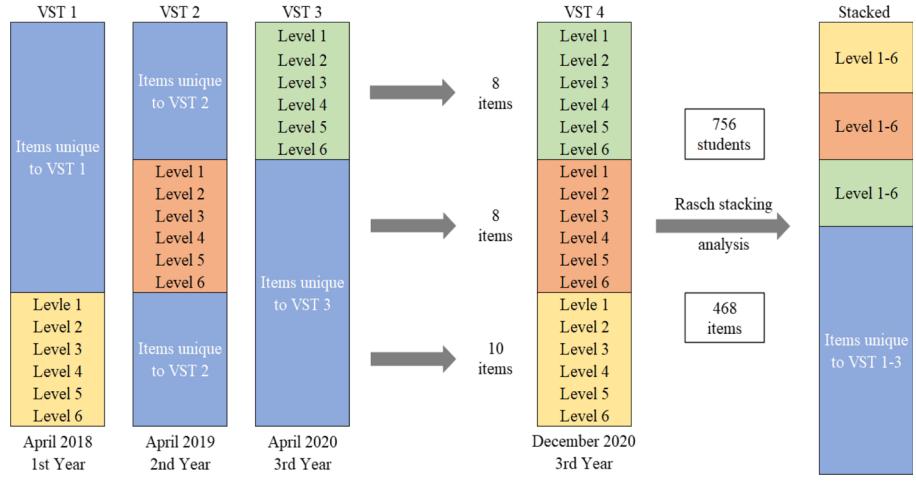

Fig. 2 Linking VST 1–3 using common items in VST 4

Table 2 Descriptive and fit statistics for VST 1–3

Measures	VST 1	VST 2	VST 3
Mean (raw score)	69.2 (44.4%)	79.7 (51.1%)	88.9 (57.0%)
SD (raw score)	12.6	16.9	17.2
Mean Rasch person measure	− 0.27	0.18	0.71
Maximum infit MNSQ	1.13	1.26	1.20
Minimum Infit MNSQ	0.86	0.76	0.86
Maximum outfit MNSQ	2.53	2.45	3.06
Minimum outfit MNSQ	0.54	0.17	0.49
Item reliability	0.98	0.98	0.98
Item separation	7.04	7.26	6.81
Person reliability	0.87	0.93	0.93
Person separation	2.54	3.71	3.57

Note. N = 189; k (number of items) = 156; Rasch measures for each test form were separately calibrated to 0 logits = mean item difficulty

original VSTs. This preliminary examination is primarily to determine whether the individual tests have adequate fit to the Rasch model and are thus suitable for linking via common items.

As shown in Table 2, each VST administration had a higher mean score. The maximum possible score on each VST was 156. Average raw scores for VST 1–3 were 69.2, 79.7, and 88.9, which were 44.4%, 51.1%, and 57.0% of the total, respectively. The standard deviation of each VST administration was also higher, which indicates that the scores were spreading out across time.

The items comprising the three forms of the VST fit a Rasch model well[3] (RQ1). Infit MNSQ for all items was within 0.7–1.3.[4] Although there were some items with high Outfit MNSQ, close examination showed that these cases were all very easy or very difficult items. Unexpected scores were probably due to guessing correctly or carelessness by a small number of test takers.

Item and person reliability[5] ranged from 0.87 to 0.98, which indicates a high degree of replicability. The person reliability of VST 1 is lower than VST 1 and 2, possibly because vocabulary knowledge was lower at Time 1 and the participants answered more items by guessing. Item separation[6], an estimate of the spread or separation of the items along the measured variable, ranged from 5.17 to 6.24. This shows that the items can be separated

[3]In Rasch measurement, the idea of model-data fit, in which a judgment of the degree to which the data under investigation meet the requirements of a model (Engelhard & Wang, 2021), is addressed by evaluation of the Infit and Outfit Mean Square statistics. Infit statistics are more sensitive to irregular patterns where persons/items are well targeted, whereas the Outfit statistics can better detect outliers throughout the whole range of the scale (Engelhard, 2013).

[4]What constitutes "acceptable" infit and outfit means square fit statistics is a matter of contention in Rasch measurement theory. Linacre (2012) has suggested a range between 0.50 and 1.50 as acceptable fit for a test or questionnaire under development. However, a more conservative range of 0.7–1.3 has been recommended for multiple-choice tests used for high stakes or substantial decisions (Wright et al., 1994; Bond et al., 2021).

[5]Reliability statistics report the reproducibility of the measures. A reliability value of 0.90 or higher is accepted as a high value. In a Rasch analysis, the person reliability estimates how likely these person measures would be reproduced using a different set of items sampled from the same domain. Likewise, the item reliability estimates how likely the item difficulty measures would be reproduced if the test were administered to a similar sample of test takers. (Linacre, 2012).

[6]The item separation index shows the number of statistically significant levels into which the items could be divided according to difficulty. A value of 3.00 or more is considered good (Linacre, 2012).

into 5 to 6 statistically distinct bands based on their difficulty. Person separation[7] ranged from 2.54 to 3.71, which means the participants can be divided into two or three statistically different groups according to their scores on the VSTs.

Based on the aforementioned results of the individual Rasch analyses for VST 1–3, items were carefully selected for VST 4 to be representative of each level band and to reflect the full range of difficulty found in each frequency band. This procedure allowed the researcher to validly link the four test administrations to a common Rasch logit scale (RQ2). For the benefit of other researchers who may wish to link other tests to the Aizawa and Mochizuki tests, the Rasch measures of all linking items (VST 4) have been included as an appendix.

Wright map of person and items measures

Figure 3 presents a Wright map generated from the stacked analysis of VST 1–4. The distribution of student abilities is shown on the left-hand side of the map, and the distribution of items by difficulty, labeled to indicate the frequency band they were drawn from, is illustrated on the right-hand side of the map. The Wright map illustrates the targeting, the match of student ability to item difficulty, of the four VST administrations as a single system of measurement.

