Vocabulary Learning and Instruction
Volume 11, Issue 2, December 2022
http://vli-journal.org

Rater Judgments and Word Difficulty: Conceptualizing the Substantive Validity of the VST

Derek N. Canning, Stuart McLean, and Joseph P. Vitta
[a]*Seigakuin University;* [b]*Momoyama University;* [c]*Kyushu University*

Abstract

The substantive component of construct validity requires a confrontation between empirical test results and content relevance. The Vocabulary Size Test (VST) has been extensively validated in terms of empirical results. Less is known, however, about expert judgments of content relevance. The VST was constructed and validated according to the principle that frequency underlies vocabulary acquisition. This does not mean, however, that the two are equivalent. To better understand the construct of word difficulty as it is measured in VSTs, the results of two Rasch analyses, one on the VST, the other a Multi Facet Rasch Model (MFRM) of a questionnaire distributed to education practitioners were correlated ($n = 80$, $r = 0.67$, $p < 0.001$, 95% CI = [0.53, 0.77]). Semi-structured interviews were then conducted to explore how practitioners understand the concept of word difficulty. Findings indicate that word difficulty is understood to encompass more than frequency, validating recent research into the predictive power of lexical sophistication variables.

Keywords: Construct Validity, Lexical Sophistication, Rasch Analysis, Word Difficulty

1 Background

The design of the vocabulary size test (VST) is predicated on the idea that the difficulty of learning a word is directly correlated to the frequency of that word (Nation & Beglar, 2007). The frequency effect underlies a great deal of our understanding of how languages are acquired in either a first (L1) or second (L2) language (Ellis, 2002). Those who have designed tests such as the VST have long acknowledged that frequency is not the only factor that may contribute to word learning difficulty (Beglar, 2010; Nation & Beglar, 2007). Recently, however, there has been increasing interest in what other variables affect word difficulty (Canning et al., 2022; Hashimoto, 2021; Hashimoto & Egbert, 2019; Stewart et al., 2022; Vitta et al., 2022).

Research in this area has shown that variables other than frequency do in fact play a key role in word learning difficulty. Examples include studies on range (Kyle & Crossley, 2015; Kyle et al., 2018), word length, (Ellis & Beaton, 1993; Willis & Ohashi, 2012), and word neighborhood (Hashimoto & Egbert, 2019), that have

Please cite this article as: Canning, D. N., McLean, S., and Vitta, J. P. (2022). Rater judgments and word difficulty: Conceptualizing the substantive validity of the VST. *Vocabulary Learning and Instruction, 11*(2), 30–37. https://doi.org/10.7820/vli.v11.2.canning

shown that these variables significantly predict word difficulty. Similarly, the status of a lexical test item as a loan word has also shown to influence word difficulty (Canning et al., 2022; Laufer & McLean, 2016; Willis & Ohashi, 2012).

Taken together, studies on variables affecting word learning difficulty have demonstrated that the construct is defined by more than frequency. There has been little discussion, however, as to the substantive construct of word difficulty. Frequency appears to explain word difficulty through analyses of test results, but frequency is not equivalent to the construct. This paper is an attempt to judge content relevance by asking practitioners to first rate a set of words derived from the VST on their perceived difficulty. These ratings were then correlated with item difficulty derived from the VST. Finally, the raters/practitioners were asked for the reasoning behind their judgments. In an attempt to demonstrate the "confrontation," in Messick's words, between empirical test results and "judged content relevance" on the other, rater judgments were correlated with the response consistencies derived from the VST (1988, p. 62). This was done to investigate two research questions:

(1) To what degree do expert ratings of word difficulty correlate with actual word difficulty as operationalized through the VST?
(2) To what degree do expert explanations of word difficulty corroborate our understanding of the construct of word difficulty as entailing more than frequency?

2 Methods

2.1 Participants

2.1.1 VST

The 80-item version of the VST was administered to 3,449 first-, second-, and third-year university students in a number of different universities in Eastern Japan (McLean et al., 2014). Of this group, only those students claiming their L1 as Japanese were considered for this study (N = 2,999) (For more details on data collection, see McLean et al.).

