Modeling vocabulary size using many-faceted Rasch measurement

Trevor Holster[1] and J. W. Lake[2]
trevholster@gmail.com
1. Fukuoka University, Fukuoka
2. Fukuoka Jogakuin University, Fukuoka
https://doi.org/10.37546/JALTSIG.TEVAL26.1-1

Abstract

Research into second-language vocabulary size has suffered from inattention to psychometric issues, with ordinal-level raw scores often analyzed as if they represented ratio-level measurement. Additionally, contextual effects have been largely ignored, leading to concern over the interpretation of research findings. This study used many-faceted Rasch measurement to analyze vocabulary data from 1,872 Japanese university students. A test of word synonymy was linked to the *Vocabulary Size Test*, and the contextual variables of item position and time of administration analyzed as measurement facets. Major findings were that data-model fit was sufficient to allow local linking of different item types and contextual variables, allowing meaningful comparison of results and score gains on a scale of vocabulary size, and that item placement within a test form had a substantive effect on item difficulty.

Keywords: Vocabulary size, many-faceted Rasch measurement, test linking, guessing correction

Read (2000) provided a detailed introduction to the nature of vocabulary knowledge, a complex construct that extends beyond simply knowing dictionary definitions. This paper therefore does not attempt to address vocabulary knowledge in its entirety, but is limited to the construct of vocabulary size as operationalized by the *Vocabulary Size Test* (VST) (Beglar, 2010; Nation & Beglar, 2007). Vocabulary size, as described by Chapelle (1994), refers to the number of content words known within a particular context of use, following Dollerup et al.'s (1989) interactionalist view that our comprehension of vocabulary will vary depending on the context in which it is encountered. Chapelle (1998) noted neglect of issues of validity in second language (L2) vocabulary assessment research, a concern that was belatedly acknowledged by Schmitt et al. (2020) over two decades later. The estimation of vocabulary size is one such area of concern.

Intuitively, estimates of vocabulary size should be invariant between repeated test administrations, but this invariance will not hold between raw percentage scores from vocabulary test forms sampling from different frequency ranges. This is because of the unavoidable presence of idiosyncratic words whose difficulty level does not align with their frequency within the target corpus, an effect seen quite dramatically in Beglar's (2010) results. One cause of such idiosyncratic items would be the inclusion of cognates between the students' L1 and L2 in a test, an issue that Read (1988) warned was a threat to the interpretation of test scores if students from different language backgrounds are tested together due to differential item functioning (DIF) of words that are cognates with the L1 of one group but not of other groups. DIF concerns also arise over differential patterns of language exposure or study between different subgroups of students sharing the same L1. Such an effect was reported by Santellices and Wilson (2010), where the different language backgrounds of Black and White American students resulted in DIF on SAT language questions. This DIF is inevitable whenever students from different language backgrounds are tested together so cannot be resolved by changing the corpus used to estimate word frequency. Researchers investigating the relationship between word frequency and test item difficulty must therefore recognize idiosyncratic knowledge as an inescapable feature of language rather than something that can be addressed through corpus sampling design.

The consequence of idiosyncratic knowledge is that vocabulary size estimates will vary between test forms sampling different frequency ranges. To illustrate the problem, if a large number of students were tested on a 5K VST form, with 10 items from each of the first 1000-word bands, students with scores of 25 out of 50 would have an estimated vocabulary size of 2,500 words, but some of those students would also know some lower frequency words. Thus, the vocabulary size estimate of 2,500 words underestimates their vocabulary size, which would be expected to increase if a 14K test was administered, and increase again if a 20K form were administered. Beglar (2010), for instance, administered the lowest group in his study a 40-item 4K form and the middle level group an 80-item 8K version so the vocabulary size estimates of those groups would have been underestimated relative to students taking 14K or 20 K test forms. In principle, any test of a practical length will always underestimate vocabulary size due to the idiosyncratic nature of vocabulary knowledge, so the theorized invariance of vocabulary size cannot be expected to be observed in practice.

Guessing effects

The underestimation due to idiosyncratic knowledge is unrelated to whether random or informed guessing is present or absent, making linking of scores between different versions of the VST necessary if they are to be compared. However, the linking of test forms does require consideration of the effects of random guessing because the 4-option selected response (S.R.) format of the VST means that random guessing alone would give an expected average score of 35 on the original 14K form. This corresponds to a vocabulary size estimate of 3,500 words if Nation's (2012) advice to multiply raw scores by a scaling constant of 100 to obtain a vocabulary size estimate. Nation emphasized that an "I don't know" option was not included in the VST because "the learners should make informed guesses" (2012, p. 4), advice that renders invalid the protocol of estimating vocabulary size by use of a simple scaling constant and has other important implications for the construct definition of vocabulary size.

As Holster and Lake (2016) discussed, guessing correction is a well-established procedure in interpreting scores from multiple-choice tests (Frary, 1988, for example). An important reason for advising students to guess unknown test items is that many students may be confused by technical explanations of guessing strategies, favoring students who adopt optimal strategies over those who do not (Budescu & Bar-Hillel, 1993). Nation's (2012) use of the term "informed guesses" reflects that knowledge is not a simple dichotomy between complete knowledge and zero knowledge. Human knowledge of anything is incomplete, so responses to test questions always reflect partial knowledge, with the probability of success increasing with a candidate's level of partial knowledge. Further, as Thissen et al. (1989) pointed out, distractors are an integral part of a test item, so S.R. vocabulary test scores represent knowledge of the stem, key, and distractors; such test items are not intended to test knowledge of a single target word. In a 4-option S.R. format, eliminating one distractor when the key is unknown increases the probability of guessing the correct response from 25% to 33%, eliminating two distractors increases it to 50%, and eliminating three distractors results in a 100% probability of success. Distractor elimination is thus a construct-relevant display of knowledge and a correct response cannot be assumed to represent knowledge of the item key. Rather than confusing students with technical explanations about when it is appropriate or inappropriate to employ informed guessing, advice which test-wise students are likely to ignore anyway, it is therefore preferable to just instruct them to guess randomly from any response options that they cannot eliminate. The use of an S.R. format coupled with Nation's (2012) endorsement of informed guessing thus has two important consequences for the validity of the VST: i) guessing correction is required to convert raw scores to vocabulary size estimates; ii) the construct is inherently limited to an estimate of how many words a student understands, not whether they understand any specific word included in the test.

Measurement invariance and test linking

The linking and rescaling of different test forms to a reference form requires measurement invariance, meaning that relative person ability is unaffected by the sample of test items used and relative item difficulty is unaffected by the sample of persons tested (Engelhard, 2013). The Rasch measurement model (Rasch, 1960; Wright & Stone, 1979) achieves this invariance through the conversion of raw percentage scores to log-odds unit, or *logits.* This logit conversion is required because raw percentage scores from different test forms or scoring protocols do not provide invariant measurement. Measurement invariance also makes Rasch generated logits useful for measuring the effect size of learning gains calculated through the subtraction of pre-test scores from post-test scores. These subtractive comparisons require an *interval level* measurement scale, following Stevens' (1946) hierarchy of measurement scales, a property of Rasch logit measures but not of raw percentage scores.

Crucially, although Beglar (2010) used Rasch analysis in his validation study, the construct of vocabulary size was defined in terms of raw scores, giving the practical advantage that classroom teachers can administer, score, and interpret the VST without needing any expertise in psychometric analysis. Under Rasch analysis, for students taking the same test form, the same raw score maps to the same logit measure. This means that all students who achieve 50% on the same test form are estimated as having the same ability, for example. This one-to-one correspondence of raw score to vocabulary size is a fundamental assumption of Nation and Beglar's (2007) definition of vocabulary size, a definition that requires each word to carry equal weight. This condition is satisfied by the Rasch model but not by more complex IRT models whose fundamental rationales are that items should not carry equal weighting (DeMars, 2010). This property of the Rasch model also simplifies linking of alternate test forms to a reference form through the use of score tables that criterion reference raw scores from each alternate form to vocabulary sizes estimates from the reference form.