Both persons and items are measured on a common logit scale that ranges from − 4 to 4 logits. A single item from level 5, the word "offspring" in fact, stands out at the top of the map as the most difficult. Along the bottom of the map, we can see that a number of items from level 1 were very easy for this group of test takers. However, a few items from Level 3 (e.g., "bean" and "balloon") were also very easy for this group of persons.

Overall, the Wright map indicates that the distribution of vocabulary item difficulties matches the distribution of student abilities during their 3-year course of study well and that there is a correspondence between frequency level band and item difficulty. However, the correspondence is not perfect. For example, a few level 4 words, such as "feast," "hinder," and "triumph," were difficult items at slightly over 3 logits, answered correctly by approximately 42% of test takers only. In addition, one level 2 word, "mend" at 2.5 logits, was located at the outer range of person ability (RQ3 and RQ4).

Results of mean item difficulties of VST

Figure 4 and Table 3 show the results of a one-way ANOVA comparing differences in mean item difficulty of VST 1–4. Although there are some differences in the mean item difficulty of the three original forms, the error bars in Fig. 4 indicate that these differences are within the error of estimation.

Only VST 4, which deliberately excluded the easiest items from VST 1–3, has a mean difficulty that appears to be outside the error bars of the other three versions. However, a one-way ANOVA indicated that after equating, mean differences among the versions were not significant (p = .45). An implication of this is that the three original forms of the VST created by Aizawa and Mochizuki (2010) do not differ significantly in difficulty and can now be regarded as having been formally equated (RQ3).

[7]The person separation index indicates the number of statistically distinct groups into which the person sample could be divided. An index of 1.50 is acceptable, 2.00 is good and 3.00 is excellent (Duncan et al., 2003).

Wright Map of VST 1-4

```
MEASURE   PERSON  -  ITEM
-----------------------------------------------------------------------------
            <more>|<rare>
 4                +
                . |  Lv5
                  |
                  |  Lv4
                  |  Lv6
                . |  Lv6
                  |  Lv4 Lv4 Lv5 Lv5 Lv5 Lv6 Lv6 Lv6 Lv6
 3              . +  Lv3 Lv4 Lv5 Lv5
                . |  Lv4 Lv4 Lv4 Lv5 Lv6 Lv6 Lv6
                . |  Lv3 Lv4 Lv5 Lv6 Lv6 Lv6
                . |  Lv2 Lv3 Lv3 Lv4 Lv4 Lv4 Lv4 Lv4 Lv5 Lv5 Lv6 Lv6
              # T|  Lv2 Lv3 Lv3 Lv4 Lv4 Lv5 Lv5 Lv5 Lv6 Lv6 Lv6 Lv6 Lv6 Lv6
             .#   |S Lv2 Lv4 Lv4 Lv4 Lv5 Lv5 Lv5 Lv5 Lv5 Lv5 Lv6 Lv6 Lv6 Lv6 Lv6
             .#   |  Lv2 Lv3 Lv3 Lv3 Lv4 Lv4 Lv4 Lv5 Lv5 Lv5 Lv5 Lv6 Lv6 Lv6 Lv6 Lv6 Lv6 Lv6 Lv6
 2          .##   +  Lv3 Lv4 Lv4 Lv4 Lv5 Lv5 Lv5 Lv5 Lv5 Lv5 Lv6 Lv6 Lv6 Lv6 Lv6
           .###   |  Lv2 Lv3 Lv3 Lv4 Lv4 Lv4 Lv4 Lv5 Lv5 Lv6
        .######   |  Lv3 Lv3 Lv4 Lv4 Lv4 Lv5 Lv6 Lv6 Lv6 Lv6 Lv6 Lv6 Lv6 Lv6
          .####   |  Lv2 Lv2 Lv3 Lv4 Lv4 Lv4 Lv5 Lv5 Lv5 Lv6 Lv6 Lv6 Lv6 Lv6 Lv6 Lv6
         .##### S|  Lv3 Lv4 Lv4 Lv4 Lv4 Lv5 Lv5 Lv5 Lv6
      .########   |  Lv2 Lv3 Lv3 Lv4 Lv4 Lv4 Lv4 Lv5 Lv5 Lv5 Lv5 Lv5 Lv5 Lv6 Lv6
      .########   |  Lv2 Lv2 Lv2 Lv3 Lv3 Lv3 Lv4 Lv4 Lv4 Lv4 Lv4 Lv5 Lv6 Lv6 Lv6 Lv6
 1    .########   +  Lv2 Lv2 Lv3 Lv3 Lv3 Lv4 Lv4 Lv4 Lv5 Lv5 Lv5 Lv6 Lv6
      #########   |  Lv2 Lv2 Lv2 Lv3 Lv3 Lv3 Lv4 Lv4 Lv4 Lv5 Lv5 Lv5 Lv5 Lv6
   .###########   |  Lv2 Lv3 Lv5 Lv5 Lv6
    .##########   |  Lv2 Lv3 Lv3 Lv3 Lv3 Lv3 Lv3 Lv4 Lv4 Lv5 Lv6 Lv6 Lv6 Lv6
  .############ M|  Lv2 Lv2 Lv2 Lv3 Lv4 Lv4 Lv4 Lv4 Lv5 Lv5 Lv5
  #############   |  Lv2 Lv2 Lv2 Lv3 Lv4 Lv4 Lv5 Lv5 Lv5 Lv6 Lv6
       ########   |  Lv2 Lv2 Lv3 Lv4 Lv5 Lv5 Lv6 Lv6
 0 .###########   +M Lv2 Lv2 Lv2 Lv4 Lv5
  .############   |  Lv2 Lv2 Lv3 Lv3 Lv3 Lv3 Lv3 Lv3 Lv3 Lv3 Lv4 Lv4 Lv6
     .#########   |  Lv2 Lv2 Lv2 Lv2 Lv3 Lv3 Lv4 Lv5 Lv6
         ######   |  Lv2 Lv2 Lv4 Lv5
         ###### S|  Lv2 Lv2 Lv2 Lv4 Lv4 Lv5 Lv6
         .#####   |  Lv2 Lv3 Lv3 Lv3 Lv4 Lv4 Lv5 Lv5
           ####   |  Lv1 Lv2 Lv3 Lv3 Lv3 Lv3 Lv3 Lv5 Lv5 Lv6 Lv6
-1          ###   +  Lv2 Lv2 Lv2 Lv3 Lv3 Lv3 Lv4 Lv4 Lv5 Lv5 Lv5
            .##   |  Lv3 Lv3 Lv4 Lv5
             .#   |  Lv1 Lv1 Lv2 Lv3 Lv3 Lv3 Lv4
              .   |  Lv1 Lv2 Lv2 Lv2 Lv2 Lv3 Lv4
             .# T|  Lv2 Lv3 Lv3
             .#   |  Lv2 Lv2 Lv2 Lv3
              .   |
-2            .   +  Lv1 Lv2
              .   |  Lv1 Lv1 Lv1 Lv2 Lv2 Lv3 Lv3
              .   |S Lv2 Lv2 Lv2 Lv3 Lv3 Lv5
              .   |  Lv1 Lv1 Lv2 Lv2 Lv2 Lv3 Lv5
              .   |  Lv1 Lv2 Lv2 Lv4
              .   |  Lv1 Lv1 Lv2 Lv2
                  |  Lv1 Lv1 Lv1 Lv2 Lv2 Lv2
-3                +  Lv1 Lv2 Lv4 Lv5
                  |  Lv1 Lv2 Lv2 Lv2
                  |  Lv1 Lv1 Lv1 Lv1 Lv1 Lv2 Lv2
                  |  Lv2
                  |  Lv1 Lv1 Lv1 Lv3
                  |  Lv1 Lv1 Lv1 Lv1
                  |  Lv1 Lv2 Lv3
-4                +  Lv1 Lv1 Lv1 Lv1 Lv1 Lv1 Lv1 Lv1 Lv1 Lv1 Lv1 Lv1 Lv1 Lv1 Lv1 Lv1
                     Lv1 Lv1 Lv1 Lv1 Lv1 Lv1 Lv1 Lv1 Lv1 Lv1 Lv1 Lv1 Lv1 Lv1 Lv1 Lv1
                     Lv1 Lv1 Lv1 Lv1 Lv1 Lv1 Lv1 Lv1 Lv1 Lv1 Lv1 Lv1 Lv1 Lv1 Lv1 Lv3
                     Lv3 Lv3 Lv3
-----------------------------------------------------------------------------
```