2.1.2 Rater judgments of difficulty

A survey on the perceived learning difficulty of the 80 words on the VST was administered to 31 teachers of English as a foreign language currently living in Japan. The average number of years teaching English as a foreign language was 16.6 years. 19 participants were L1 English speakers, and 12 participants listed their L1 as other than English.

2.2 Instruments

2.2.1 VST

Data collected from the 80-item version of the VST (Nation and Beglar, 2007) was used for this study. For details on its construction and administration,

see McLean et al. (2014). Results were derived from a dichotomous Rasch analysis conducted via Winsteps (Linacre, 2022). Word difficulty as a dependent variable was calculated in logit scores from this analysis.

2.2.2 Rater judgments of difficulty

A rating scale was administered online to elicit teacher judgments of word difficulty. The Likert-type scale ranged from 1: This word is very difficult to learn, to 6: This word is very easy to learn (See supplementary online materials at osf.io/k28my for the complete scale). Results of the questionnaire were submitted to a Rasch analysis along three facets: rater, item, and L1 of the rater. The third facet was included to determine if the L1 of the rater contributed to a bias in judgments of word difficulty. Rasch analysis was used to determine the fair average of lexical difficulty, that is, a difficulty measure that "corrected" for the relative severity or leniency of the raters (Bond et al., 2021, p. 177). In this Multi Facet Rasch Model (MFRM) analysis, raters are treated as "independent experts," whose understanding of the relative difficulties of the words can be measured on a common logit scale that accounts for individualized interpretations of the Likert scale (Bond et al., 2021, p. 175).

Brief online interviews were conducted with 21 of the 31 respondents. In these semi-structured interviews, one word was chosen at random from each of the categories the raters had assigned. The participants were asked about their reasoning for assigning a given category to the chosen word.

3 Results

3.1 VST: Substantive Aspect of Construct Validity

The substantive aspect of construct validity of the VST is predicated on the theory that frequency is a correlate for word-learning difficulty. This is then corroborated by observing that the difficulty continuum, as determined by the Rasch analysis, is in line with frequency levels (Beglar, 2010). Design and validation studies (Beglar, 2010; Nation & Beglar, 2007) of the VST make no claims that frequency and word difficulty are one and the same construct, only that vocabulary acquisition is likely to follow exposure to input in the language, and that the frequency effect then accounts for why words encountered more often are "easier," in terms of test scores, than more "difficult" words encountered less frequently (see Ellis, 2002, for a discussion of the frequency effect in second language acquisition).

Word difficulty was calculated from the Rasch analysis of the VST and represented in logit scores. Representativeness, responsiveness, and technical quality of the instrument were verified. Similarly, the technical quality, responsiveness, and representativeness of the Rasch analyses of the rating judgments were assessed. Details of these analyses are available in the online supplementary materials at osf.io/k28my. To determine if rater judgments of difficulty reflected the construct being measured by the VST, the logit scores of both analyses were correlated.

Teacher judgments of word difficulty item logit scores correlated strongly with VST logit scores at $r = 0.67$, $p < 0.001$, 95% CI = (0.53, 0.77) (assumptions of parametric testing met; see Appendix A). This observed correlation value converges well with values involving frequency and word difficulty, e.g., $r = 0.50$ (Hashimoto, 2021) and $r = 0.78$ (Stewart et al., 2022). The inclusion of L1 as a facet in the MFRM, additionally, indicated no bias in the determination of word difficulty. Teacher-raters were then asked what it was about a subset of those words that made them more or less difficult to learn. These qualitative findings are discussed below and suggest that their theoretical understanding of the word difficulty construct involves more than frequency.

3.2 Rater Judgments of Difficulty: Structural Aspect of Construct Validity

Positive and negative residual loadings of a principal components analysis were scrutinized for potential patterns. Positive loadings above 0.50 were noted to be difficult items, ranging from 0.72 to 2.11 logits. Negative loadings below −0.49 were easy items, ranging in logit difficulty from −2.99 to 0.92. Crucially, most negative loadings were English words with loanword counterparts in Japanese (Canning et al., 2022; Daulton, 2008). Although the scree plot did not appear to suggest a second dimension, there was a clear pattern to the positive and negative loadings. These findings are summarized in the supplementary online materials at osf.io/k28my.