Contextual effects and many-faceted Rasch measurement

Henning (1992) distinguished between *psychological* and *psychometric* unidimensionality. The former means that scores are interpretable in terms of the intended construct and the latter reflects homogeneity of item variances. Chapelle (1998) identified *trait, behaviorist,* and *interactionalist* models of knowledge. Trait models attribute knowledge to learner factors without specification of context. Nation (2012), for example, asserted that the VST tested vocabulary without context. Behaviorist views hold that knowledge can only be defined with reference to the context of use, while interactionalist models hold that both traits and contexts of use must be defined. Investigations of the effect of task type on item difficulty implicitly assume an interactionalist model of knowledge, where item difficulty derives from interaction of the word (the trait component) with the task type (the context), echoing Oller's argument that "knowing a word is knowing how to use it in a meaningful context" (1979, p. 189). One concern that arises here is that some VST item stems used a definitional sentence, requiring syntactic parsing, while others used a single word synonym. Nation recognized that "the difficult grammar of English definitions" (2012, p. 4) was problematic for the construct definition, so recommended the use of bilingual test forms. However, bilingual forms are problematic for any groups of students with varied L1s because they will not function as parallel test forms unless all the items, including the distractors, function identically in every test form. The mixing of multi-word and single word items in the VST thus raises questions about the psychological unidimensionality of the two item types, but Beglar (2010) reported sufficient psychometric unidimensionality that any sub-dimension related to syntactic parsing was not of major concern. Nation's (2012) advocacy of bilingual test forms is thus both unnecessary and undesirable if Beglar's (2010) analysis is accepted.

A further issue relating to unidimensionality concerns nuisance dimensions; small sub-dimensions that manifest differently in different contexts or at different times (Luecht & Ackerman, 2018). For example, Japanese language proficiency would constitute a nuisance dimension if foreign students in Japan were administered a bilingual vocabulary test that tested the synonymy of English and Japanese words. Test scores would represent a multidimensional trait of knowledge of both English and Japanese rather than a unidimensional trait of English knowledge. Nuisance dimensions are of particular concern for longitudinal studies, where pre-test and post-test scores may represent different composite constructs because of context related changes in nuisance dimensions. For example, foreign students studying at Japanese universities are often required to take both English and Japanese language classes so score gains on a bilingual vocabulary test administered at the beginning and end of a semester might represent improved Japanese proficiency as well as improved English proficiency.

A common approach to longitudinal datasets is to "rack" the data so that the pre-test and post-test responses for each item are analyzed as two separate items within a single analysis (Wright, 2003, p. 905). Tests of dimensionality and data-model fit can then be performed to investigate possible nuisance dimensions. However, this procedure violates the requirement of local item independence, so many-faceted Rasch measurement (MFRM) (Linacre, 1994) addresses this by allowing contextual variables to be modeled as measurement *facets* in addition to the familiar facets of *items* and *persons.* Rather than treating multiple responses by the same person to the same item as representing two items, MFMR treats them as one person responding to one item in different contexts. Although commonly used to model the effect of human raters in performance tests, such as in McNamara's (1996) seminal work, MFRM is applicable to any dataset where each student can interact with each item under different contextual conditions.

Background to this study

This study reanalyzed data collected at two Japanese universities, a public women's university and a private co-educational university. The introduction of a new Academic English Program (AEP) at the women's university led to disappointment when the expected TOEFL score improvements were not achieved, leading to curriculum reform and placement test development projects. One major issue was determining a suitable lexical level for both instructional and assessment content, consistent with Nation's (2012) recommended uses of the VST. It was also desirable to gather longitudinal data to determine whether the revised curriculum led to the intended improvements in language ability. A similar situation occurred at the co-educational university, where a proposed new language program raised questions about an appropriate level of content. The existing official course objectives assumed a level of proficiency that both Japanese and non-Japanese teachers considered unrealistic so criterion-referenced measures of the range of student ability were desirable to make recommendations for the proposed new program. To avoid detailed technical explanation about the use of logits and Rasch analysis, results were rescaled to vocabulary size estimates. The vocabulary sections of classroom and semester final tests used an item format based on the *Test of Vocabulary Synonymy* (TVS) used by Holster and Lake (2016), so these tests were linked and rescaled

to the VST vocabulary size scale. As the test linking was conducted through Rasch analysis of concurrently administered items from the VST and TVS item banks, the essential research questions revolve around whether the requirements of the Rasch model were satisfied and claims of measurement invariance warranted.

Research questions

RQ1: Do rescaled logit scores and guessing-corrected raw scores provide invariant estimates of vocabulary size?

RQ2: Do the TVS and VST items measure a unidimensional construct?

RQ3: Are item difficulties from longitudinal datasets sufficiently invariant to support measurement of learning gains?

RQ4: Does item position within a test form affect measurement invariance?

Method

Participants

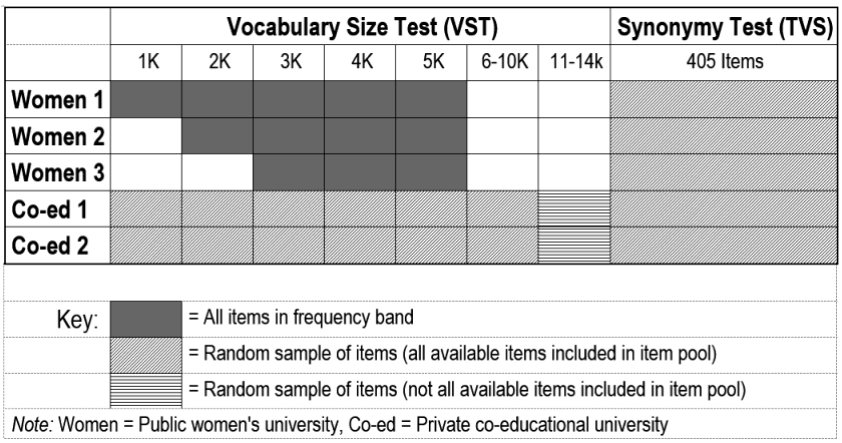
Tests were administered to a convenience sampling of 1,872 first-year students (typically 18 or 19 years old) taking compulsory English classes at a public women's university and a private co-educational university. Students came from a range of departments at each institution. Consistent with Beglar's (2010) sample of Japanese undergraduate students, students predominantly had vocabulary sizes below the 5K level.

Instruments

The 4-option VST provided a reference form for test linking and rescaling. A 50-item VST test was used at the women's university, limited to items in the 1K to 5K range, reflecting the typically low vocabulary sizes of Japanese students also noted by Beglar (2010). Some students were tested on all 50 items while others were administered 30-item or 40-item tests due to constraints on class time. Fifteen 50-item VST forms were created for use at the co-educational university, using items from the 1K to 14K range. Microsoft Excel was used to randomize item placement but biased to favor high-frequency words and to place them earlier in the test form. This algorithm resulted in inclusion of all items from the 1K to 10K bands, but gaps in the 11K to 14K range. The test administration pattern is shown schematically in Figure 1.

Figure 1

Test administration pattern



Note. Multiple test forms were created for both the VST and TVS, with quasi-random item placement. Students were administered 30 to 50 VST items and 108 TVS items as a pre-test and 108 TVS items as a post-test. Some students took a 54 item TVS test as a mandated final exam.

The 5-option TVS items used single-word synonyms rather than the definitional sentences of the VST to eliminate the syntactic parsing that Nation (2012) reported as problematic in the VST. TVS specifications (Holster & Lake, 2016) were used to develop additional items based on classroom materials by substituting synonyms for target words in listening and reading texts. For example, students heard the following sentences in one of the dialogues from a coursebook:

A: You're working? What do you do?

B: I'm a tutor.

However, in the transcript given to students, the target word *tutor* was changed to *teacher* and students were required to highlight any such discrepancies while they listened. Typically, 10 to 15 synonyms were presented each week. These were then tested in weekly written classroom review tests and semester tests, both contributing to a significant proportion of course grades. Based on the word frequencies published by Davies and Gardner (2010), the higher frequency synonym was used as the item stem and the lower frequency synonym as the key. Four distractors were selected from Davies and Gardner's (2010) list by finding the two next more frequent and two next less frequent words of the same part of speech as the stem and key, with any potentially problematic distractors skipped in favor of the next more or less frequent word. Students were instructed to read the item stem and identify a synonym from the five answer choices.

A sample test item is:

Teacher

A) Fee B) Tutor C) Sense D) Market E) Nation

Procedure

The VST was administered at the beginning of the first semester at the women's university to calibrate placement tests and at the end of the semester at the co-educational university to calibrate achievement tests. The original 50-item TVS formed the vocabulary section of the placement test used at the women's university. The entire placement test was administered again at the end of the semester, one week before the semester final test. At the co-educational university, new textbook derived TVS items were administered as weekly review tests, with the primary intention of rewarding students for being engaged in class and reviewing class materials each week. Weekly review test data was not included in this study. In order to familiarize students with the item formats used in the weekly review tests, a practice test was administered in the first or second week of class, with TVS items forming the vocabulary section. TVS items were administered again as the vocabulary section of the semester final test in the final week of class. Some courses were required by the co-educational institution to be administered an official final exam, typically two weeks after the end of the 15-week semester. These official final exams were administered by university staff under test conditions, were limited to one side of an A4 sheet of paper for administrative convenience, and could not include listening tasks because students taking different courses with different teachers were combined within each test room. These constraints limited the final exam to 54 TVS items, in contrast to the classroom administrations which contained 108 items on two A4 pages. Each TVS form was generated from an Excel workbook that randomized item placement. The 658 students in these courses thus took three administrations of the TVS. Test forms were scanned using Remark Office OMR version 8.4 and data analyzed with Winsteps version 4.0.0 and Facets version 3.8.04 using the default settings for the Rasch dichotomous model.

