



Rigor and Self-Awareness in L2 Instructed Vocabulary Research: A Commentary on Isbell and colleagues (2024), Stoeckel and Ishii (2024), and Aizawa (2024)

Joseph P. Vitta

Waseda University – Tokyo, Japan

vittajp@waseda.jp

Abstract

In this discussion paper, I examine three exceptional articles presented at the 2023 Symposium of the *Japan Association for Language Teaching's* Vocabulary Special Interest Group. These papers are analyzed from the standpoint of providing positive examples for future vocabulary researchers to follow in order to promote rigor in their work. I also discuss how these reports demonstrate a clear self-awareness regarding the conclusions their inquiries can and cannot support. Taken together, these three reports are invaluable for promoting better vocabulary research.

Keywords: L2 methods reform, L2 vocabulary, Testing and measurement.

Introduction

It was an honor to be invited to serve as a discussant for the 2023 Vocabulary Symposium hosted by the Vocabulary Special Interest Group of the *Japan Association for Language Teaching*. To join the ranks of such esteemed scholars as Kyle (2022), Larson-Hall (2021), and Nakata (2020) is both humbling and gratifying. In the strand for which I was tasked to serve as a discussant, there were three exceptional quantitative research

papers relating to vocabulary testing and learning issues: Isbell et al. (2024), Stoeckel and Ishii (2024), and Aizawa (2024). When reviewing these papers and crafting a theme for my discussion presentation (and this commentary piece), it became clear that these studies were useful for engaging with two potential areas for improvement within L2 vocabulary (quantitative) research: rigor and self-awareness.

Rigor and Self-Awareness in L2 Vocabulary Quantitative Research

The call for L2 researchers to improve the demonstrated rigor in their work dates back several decades (Gass et al., 2021). An entire strand of L2 research has been established with the aim of understanding quantitative research quality in the field while providing recommendations for improvement (Plonsky, 2013, 2014; Al-Hoorie & Vitta, 2019). The call for reform is broad, with research synthesis and methodological commentary work focusing on various issues such as the handling of outliers (Nicklin & Plonsky, 2020), the ethical management of data (Isbell et al., 2022), and the need to promote research with populations outside of the WEIRD (White, Educated, Industrialized, Rich, and Democratic) circle (Andringa & Godfroid, 2020). As Gass and colleagues (2021) argued bluntly, the field will fail to achieve its true potential until such a methods reform is truly realized. L2 vocabulary quantitative research is likewise constrained until such reform occurs (see arguments in Vitta et al., 2022).

Language motivation research offers a cautionary tale for when L2 researchers fail to heed the call to improve and refine their measures. Al-Hoorie et al. (2023) showed that popular surveys used to operationalize the constructs of the L2 motivational self system (L2MSS) (e.g., You et al., 2016) produced measurements that failed the most checks for divergent validity, i.e., the constructs that the L2MSS theory proposed as being distinct were actually observed to potentially measure the same ‘thing.’ Al-Hoorie et al.’s observations fit into a growing trend of research where actual research findings and what L2MSS theory would predict were not aligned. When discussing the ubiquity and somewhat unsubstantiated trust in L2MSS theory, Al-Hoorie and colleagues pointed to the overreliance on precedence and trust in individual scholars on a personal level. Ideally, motivation researchers would have the expectation and practice of considering the “empirical demonstration of measure validity” (Al-Hoorie et al., 2023, p. 324) when reviewing and publishing L2MSS research. The conclusions of Al-Hoorie and colleagues were mostly accepted by Henry and Liu (2024) with the caveat that the construct might not have been operationalized relative to theory in the prior seminal work of the past decades. What this cautionary tale can highlight for vocabulary researchers is that the onus is on us to consistently consider and challenge the assumptions and trends that the preceding body of work has established for us. After all, validity and psychometrics constitute a ‘long game’ where a one-off measurement can never ensure that a measurement and the inferences made from it are valid (Vitta et al., 2023).

