Vocabulary Learning and Instruction
Volume 10, Issue 2, December 2021
http://vli-journal.org

Developing a Measure of Proper Name Familiarity for Japanese University Students

Christopher Nicklin
Rikkyo University

Abstract

In this study, an instrument for measuring proper name (PN) familiarity was developed for a psycholinguistic experiment investigating the effect of PNs on Japanese university students' English reading fluency. Familiarity has previously been operationalized in disparate ways, producing contradictory results. Furthermore, authors of previous studies did not conduct validation analyses on their familiarity instruments. To address this issue, a four-point Likert-type scale instrument was constructed to assess Japanese university students' familiarity with a set of 100, two-syllable PNs. The responses of 216 participants from 2 Japanese universities were subjected to Rasch analysis with the rating scale model to determine whether the resulting data fit the expectations of the model. The results suggested that a dichotomous response instrument was more appropriate than the scale-based instruments utilized in previous studies.

Keywords: Proper nouns, Familiarity, Rasch analysis

1 Background

Proper names (PNs) constitute a lexical class comprising the names of people, places, facilities and institutions, objects, and works of art that might be considered unique (Valentine et al., 1996). Despite accounting for approximately 1 and 5% of text in novels and newspapers, respectively, (Nation, 2006), PNs are essentially ignored in second language (L2) research. Because they are distinguished through the use of initial capital letters, it is assumed that learners will know PNs and they are thus included in text coverage counts as known items (e.g., Nation, 2006; Schmitt, 2008; Webb & Chang, 2015; Webb & Macalister, 2013). This assumption contrasts with research suggesting that advanced English learners' reading and listening comprehension is disrupted by unfamiliar PNs (e.g., Erten & Razi, 2009; Kobeleva, 2012).

The omission of PNs from text coverage counts can be legitimately questioned when considering that L2 researchers and pedagogists traditionally quantify the likelihood that readers will know a word, and in turn lexical 'difficulty', through corpus-based frequency. Proper names such as *Holden* and *Hartright,* with merely 3,135 and 32 appearances in the Corpus of Contemporary American English (COCA; Davies, 2008), respectively, are assumed to be unproblematic

Please cite this article as: Nicklin, C. (2021). Developing a measure of proper name familiarity for Japanese university students. *Vocabulary Learning and Instruction, 10*(2), 91–100. https://doi.org/10.7820/vli.v10.2.nicklin

because they are capitalized. This assumption becomes more problematic with the case of extensive reading, which involves learners reading massive amounts of simplified text with high speed and comprehensibility (Waring & McLean, 2015). One of the main benefits of extensive reading is its propensity to develop learner's reading fluency (e.g., Beglar & Hunt, 2014). However, this quality is potentially hindered if the assumption regarding PNs' unproblematic nature is inaccurate. To determine whether the assumption holds true, the effect of PNs on L2 learner's reading fluency warrants research explicitly addressing the issue.

One way to approach the issue involves reliance on the traditional corpus frequency-based approach. However, it is debatable as to how pertinent the material gathered in COCA are to a population such as Japanese English as a foreign language (EFL) university students. Another way would be to assess how *familiar* a sample of the target population are with the target PNs. Familiarity has been operationalized by L1 and L2 vocabulary researchers in disparate ways, such as through the use of post-experimental interviews (e.g., Schmitt & Underwood, 2004), and varying question prompts along with five- (e.g., Libben & Titone, 2008; Titone et al., 2019) or seven-point scales (e.g., Carroll & Conklin, 2020; Valentine et al., 1991). Furthermore, the results produced with the instruments did not undergo any validation.

The present study reports on the development of PN familiarity ratings, measured in Rasch logits, for inclusion as an independent variable in an experiment conducted to assess the effect of PNs on Japanese EFL university students' English reading fluency. The familiarity variable will eventually be included in linear-mixed effects model as a predictor of reading times in a self-paced reading experiment. With this in mind, the following research question was addressed.