Results

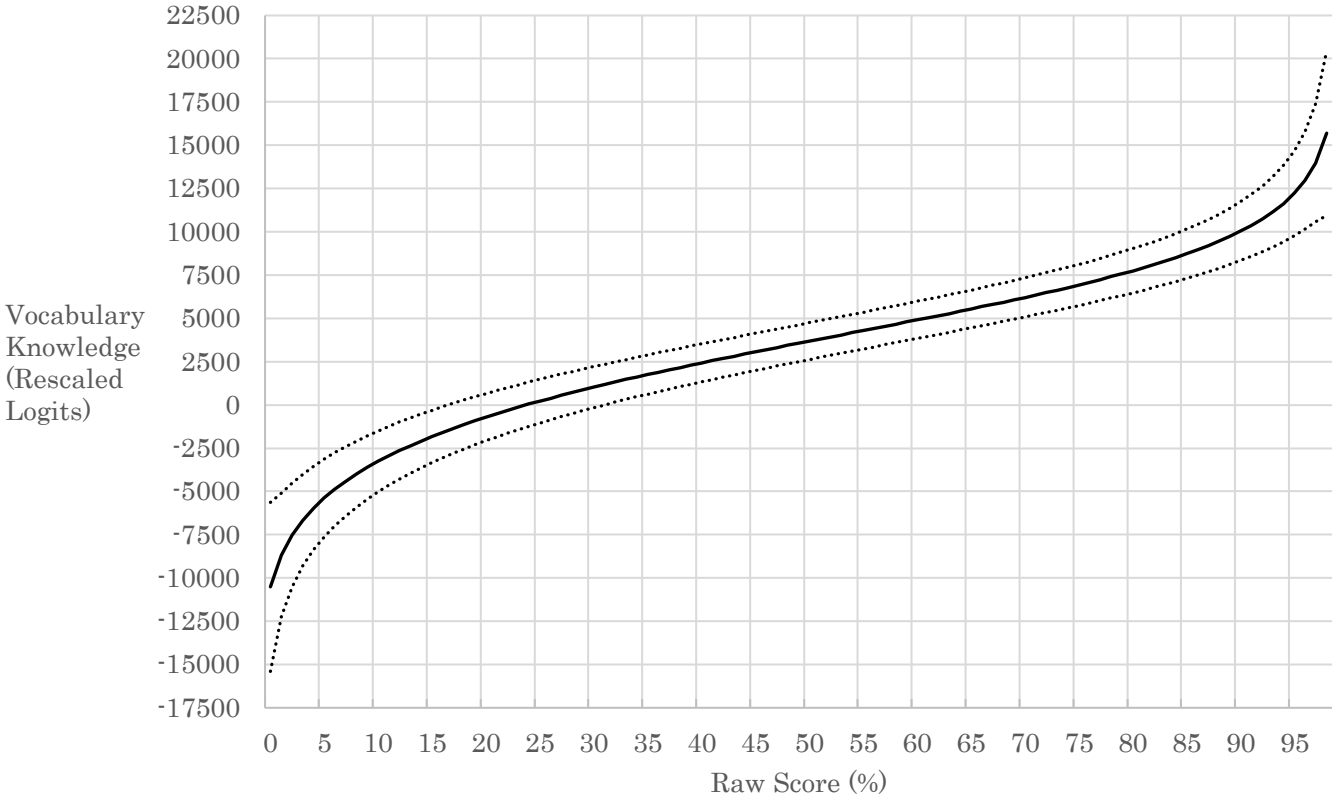
Rescaling logit measures to vocabulary size

The first stage of analysis focused on rescaling item difficulty to a vocabulary size scale. Winsteps was used to produce a score table matching raw scores to logit measures for a VST reference form containing all 100 items in the 1K to 10K bands (hereafter VST10). Scaling of logit measures to vocabulary size was based on the following assumptions: 1) Mean item difficulty should be approximately 3,333 words, equaling the guessing-corrected vocabulary size of a person scoring 50%; 2) 1 logit should be scaled to 2,300 words, giving a 4 logit range from -1 logit (27%) to 3 logits (95%) corresponding to guessing-corrected vocabulary sizes of 67 words to 9,333 words. In practice, the relationship between logit measures and raw percentages was found to be approximately linear from raw scores of 25% to 80%, but increasingly non-linear beyond that. Empirical results showed that scaling 1 logit to 2,400 words, with mean difficulty of 3,300 words, produced a score

table with close approximations between guessing-corrected raw scores and rescaled logits between 25% and 80%. These results are shown in Figure 2, with raw scores of 25% and 80% respectively producing VST sizes of approximately zero and 7,500 words, very close to the expected values. Rescaled logit scores are thus usefully invariant with vocabulary size estimates within the range of 0 to 7,500 words, allowing learning gains for students within this range to be expressed in terms of word families known.

Figure 2

Raw VST10 score versus vocabulary knowledge

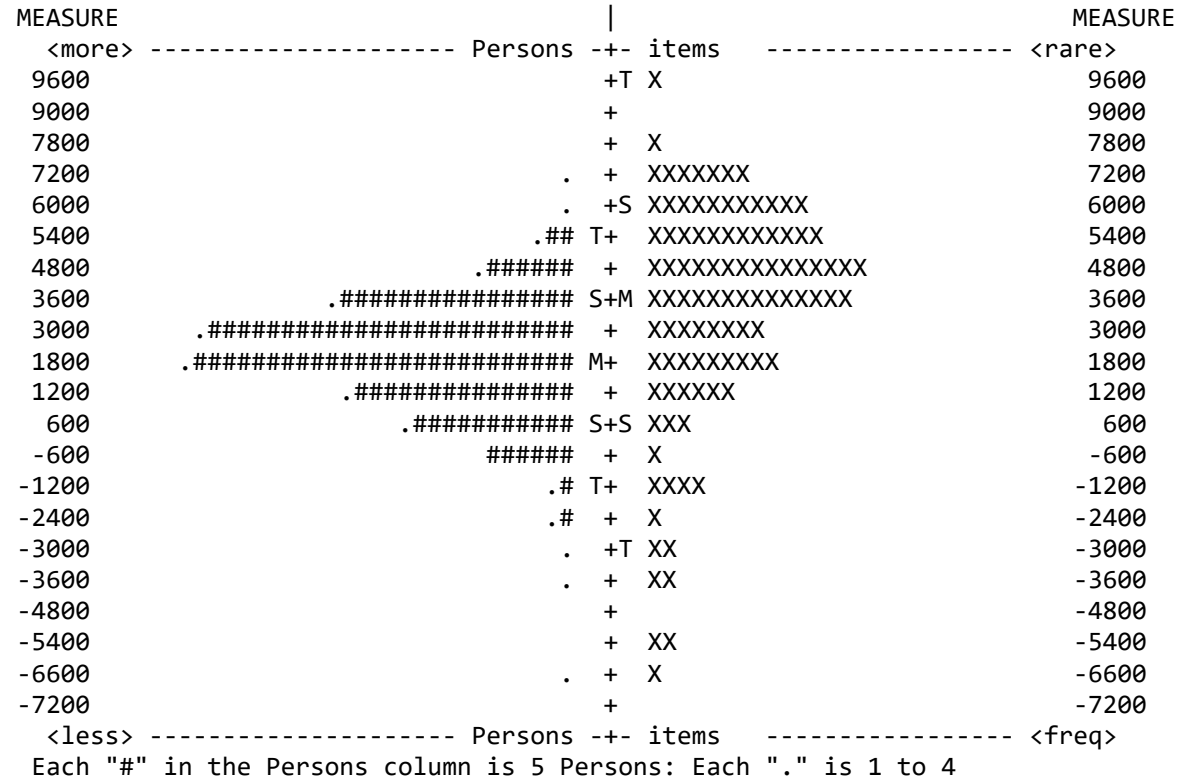


Note. The upper and lower dashed lines show the 95% confidence intervals. The vertical axis shows logit scores rescaled to estimated vocabulary size, with 1 logit = 2 400 words.

