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

Some Trends in Vocabulary Research: A Discussion of Three Papers Presented at the JALT Vocabulary SIG

Kristopher Kyle
University of Oregon

Abstract

This paper represents a summary and discussion of the three studies presented at the morning session of the Fall 2022 JALT Vocabulary SIG. The papers, which were written by Ali Al-Hoorie, Masaki Eguchi, Derek Canning, Stuart McLean, Christopher Nicklin, and Joseph Vitta, represent a range of topics and study designs that are common in recent vocabulary research. These include a systematic scoping review, a vocabulary assessment validation study, and a study of the relationship between features of productive lexical use and oral proficiency. For each study, a brief summary, followed by a discussion of the particularly admirable qualities of each study and suggestions for future research have been provided. Furthermore, the studies are discussed in light of the open science movement.

Keywords: lexical sophistication, lexical richness, research synthesis, scoping review, vocabulary size test, substantive validity, Rasch analysis, open science

1 Introduction

The Japanese context is a particularly rich area for a wide range of vocabulary research. I was therefore quite honored (and humbled) to be asked to serve as one of the discussants for the 2022 JALT Vocabulary SIG. The morning session of the Fall 2022 Vocabulary SIG included three papers that represent a range of topics that are common in vocabulary research. The first study (Eguchi, 2022) represents recent research on multivariate approaches to lexical sophistication, and highlights the importance of index selection, statistical choices, and the consideration of register when interpreting results. The second study (Al-Hoorie et al., 2022) highlights the wealth of vocabulary learning research that has been conducted in the last 30 years through a scoping review of research syntheses. The third study (Canning et al., 2022) represents an evaluation of one aspect of a substantive validity argument for a well-known receptive vocabulary knowledge test. The study highlights the idea that validity arguments are multi-faceted and that the evaluation of the validity of an assessment is an iterative, context-specific process, and is not represented by a one-time, one-size-fits-all study. Below, a brief summary of each study is provided, highlighting the particularly admirable qualities of each followed by suggestions for future research.

Please cite this article as: K. Kyle. (2022). Some trends in vocabulary research: A discussion of three papers presented at the JALT Vocabulary SIG. *Vocabulary Learning and Instruction, 11*(2), 38–47. https://doi.org/10.7820/vli.v11.2kyle

1.1 Open Science and Second Language (L2) Vocabulary Research

The open science (OS) movement broadly represents a turn toward transparency in research practices, and has been bolstered by the increase in tools (e.g., R and Rstudio) and repositories (e.g., IRIS and OSF) that facilitate the sharing of analytic methods and data. While a brief overview of OS practices are provided below, the readers are referred to in-depth treatments of OS in applied linguistic research such as Marsden (2012, 2018, 2019), In'Nami et al. (2022), and Marsden and Mackey (2014), among others. Importantly, OS practices do not represent a single (and comprehensive) practice but instead represents a spectrum of research practices. At the more comprehensive end of the spectrum, all data collection and analytical methods, along with the raw and processed data, is open and freely available in an online repository. At this end of the spectrum, subsequent researchers could follow each step that the original researchers took and end up with the same results. Alternatively (for example), subsequent researchers could also make changes in the analytical procedures to determine if particular aspects of the analysis substantively affected the results. However, as a participant in the JALT Vocabulary SIG rightly pointed out, there are often insurmountable barriers that make the sharing of all materials and/or data infeasible. Consent for the sharing of raw student production data, for example, may be difficult (or impossible) because of (for example) institutional policies. However, much data collection and analysis procedures and most numerical data can be shared, even in more restrictive environments. Even if only numerical data and the statistical analysis code (e.g., in R or Python, though this is also possible if unwieldy with programs such as SPSS) in a quantitative study can be shared, this represents OS. Given the growing importance of OS in the field of L2 vocabulary studies and the fact that each author team independently made the choice to engage with OS in various ways, I will also comment on the positive aspects of the studies with regard to OS and provide some suggestions for future research endeavors.