Note. Each "#" is 4 persons; each "." is 1-3 persons. $N = 756$, 468 items.

Fig. 3 Wright Map of VST 1–4. *Note.* Each "number sign" is 4 persons; each "full stop" is 1–3 persons. $N = 756$, 468 items

Estimated mean vocabulary size

Figure 5 shows the estimated mean vocabulary size at each administration of the VST. Both the estimated vocabulary size using the raw score method (score EVS) and the estimated vocabulary size using the Rasch method (Rasch EVS) indicate growth at each consecutive administration.

Although there are some slight differences between the methods, both indicate consistent growth in VS across time. Because VST 2 was slightly easier than VST 1, it is probable that the raw score EVS slightly overestimated mean growth in VS during the first year. The difference between the two methods is most pronounced at Time 4. This is most likely because VST 4 by design omitted the easiest items because items answered correctly by all test takers are not suitable for Rasch equating. The score EVS thus underestimates

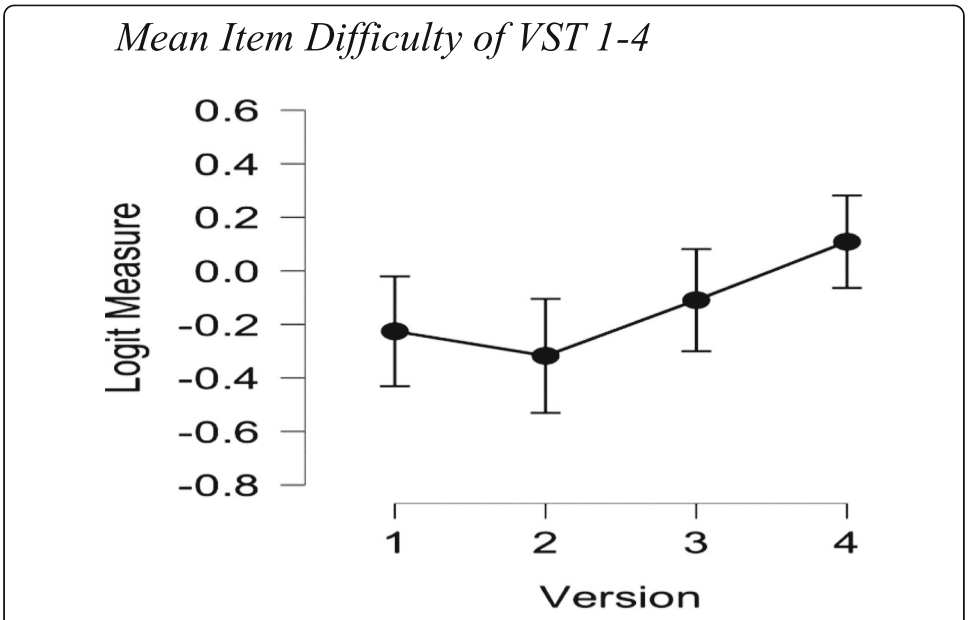

Fig. 4 Mean item difficulty of VST 1–4

growth in VS because it is a slightly more difficult test and the raw scores cannot be regarded as equivalent to those of the original three versions of the VST.

Item difficulty ranges of frequency level bands

Figure 6 shows the distribution of item difficulty within the vocabulary frequency bands. Frequency band Level 1 is clearly the easiest, and in fact, the 26 items in this level band were answered correctly by about 95 to 98% of the test takers in each VST. In contrast, other frequency bands overlap considerably in difficulty (RQ4).