Figure 2 also shows 95% confidence intervals, typically spanning a range of about 2,500 words, evidence that the VST10 is unsuitable for measurement of individual student learning unless very large learning gains have been achieved. This is not a reflection on the VST's validity as a general measure of vocabulary sizes, it simply reflects that it was not intended to be precise enough to measure small learning gains by individual students. Figure 3, mapping person ability against item difficulty, confirms this, with mean person ability of 2,113 words and a standard deviation of 1,654 words. The confidence interval is thus about 1.5 standard deviations of this sample of persons, meaning that more items are required to reduce the measurement error. Figure 3 also shows many items that were far too difficult for any student, so measurement quality would be improved by removing items above the 5K level and replacing them with 1K and 2K items. Additionally, although the VST sampled equally across frequency bands, the distribution of item difficulties did not reflect this, confirming the presence of many idiosyncratic items observed by Beglar (2010). An important implication of this finding is that the suitability of items for many classroom testing purposes will be determined by the empirically derived logit difficulty rather than the BNC frequency band, whereas researchers may prefer to select items based on frequency to simplify estimation of vocabulary size. Comparison with Figure 2 shows that the highest density of item difficulty aligns with the range of vocabulary sizes that show the most linear relationship with logit measures. Clearly, many more items were required in the 1K and 2K bands and many fewer items above the 5K level were needed, a limitation the TVS items were developed to address.

Figure 3

Person-item map of VST10 results



Note. Persons and items are mapped against a common scale of vocabulary knowledge expressed as words known. Higher placement on the map indicates higher person ability or higher item difficulty.

Linking VST and TVS test forms

RQ2 addressed the unidimensionality of the VST and TVS items, a fundamental requirement for linking the two tests. Table 1 shows principal components analysis of residuals (PCAR) from the combined VST and TVS dataset. The Rasch dimension explained 35.6% of total variance, exceeding Reckase's (1979) guideline of a minimum of 20% variance explained, with the largest subdimension accounting for 0.6% of variance. These results should be treated cautiously due to the low data density but are consistent with the TVS and VST measuring a unidimensional construct.

Table 1

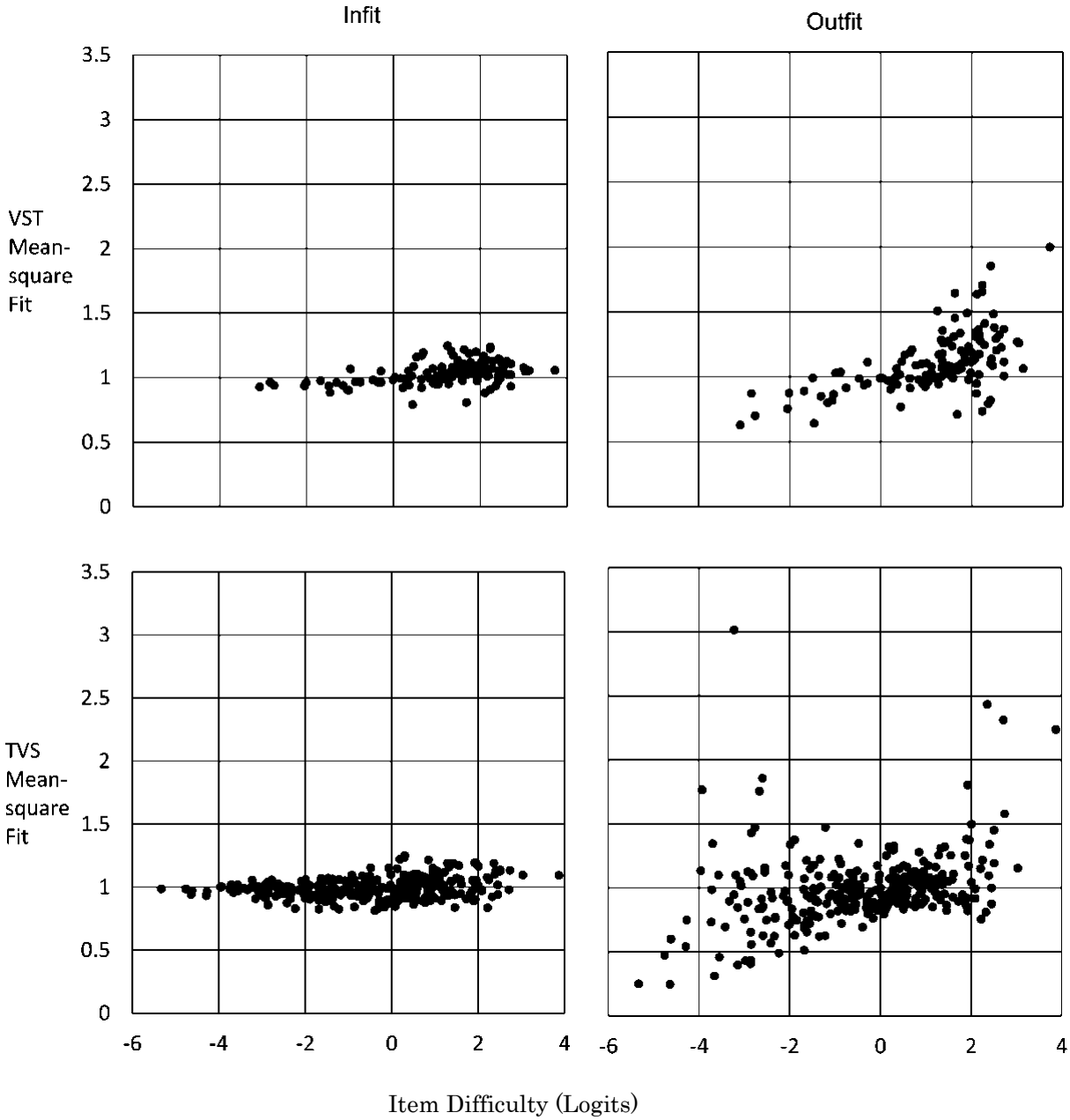
Variance explained by measures

Variance		Eigenvalue	Observed %	Expected%
Total:		655.7	100.0%	100.0%
Rasch:	Measures	233.7	35.6%	35.6%
	Persons	61.7	9.4%	9.4%
	Items	172.0	26.2%	26.2%
Unexplained:	Total	422.0	64.4%	64.4%
	1st contrast	4.2	0.6%	
	2nd contrast	3.8	0.6%	
	3rd contrast	3.5	0.5%	

Dimensionality can also be investigated by checking for systematic patterns in mean-square fit statistics for the VST items and TVS items, as shown in Figure 4. The VST items, shown in the two upper panels, were more difficult on average than the TVS items, shown in the two lower panels, with very few VST items below 0.00 logits compared with many TVS items. Mean-square infit, reflecting information weighted responses, is shown in the two left-hand panels, with all values comfortably below Linacre's (2009) 1.50 rule-of-thumb guideline for concern. Mean-square outfit, reflecting unweighted response, is shown in the two right-hand panels, with difficult items tending to misfit for both item types. Although easy items of both types showed a tendency to overfit, this pattern was very pronounced for the VST items. The easy TVS items were somewhat less consistent, with an extreme range of outfit including some highly overfitting items and some highly misfitting items, but there were insufficient easy VST items to draw firm conclusions. The information-weighted infit mean-square value is a crucial indicator of measurement quality (Linacre, 2009), and all items performed acceptably. The outfit mean-square value indicates unexpected outlying responses, with 17 items exceeding the 1.50 threshold of concern, including four TVS items with values exceeding 2.0. In a battery of 422 items, 17 misfitting items constitutes about 4% of total items and all the misfitting items were at the extremes of the measurement range so these do not pose a substantive threat to test linking.

Figure 4

Mean-square item fit



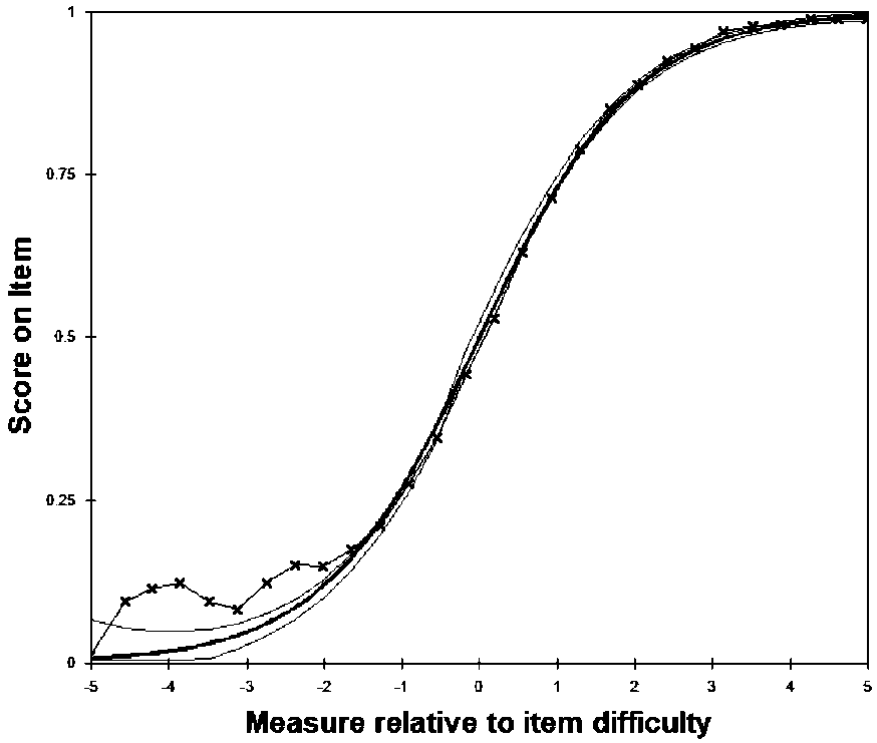
Note. The two upper panels show VST items, the two lower panels show TVS items. The horizontal scale shows item difficulty in logits, the vertical scale shows mean-square fit, with infit shown in the two left-hand panels and outfit in the two right-hand panels. Mean-square values lower than 1.50 indicate acceptable item functioning.