2 Study 1: Modeling Lexical and Phraseological Sophistication in Oral Proficiency Interviews: A Conceptual Replication

A number of studies over the past 10 years have expanded our understanding of the construct of lexical sophistication (e.g., Crossley et al., 2011; Kyle et al., 2018; Kyle & Crossley, 2015; McNamara et al., 2010), and have suggested that lexical sophistication is a multifaceted construct. One downside to the flood of lexical sophistication indices that have been proposed based on research in psycholinguistics and educational psychology and introduced in tools such as the Tool for the Automatic Analysis of Lexical Sophistication (Kyle & Crossley, 2015; Kyle et al., 2018) is that it can be particularly difficult to choose which indices to include in an analysis. One approach to dealing with this issue is to identify the subconstructs of lexical sophistication that are pertinent to one's research questions and then select a representative index for each subconstruct (e.g., based on theoretical rationale and previous empirical findings). Another is to use a technique such as exploratory factor analysis (EFA) to combine indices into latent factors (e.g., Eguchi & Kyle, 2020; Kim et al., 2018). In this study, Eguchi (2022) combines both of these techniques in a principled replication of Eguchi and Kyle (2020) by using a refined set of indices (based on previous research), and an EFA as a variable reduction technique. The resulting factors are then used to investigate the relationship between oral proficiency interview (OPI) scores (*n* = 85) and lexical/phrasal sophistication. The results of a Bayesian ordinal mixed-effect regression, which achieved a prediction accuracy that was on par with human inter-rater reliability, indicated that six lexical sophistication factors were meaningfully related to OPI scores. The results supported some aspects of previous findings. For example, the results indicated that more proficient L2 speakers tended to use words that are considered to be learned later (based on Age of Acquisition ratings, see Kuperman et al., 2012) but also tend to use more frequent words (e.g., verbs that serve important communicative functions in an OPI setting). These results support a number of previous studies (including Eguchi & Kyle, 2020) which have found that more proficient L2 users tend to use more frequent lexical items than less proficient L2 users when completing spoken tasks. The results also diverged in some ways from previous research. For example, previous research has consistently indicated that more proficient L2 writers and speakers tend to use word combinations (broadly construed) that are more strongly associated than less proficient users (e.g., Kyle & Crossley, 2017; Kyle & Eguchi, 2021; Paquot, 2019; Rubin et al., 2021). In this study, however, the opposite trend was observed, which (as explained by Eguchi) may be because of the reliance on pre-fabricated chunks (which are strongly associated in corpus data) by lower proficiency users.

2.1 Particularly Admirable Qualities of the Paper

There are a number of aspects of this paper that are appreciable. First, instead of using common (but less appropriate) approaches to statistical analysis of categorical outcome data, Eguchi (2022) used Bayesian mixed effects models. Importantly, despite the limited space available in this manuscript format, Eguchi (2022) provided clear, rationale for each statistical choice, and also clearly and competently interpreted the results (with helpful visualizations). Second, the author conducted a principled replication that was conceptually related to previous work (e.g., Eguchi & Kyle, 2020; Kim et al., 2018), but included methodological choices that reflect the current state of the field (e.g., the inclusion of dependency bigram indices). Finally, as discussed in more detail below, Eguchi (2022) included general audience-friendly description of the methods and results in the manuscript while also including more technical methodological details in online supplementary materials.

2.2 Some Considerations for Future Research

One area that could be improved in future research is an explanation of why each index was included in the analysis and what the expected relationships between these indices and the OPI scores might be (based on theoretical perspectives and previous empirical research). While much of the methodological choices (e.g., statistical analyses) were clearly outlined, readers are presumed to be reasonably literate in the indices of lexical sophistication produced by TAALES. While this issue is likely because of length requirements, it would have been possible to add further descriptions in supplementary materials. Additionally, it would have been helpful to more comprehensively outline the ways in which the dataset used in the current study was similar to (and differed from) the study that it most closely replicates (e.g., Eguchi & Kyle, 2020).