1. To what extent are the Rasch model expectations met by an instrument designed to measure how familiar a set of English proper nouns are to a group of Japanese university students?

2 Method

In accordance with recent calls for multisite samples (Vitta et al., 2021), participants (N = 263) were recruited from six intact English classes at two Japanese universities, Site A (n =207) and Site B (n = 56). Permission was granted to conduct the research from both universities and all participants signed a consent form. Both universities prohibited reporting standardized proficiency measures, which constitutes a limitation of the study. However, all participants had received at least 6-years of pre-university classroom English instruction.

The target PNs were extracted from a small 256,956-word corpus constructed from 15 Oxford Bookworms graded readers. Three books were randomly selected from each of Stages 2 through 6 of the Bookworms series, and each of the 15 books was converted into .txt file, tagged with TagAnt (Anthony, 2015), and analyzed with AntConc (Anthony, 2014). In total, 127 two-syllable PNs that appeared five or more times were extracted from the corpus. Twenty-seven PNs were omitted based upon graded-reader corpus frequency (i.e., the PNs with fewest occurrences were removed), which left 100 target PNs remaining.

In addition to the 100 target PNs, 65 control items were included on the instrument, consisting of the 33 most common two-syllable Japanese family names as of 2009[1] and 32 non-names. The non-names were constructed from a list of 32 two-syllable location PNs (e.g., *London*) that were extracted from the graded-reader corpus and had the first letter (or sound) of each item substituted for next consonant or vowel in the alphabet, relative to the letter being substituted. For instance, *London* became *Mondon*, and *Iran* became *Oran*. The 165 items were presented to participants in random order on a Google Form. Test takers were instructed in Japanese with the following Japanese prompt:

> *"世界中から集められた名前が表示されます。例えば、英語、日本語、その他の言語のものがあります。それぞれの名前に馴染みがあるかどうか、1～4で評価してください。 ※1 = この名前に馴染みがない 4 = この名前に非常に馴染みがある"*
>
> *[You will see a group of names from around the world. For example, some will be English, some will be Japanese, some will be from other languages. Please rate on a scale of 1 to 4 how familiar each name is to you: 1 = Not familiar at all and 4 = Very familiar"]*

Test takers were prompted (in Japanese) to answer [lit.] *How familiar do you feel with this name* and were asked to respond by selecting one of the following options: *I'm not familiar with this name*, *I'm a little familiar with this name*, *I'm familiar with this name*, and *I'm very familiar with this name*. All responses were automatically recorded on a Google spreadsheet for analysis.

The data preparation process comprised three stages. Firstly, the responses were recoded as numerical values ranging from 1 to 4, where 1 = *I'm not familiar with this name* and 4 = *I'm very familiar with this name*. Secondly, items and participants with high false-alarm (FA) rates were removed from the dataset. False-alarm rate related to the number of times one of the 32 non-names, such as *Zorkshire*, was responded to as being *very familiar*, *familiar*, or *a little familiar*, while person FA related to the number of times a participant responded to a non-name as being *very familiar*, *familiar*, or *a little familiar*. A FA rate cut-off of 10% was utilized to ensure that participants were not overestimating their PN knowledge and to exclude participants who were perhaps not concentrating.

The FA-rate check revealed that five items received 27 or more *a little familiar, familiar,* or *very familiar* responses and were removed from further analysis for being too endorsable. Based on responses to the remaining 27 non-names, 47 participants with an FA rate > 10% (i.e., three or more [(27/100)*10 = 2.7] *a little familiar, familiar,* or *very familiar* responses to non-names) were excluded from further analysis. Finally, a spreadsheet was constructed that contained only the responses to the PNs. The non-names were removed because their only function was for FA-rate calculations, and the Japanese PNs were removed because they functioned purely as filler items.