One level 2 item, "mend," was more difficult than the average difficulty of other frequency bands. The words "bean" and "balloon" were the easiest items in level 3 and, in fact, were also below the difficulty measures of most level 1 words. The word "offspring," which is a level 5 word, was the most difficult item on the test. Similar examples can be observed in other level bands. In some cases such as "balloon," "garbage," "hydrogen," and perhaps "economically," the easiness can be explained by the school curriculum, which is science and technology oriented. It is possible that 3rd year students encounter these words outside of their English classes.

Mastery of frequency level bands across time

Figure 7 shows the mastery of level bands across four time periods. Overall, the growth of VS within frequency bands tended to be consistent and parallel. Students made

Table 3 Differences in mean item difficulty of VST 1–4 (stacked)

Cases	Sum of squares	*df*	Mean square	*F*	*p*
Version	15.827	3	5.276	0.882	0.45
Residuals	3710.102	620	5.984		

Note. Each VST form contains 156 items, type III sum of squares

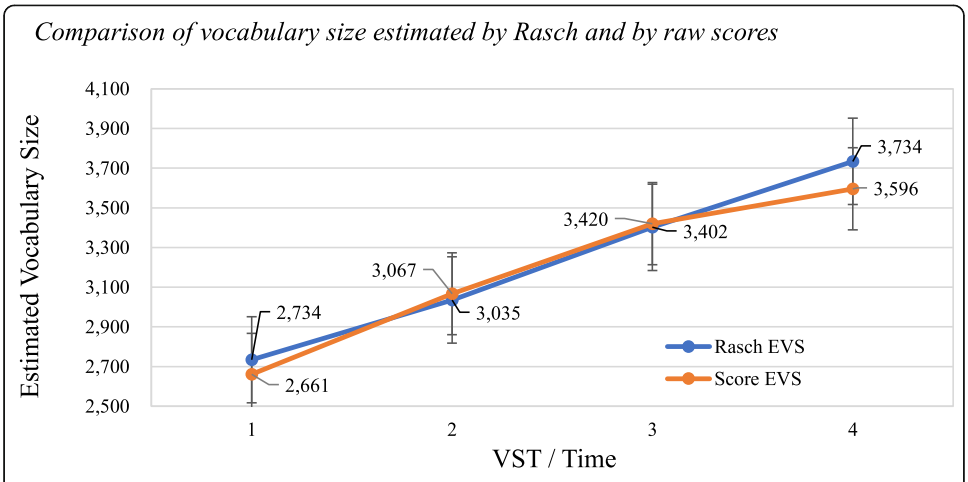

Fig. 5 Comparison of vocabulary size estimated by Rasch and by raw scores

progress in all frequency bands above level 1 at a similar rate. The annual growth rates of levels 2–6 were consistently between 6 and 8%, and the growth rates of level 3 and level 4 were so close as to be overlapping in the figure. Growth rates tended to be slightly higher between the 2nd and 3rd VST administrations (RQ5).

Discussion

This study sought to answer five RQs related to the validating and equating of the three forms of the VST created by Aizawa and Mochizuki (2010). A Rasch analysis was used to equate and compare the original three forms of the VST using a fourth form comprised of common items as a linking test. All items showed good fit to the Rasch model, which further establishes the validity and reliability of the VSTs. A Wright map based on a stacking analysis demonstrated that the four test forms, as a measurement system, were well targeted for measuring vocabulary growth during three years of study at NIT. In addition, an analysis of the relative difficulty of the three original forms based on the equated Rasch item difficulties found that they were not significantly

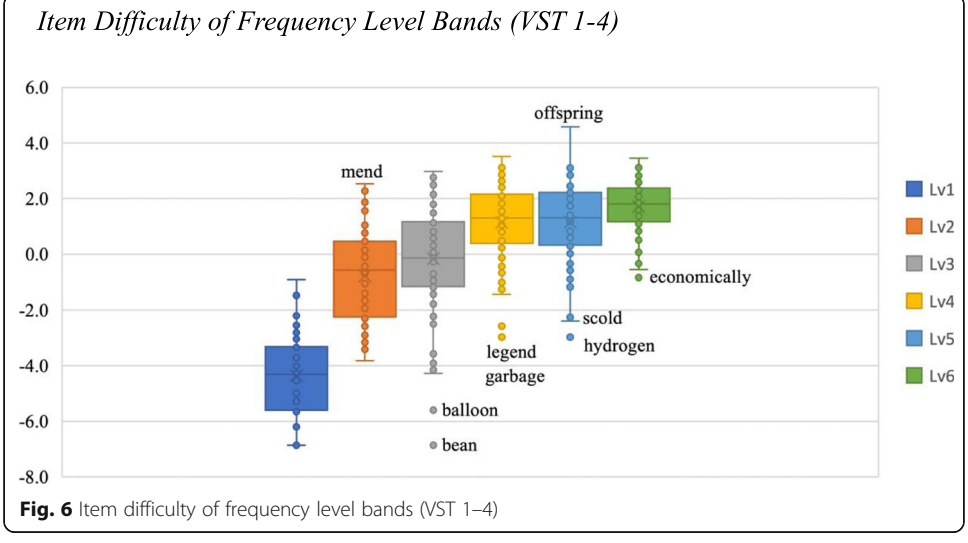

Fig. 6 Item difficulty of frequency level bands (VST 1–4)