Figure 5 shows the modelled and empirical test characteristic curves for the combined VST and TVS analysis. The empirical results closely match the Rasch model above a probability of success of approximately 20%, below which the results misfitted the model. These results are consistent with low-ability persons succeeding on difficult items through random guessing, with odds of random guessing of 25% on the VST items and 20% on the TVS items. This illustrates the importance of S.R. test items being well matched to the ability of test takers. Figure 5 supports the view that the misfit associated with difficult items in Figure 4 arose due to random guessing but does not resolve the cause of the misfit of easy items. Item dependency was analyzed through correlations between standardized item residuals, the standard Rasch procedure (Aryadoust et al., 2021), with the 10 largest values shown in Table 2. One pair of items showed a correlation of .76, meaning

that shared variance exceeded 50%, the level at which inter-item dependency exceeds random variance (Linacre, 2020). These two items tested the synonymy of *good/nice* and *child/youngster*, the two items having no obvious semantic relationship. One other pair of items showed a correlation of .66, indicating 44% shared variance. These items tested *talk/speak* and *seafood/fish*, which also have no obvious semantic connection. Given the lack of semantic connection, these four items do not threaten the requirement of independence in a test battery of over 400 items.

Figure 5

Empirical versus modelled test characteristic curve for combined VST and TVS items



Note. The solid central line shows the modelled expectation of success for persons of different ability, with each X showing observed probabilities and the upper and lower solid lines showing confidence intervals.

Table 2

Standardized residual correlations

Item Number	*Item Number*	*Correlation*
196	400	.76
151	411	.66
130	368	.52
122	126	.52
130	403	.51
343	365	.45
126	130	.41
71	89	.40
256	262	.40
183	267	.40
368	403	.39
106	119	.39
115	120	.37

Table 3 shows fit statistics for the 17 items with mean-square values exceeding 1.50. All had low or negative point-measure correlations, indicating an inability to discriminate between high and low-proficiency persons. Twelve of the items were

extremely difficult, with logit values exceeding 1.90, corresponding to vocabulary sizes exceeding 7,500, and raw scores within the range of random guessing. Three misfitting items were extremely easy, with five incorrect responses or fewer, meaning that a single careless response would be sufficient to cause misfit. The remaining two misfitting items were the VST items *Gimmick* and *Upbeat*, with respectively 11/38 and 18/50 correct responses and outfit mean-square values of 1.64 and 1.51. Table 3 provides further evidence that misfit resulted from a very small number of responses so a larger sample of persons would likely have resulted in better fit (and point-measure correlations). This small number of misfitting responses does not pose a substantive threat to test linking because the large number of well-fitting responses included all the items matched to the range of person ability. These well-matched items provide much more information than the outlying items, reflected in the much lower levels of infit than outfit. In response to RQ2, PCAR analysis and data-model fit indicated sufficient unidimensionality to map VST and TVS items into a common measurement scale for the purpose of measuring score gains across a semester of instruction.

Table 3

Most misfitting items

Item	Freq.	Synonyms					Infit		Outfit		Pt-M
Number	Level	Tested	Count	Score	Logits	*SE*	*MS*	*ZStd*	*MS*	*ZStd*	Corr
323	S3:	Help-Assist	115	112	-3.23	0.59	1.05	0.28	3.01	2.03	-.08
417	S5:	Potential-Implied	118	24	2.36	0.24	1.19	1.39	2.43	5.04	-.10
343	S3:	Gradually-Slowly	114	16	2.71	0.28	0.98	-0.05	2.31	3.58	.11
357	S4:	Compose-Write	231	12	3.87	0.30	1.09	0.44	2.23	2.67	-.11
49	V5:	Fracture-Break	207	12	3.72	0.30	1.06	0.31	2.00	2.46	-.08
80	V8:	Mumble-Speak	88	15	2.42	0.29	1.15	0.83	1.85	2.93	-.16
183	S1:	Girl-Daughter	112	107	-2.60	0.46	1.02	0.17	1.85	1.40	.00
365	S4:	Genuine-Actual	114	29	1.92	0.22	1.19	1.74	1.80	3.95	-.04
144	S1:	School-University	335	331	-3.93	0.50	1.00	0.17	1.76	1.22	.03
148	S1:	Look-Watch	114	109	-2.67	0.47	1.00	0.13	1.75	1.29	.06
117	V12:	Coven-Society	13	2	2.23	0.79	1.23	0.58	1.70	1.09	-.22
108	V11:	Hutch-Cage	52	10	2.23	0.36	1.24	1.12	1.66	2.12	-.30
68	V7:	Gimmick-Trick	38	11	1.63	0.37	1.08	0.57	1.64	2.47	-.01
77	V8:	Locust-Insect	75	15	2.12	0.30	1.13	0.79	1.63	2.45	-.11
319	S3:	Column-Tower	230	34	2.74	0.19	1.13	1.06	1.58	2.66	.02
95	V10:	Upbeat-Good	50	18	1.25	0.31	1.25	2.20	1.51	2.94	-.12
305	S3:	Purchase-Invest in	113	25	2.01	0.24	1.16	1.27	1.50	2.49	-.02

Note: VST items are coded "V" followed by frequency band. TVS items are coded "S" followed by frequency band. Count = number of responses recorded; Score = number of correct responses; Pt-M Corr = Point-measure correlation.

Linking longitudinal data using MFRM

Measuring learning gains through pre-tests and post-tests introduces a potential problem of multi-dimensionality due to nuisance dimensions, which may not be detected by tests of unidimensionality commonly used in IRT analysis (DeMars, 2010). MFRM allows time of administration to be isolated as a separate measurement facet and fit statistics to be analyzed for evidence of measurement distortion due to contextual effects. Longitudinal data was analyzed using a 4-faceted model using Facets version 3.80.0, with the facets of *Time* and *Position* added to the usual facets of *Persons* and *Items. Time* refers to the time of administration of the test; the beginning of the course (Week 1), the final class (Week 15), or during the official exam period (Final Exam). *Position* refers to the location of the item in the test form, ranging from 1 (the first item) to 108 (the final item). Responses from all 1,872 persons were used to calibrate the TVS items to the VST10 scale, including 24 VST items from the 11K to 14K bands, giving 529 items in total. This calibration was achieved through concurrent equating, with mean item difficulty adjusted empirically so the average difficulty of the VST items remained constant. Item difficulties from the combined analysis were then compared with those from the VST and TVS datasets analyzed in isolation. Summary statistics are shown in Table 4, with the mean of all items found to be -366 words and the TVS items to be -1737

words. This range represented a difference between the mean VST10 and TVS item of 2.10 logits, or 5,036 words on the VST10 scale.