When I use the term ‘self-awareness’ in the context of L2 vocabulary research, I refer to the researcher(s) having a clear understanding of how the aspects of their design and data influence their presentation and analysis of results. Sampling practices provide a good example of how researchers can be aware and unaware in this regard. When inferential testing is performed on data from single sites, the inferences from

this data only generalize to the site from which the sample has been drawn (Vitta & Al-Hoorie, 2021). To wit, single-site designs have been taken to task as having severe external validity limitations (Moranski & Ziegler, 2021). When reviewing vocabulary experiment designs, Vitta and colleagues (2022) observed that approximately 85% of the samples were single-site; trends drawn from this body of research should be interpreted with caution. Where the concept of ‘self-awareness’ comes into play is that a research design defect is not necessarily a negative as long as the researcher acknowledges it and can still argue for the relevance of their work. Zhang (2022), for instance, clearly noted its single-site design as a limitation but was still able to demonstrate the research’s relevance in relation to vocabulary development through engagement with teachers’ spoken instruction.

To sum up, there is a need for L2 vocabulary quantitative researchers to be as committed to rigor as much as possible while being self-aware in relation to what their work can and cannot do when it comes to implications. In this commentary article, I discuss Isbell et al. (2024), Stoeckel and Ishii (2024), and Aizawa (2024) through this lens.

Commentary

Each study is discussed in a systematic manner where a summary is first presented. Then, I comment on the rigor and (researcher) self-awareness observed in the study. Finally, additional comments and a consideration of how the report can act as a positive example for L2 vocabulary researchers is presented.

Isbell et al. (2024)

Summary. Isbell et al. (2024) presented the construction processes for two L2 Korean vocabulary tests: one for meaning recall and another for meaning recognition. The researchers described this process in an engaging manner, adroitly interweaving empirical reliability and validity evidence with narratives of the creation process. In general, the researchers presented evidence for two psychometrically sound vocabulary tests.

Rigor and Self-Awareness. Regarding rigor, there are many points to admire in this work. The researchers correctly juxtaposed construct validity evidence (e.g., internal reliability and arguments for item selection) with predictive (or criterion) validity evidence (e.g., how vocabulary measures from the tests predicted L2 proficiency) to argue for the trustworthiness of their tests’ measures (see Schimmack & Diener, 2003). In the narrative sections of the text about how they created these tests, the authors were clear about how they borrowed from the processes of already established tests. Additionally, the researchers provided clear visuals of how the tests would appear to test takers. In sum, the paper demonstrated a clear commitment to rigor and transparency.

Regarding self-awareness, I was impressed by the researchers’ clear description of why they were presenting this work. When the researchers stated that “we came to the realization that no one was going to build Korean vocabulary tests for us” (Isbell et al., 2024, p. 8), it helped me to understand the validation work presented while also

being more accepting of the limitations of the work that I perceived. Another aspect of self-awareness that the researchers demonstrated was that they clearly acknowledged that they were borrowing from procedures and materials originally designed for L2 English (e.g., the Vocabulary Levels Test: Webb et al., 2017). While the authors did not explicitly state this as a limitation, the reader is made fully aware of the crossover from L2 English to L2 Korean. As a final note, the researchers had the good sense and thereby self-awareness not to comment on the debate of meaning recall versus meaning recognition modality as the better predictor of L2 proficiency (Stewart et al., 2024), as their data was drawn from samples that were not clearly labeled as multi-site and the project was not undertaken to answer this question.

Further Comments and Future Considerations. The authors' interweaving of their 'story' of developing the test with the empirical presentation of validity evidence is something that can enhance future presentations of test development research. As highlighted in the preceding text, the authors present a robust and comprehensive defense of their two vocabulary tests. Future researchers developing vocabulary tests would do well to follow this example. One thought I had while reviewing the empirical presentation of validation evidence was how assumptions were considered. Cronbach's alpha, for instance, was used in the paper, and this test assumes unidimensionality but cannot check for it (Al-Hoorie & Vitta, 2019). While the researchers did not present such unidimensionality checks, future researchers might consider doing this. In fairness to the researchers, however, the factor analysis procedures for such checks require more subjects than the researchers apparently had access to (see heuristics and power considerations in Vitta et al., 2023). The assumptions of the parametric testing presented in the paper could likewise have been expanded on, but it is important to remember that the general linear model is very robust against violations (Rasch & Guiard, 2004). One area to also consider is better sampling procedures in terms of power and representativeness (Vitta et al., 2022), but such criticisms are difficult to make when the target language is Korean, as the researchers probably do not have access to large participant pools. These small areas for future improvement notwithstanding, Isbell et al. (2024) provide a strong example and framework for L2 vocabulary test developers, especially those seeking to develop tests for non-English L2s.