2.3 Discussion of Open Science Principles

As discussed previously, OS represents a variety of practices ranging from preliminary transparency in analytical methods used to the provision of all data collection techniques, raw data, and analysis techniques. In this study, Eguchi (2022) uses a publicly available text analysis tool (and reports the version used, which is essential for replication), and provides the statistical analysis code used in an easy-to-read format that includes all results of the analyses in an Rmarkdown file rendered as .html. This file is hosted in an OSF repository that is publicly available (and was available to reviewers). However, because of data use agreements, Eguchi (2022) was not able to share the raw corpus or numerical data. Furthermore, the author did not provide the analysis code used to calculate the dependency bigram strength of association indices due to use agreements related to the distribution of COCA- derived frequency lists.

3 Study 2: Research Syntheses in L2 Vocabulary Research: A Scoping Review

In the second study, Al-Hoorie et al. (2022) conduct a “scoping” review of research synthesis studies (meta-analyses and systematic reviews) over the past 30 years. As they explain, a “scoping” review has a wider scope than a typical review of research, therefore in this study, the focus is on wider trends in vocabulary research (as opposed to a narrower focus on a particular topic). Their study highlights the wealth of L2 vocabulary research given that over 31 research syntheses that fit their search criteria were conducted during that period. Further, these research syntheses represented over 1,000 studies and over 85,000 participants (though there is likely overlap in studies and participants across the represented syntheses).

The findings of the scoping review highlight trends in L2 vocabulary research and indicate that more meta-analyses (25) have been conducted than systematic reviews (7). Within meta-analyses, the majority (21) were experimental in nature (e.g., pedagogical intervention designs) while only a few (4) were correlational. Within systematic reviews, a majority (4) were related to methodological issues, while the remaining three had various other foci. The review also highlighted topical trends within research syntheses in L2 vocabulary research over the past 30 years. Using a bottom-up coding system, Al-Hoorie et al. (2022) found that a little over half of the syntheses represented three topical categories. The most common topic was technology use and vocabulary learning (k = 9), followed by the effects of glossing vocabulary items (k = 5), and vocabulary knowledge as a predictor of reading ability (k = 5). Perhaps the most important finding of the study was that bulk of the syntheses focused on first-generation research questions (e.g., "is a pedagogical treatment effective?"; see Zanna & Fazio, 1982). For example, the meta-analyses on the aforementioned topics broadly indicated that technology and glossing enhance vocabulary learning, and that vocabulary knowledge is a predictor of reading (and listening) ability. Importantly, Al-Hoorie et al. (2022) indicate that it is time for L2 vocabulary research synthesizers to focus on second generation questions (e.g., "under what circumstances is a treatment effective?") and third generation questions (e.g., "how does a treatment promote learning?").

3.1 Particularly Admirable Qualities of the Paper

There are a number of aspects of this study that are to be applauded. First, it is very helpful to take a large step back in order to get a fuller view of the trends, strengths, and weaknesses of a particular field, which can be accomplished via a scoping review. In this case, the scoping review allows us as a field to identify where we have been putting our effort (e.g., in experimental, interventionist studies focused on technology and vocabulary learning) and some areas to which we might want to turn (such as addressing second and third generation questions). The methodological detail that was given in the paper is to be applauded (and emulated) by other researchers. Al-Hoorie et al. (2022) give a clear and concise description of their study selection and study analysis methods. This includes a detailed account of the manuscript search, the inclusion/exclusion criteria, and the number of manuscripts that were considered at each decision point. Helpfully, they also share their criteria in the IRIS database.

3.2 Some Considerations for Future Research

As future research is conducted in this area, there are some aspects that might be considered. Al-Hoorie et al. (2022) make a number of suggestions for future research syntheses including a focus on second and third generation questions, a greater focus on correlational studies, and (for meta-analytical studies) the use of *Q*-tests and/or related inferential tests. Another aspect of research syntheses (including meta-analyses, systematic reviews, scoping reviews) that might be considered is how methodological choices during the manuscript selec-tion process may impact the findings. For example, Al-Hoorie et al. (2022) chose to limit their search to manuscripts that were published in the Web of Science (WoS) Core collection. While there are many advantages to this methodolog-ical choice (well-known and arguably high-quality journals are represented, the search is objective and replicable, no duplicates were found), there are also possible disadvantages. For example, via this criterion some researchers (and research contexts) are silenced/overlooked. For example, any research synthe-sis published in this journal would have been excluded because it is not part of the WoS Core collection. One possible way to merge the two approaches would be to primarily rely on a particular publication repository (such as WoS Core) and then supplement this search with additional important vocabulary research output venues (perhaps identified via a survey of a wide range of L2 vocabulary researchers).