Table 4

Summary statistics of VST and TVS items

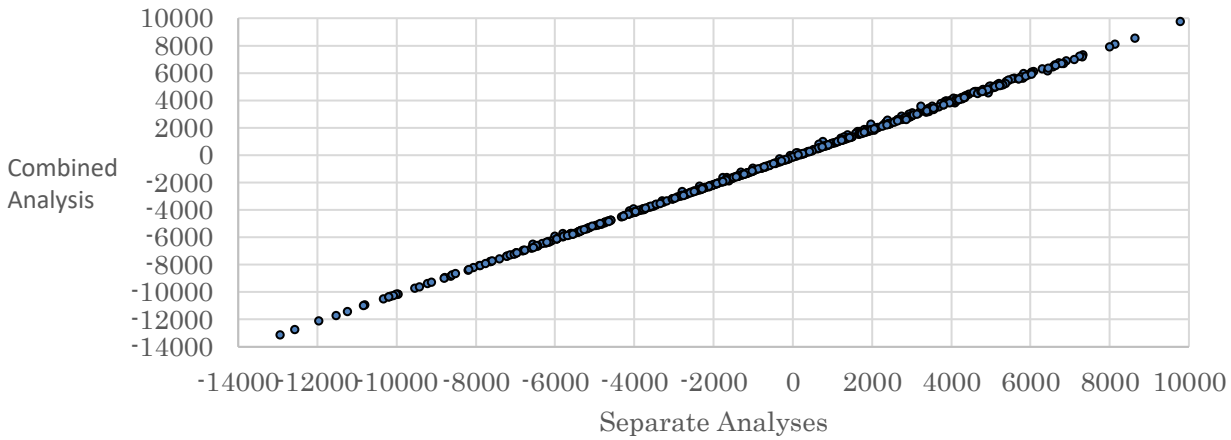
Item			Difficulty		Mean	Fair		Infit	Outfit	Pt-M	Item
Subset		*n*	(Words)	*SE*	Score	Ave	Count	*MS*	*MS*	Corr	Rel
All:	*M*	529	-366	598	0.64	0.61	549.2	1.01	1.02	.25	.96
	S.D.		4436	650	0.27	0.30	562.4	0.09	0.29	.16	
VST:	*M*	124	3299	519	0.41	0.39	236.4	1.01	1.03	.21	.97
	S.D.		3156	248	0.23	0.24	183.7	0.07	0.17	.15	
TVS:	*M*	405	-1737	571	0.72	0.70	657.3	0.99	0.98	.28	.95
	S.D.		4055	698	0.24	0.25	594.1	0.09	0.29	.14	

Note: Count = number of responses recorded; *Mean Score* = proportion of correct responses: *Fair Ave* = Probable mean score if all persons attempted all items; *Pt-M Corr* = Point-measure correlation; *Item Rel* = Reliability of item separation.

Figure 6 compares item difficulties for the combined and separate analyses of VST and TVS items, with deviations from the linear trendline much smaller than the typical measurement errors of 500 words shown in Table 4. Item reliability of all three analyses exceeded .95, indicating a stable hierarchy of item difficulty. The mean score column in Table 4 shows the observed average score, while the fair-average column shows the expected score if all students had taken all items. Clearly, the TVS items were much easier than the VST items, with respective fair-average scores of 70% versus 39%. This is consistent with the TVS items being targeted at the 5K frequency band and lower. Noteworthy is that the TVS items were slightly overfitting on average and had a higher point-measure correlation at .28 compared with .21 for the VST items, reflecting the better match of item difficulty to student ability. These results provide evidence that the combination of cross-sectional VST data and longitudinal TVS data was not a threat to measurement invariance, addressing RQ3.

Figure 6

Item difficulty for combined analysis versus separate analyses



Note. All 529 items were analyzed together for the combined analysis. For the separate analyses, the mean item difficulty of each subset of items was set to the value obtained from the combined analysis.

Measuring learning gains

The VST10 anchored item difficulties were then used to analyze learning gains for the 658 students at the co-educational university who were administered the TVS as an official final exam. This calibration allowed comparison between the classroom test in the final week with the final exam one or two weeks later. TVS items with respective infit and outfit mean-square fit values below 1.20 and 1.30 were anchored to the VST10 scale, with less-fitting items unanchored to avoid measurement distortion. Learning gains were then measured against this anchored scale, shown in Figure 7, with the estimated VST10 vocabulary size shown on the left. Students showed a substantive gain between Week 1 and Week 15, and also between Week 15 and the final exam. Summary statistics for all four facets are shown in Table 5. Although average mean-square statistics were close to the expected value of 1.00, *Persons* and *Items* had outfit standard deviations of 0.35 and 0.32 respectively, consistent with the misfit to outlying responses discussed earlier.

Table 5

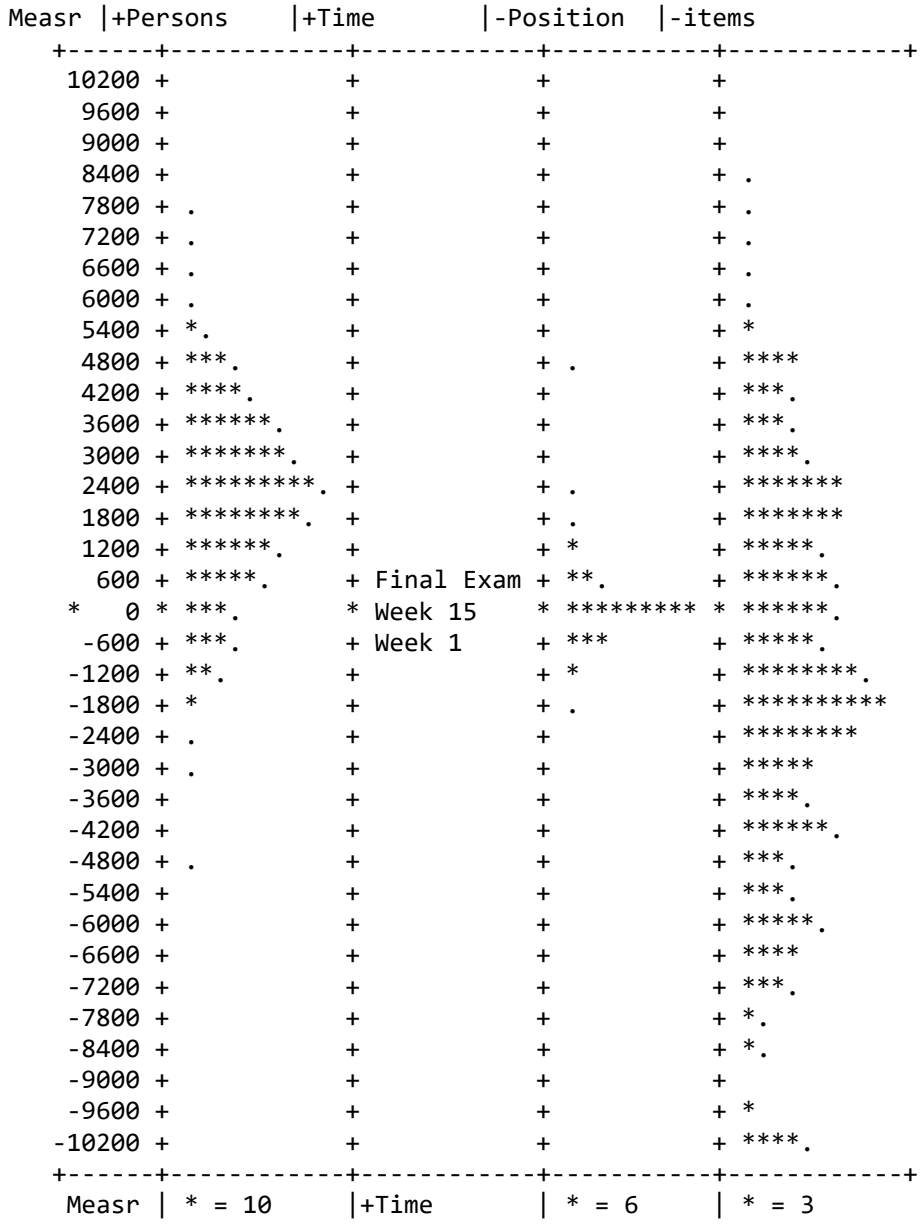
Summary statistics of measurement facets

Facet		*N*	Rel	Sep	Count	Mean Score	Fair Ave	Vocab Size	*SE*	Infit *MS*	Outfit *MS*
Persons:	*M*	658	.95	4.48	314.60	.72	.79	2088	395	1.01	1.01
	SD				94.70	.11	.12	1860	92	0.11	0.35
Time:	*M*	3	1.00	28.56	69013.00	.74	.81	0	28	1.01	0.98
	SD				30320.10	.07	.05	823	10	0.02	0.08
Position:	*M*	108	.96	5.20	1917.00	.72	.81	0	154	1.01	0.99
	SD				76.50	.11	.06	824	24	0.05	0.14
Items:	*M*	380	.95	4.16	544.80	.71	.73	-1518	594	0.98	0.97
	SD				484.20	.24	.23	3968	714	0.17	0.32

Note: Count = number of responses recorded; *Mean Score* = proportion of correct responses: *Fair Ave* = Probable mean score if all persons attempted all items.

Figure 7

TVS facets measurement rulers



Note. The measurement scale on the left is calibrated to VST10 vocabulary sizes. The Time facet shows gains through a 15-week semester. The position facet shows a substantively large effect of item position within the test forms.

Table 6 provides the measurement report for the *Time* facet, with a gain of 0.24 logits (570 on the VST10 scale) between Week 1 and Week 15, and a further gain of 0.44 logits (1,050 on the VST10 scale) between Week 15 and the final exam.