Stoeckel and Ishii (2024)

Summary. Stoeckel and Ishii (2024) aimed to evaluate learners' ability to recall the meanings of words within a reading passage. Given the variety of existing meaning-recall test formats, their small-scale study, involving 10 participants, sought to identify which of three test formats, each with 30 target items, most closely aligned with a criterion interview in terms of mean score and the consistency of correct/incorrect classifications. In Test 1, participants were prompted with only the target word and asked for its written translation. Test 2 presented a short sentence with the target word highlighted, requesting a written translation of just that word. Test 3 used the same sentence as Test 2 but did not highlight the target word, and participants were asked to translate the entire sentence. The criterion interview required participants to

demonstrate their understanding of the target words within the same sentences used in Tests 2 and 3. The findings revealed that Test 3 yielded mean scores and match rates most similar to the interview, followed by Test 2, with Test 1 showing the least similarity. This order of accuracy (Test 3: best; Test 1: worst) was contradicted by the observed internal reliability values for each test, whose utility the researchers rightly questioned given the very small sample.

Rigor and Self-Awareness. The motivation for the study was impressive in trying to further probe the best way to operationalize the meaning-recall modality with the aim of using the data for larger-scale research. In other words, the researchers were aware of the need to conduct multifaceted small-scale yet rigorous research to inform their future large-scale work. The authors, in a general sense, have been at the forefront of refining and honing meaning recall items (see e.g., Stoeckel et al., 2023), and this research adds to their body of work on vocabulary testing modality, especially meaning recall test items. Regarding rigor in a specific sense, the paper presents strong arguments and connections with the relevant literature to highlight the motivations for their three meaning-recall test formats and why there is a need to investigate the differences among them in terms of accuracy and utility (given $N = 10$, I will avoid using psychometric terms). Most refreshingly, the researchers used interview data to infuse their inquiry with a robustness and nuance that quantitative validation work alone cannot provide. This interview data proved useful in analyzing the performance of the three different iterations of their meaning-recall test. The paper's strong methodological design within the parameters of a clearly presented small-scale study (i.e., the researchers were open and self-aware of the design) demonstrates how being cognizant of one's research design and its limitations and advances can enhance the research's rigor.

Further Comments and Future Considerations. Stoeckel and Ishii (2024) provide a needed model for researchers with interesting questions constrained by practical limitations. It is indeed possible to probe new ground with small-scale research. The trick is to be clever with the resources and methods at one's disposal while being open and honest about limitations. Just as the researchers of this report were able to have readers rethink the way we test meaning-recall knowledge with an adroit small-scale study, future vocabulary researchers can emulate this report to present small-scale yet impactful research.

Aizawa (2024)

Summary. This study explored the contrast between loanwords, i.e., words that exist in the participants' first language as well as the target language, and non-loanwords in the VST (Nation & Beglar, 2007). In line with past research into the area, Aizawa observed that loanwords were easier across all 14 levels (1K bands) on the VST with a large effect observed, $d_{av} = 5.22$. Classical test theory psychometric checking was adequate to accept the validity of the results. One interesting finding in this regard was that the levels did not disturb the 'elbow distribution' suggesting unidimensionality across the three levels.

Rigor and Self-Awareness. The observed rigor in this study was indeed exceptional, especially as this work was derived from a master's thesis. Regarding sampling, there were multiple sites, and the researcher correctly defined the population as encompassing high-proficiency Japanese learners of English. The reporting of results was clear and transparent. When appropriate, psychometric checks were presented in a complete and accurate manner. For instance, unidimensionality checks complemented presentations of Cronbach's alpha. The researcher also exhibited a rigorous grounding of the study in the literature by citing West (1930), one of the first articles to present and support the frequency principle and the need to teach vocabulary in a deliberate manner—a seminal work that even Laufer and Nation (1995) failed to cite.

Regarding self-awareness, the researcher presented clear constraining text about the implications of her work. In the presentation of the participants, one reads: "It can be said that the participants are advanced English users. Therefore, the results of this study cannot represent a wider range of L1 Japanese university students and need to be interpreted with caution" (Aizawa, 2024, p. 4). Such overt acknowledgment of a study's limitations due to sample construction is something I have rarely seen, having reviewed hundreds of primary quantitative reports in both L2 vocabulary research syntheses (e.g., Nicklin & Vitta, 2021) and syntheses in other areas such as self-determination theory (e.g., Al-Hoorie et al., 2022). Another instance of self-awareness on the part of the researcher was her clear statement that alternative vocabulary testing item modalities might yield different results from those she was reporting.