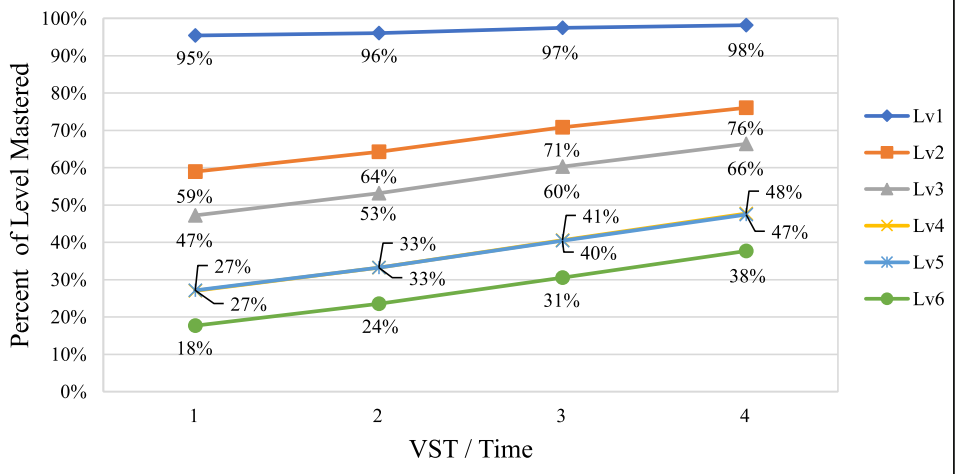

Fig. 7 Mastery of level bands across VST/time

different. This point could be important to teachers who wish to monitor growth in vocabulary size during high school, but feel they do not have the knowledge or means to link the test forms using Rasch analysis. This study has formally equated the original three forms and teachers or future researchers may now regard the raw score counts as equivalent. Estimates of growth in VS derived from these forms should be free from both practice effects and testing effects.

Regarding estimating VS based on test performance, the score method and the Rasch method were introduced and compared. The two methods produced similar but slightly different estimates of VS. It is hypothesized that differences might arise because the Rasch method incorporates information from all items to estimate the probability that a learner will know any specific word. The score method, in comparison, is derived from the total number correct in a single frequency band only and is not influenced by words outside of that band. This could be problematic in view of the fact that item difficulty is not determined entirely by word frequency. Consequently, it is suggested that future researchers employ the Rasch method for estimating VS as a more precise and valid estimate.

The analysis of word difficulty by frequency band (Fig. 6) indicated that word difficulty is not entirely determined by frequency. Although students can generally be expected to know a greater number of high-frequency words than low-frequency words, their occurrence in the school curriculum as well as intralexical factors (Schmitt & Schmitt, 2020) almost certainly influence students' learning of vocabulary. Due to the demand for engineers to have English communication skills, students in science and engineering fields are also increasingly required to learn some English words related to their areas of expertise. Thus, it is possible that the frequency of encounter is not the same as that found in the HUVEL corpus (Sonoda, 1996), which was based on current events and science articles from the 1990s. Future research may need to reexamine the consequences of using HUVEL or any other specific corpora to estimate word frequency in the Japanese EFL context. Although this result was not unexpected, it does emphasize the need for further research in the interlexical factors that make words

difficult, as well as the efficacy of the order of vocabulary presentation in the school curriculum.

The mastery of level bands across the four time periods (Fig. 7) demonstrates that apart from the Level 1 words, learners do *not* learn the words comprising the frequency bands in succession. Rather, they seem to be learning a portion of words from Level 2 to 6 each year at a gain of roughly 5–8% per band per year. This is possibly because in the Japanese EFL context, new vocabulary is acquired by deliberate study of textbook passages and assigned word lists rather than from graded readers. As noted above, further research into the relationship between word difficulty and the order in which new words are introduced in the school curriculum could have useful pedagogical implications.

Another point worth mentioning is that in Kasahara's (2006) modified version of the VST, the 1000-word level band was left unchanged. Furthermore, he did not include the Level 1 words in a study conducted with university students, assuming they would know most of the words. However, easy items known by nearly all students are necessary for accurate estimation of VS. Furthermore, the small gains of 95 to 98% that these high school students made in the 1000–word level could be pedagogically important, as some researchers (e.g., Hu & Nation, 2000; Laufer, 2005) have suggested that 98% coverage of a foreign language text is required for good comprehension and guessing from context. If so, it would be especially important that learners master a minimum of 98% of the very high-frequency level 1 words.

Conclusion and future research

This study has completed a formal validation and linking of three forms of a VST designed to measure growth in VS and tested their efficiency in the context of a three-year high school program at NIT. A Rasch analysis both established their equivalence and demonstrated that observed gains in VS could validly be attributed to changes in learner ability, rather than differences in the difficulty of the tests. The study also demonstrated that in the Japanese EFL context, word frequency is not the sole determiner of whether a student will know a word. Furthermore, learners in this study were shown to make similar gains in all frequency bands during the course of the study, an indication that for better or worse, word frequency may not be the primary factor determining the order of word presentation in the Japanese school curriculum.

A limitation that could be addressed in future studies is whether these results would generalize to other cohorts of learners of various academic backgrounds. Since this study was based entirely on science and engineering students, high school students who opt for humanities courses may show different patterns of development in their VS. Such studies could shed further light on how high school students develop their knowledge of English vocabulary.

Future research might make use of these now equated test forms to investigate other factors that influence growth in VS, such as learning strategies, motivation, and other affective factors. Further examination of which individual differences influence growth and rate of growth in VS would lead to a better understanding of vocabulary acquisition among Japanese EFL high school students. Finally, this study has demonstrated the utility of using Rasch measurement to link test forms and separate true gains in VS

from differences in test difficulty. Furthermore, a method for more accurately estimating VS in each level band using Rasch measures was demonstrated. It is recommended that future researchers investigating growth in VS consider using this approach for maximum validity and accuracy.