Table 6

Time measurement report

Time	Count	Mean Score	Fair Ave	Vocab Size	*SE*	Infit *MS*	Infit *ZStd*	Outfit *MS*	Outfit *ZStd*	Disc	Pt-M Corr
Final	34452	0.82	0.87	891.8	39.0	1.01	1.3	0.91	-2.1	1.00	.44
Week 15	81450	0.72	0.81	-161.0	22.9	0.98	-3.6	0.95	-2.6	1.03	.53
Week 1	91137	0.69	0.77	-730.7	21.1	1.03	6.7	1.06	4.0	0.95	.53
M	69013.0	0.74	0.81	0.0	27.7	1.01	1.5	0.98	-0.3		.50
SD	30320.1	0.07	0.05	823.2	9.9	0.02	5.2	0.08	3.7		.05

Pop: RMSE 28.80 S.D. 671.5 Separation 23.31 Strata 31.42 Reliability 1.00
Samp: RMSE 28.80 S.D. 822.7 Separation 28.56 Strata 38.42 Reliability 1.00
Fixed (all same) chi-square: 1387.3 d.f.: 2 significance (probability): .00
Random (normal) chi-square: 2.0 d.f.: 1 significance (probability): .16

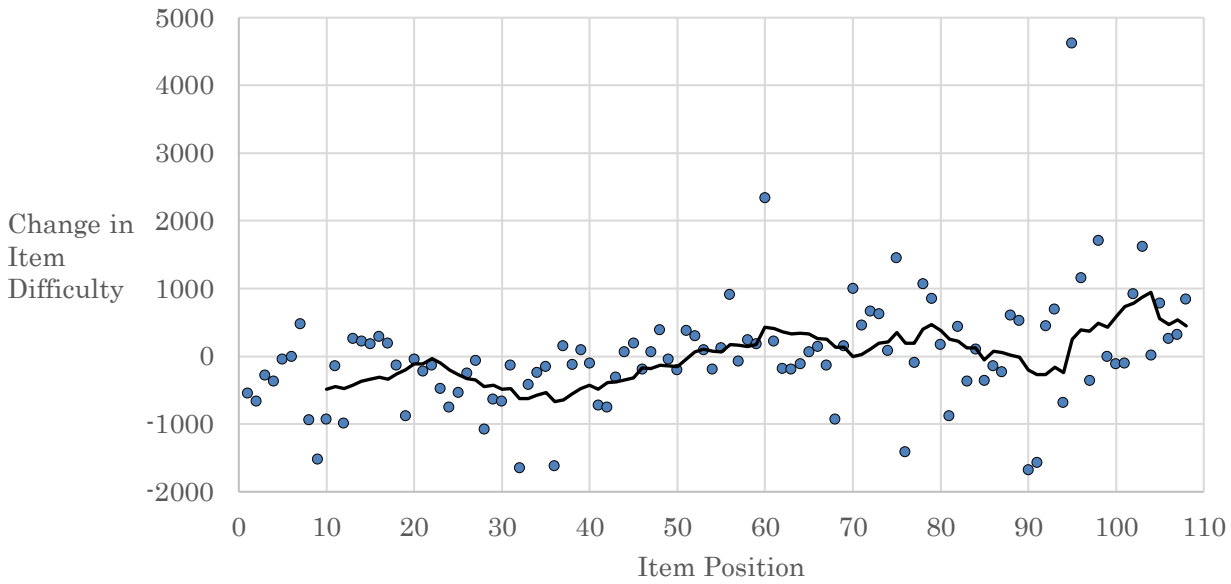
Note: Count = number of responses recorded; *Mean Score* = proportion of correct responses: *Fair Ave* = Probable mean score if all persons attempted all items; *Disc* = Discrimination; *Pt-M Corr* = Point-measure correlation.

Invariance of item position

Also of note from Figure 7 is that the *Position* facet has a substantively large range, with a standard deviation of 824 words (0.34 logits) reported in Table 5. Figure 8 shows item difficulty versus position, with each point on the solid trendline showing the mean of the preceding 10 items, allowing the general trend to be visible though the fluctuations in the data. Although the trend is quite noisy, moving an item from the beginning of a test form to the end would typically result in item difficulty increasing by the equivalent of 1,000 words. In a test such as the VST, with items ordered by frequency band, the difficulty of high-frequency items would be substantively under-estimated and the difficulty of low-frequency items overestimated, so research into the relationship between frequency and difficulty should take this effect into consideration.

Figure 8

The effect of item position on difficulty

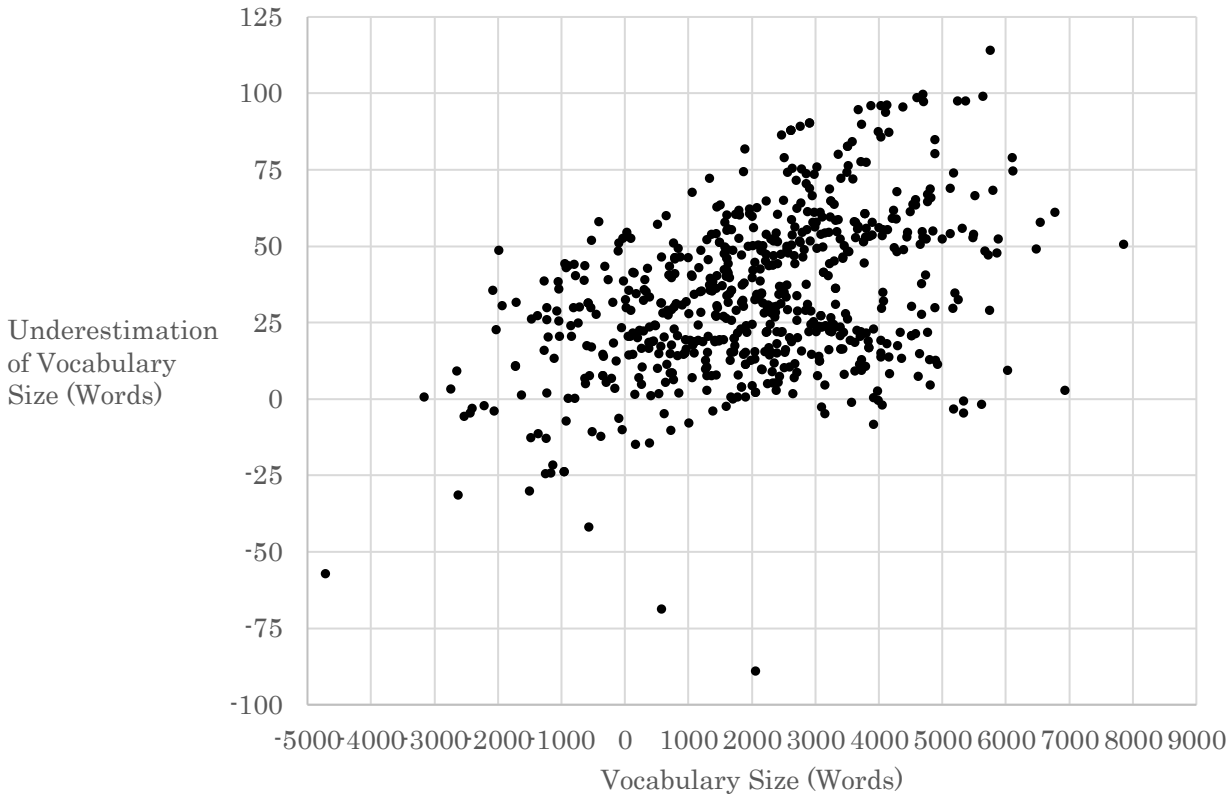


Note. The solid trendline shows the moving average of 10 items, with placement near the end of the test associated with a substantive increase in item difficulty.

In this study, however, the objective was the measurement of person ability rather than item difficulty and the use of a randomization algorithm greatly weakened the relationship between frequency and item position. Figure 9 shows the effect on vocabulary size estimates of including item position as a measurement facet, with the vertical axis scale exaggerated for emphasis. Person ability increased by an average of about 34 words when item position was included, with a greater effect on higher-ability persons. However, the substantive size of the effect is very small compared with the SE of 395 words reported in Table 5. The effect of item placement on the estimated vocabulary size of an individual student is thus an order of magnitude smaller than the measurement error, so not of concern to classroom teachers. A qualified answer to RQ4 is thus that item position within a test form has an effect too small to substantively affect the measurement of individual persons, but large enough to be of concern to researchers investigating the relationship between word frequency and item difficulty. It is therefore recommended that researchers include item position as a measurement variable.

Figure 9

The effect of including item position on estimates of vocabulary size



Note. The vertical axis shows the difference in vocabulary size after including item position as a measurement facet. Note that the scale of the two axes differs by two orders of magnitude.