To address the research question, the responses of the 100 target items and remaining 216 persons were subjected to Rasch analysis (Rasch, 1960) with the rating scale model (RSM; Andrich, 1978), which allows for polytomous data to be

[1] Retrieved from https://www.japantimes.co.jp/life/2009/10/11/lifestyle/japans-top-100-most-common-family-names/

fit to a single rating scale (Bond et al., 2020). The analysis was conducted with R (R Core Team, 2019), and the Test Analysis Modules (TAM; Robitzsch et al., 2020) package. A confirmatory analysis was conducted with the Extended Rasch Measurement package (eRm; Mair & Hatzinger, 2007), based upon Linacre's (in press) recommendation to estimate parameters with at least two packages when conducting Rasch analysis with R.

The Rasch misfit statistics for each item and person were analyzed to determine whether they were conducive to measurement. Misfit was determined via Wright and Linacre's (1994) thresholds for productive measurement, whereby infit and outfit mean square (MNSQ) values outside of 0.50 to 1.50, and *t*-scores outside of −2.00 to 2.00 are considered detrimental to measurement. Items and persons with fit statistics below 0.50 were considered acceptable because they are unlikely to have practical implications in human science research (Bond et al., 2020). Misfitting items and persons were examined individually to assess why they failed to conform to the model's expectations. For instance, an item might misfit the model's expectations because of an error by a participant who produced several unusual answers (e.g., responded not only to several low familiarity target items as *Very familiar* but also responding to several high familiarity items as *Not familiar*). In this instance, it might be better to remove the person as opposed to the item and then re-run the analysis. The result of this iterative process was a set of logits representing a measure of the familiarity of each target item.

3 Results and Discussion

Despite the precedent in the literature for measuring familiarity on a scale, the results from the four-point scale utilized in the present study failed to meet the expectations of the RSM. When misfitting persons were removed, large numbers of items were also required to be removed due to insufficient responses to all four category levels. The middlemost categories failed to distinguish between *Familiar* and *A little familiar* across the sample suggesting that the meaning of each category was not invariant across the participants. However, the Guttman plot (see Figure 1) did suggest that a binary choice of either *Familiar* or *Not familiar*, might be more successful. This was based upon the observation that the top-left corner cells were generally all lighter than the bottom-right cells and that a diagonal line was visible stretching from the bottom-left to top-right corner. Consequently, the *Very familiar, Familiar*, and *Slightly familiar* categories were collapsed into a single category representing simply *Familiar* (*to some degree*). This resulted in a recoded, binary dataset, whereby 0 = *I'm not familiar with this name* and 1 = *I'm familiar with this name (to some degree)*, which was fit with a dichotomous Rasch model using TAM.

Although collapsing categories might seem controversial because the new dataset represents answers that were incongruent with the response categories presented to the participants, it is an acceptable practice when ascribing to a school of thought that considers the model as subject to exploration (Wright & Linacre, 1992). Under this school of thought, the analyst is responsible for extracting the maximum amount of meaning from the observed responses. Wright and Linacre explained that collapsing categories with RSM frequently results in equivalent fit and results, and that such behavior is acceptable provided the decision can be