3.3 Interview Findings

To determine how raters conceptualized the substantive construct of word difficulty, 21 of the 31 participants were interviewed. These interviews were coded on emergent themes. In the semi-structured interviews, raters were asked seven questions, the first six about what made a word they had assigned to each difficulty level easy or difficult to learn. The seventh question asked the raters if there was anything else they would like to say about word-learning difficulty that they had not covered earlier. Any given code was only used once per question, although multiple different codes were permitted for a single question. A summary and description of the codes can be found in the supplementary online materials at osf.io/k28my.

The results of this coding process are summarized in Figure 1. They show that education practitioners understand the construct of word difficulty as multifaceted. Frequency is the second most mentioned factor, following the conceptual difficulty of a word. That a word has a parallel in the learner's L1 was thought to play a key role in determining its learning burden. Importantly, this code includes loanword status.

4 Discussion and Conclusion

The findings in this study demonstrate that the substantive aspect of construct validity of the VST is understood by education practitioners to be multifaceted.

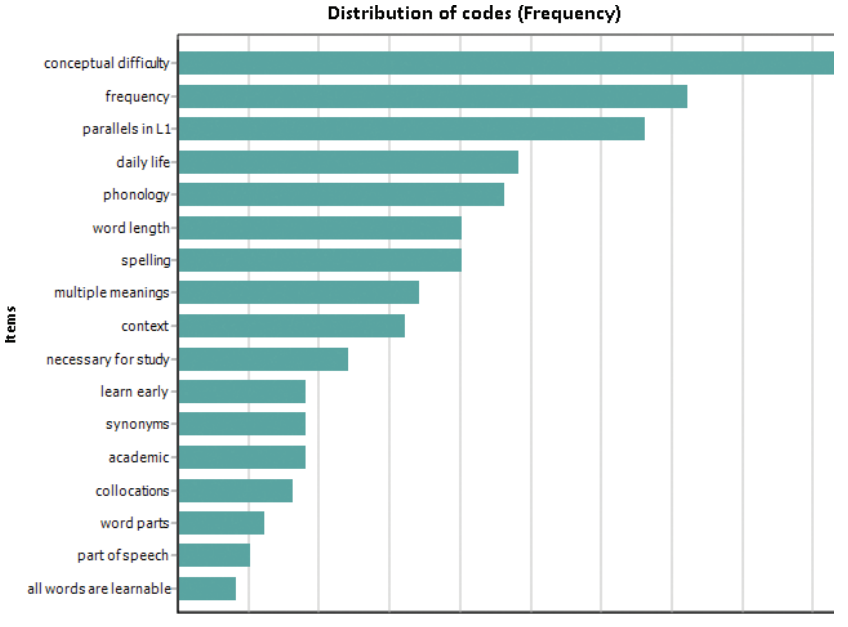

Figure 1. Frequency of Reasons Given for Word-Learning Difficulty.

As hypothesized, rater judgments of word difficulty correlated with VST logit scores. Semi-structured interviews with raters indicated that they conceived of word difficulty, presumably the underlying construct reflected in the VST, as made up of more than frequency. Factors such as the conceptual difficulty of a word, its parallels in L1 (including loanword status), and the utility of the word in daily life were all identified as important to its learnability. These findings are consistent with recent research showing that word difficulty is predicated on more than frequency (Canning et al., 2022; Hashimoto, 2021; Hashimoto & Egbert, 2019; Stewart et al., 2022; Vitta et al., 2022).

Certain words present lower learning burdens to specific populations of learners. A L2 word that corresponds to a L1 loanword can facilitate its acquisition in terms disproportionate to its frequency. As a variable, loanwords have been shown to predict word learning difficulty (Willis & Ohashi, 2012) and to affect VST scores (Laufer & McLean, 2016). In Canning et al. (2022), a multiple linear regression model found that in addition to frequency, loanword status was a significant predictor of word difficulty on the VST difficulty scores analyzed in the current study.

The conclusions of the present study are tempered by several limitations. First, the interview data was coded by a single rater. Future mixed methods studies in this area should adopt more stringent approaches to the analysis of qualitative data. Finally, not all avenues of validation made available by the Rasch model were utilized. To better understand how elements of word difficulty contribute to VST scores, concurrent validity studies will need to match standards set in previous research.