Further Comments and Future Considerations. Aizawa (2024) demonstrated a strong positive example for vocabulary researchers to follow, and the attention to detail, despite the short nature of the report, was indeed impressive. As with Isbell et al. (2024), more could have been done to report assumption checks of the underlying parametric inferential tests. Future researchers might also use sensitivity power analyses to determine the minimal effect size that the already constructed sample can detect (see Lakens, 2022; Vitta et al., 2022). Nevertheless, Aizawa provides a strong framework for presenting trustworthy research that I hope other Japan-based vocabulary researchers will use in their work.

References

- Aizawa, A. (2024). The impact of loanwords on the English-Japanese version of vocabulary size test. *Vocabulary Learning and Instruction*, 13(1), 1–8. <https://doi.org/10.29140/vli.v13n1.1234>
- Al-Hoorie, A. H., Hiver, P., & In'nami, Y. (2023). The validation crisis in the L2 motivational self system tradition. *Studies in Second Language Acquisition*, 46(2), 307–329. <https://doi.org/10.1017/s0272263123000487>
- Al-Hoorie, A. H., Oga-Baldwin, W. Q., Hiver, P., & Vitta, J. P. (2022). Self-determination mini-theories in second language learning: A systematic review of three decades of research. *Language Teaching Research*, 136216882211026. <https://doi.org/10.1177/13621688221102686>

- Al-Hoorie, A. H., & Vitta, J. P. (2019). The seven sins of L2 research: A review of 30 journals' statistical quality and their CiteScore, SJR, SNIP, JCR impact factors. *Language Teaching Research*, 23, 727–744. <https://doi.org/10.1177/1362168818767191>.
- Andringa, S., & Godfroid, A. (2020). Sampling bias and the problem of generalizability in applied linguistics. *Annual Review of Applied Linguistics*, 40, 134–142. <https://doi.org/10.1017/S0267190520000033>.
- Gass, S., Loewen, S., & Plonsky, L. (2021). Coming of age: The past, present, and future of quantitative SLA research. *Language Teaching*, 54, 245–258. <https://doi.org/10.1017/s0261444819000430>.
- Henry, A., & Liu, M. (2024). Jingle–Jangle Fallacies in L2 Motivational Self System Research: A Response to Al-Hoorie et al. (2023/2024). *Applied Linguistics*, 45(4), 738–746. <https://doi.org/10.1093/applin/amae041>
- Isbell, D. R., Brown, D., Chen, M., Derrick, D. J., Ghanem, R., Arvizu, M. N. G., Schnur, E., Zhang, M., & Plonsky, L. (2022). Misconduct and questionable research Practices: The ethics of quantitative data handling and reporting in applied Linguistics. *Modern Language Journal*, 106(1), 172–195. <https://doi.org/10.1111/modl.12760>
- Isbell, D. R., Park, L., & Delaney, M. (2024). Developing measures of L2 Korean vocabulary knowledge. *Vocabulary Learning and Instruction*, 13(1), 1–10. <https://doi.org/10.29140/vli.v13n1.1227>
- Kyle, K. (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>
- Lakens, D. (2022). Sample size justification. *Collabra Psychology*, Article 33267, 8(1). <https://doi.org/10.1525/collabra.33267>
- Larson-Hall, J. (2021). Discussion paper: Using statistics to solve practical vocabulary problems. *Vocabulary Learning and Instruction*, 10(2), 101–113. <https://doi.org/10.7820/vli.v10.2.larson-hall>
- Laufer, B., & Nation, 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>
- Moranski, K., & Ziegler, N. (2021). A case for multisite second language acquisition research: Challenges, risks, and rewards. *Language Learning*, 71, 204–242. <https://doi.org/10.1111/lang.12434>.
- Nakata, T. (2020). Vocabulary and computer technology: A commentary on four studies for JALT vocabulary SIG. *Vocabulary Learning and Instruction*, 9(2), 39–47. <https://doi.org/10.7820/vli.v09.2.nakata>
- Nation, I., & Beglar, D. (2007). A Vocabulary Size Test. *The Language Teacher*, 31, 9–13.
- Nicklin, C., & Plonsky, L. (2020). Outliers in L2 research in applied linguistics: A synthesis and data re-analysis. *Annual Review of Applied Linguistics*, 40, 25–55. <https://doi.org/10.1017/S0267190520000057>
- Nicklin, C., & Vitta, J. P. (2021). Effect-driven sample sizes in second language instructed vocabulary acquisition research. *The Modern Language Journal*, 105, 218–236. <https://doi.org/10.1111/modl.12692>.