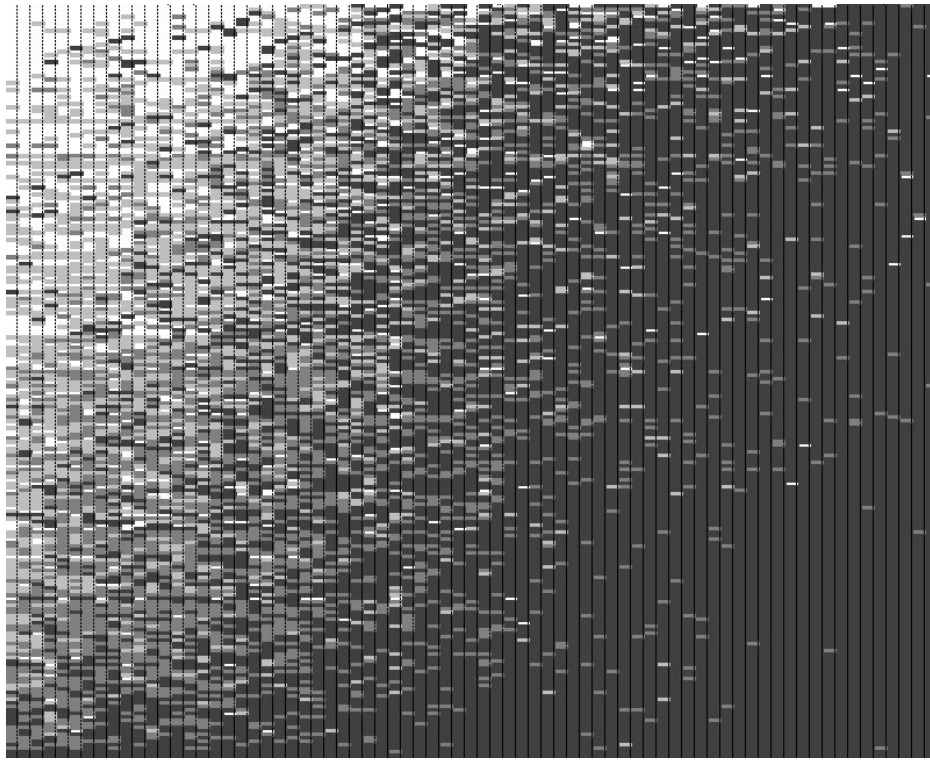

Figure 1. Guttman Plot of the Initial Dataset.

The initial iteration of the dichotomous model resulted in misfitting items with outfit mean square (MNSQ) values > 1.5 and outfit and infit *t*-scores > 2.00. Thus, the initial analysis was followed-up with nine iterations until all remaining items and persons satisfactorily fit the expectations of the Rasch model. The final iteration comprised 92 items and 185 persons, all displaying MNSQ fit statistics within Wright and Linacre's thresholds, but with three items (*Baby, Hatta, Sunset*) and two persons displaying infit *t*-scores above 2.00. However, this was deemed acceptable because 5% of the items or persons are expected to misfit by chance (Beglar, 2010). Reliability was measured with expected a posteriori (EAP) reliability, which is a measure that, although not the same, can be interpreted in the same way as Cronbach's α (Neumann et al., 2011). The EAP reliability of 0.91 was high, indicating that the results produced by the reduced sample were reliable. Figure 2 illustrates that the TAM person parameters' distribution was Gaussian, which is an assumption of the marginal maximum likelihood estimation equation utilized in the TAM analysis. The R script containing details of each iteration (and also the Supplementary Materials) is available online at https://github.com/nicklinc/vli2022. Supplementary Materials A contains a brief report of the confirmatory eRm analysis, which produced almost identical results. The descriptive statistics for the final iteration are displayed in Table 1.

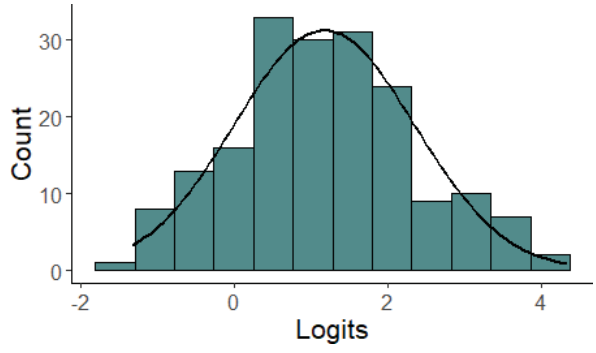

Figure 2. Distribution of the Person Parameters.