References

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. G., Yan, Z., & Heene, M. (2021). *Applying the Rasch model: Fundamental measurement in the human sciences*. Routledge.

Canning, D. N., Vitta, J. P., & McLean, S. (2022). Relative complexity in a model of word difficulty: The role of loanwords in vocabulary size tests. [Manuscript in preparation]. Graduate College of Education, Temple University Japan.

Daulton, F. (2008). *Japan's built-in lexicon of English-based loan words*. Multilingual Matters.

Ellis, N. C. (2002). Frequency effects in language processing. *Studies in Second Language Acquisition, 24*(2), 143–188. https://doi.org/10.1017/S0272263102002140

Ellis, N. C., & Beaton, A. (1993). Psycholinguistic determinants of foreign language vocabulary learning. *Language Learning, 43*(4), 559–617. https://doi.org/10.1111/j.1467-1770.1993.tb00627.x

Hashimoto, B. J. (2021). Is frequency enough?: The frequency model in vocabulary size testing. *Language Assessment Quarterly, 18*(2), 171–187. https://doi.org/10.1080/15434303.2020.1860058

Hashimoto, B. J., & Egbert, J. (2019). More than frequency? Exploring predictors of word difficulty for second language learners. *Language Learning, 69*(4), 839–872. https://doi.org/10.1111/lang.12353

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly, 49*(4), 757–786. https://doi.org/10.1002/tesq.194

Kyle, K., Crossley, S. A., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods, 50*(3), 1030–1046. https://doi.org/10.3758/s13428-017-0924-4

Laufer, B., & McLean, S. (2016). Loanwords and vocabulary size test scores: A case of different estimates for different L1 learners. *Language Assessment Quarterly, 13*(3), 202–217. https://doi.org/10.1080/15434303.2016.1210611

Linacre, J. M. (2022). *Winsteps Rasch measurement computer program (Version 5.2.3)*. Winsteps.com.

McLean, S., Hogg, N., & Kramer, B. (2014) Estimations of Japanese university learners' English vocabulary sizes using the vocabulary size test. *Vocabulary Learning and Instruction, 3*(2), 47–55. https://doi.org/10.7820/vli.v03.2.mclean.et.al

Messick, S. (1988). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 3–209). Macmillan.

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

Stewart, J., Vitta, J. P., Nicklin, C., McLean, S., Pinchbeck, G. G., & Kramer, B. (2022). The relationship between word difficulty and frequency: A response

to Hashimoto (2021). *Language Assessment Quarterly, 19*(1), 90–101. https://doi.org/10.1080/15434303.2021.1992629

Vitta, J. P., Nicklin, C., & Albright, S. (2022). *Academic word difficulty and multidimensional lexical sophistication: An EAP-focused conceptual replication of Hashimoto & Egbert (2019)*. Paper presented at the 2022 American Association of Applied Linguistics Convention.

Willis, M., & Ohashi, Y. (2012). A model of L2 vocabulary learning and retention. *Language Learning Journal, 40*(1), 125–137. https://doi.org/10.1080/09571736.2012.658232

Appendix A

Assumption Check of Correlation between Aggerated Teacher Judgments and Word Difficulty Logits

The assumptions of the correlation were checked using the parametric testing checklist found in Vitta et al. (2022) and Stewart et al. (2022) which referenced relevant applied statistics theory. The assumptions of linearity and homoscedasticity appeared to be met after inspecting the standardized residual and predicted values scatter plot. The F test for heteroscedasticity was nonsignificant (p = 0.289; F [1, 78] = 1.14) and thus there was further empirical evidence that the errors had equal variance across the model. The observed Durbin-Watson value (= 1.36) was within the acceptable range (1 to 3) and thus autocorrelation (or natural ordering in cross-sectional designs) was not a concern. The observed maximum Cook's distance (max = 0.23) and centered leverage (max = 0.09) values were under 1 and thus no case had undue influence in the model. Finally, the residuals displayed an approximate normal distribution as the absolute z-scores for both skew (z = 2.52) and kurtosis (z = 1.78) were under the 3.29 threshold. This implied that the parameters of the model could be generalized with confidence to the population.