3.3 Discussion of Open Science Principles

As highlighted above, Al-Hoorie et al. (2022) made a considerable effort to clearly and transparently report their analytical methods. The authors clearly outlined their search domain (i.e., articles in WoS Core) and also included their exact search string (which helps reduce ambiguity in how the search was conducted). They also shared their manuscript coding criteria via the IRIS database, and indicated in their manuscript how many articles were included/excluded at each step in the process. One possible way in which the materials could be even more aligned with OS principles would be to include a list of each manuscript considered and why each manuscript was included or excluded. This would be particularly helpful given the fact that I was not able to perfectly replicate the initial WoS Core search (even after consultation with the authors). While the inconsistency is because of behind-the-scenes issues with the WoS Core search engine and is not the fault of the authors, determining whether the inconsistency would have affected the results is impossible since the raw list of identified articles was not shared.

4 Study 3: Rater Judgments and Word Difficulty: Conceptualizing the Substantive Validity of the VST

The frequency of a word in a language learner's input (which is often estimated using reference corpus frequencies) is an important factor in the difficulty/ ease with which an individual learns that word. Words that are more frequent in the input are more likely to be learned earlier, while words that occur less frequently are more likely to be learned later (see, e.g., Ellis, 2002 inter alia). When input frequency is estimated using reference corpora (e.g., Brysbaert & New, 2009; Laufer & Nation, 1995; Nation & Beglar, 2007), frequency is an intuitive measure of word difficulty that is psycholinguistically defensible, and is relatively easy to measure and to implement in pedagogical materials. As such, word frequency has played a major role in applied linguistics research, including the estimation of a learn-er's vocabulary size via assessment tools such as the Vocabulary Size Test (VST) (Nation & Beglar, 2007). Despite the importance of frequency, research in psy-cholinguistics (e.g., Balota et al., 2007), word difficulty assessment tool validation (e.g., Hashimoto & Egbert, 2019), and lexical sophistication (e.g., Kyle & Crossley, 2015) has indicated that word characteristics beyond frequency (e.g., concreteness, contextual distinctiveness) contribute to word difficulty estimation models.

In this study, Canning et al. (2022) contribute to the discussion of word difficulty and the validity argument of the VST (Beglar, 2010; Nation & Beglar, 2007) by investigating evidence related to substantive validity (Messick, 1989) using both quantitative and qualitative methods. Specifically, they investigate the degree to which experienced English as a foreign language (EFL) instructors' (n = 31) judgements of word difficulty align with word difficulty based on a large sample (n = 2,999) of VST results from Japanese L1 speakers of L2 English. They then conducted follow-up semi-structured interviews with 21 of the EFL instructors to investigate why particular difficulty ratings were assigned to a word. The quantitative results indicated that instructor judgments were strongly (but not perfectly) correlated with difficulty scores based on VST results (r = 0.67; R^2 = 0.45). The results of the semi-structured interviews indicated that EFL instructors' ratings were based on a wide range of factors. The most frequently mentioned factor was conceptual difficulty (see Brysbaert et al., 2014 for a treatment of concreteness, which is one objective measure of conceptual difficulty), which was followed by frequency and parallels in L1 (which was inclusive of cognates). As Canning et al. (2022) explain in their discussion, the results provide some positive evidence for the substantive validity of the VST given the strong correlation between word difficulty in the VST and instructor judgments of word difficulty (though see caveats below). The semi-structured interviews also provide further evidence (in this case from a rater cognition perspective) that the construct of word difficulty is multifaceted. For an overview of the possible factors beyond word frequency that contribute to word difficulty, I point readers to extant conceptual (e.g., Ellis, 2002; Nation, 2001) and empirical (e.g., Balota et al., 2007; Kyle et al., 2018) overview papers.