Table 1. Descriptive Statistics for the Final Iteration of the Dichotomous Rasch Model

Parameter	Measure	Mean	SD	Min	Max	Skew	Kurt
Items	Raw	60.61	60.00	1.00	177.00	0.67	-1.13
(*N* = 92)	Logit	1.48	2.53	-3.65	5.76	-0.30	-1.04
	SE	0.31	0.18	0.16	1.01	1.82	3.36
	Infit MNSQ	1.00	0.07	0.87	1.24	1.17	2.28
	Outfit MNSQ	0.89	0.25	0.29	1.38	-0.32	-0.40
	Infit-*t*	0.14	0.70	-1.68	2.93	1.58	4.71
	Outfit-*t*	0.01	0.76	-1.36	2.79	0.89	1.24
Persons	Raw	30.14	10.37	7.00	55.00	0.06	-0.50
(*N* = 185)	Logit	1.18	1.17	-1.31	4.35	0.28	-0.32
	SE	0.34	0.03	0.30	0.47	1.59	3.69
	Infit MNSQ	0.99	0.19	0.56	1.45	0.09	-0.71
	Outfit MNSQ	0.80	0.31	0.24	1.44	0.28	-0.99
	Infit-*t*	-0.03	1.00	-2.68	2.13	-0.15	-0.63
	Outfit-*t*	-0.12	0.60	-1.51	0.99	-0.33	-0.80

Note: Min, Minimum; Max, Maximum.

The final selection process involved selecting 30 PNs for the psycholinguistic experiment assessing the effect of PNs on Japanese EFL university students' L2 English reading fluency. The TAM-derived logits were reverse signed for ease of interpretability (i.e., the logit for *Tony* [−1.96] became 1.96, thus larger values represented *more familiar* according to the target group), resulting in a spread of 92 PNs from the least familiar PN, *Halcombe* (−5.76) to the most familiar, *William* (3.65). The 15 least familiar PNs with logits < −4.00 were removed from the list, leaving a spread from approximately −4.00 through 4.00. The 15 excluded PNs were also the items with the largest standard errors (SEs), indicating that they were estimated with the least precision. Furthermore, all PNs with alternative meanings, which constituted potential confounds, were removed (*Baby, Rosso,* and *Sunset*). From the remaining 74 PNs, 15 PNs with logits above and below zero were selected, with the aim of achieving an even spread of values from −4.00 to 4.00. The final 30 PNs

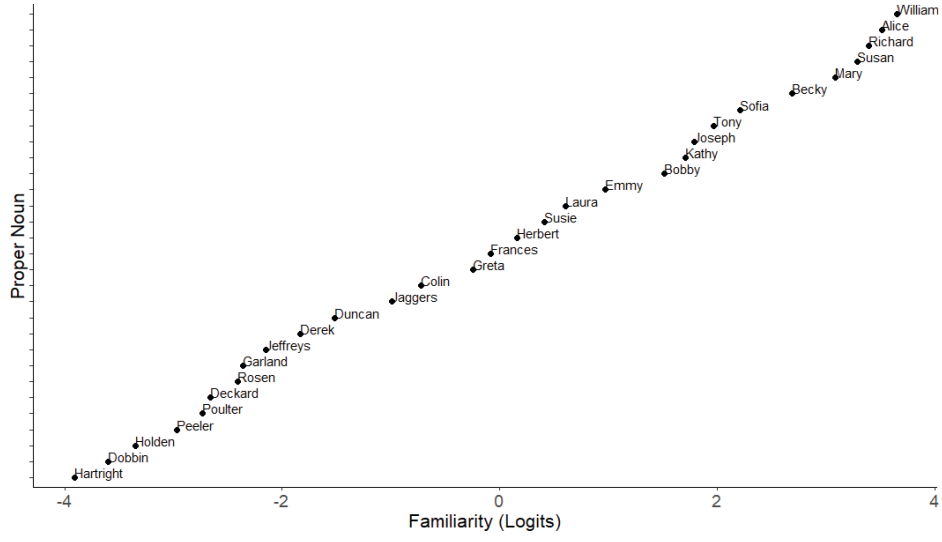

Figure 3. Final 30 Proper Names Plotted by Familiarity Logit.

are plotted by familiarity logit in Figure 3, while the logit, SE, and fit statistics for each PN are presented in Supplementary Materials B.