Appendix. Rasch item difficulties of VST 4 (linking items)

Item	Form	Lv.	No.	Measure	Item	Form	Level	Lv.	Measure
pot	1	1	1	− 1.47	flood	1	2	7	1.24
sofa	1	1	2	− 4.34	equipment	1	2	8	0.87
meal	1	1	7	− 3.34	discipline	1	2	9	2.46
piece	1	1	8	− 4.15	coast	1	2	42	1.87
mystery	1	1	13	− 1.31	mend	1	2	15	2.53
exam	1	1	14	− 2.48	contain	1	2	16	1.14
nice	1	1	21	− 3.04	curious	1	2	23	0.46
large	1	1	22	− 5.28	raw	1	2	24	− 0.30
hers	1	1	25	− 5.28	separate	1	2	25	0.26
my	1	1	26	− 4.34	urgent	1	2	26	1.56
orange	2	1	1	− 3.65	reward	2	2	7	0.32
corn	2	1	2	− 3.25	enemy	2	2	8	− 1.47
welcome	2	1	17	− 4.28	promise	2	2	15	− 2.30
work	2	1	18	− 4.52	discuss	2	2	16	− 2.69
notice	2	1	19	− 2.07	suffer	2	2	19	0.88
become	2	1	20	− 4.52	argue	2	2	20	0.56
pull	2	1	23	− 1.26	necessary	2	2	23	− 1.77
look	2	1	24	− 4.52	asleep	2	2	24	− 3.16
ship	3	1	3	− 2.73	author	3	2	7	− 1.94
hand	3	1	4	− 2.81	blood	3	2	8	− 3.07
uncle	3	1	7	− 2.81	expect	3	2	15	− 0.29
glove	3	1	8	− 3.71	win	3	2	16	− 3.41
question	3	1	13	− 2.16	complain	3	2	19	0.17
record	3	1	14	− 3.17	indicate	3	2	20	1.04
drop	3	1	19	− 2.20	dirty	3	2	23	− 0.91
hold	3	1	20	− 2.73	military	3	2	24	− 1.29
campaign	1	3	1	1.56	admission	1	4	3	− 0.28
flesh	1	3	2	2.17	feast	1	4	4	3.51
ceremony	1	3	7	− 0.87	craft	1	4	9	1.69
emergency	1	3	8	− 1.61	portion	1	4	42	1.16
revise	1	3	15	0.91	reconcile	1	4	17	1.36
decay	1	3	16	2.76	hinder	1	4	18	2.85
aware	1	3	19	0.61	purchase	1	4	19	1.29
upright	1	3	20	− 0.21	resume	1	4	20	− 0.44
actually	1	3	25	− 2.09	singular	1	4	23	0.23
anyhow	1	3	26	0.61	linguistic	1	4	24	− 0.09
stranger	2	3	5	− 0.27	bulb	2	4	1	1.15
governor	2	3	6	− 1.33	orbit	2	4	2	1.54
fraction	2	3	7	1.66	investigation	2	4	7	2.62

Appendix. Rasch item difficulties of VST 4 (linking items) *(Continued)*

Item	Form	Lv.	No.	Measure	Item	Form	Level	Lv.	Measure
troop	2	3	8	2.58	exhibition	2	4	8	− 1.43
employment	2	3	9	0.57	intensity	2	4	9	2.09
clerk	2	3	42	0.81	geography	2	4	42	0.42
household	2	3	13	− 1.18	assign	2	4	15	1.40
collar	2	3	14	1.31	mutter	2	4	16	2.60
harbor	3	3	1	− 0.71	dignity	3	4	1	0.94
slave	3	3	2	− 0.19	award	3	4	2	− 1.26
creature	3	3	3	− 0.83	triumph	3	4	5	3.17
wequence	3	3	4	1.19	prestige	3	4	6	2.20
motive	3	3	11	− 0.17	volcano	3	4	9	− 1.17
ladder	3	3	12	− 0.20	acquaintance	3	4	42	1.82
survey	3	3	13	1.97	bold	3	4	23	0.43
wisdom	3	3	14	− 1.21	rigid	3	4	24	2.29
excellence	1	5	7	− 0.85	sidestep	1	6	11	0.52
swarm	1	5	8	1.98	deepen	1	6	12	0.08
mischief	1	5	13	1.24	envelop	1	6	13	1.65
restoration	1	5	14	0.02	flush	1	6	14	1.24
whirl	1	5	15	1.57	disrupt	1	6	17	2.00
repel	1	5	16	2.04	prosecute	1	6	18	2.08
genetic	1	5	21	0.83	tidal	1	6	23	2.82
frantic	1	5	22	1.26	feeble	1	6	24	2.63
statistical	1	5	25	1.17	economically	1	6	25	− 0.80
profitable	1	5	26	1.89	fundamentally	1	6	26	1.84
garment	2	5	5	2.21	pope	2	6	3	3.22
ornament	2	5	6	0.20	crest	2	6	4	2.58
intent	2	5	13	0.81	merger	2	6	5	2.09
nationality	2	5	14	− 1.03	sewer	2	6	6	1.76
bloom	2	5	15	− 2.26	cautiously	2	6	23	1.73
scold	2	5	16	− 2.39	alternately	2	6	24	1.76
diagnose	2	5	19	2.53	continually	2	6	25	0.62
evaporate	2	5	20	2.84	mentally	2	6	26	− 0.17
basin	3	5	1	1.40	diabetes	3	6	3	3.19
seam	3	5	2	2.34	abortion	3	6	4	3.45
witch	3	5	7	− 1.17	swamp	3	6	11	1.52
offspring	3	5	8	4.58	legislature	3	6	12	1.21
commodity	3	5	11	0.67	wag	3	6	15	1.09
bruise	3	5	12	1.53	forge	3	6	16	2.37
divert	3	5	15	2.66	illegitimate	3	6	25	0.67
exert	3	5	16	2.27	premature	3	6	26	2.17

Abbreviations

VST: Vocabulary size test; VS: Vocabulary size; HUEVL: The Hokkaido University English Vocabulary List; JACET: Japan Association of College English Teachers; NIT: National Institute of Technology; CEFR: Common European Framework of Reference for Languages; STEP: The Society for Testing English Proficiency; TOEIC: Test of English for International Communication; MNSQ: Mean square; EVS: Estimated vocabulary size

Acknowledgements
The author thanks all the students and teachers who helped in collecting data. This study would not have been possible without the support of my supervisor, Dr. James Sick at Takushoku University Graduate College of Language Education. The author would also like to express my sincere gratitude for his helpful discussions and insightful comments on the manuscript.