Discussion

This study investigated the rescaling of classroom vocabulary tests to the VST scale using Rasch modelling. Although the VST was developed to provide a general indication of students' vocabulary sizes (Nation, 2012), it provides only 10 items per 1K frequency band. As the majority of students in this study had vocabulary sizes below the 3K level, relatively few VST items were matched to students' levels, limiting measurement precision and making it unsuitable as an instrument to measure learning gains. Synonymy test items based on textbook content were therefore developed to provide a much larger pool of items below the 5K level. However, the synonymy test was not designed to sample equally across all relevant frequency levels, a necessity for the estimation of vocabulary size using the protocol established by Nation and Beglar (2007). The VST was therefore administered as a reference test in order to calibrate the classroom tests to the VST scale using Rasch analysis.

The estimation of vocabulary size is based on raw scores which do not provide invariant interval level measurement, a major limitation on the potential use of scores. RQ1, the major research question, investigated invariance between guessing-corrected estimates of vocabulary size and logit scores, finding sufficient invariance for the purpose of test linking. This linking demonstrates how a measure of vocabulary size can be rescaled to a VST derived scale using vocabulary tests developed to different specifications.

RQ2 investigated the requirement that a unidimensional construct underlies both the VST and TVS despite the very different interpretations of the resulting scores. Unidimensionality is a requirement for the analysis of raw percentage scores as well as Rasch analysis. Although both the VST and TVS required students to match synonymous expressions, the VST included definitional phrases whereas the TVS used only single-word synonyms. The VST and TVS items were found to be consistent with a strongly unidimensional trait of vocabulary knowledge, supporting the appropriacy of test linking.

RQ3 investigated whether invariance was maintained across longitudinal data, a requirement for the measurement of learning gains. Item difficulty was found to be usefully invariant, evidence that any nuisance dimensions related to time of test administration were too small to effect test linking.

RQ4 investigated the effect of item position on difficulty. This study found a statistically significant effect whereby placement near the end of the test increased item difficulty. Although too small to be of concern for testing person ability, this effect threatens the validity of research into the relationship between word frequency and item difficulty. This is because it is standard practice to arrange test items in order of decreasing frequency, such as in the VST forms published by Nation and Beglar (2007) and Nation (2012). Although Beglar (2010) found a general tendency for lower frequency items to be more difficult, this effect was much more pronounced for very high-frequency items, with a very small effect above the 10K level. The effect of item position on difficulty observed in this study makes it plausible that Beglar's results exaggerated the effect of frequency on difficulty and that the very small increases above the 10K level actually reflected item position, not item difficulty itself. Although this is speculative given the different sampling of students and test administration protocols, it is a plausible hypothesis given the results found in this study. Researchers investigating the relationship between item difficulty and frequency need to either empirically demonstrate that item position does not affect item difficulty or use multiple test forms with randomized item placement.

Conclusion

This study demonstrated the use of Rasch modelling of vocabulary size, vocabulary size being an ordinal scale of measurement based on the protocol of assigning the same vocabulary size to students with the same raw score on the same test form. Conceptualizing vocabulary size as invariant carries the implication that scores from different test forms can be linked and calibrated to a common scale. The Rasch model provides this invariance and also supports the one-to-one mapping of raw score to vocabulary size that underlies the concept of vocabulary size. However, measurement invariance requires psychometric unidimensionality and acceptable data-model fit. This study found sufficient unidimensionality to rescale scores from a test of vocabulary synonymy and to measure gains over a semester. Raw scores are fundamentally unable to provide invariant vocabulary size estimates because of practical limits on test length. Decreasing test length by removing low-frequency items will cause underestimation of vocabulary size, a problem addressed through Rasch linking of test forms. Contextual effects are a further threat to the invariance of vocabulary size estimates, with item position shown to cause substantive misestimation of item difficulty.

References

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

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

Budescu, D., & Bar-Hillel, M. (1993). To guess or not to guess: A decision-theoretic view of formula scoring. *Journal of Educational Measurement*, *30*(4), 277-291. https://doi.org/10.1111/j.1745-3984.1993.tb00427.x

Chapelle, C. A. (1994). Are C-tests valid measures for L2 vocabulary research? *Second Language Research*, *10*(2), 157-187. https://doi.org/10.1177/026765839401000203

Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In L. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 32-70). Cambridge University Press.

Davies, M., & Gardner, D. (2010). *A frequency dictionary of contemporary American English*. Routledge.

DeMars, C. (2010). *Item response theory*. Oxford University Press.

Dollerup, C., Glahn, E., & Hansen, C. R. (1989). Vocabularies in the reading process. *International Association of Applied Linguistics Review*, *6*, 21-33.

Engelhard, G. (2013). *Invariant measurement*. Routledge.

Frary, R. B. (1988). Formula scoring of multiple-choice tests (correction for guessing). *Educational Measurement: Issues and Practice*, *7*(2), 33-38. https://doi.org/10.1111/j.1745-3992.1988.tb00434.x

Henning, G. (1992). Dimensionality and construct validity of language tests. *Language Testing*, *9*(1), 1-11. https://doi.org/10.1177/026553229200900102

Holster, T. A., & Lake, J. W. (2016). Guessing and the Rasch model. *Language Assessment Quarterly*, *13*(2), 124-141. https://doi.org/10.1080/15434303.2016.1160096

Linacre, J. M. (1994). *Many-facet Rasch measurement* (2nd ed.). MESA Press.

Linacre, J. M. (2009). *Misfit diagnosis: infit outfit mean-square standardized*. Retrieved 18 January from http://www.winsteps.com/winman/index.htm?globalfitstatistics.htm

Linacre, J. M. (2020). *Table 23.99 Largest residual correlations for items*. Retrieved 12 July from https://www.winsteps.com/winman/table23_99.htm

Luecht, R., & Ackerman, T. A. (2018). A technical note on IRT simulation studies: Dealing with truth, estimates, bbserved data, and residuals. *Educational Measurement: Issues and Practice*, *37*(3), 65-76. https://doi.org/doi:10.1111/emip.12185

McNamara, T. F. (1996). *Measuring second language performance*. Pearson Education.

Nation, I. S. P. (2012). *The vocabulary size test*. Retrieved 22 August from http://www.victoria.ac.nz/lals/about/staff/publications/paul-nation/Vocabulary-Size-Test-information-and-specifications.pdf

Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, *31*(7), 9-13. http://jalt-publications.org/files/pdf/the_language_teacher/07_2007tlt.pdf

Oller, J. W. (1979). *Language tests at school: A pragmatic approach*. Longman.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Denmark Paedogiske Institut.

Read, J. (1988). Measuring the vocabulary knowledge of second language learners. *RELC Journal*, *19*(2), 12-25. https://doi.org/10.1177/003368828801900202

Read, J. (2000). *Assessing vocabulary*. Cambridge University Press. https://doi.org/DOI: 10.1017/CBO9780511732942

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, *4*, 207-230. https://doi.org/10.2307/1164671

Santellices, M. V., & Wilson, M. (2010). Unfair treatment?: The case of Freedle, the SAT, and the standardization approach to differential item functioning. *Harvard Educational Review*, *80*(1), 106-133. https://doi.org/10.17763/haer.80.1.j94675w001329270

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

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 677-680. https://doi.org/10.1126/science.103.2684.677

Thissen, D., Steinberg, L., & Fitzpatrick, A. R. (1989). Multiple-choice models: The distractors are also part of the item. *Journal of Educational Measurement*, *26*(2), 161-176. https://doi.org/10.1111/j.1745-3984.1989.tb00326.x

Wright, B. D. (2003). Rack and stack: Time 1 vs. time 2 or pre-test vs. post-test. *Rasch Measurement Transactions*, *17*(1), 905-906. https://www.rasch.org/rmt/rmt171a.htm

Wright, B. D., & Stone, M. H. (1979). *Best test design*. MESA Press.

Appendix

Table A1

Randomization algorithms for VST test forms

Frequency	Randomization Algorithm
1K	=RANDBETWEEN(1,20000)
2K	=RANDBETWEEN(1000,20000)
3K	=RANDBETWEEN(2000,20000)
4K	=RANDBETWEEN(3000,20000)
5K	=RANDBETWEEN(4000,20000)
6K	=RANDBETWEEN(5000,20000)
7K	=RANDBETWEEN(6000,20000)
8K	=RANDBETWEEN(7000,20000)
9K	=RANDBETWEEN(8000,20000)
10K	=RANDBETWEEN(9000,20000)
11K	=RANDBETWEEN(10000,20000)
12K	=RANDBETWEEN(11000,20000)
13K	=RANDBETWEEN(12000,20000)
14K	=RANDBETWEEN(13000,20000)