With regard to the research question, the results of the analysis suggest that the results of the proper noun familiarity instrument met the expectations of the Rasch model. The final 30 items comprise a spread of logits ranging from −3.91 to 3.65, with a maximum SE of 0.42. The MNSQ fit statistics indicated that the items fit the expectations of the Rasch model well, with all values within Wright and Linacre's (1994) 0.50 to 1.50 thresholds for productive measurement. Furthermore, the order of familiarity ratings makes logical sense. For instance, it is understandable that *William* and *Alice* are the most familiar names from the list and that they are more familiar to Japanese university students than *Tony* and *Sophia*, who are located several places lower. *William* is the name of an English Prince who features regularly in the Japanese news, while the name *Alice* is embedded in popular culture through the Lewis Carroll novels, *Alice in Wonderland* and *Through the Looking Glass*, and their Disney interpretations. It is also understandable that *Holden, Dobbin,* and *Hartright* are the least familiar names, as these can hardly be said to be typical names.

The result of this short study has one main implication for the use of familiarity as a variable in L2 research. Although familiarity has been utilized in previous research, the authors of those studies operationalized the variable in various guises, such as five- and seven-point Likert scales. The present study is the only one of these studies that has investigated a scale-derived approach to familiarity with Rasch analysis, and the result suggested that such an approach might be inappropriate for L2 learners. The participants failed to use the categories of the scale in a consistent manner, thus the categories failed to separate. Collapsing the categories and constructing a dichotomous model solved this issue, thus a dichotomous instrument should be utilized if this process is replicated with another set of PNs. It is possible that this conclusion is relevant for familiarity measures in

L1 psycholinguistic research, but Rasch analysis of such instruments is required for confirmation.

4 Conclusion

In the present study, the development of a PN familiarity measure for inclusion as a variable in a model of Japanese EFL university students' reading fluency was reported upon. Although previous researchers have utilized familiarity variables by collecting information from the target population, none have reported the results of a validation analysis to determine whether the instruments were measuring what they were designed to measure. Rasch analysis with the rating scale model indicated that the scale-derived approach to familiarity adhered to in previous L1 research might be inappropriate for L2 research because the meaning of each category was not invariant across the participants. It is possible that alternative wording of the instrument could result in more consistent responses, but further research would be required. However, once the categories were collapsed to allow a binary, "Yes" or "No" approach to familiarity, the model fit the expectations of the dichotomous Rasch model. The main implication of this short study is that the construction of a familiarity measure in L2 research, and perhaps even L1 research, should involve an instrument with a dichotomous response as opposed to four-, five-, or seven-point scales, because the difference between the categories are unlikely to be consistent across all participants.

References

Andrich, D. (1978). A rating formula for ordered response categories. *Psychometrika, 43*, 561–573. https://doi.org/10.1007/BF02293814

Anthony, L. (2014). *AntConc (Version 3.4.1w) [Computer Software]*. Waseda University. http://www.antlab.sci.waseda.ac.jp/

Anthony, L. (2015). *TagAnt (Version 1.2.0) [Computer Software]*. Waseda University. http://www.antlab.sci.waseda.ac.jp/

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

Beglar, D., & Hunt, A. (2014). Pleasure reading and reading rate gains. *Reading in a Foreign Language, 26*(1), 29–48. https://doi.org/10125/66684

Bond, T., Yan, Z., & Heene, M. (2020). *Applying the Rasch model: Fundamental measurement in the social sciences* (4th ed.). Routledge. https://doi.org/10.4324/9780429030499

Carroll, G., & Conklin, K. (2020). Is all formulaic language created equal? Unpacking the processing advantage for different types of formulaic sequences. *Language and Speech, 63*(1), 95–122. https://doi.org/10.1177/0023830918823230

Davies, M. (2008). *The corpus of contemporary American English (COCA)*. https://www.english-corpora.org/coca/