Author's contributions
I am the sole author of the manuscript. The author read and approved the final manuscript.

Authors' information
Masaki Akase is an Associate Professor in the Division of General Education at National Institute of Technology, Nagano College. His research interests include second language vocabulary acquisition, learning strategies, and individual differences. He also collaborates with other researchers on comparative analysis of English textbooks in Japan and other Asian countries.

Funding
The author would like to acknowledge the Japan Society for Promotion of Science (JSPS) for providing financial support (Grant-in-Aid, Project Number: 20K00882) for this study.

Availability of data and materials
A copy of the linking test (VST 4) as well as summarized Rasch difficulty calibrations for the 156 items comprising that form will be provided as supplementary materials. Individual participant responses to all test items cannot be provided at the present time because permission to publish or provide raw data (non-summarized) was not granted by the institutional ethics committee.

Declarations

Ethics approval and consent to participate
Full consent to conduct the research and publish results was obtained from the National Institute of Technology, Nagano College, the institution where it was conducted. The NIT Ethics Oversite committee reviewed and approved the research in accordance with institutional and official policies regarding research involving high school students below the age of majority. Participants were informed that their participation was voluntary and test results would not affect their grades in any way.

Consent for publication
The institutional ethics committee gave permission for de-identified, summarized data to be published, on behalf of the participants. Approval from the institution, rather than the parents, is customary in Japan in the case of high school, junior high school, or elementary school students.

Competing interests
The author declares that he has no competing interests or other conflicts of interest related to this research.

Received: 21 October 2021 Accepted: 20 January 2022
Published online: 02 March 2022

References

Aizawa, K. (1998). Developing a vocabulary size test for Japanese EFL learners. *ARELE*, *9*, 75–85. https://doi.org/10.20581/arele.9.0_75.

Aizawa, K. & Mochizuki, M. (2010). Eigo goishidouno jissen idea shu: Katsudourei kara tesuto sakusei made [Practical handbook for English vocabulary teaching]. Taishukan Shoten.

Akase, M., & Uenishi, K. (2015). A longitudinal study of progress in vocabulary size of Japanese EFL senior high school learners: A comparison of the general and commercial courses. *Journal of Pan-Pacific Association of Applied Linguistics*, *19*(1), 163–182 Retrieved from https://files.eric.ed.gov/fulltext/EJ1085342.pdf.

Anderson, R. C., & Freebody, P. (1981). Vocabulary knowledge. In J. T. Guthrie (Ed.), *Comprehension and teaching: Research reviews*, (pp. 77–117) International Reading Association.

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., Yan, Z., & Heene, M. (2021). *Applying the Rasch model: Fundamental measurement in the human sciences*, (4th ed.,). New York,: Routledge.

de Groot, A. M. B. (2006). Effects of stimulus characteristics and background music on foreign language vocabulary learning and forgetting. *Language Learning*, *56*(3), 463–506. https://doi.org/10.1111/j.1467-9922.2006.00374.x.

de Groot, A. M. B., & van Hell, J. G. (2005). The learning of foreign language vocabulary. In J. F. Kroll, & A. M. B. de Groot (Eds.), *Handbook of bilingualism*. Oxford University Press.

Duncan, P. W., Bode, R. K., Lai, S. M., & Perera, S. (2003). Rasch analysis of a new stroke-specific outcome scale: The stroke impact scale. *Archives of Physical Medicine and Rehabilitation*, *84*(7), 950–963. https://doi.org/10.1016/s0003-9993(03)00035-2.

Engelhard, G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. Thousand Oaks: Routledge. https://doi.org/10.4324/9780203073636.

Engelhard Jr., G., & Wang, J. (2021). *Rasch models for solving measurement problems*. Sage.

Gibson, A., & Stewart, J. (2014). Estimating learners' vocabulary size under item response theory. *Vocabulary Learning and Instruction*, *3*(2), 78–84. https://doi.org/10.7820/vli.v03.2.gibson.stewart.