- Plonsky, L. (2013). Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research. *Studies in Second Language Acquisition*, 35, 655–687. <https://doi.org/10.1017/S0272263113000399>
- Plonsky, L. (2014). Study quality in quantitative L2 research (1990–2010): A methodological synthesis and call for reform. *The Modern Language Journal*, 98, 450–470. <https://doi.org/10.1111/j.1540-4781.2014.12058.x>
- Rasch, D., & Guiard, V. (2004). The robustness of parametric statistical methods. *Psychology Science*, 46(2), 175–208.
- Schimmack, U., & Diener, E. (2003). Predictive validity of explicit and implicit self-esteem for subjective well being. *Journal of Research in Personality*, 37(2), 100–106. [https://doi.org/10.1016/S0092-6566\(02\)00532-9](https://doi.org/10.1016/S0092-6566(02)00532-9)
- Stewart, J., Gyllstad, H., Nicklin, C., & McLean, S. (2023). Establishing meaning recall and meaning recognition vocabulary knowledge as distinct psychometric constructs in relation to reading proficiency. *Language Testing*, 41(1), 89–108. <https://doi.org/10.1177/02655322231162853>
- Stoeckel, T., & Ishii, T. (2024). An exploratory criterion validation of three meaning-recall vocabulary test item formats. *Vocabulary Learning and Instruction*, 13(1), 1–10. <https://doi.org/10.29140/vli.v13n1.1233>
- Stoeckel, T., Ishii, T., Kim, Y. A., Hung, T. H., Ho, N. T. P., & McLean, S. (2023). A comparison of contextualized and non-contextualized meaning-recall vocabulary test formats. *Research Methods in Applied Linguistics*, 2(3), 100075. <https://doi.org/10.1016/j.rmal.2023.100075>
- Vitta, J. P., & Al-Hoorie, A. H. (2021). Measurement and sampling recommendations for L2 flipped learning experiments: A bottom-up methodological Synthesis. *Journal of Asia TEFL*, 18(2), 682–692. <https://doi.org/10.18823/asiatefl.2021.18.2.23.682>
- Vitta, J. P., Leeming, P., McLean, S., & Nicklin, C. (2023). Initial development of a second language Discussion Task-Specific Self-Efficacy Instrument: an illustration for Frontline researchers. *RELC Journal*, Advance Online Publication. <https://doi.org/10.1177/00336882231195102>
- Vitta, J., Nicklin, C., & McLean, S. (2022). Effect-size driven sample size planning, randomization, and multi-site use in L2 instructed vocabulary acquisition experimental samples. *Studies in Second Language Acquisition*, 44, 1424–1448. Advance online publication. <https://doi.org/10.1017/S0272263121000541>
- Webb, S., Sasao, Y., & Ballance, O. (2017). The updated Vocabulary Levels Test: Developing and validating two new forms of the VLT. *ITL – International Journal of Applied Linguistics*, 168(1), 34–70. <https://doi.org/10.1075/itl.168.1.02web>
- West, M. (1930). Speaking-vocabulary in a foreign language. *The Modern Language Journal*, 14(7), 509–521. <https://doi.org/10.1111/j.1540-4781.1930.tb05634.x>
- You, C., Dörnyei, Z., & Csizér, K. (2016). Motivation, vision, and gender: A survey of learners of English in China. *Language Learning*, 66, 94–123. <https://doi.org/10.1111/lang.12140>
- Zhang, P. (2022). How does repetition affect vocabulary learning through listening to the teacher's explicit instruction? The moderating role of listening proficiency and preexisting vocabulary knowledge. *Language Teaching Research*, Advance Online Publication. <https://doi.org/10.1177/13621688221140521>