4.1 Particularly Admirable Qualities of the Paper

There are a number of aspects of this paper that should be applauded (and emulated). First, Canning et al. (2022) used advanced statistical analysis techniques (multi-faceted Rasch analysis; principal components analysis) and a qualitative approach (semi-structured interviews) in a complementary manner to address their research questions. While many researchers will often stop at the quantitative results (or conductive further quantitative analyses), very few will turn to grounded qualitative approaches despite the rich data (and explanatory detail) that can be gained. Second, the way in which the study was situated both within the auspices of validity argument work (e.g., Chapelle et al., 2008; Messick, 1989) and within current discussions of the L2 word difficulty (e.g., Hashimoto, 2021; Hashimoto & Egbert, 2019; Stewart et al., 2022) is highly appreciable.

4.2 Some Considerations for Future Research

As with any paper, there are some aspects that could be addressed in future work. First, as Canning et al. (2022) highlight, the coding of the qualitative data was only conducted by a single individual, which leaves some questions with regard to the reliability of the results. Ideally, the data would be coded by at least two annotators (and inter-coder reliability reported or the adjudication process discussed in detail). Second, although frequency and other factors related to word difficulty were discussed in relation to the qualitative results, they were not formally investigated quantitatively. While previous research has indicated that frequency and VST item difficulty are correlated (e.g., Beglar, 2010; Hashimoto, 2021; Hashimoto & Egbert, 2019; Stewart et al., 2022), it is unclear how strong this relationship is in this particular dataset. The quantitative results simply indicated that instructors' judgments of difficulty were strongly correlated with item-level word difficulty on the VST (which were not necessarily connected to specific frequency bands). As the authors note, a fruitful area for future research would be a multivariate (and theory-driven) approach to predicting word difficulty. Finally, it is unclear how generalizable the results of this study are outside of the Japanese EFL context. It would be helpful if future researchers replicated this study in other EFL and English as a Second Language (ESL) contexts.

4.3 Discussion of Open Science Principles

As noted earlier, all three papers made an effort to contribute to the OS movement in some way. Canning et al.'s (2022) contribution was the inclusion of online supplementary materials that contributed to transparency of the collection of rater judgments (e.g., by providing the rating scale used and providing a screenshot of the rating interface) and of the results of the multifaceted Rasch analysis (i.e., the inclusion of vertical rulers for each analysis and the fit analysis). While it may not be feasible to release all of the raw data because of use restrictions, one step that the authors might consider toward a fuller embrace of OS principals would be to (a) conduct analyses in R (including the Rasch analysis, which is possible in R) and (b) share their analysis code both via an .R script and as an RMarkdown (.rmd) file rendered as .html that includes the analysis code AND all data outputs and visualizations. This option allows for complete transparency in a quantitative data analysis without necessitating sharing raw data. Such files can be shared via an online repository such as osf.io, which allows users to share anonymized links to their repository during the review process. After a paper is accepted, the repository can be made public.

5 Conclusion

The papers included in the morning session of the 2022 JALT Vocabulary SIG are representative of the high-quality research that is characteristic of the vocabulary research community in the Japanese context. They represent three important methodological and topical trends in vocabulary research (systematic reviews, assessment tool validation, lexical sophistication) and include a number of features that should be emulated in future research studies. These papers also highlight a number of areas for future researchers to consider in upcoming projects. Finally, each paper includes an acknowledgment of the growing importance of OS principles and provided an opportunity to discuss some ways in which the field can increasingly embrace the OS movement.

References

Al-Hoorie, A. H., Vitta, J. P., and Nicklin, C. (2022). Research syntheses in L2 vocabulary research: A scoping review. *Vocabulary Learning and Instruction, 11*(2), 17–29. https://doi.org/10.7820/vli.v11.2.al-hoorie

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, *39*(3), 445–459. https://doi.org/10.3758/BF03193014

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

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911. https://doi.org/10.3758/s13428-013-0403-5

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

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). *Building a validity argument for the Test of English as a Foreign Language*. Routledge.

Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011). Predicting lexical proficiency in language learner texts using computational indices. *Language Testing*, *28*(4), 561–580. https://doi.org/10.1177/0265532210378031

Eguchi, M. (2022). Modeling lexical and phraseological sophistication in oral proficiency interviews: A conceptual replication. *Vocabulary Learning and Instruction, 11*(2), 1–16. https://doi.org/10.7820/vli.v11.2.Eguchi

Eguchi, M., & Kyle, K. (2020). Continuing to explore the multidimensional nature of lexical sophistication: The case of oral proficiency interviews. *The Modern Language Journal*, *104*(2), 381–400. https://doi.org/10.1111/modl.12637

Ellis, N. C. (2002). Frequency effects in language processing. *Studies in Second Language Acquisition*, *24*(02), 143–188. https://doi.org/10.1017/S0272263102002024

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

In'nami, Y., Mizumoto, A., Plonsky, L., & Koizumi, R. (2022). Promoting computationally reproducible research in applied linguistics: Recommended practices and considerations. *Research Methods in Applied Linguistics*, *1*(3), 100030. https://doi.org/10.1016/j.rmal.2022.100030

Kim, M., Crossley, S. A., & Kyle, K. (2018). Lexical sophistication as a multidimensional phenomenon: Relations to second language lexical proficiency, development, and writing quality. *The Modern Language Journal*, *102*(1), 120–141. https://doi.org/10.1111/modl.12447

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*(4), 978–990. https://doi.org/10.3758/s13428-012-0210-4

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. (2017). Assessing syntactic sophistication in L2 writing: A usage-based approach. *Language Testing*, *34*(4), 513–535. https://doi.org/10.1177/0265532217712554

Kyle, K., Crossley, S. A., & Berger, C. M. (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

Kyle, K., & Eguchi, M. (2021). Automatically assessing lexical sophistication using word, bigram, and dependency indices. In *Perspectives on the L2 Phrasicon: The view from learner Corpora* (pp. 126–151). Multilingual Matters.

Laufer, B., & Nation, I. S. P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, *16*(3), 307–322. https://doi.org/10.1093/applin/16.3.307

Marsden, E. (2012). Open science and transparency in applied linguistics research. *The Encyclopedia of Applied Linguistics*, 1–10. https://doi.org/10.1002/9781405198431.wbeal1493

Marsden, E. (2019). Methodological transparency and its consequences for the quality and scope of research. In J. McKinley & H. Rose (eds.), *The Routledge handbook of research methods in applied linguistics* (pp. 15–28). Routledge.

Marsden, E., & Mackey, A. (2014). IRIS: A new resource for second language research. *Linguistic Approaches to Bilingualism*, *4*(1), 125–130. https://doi.org/10.1075/lab.4.1.05mar

Marsden, E. J. (2018). Open science and transparency in applied linguistics. In *The concise encyclopedia of applied linguistics*, C. Chapelle (ed.), Wiley.

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. *Written Communication*, *27*(1), 57–86. https://doi.org/10.1177/0741088309351547

Messick, S. (1989). Validity. In *Educational measurement* (3rd ed., pp. 13–103). American Council on Education, R.L. Linn (ed.).

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge University Press.

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

Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. *Second Language Research*, *35*(1), 121–145. https://doi.org/10.1177/0267658317694221

Rubin, R., Housen, A., & Paquot, M. (2021). *Phraseological complexity as an index of L2 Dutch writing proficiency: A partial replication study*. In S. Granger (Ed.), Perspectives on the second language phrasicon: The view from learner corpora (pp. 101–125). Multilingual Matters.

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

Zanna, M. P., & Fazio, R. H. (1982). *The attitude–behavior relation: Moving toward the third generation of research*. In M. P. Zanna, E. T. Higgins, & C. P. Herman (Eds.), Consistency in social behavior: The Ontario Symposium (Vol. 2, pp. 283–301). Hillsdale, NJ: Erlbaum.