Erten, I. H., & Razi, S. (2009). The effects of cultural familiarity on reading comprehension. *Reading in a Foreign Language, 21*(1), 60–77. https://doi.org/10125/66632

Kobeleva, P. (2012). Second language listening and unfamiliar proper names: Comprehension barrier? *RELC Journal, 43*(1), 83–98. https://doi.org/10.1177/0033688212440637

Libben, M. R., & Titone, D. A. (2008). The multidetermined nature of idiom processing. *Memory & Cognition, 36*(6), 1103–1121. https://doi.org/10.3758/MC.36.6.1103

Linacre, J. M. (in press). Advancing the metrological agenda in the social sciences. In S. Cano, P. Marquis, A. Regnault, & W. P. Fisher Jr. (Eds.), *Person-centered outcome metrology*. Springer.

Mair, P., & Hatzinger, R. (2007). Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software, 20*(9), 1–20. https://doi.org/10.18637/jss.v020.i09

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review, 63,* 59–82. https://doi.org/10.1353/cml.2006.0049

Neumann, I., Neumann, K., & Nehm, R. (2011). Evaluating instrument quality in science education: Rasch-based analyses of a Nature of Science Test. *International Journal of Science Education, 33*(10), 1373–1405. https://doi.org/10.1080/09500693.2010.511297

R Core Team. (2019). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. https://www.R-project.org/

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

Robitzsch, A., Kiefer, T., & Wu, M. (2020). TAM: Test analysis modules. R package version 3.5–19. https://CRAN.R-project.org/package=TAM

Schmitt, N. (2008). Instructed second language vocabulary learning. *Language Teaching Research, 12*(3), 329–363. https://doi.org/10.1177/1362168808089921

Schmitt, N., & Underwood, G. (2004). Exploring the processing of formulaic sequences through a self-paced reading task. In N. Schmitt (Ed.), *Formulaic sequences: Acquisition, processing, and use* (pp. 173–190). Amsterdam: Benjamins. doi: 10.1075/lllt.9

Titone, D., Lovseth, K., Kasparian, K., & Tiv, M. (2019). Are figurative interpretations of idioms directly retrieved, compositionally built, or both? Evidence from eye movement measures of reading. *Canadian Journal of Experimental Psychology, 73*(4), 216–230. https://doi.org/10.1037/cep0000175

Valentine, T., Brédart, S., Lawson, R., & Ward, G. (1991). What's in a name? Access to information from people's names. *European Journal of Cognitive Psychology, 3*(1), 147–176. https://doi.org/10.1080/09541449108406224

Valentine, T., Brennen, T., & Brédart, S. (1996). *The cognitive psychology of proper names: On the importance of being Ernest.* Routledge. https://doi.org/10.4324/9780203285763

Vitta, J. P., Nicklin, C., & McLean, S. (2021). Effect-size driven sample size planning, randomization, and multi-site use in L2 instructed vocabulary acquisition experimental samples. *Studies in Second Language Acquisition.* Advance online publication. https://doi.org/10.1017/S0272263121000541

Waring, R., & McLean, S. (2015). Exploration of the core and variable dimensions of extensive reading research and pedagogy. *Reading in a Foreign Language, 27*(1), 160–167. https://doi.org/10125/66708

Webb, S., & Chang, A. C.-S. (2015). Second language vocabulary learning through extensive reading with audio support: How do frequency and distribution of occurrence affect learning? *Language Teaching Research, 19*(6), 667–686. https://doi.org/10.1177/1362168814559800

Webb, S., & Macalister, J. (2013). Is text written for children appropriate for L2 extensive reading? *TESOL Quarterly, 47*(2), 300–322. https://doi.org/10.1002.tesq.70

Wright, B., & Linacre, J. M. (1992). Combining (collapsing) and splitting categories. *Rasch Measurement Transactions, 6*(3), 233–235. https://www.rasch.org/rmt/rmt63f.htm

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370. https://doi.org/10.1177/0013164410387335