Ha, H. T. (2021). A Rasch-based validation of the Vietnamese version of the Listening Vocabulary Levels Test. *Language Testing in Asia*, *11*(16):1–19. https://doi.org/10.1186/s40468-021-00132-7.
Henriksen, B. (1999). Three dimensions of vocabulary development. *Studies in Second Language Acquisition*, *21*(2), 303–317. https://doi.org/10.1017/S0272263199002089.
Hu, M., & Nation, I. S. P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, *13*(1), 403–430. https://doi.org/10.26686/wgtn.12560354.
JACET (2003). *JACET list of 8000 basic words: JACET 8000.* JACET.
Kasahara, K. (2006). Producing a revised version of the Mochizuki Vocabulary Size Test. *JLTA Journal*, *9*(0), 55–72. https://doi.org/10.20622/jltaj.9.0_55.
Katagiri, K. (2009). A three-year longitudinal study of vocabulary size in Japanese SHS Students and a description of their developmental patterns. *ARELE*, *20*, 131–140. https://doi.org/10.20622/jltajournal.15.0_43.
Kosuge, A. (2003). Goi saizu tesuto kara mita goi no shuutoku [Looking into vocabulary acquisition, utilizing the results of vocabulary size tests]. In H. Ohta, K. Kanatani, A. Kosuge, & S. Hidai (Eds.), *Eigoryoku wa donoyounishite nobiteyukuka [How does Japanese EFL learners' English ability develop?]*. Tokyo: Taishukan Shoten.
Kremmel, B. (2016). Word families and frequency bands in vocabulary tests: Challenging conventions. *TESOL Quarterly*, *50*(4), 976–987. https://doi.org/10.1002/tesq.329.
Laufer, B. (1997). What's in a word that makes it hard or easy? Intralexical factors affecting the difficulty of vocabulary acquisition. In N. Schmitt, & M. McCarthy (Eds.), *Vocabulary: Description, acquisition, and pedagogy*, (pp. 140–155). Cambridge: Cambridge University Press.
Laufer, B. (2005). Focus on form in second language vocabulary learning. *EUROSLA Yearbook*, *5*, 223–250. https://doi.org/10.1075/eurosla.5.11lau.
Laufer, B., Elder, C., Hill, K., & Congdon, P. (2004). Size and strength: Do we need both to measure vocabulary knowledge? *Language Testing*, *21*(2), 202–226. https://doi.org/10.1191/0265532204lt277oa.
Linacre, J. M. (2005). *Facets: Rasch-measurement computer program (Version 3.57.0) [computer software]*. MESA Press.
Linacre, J. M. (2012). *Rasch-Winsteps-Facets online Rasch tutorial pdfs*. Retrieved from https://www.winsteps.com/tutorials.htm
Linacre, J. M. (2019). Winsteps® Rasch measurement computer program (Version 4.7). Winsteps.com
McLean, S. (2018). Evidence for the adoption of the flemma as an appropriate word counting unit. *Applied Linguistics*, *39*(6), 823–845. https://doi.org/10.1093/applin/amx003.
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.
McNamara, T. F. (1996). *Measuring second language performance*. United States; Longman.
Meara, P. (1996). The dimensions of lexical competence. In G. Brown, K. Malmkjaer, & J. Williams (Eds.), *Performance and competence in second language acquisition*, (pp. 35–53). Cambridge: Cambridge University Press Retrieved from https://www.lognostics.co.uk/vlibrary/meara1996a.pdf.
Meara, P. (2002). The rediscovery of vocabulary. *Second Language Research*, *18*(4), 393–407. https://doi.org/10.1191/026765830202sr211xx.
Meara, P., & Jones, G. (1990). Eurocentres Vocabulary Size Test (version E1.1/K10,MSDOS). Eurocentres Learning Service.
Mochizuki, M. (1998). Nihonjin eigo gakushusha no tameno goi saizu tesuto [A Vocabulary Size Test for Japanese learners of English]. *The Institute for Research in Language Teaching Bulletin*, *12*, 27–53.
Nation, I. S. P. (1990). *Teaching and learning vocabulary*. New York: Newbury House.
Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge University Press. https://doi.org/10.1017/CBO9781139524759.
Nation, I. S. P. (1994). *New ways in teaching vocabulary*. TESOL.
Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, *31*(7), 9–13 Retrieved from https://www.lextutor.ca/tests/nation_beglar_size_2007.pdf.
Nation, P. (2020). The different aspect of vocabulary knowledge. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies*, (pp. 15–29). Routledge https://www.routledge.handbooks.com/doi/10.4324/9780429291586-2.
Nation, P., & Anthony, L. (2016). Measuring vocabulary size. In *Handbook of Research in Second Language Teaching and Learning*, (vol. 3, pp. 355–368). Taylor and Francis. https://doi.org/10.4324/9781315716893.
Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*, (Expanded ed.,). University of Chicago Press.
Read, J. (2000). *Assessing vocabulary*. Cambridge University Press. https://doi.org/10.1017/CBO9780511732942.
Read, J. (2004). Research in teaching vocabulary. *Annual Review of Applied Linguistics*, *24*, 146–161. https://doi.org/10.1017/S0267190504000078.
Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge University Press.
Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Palgrave MacMillan. https://doi.org/10.1057/9780230293977.
Schmitt, N., & Schmitt, D. (2020). *Vocabulary in language teaching*, (2nd ed.,). Cambridge University Press. https://doi.org/10.1017/9781108569057.
Sick, J. (2008a). Rasch measurement in language education: Part 1. *Shiken: JALT Testing and Evaluation SIG Newsletter*, *12*(1), 1–6 Retrieved from https://hosted.jalt.org/test/PDF/Sick1.pdf.
Sick, J. (2008b). Rasch measurement in language education: Part 2. *Shiken: JALT Testing and Evaluation SIG Newsletter*, *12*(2), 26–31 Retrieved from https://hosted.jalt.org/test/PDF/Sick2.pdf.
Sick, J. (2010). Rasch measurement in language education: Part 5. *Shiken: JALT Testing and Evaluation SIG Newsletter*, *14*(2), 23–29 Retrieved from https://hosted.jalt.org/test/PDF/Sick5.pdf.
Sonoda, K. (1996). *Daigakuseiyou eigo goihyou no tameno kisoteki kenkyuu [Basic research for a vocabulary list for college students]. Gengo bunkabu kenkyuu houkoku sousho*: Hokkaido University.
Stewart, J. (2014). Do multiple-choice options inflate estimates of vocabulary size on the VST? *Language Assessment Quarterly*, *11*(3), 271–282. https://doi.org/10.1080/15434303.
Webb, S., & Nation, P. (2017). *How vocabulary is learned*. Oxford University Press.

Webb, S., & Paribakht, T. S. (2015). What is the relationship between the lexical profile of test items and performance on a standardized English proficiency test? *English for Specific Purposes*, *38*(1), 34–43. Retrieved from https://daneshyari.com/article/preview/355379.pdf. https://doi.org/10.1016/j.esp.2014.11.001.
Wesche, M., & Paribakht, T. S. (1996). Assessing L2 vocabulary knowledge: Depth versus breadth. *Canadian Modern Language Review*, *53*(1), 13–40. https://doi.org/10.3138/CMLR.53.1.13.
Wilkins, D. (1972). *Linguistics in language teaching*. Edward Arnold.
Wright, B. D. (1996). Time 1 to Time 2 (Pre-Test to Post-Test) Comparison and equating: Racking and stacking. *Rasch Measurement Transactions*, *10*(1), 478.
Wright, B. D., Linacre, J. M., Gustafson, J. E., & Martin-Löf, P. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(3), 370 Retrieved from http://www.rasch.org/rmt/rmt83b.htm.
Yashima, H. (2002). Nihonjin koukousei no goi saizu [The vocabulary size of Japanese EFL senior high school students]. *KATE Bulletin*, *16*, 29–41. https://doi.org/10.20806/katejo.16.0_29.

Publisher's Note