Research Syntheses in L2 Vocabulary Research: A Scoping Review

Ali H. Al-Hoorie, Joseph P. Vitta, and Christopher Nicklin
Royal Commission for Jubail and Yanbu, Saudi Arabia;
Kyushu University; Rikkyo University

Abstract

In this brief report, we present a scoping review of 31 second language (L2) vocabulary research syntheses published in the Web of Science from January 1990 to May 2022. The purpose of this undertaking was to understand what L2 vocabulary research synthesizers have aggregated and the investigative foci they intended to address. It was observed that most of the cataloged reports were experimental design meta-analyses, and thus there was room for future research syntheses in the vocabulary space to consider correlational research questions and hypotheses. Because systematic reviews were relatively scarce, there is also a need for these types of syntheses to be conducted in the future, especially those focused upon the observed quality of aggregated L2 vocabulary research.

Keywords: L2 Vocabulary, Scoping Review, Research Synthesis, Meta-analysis, Systematic Review

1 Introduction

In the 1990s, the importance of vocabulary in both second language acquisition (SLA) research and in teaching began to take form. It would not be an overstatement to claim that vocabulary research has come to comprise one of the major strands of both instructed SLA research and foreign language teaching guidance (see review in Nicklin & Vitta, 2021). As the focus on vocabulary developed, the number of second language (L2) researchers conducting research syntheses has also grown. As noted by Raeisi-Vanani et al. (2022), research syntheses have gained popularity in many academic areas, including applied linguistics, to become a standard by which researchers aggregate the findings of past studies addressing a common research question (RQ) or hypothesis. In this brief report, we present a scoping review of the L2 vocabulary-focused research syntheses published over the last three decades with a subsequent consideration of future directions.

2 Background and Research Questions

When properly conducted, synthetic reviews, subsuming focused systematic reviews and broader scoping reviews, are grounded in literature-driven RQs (O'Connor et al., 2008; Vitta et al., 2021). To that end, the following RQs drove this current review and are presented with a brief summary of the theory and/or literature in which they are situated.

RQ1. What types of vocabulary-based research syntheses were conducted between January 1990 through May 2022?

1. Meta-analysis vs. Systematic reviews reviews?
2. For meta-analyses: Correlational vs. experimental?
3. For systematic reviews: What type (e.g., methodological synthesis)?

Regarding RQ1 and its sub-questions, a meta-analysis is the aggregation of effect sizes (e.g., Pearson's *r*) from studies addressing a common RQ or focus (Ahn & Kang, 2018). Lin and Lin (2019), for instance, aggregated effect sizes for experimental studies where the treatment group learned target vocabulary via mobile applications for language learning (MALL), and the comparison group studied these words without this technology. This meta-analysis is an example of an experimental meta-analysis in that the reports featured experimental designs where researchers manipulated the experiences of participants to ascertain the relative effects of vocabulary study with MALL. Correlational meta-analyses, such as Zhang and Zhang (2020), aggregate effect sizes from correlation studies focusing on a common research focus, which in their case was vocabulary's association with reading and listening.

As the name implies, systematic reviews feature a systematic process where past reports are selected and analyzed according to a predetermined process to address a RQ or set of RQs (O'Connor et al., 2008). What differentiates a systematic review from a meta-analysis is that the aggregation of effect sizes is not of primary interest, but rather the features and characteristics of the selected reports. Sometimes reviews are scoping in nature, which implies that they cover a broader "scope" than an ordinary, more focused systematic review (Munn et al., 2018). Scoping reviews are helpful in proving a general overview of a body of knowledge and in identifying gaps in it. Elgort (2018), for instance, reviewed 82 technology-mediated vocabulary development (TMVD) studies to investigate the methodology and approaches of researchers in this domain. Systematic reviews can also be purely methodology focused; in which case they are called methodological syntheses. For example, 110 experimental instructed vocabulary studies were examined by Vitta et al. (2021) to determine the extent to which researchers in this domain adhered to three basic methodological requirements of generalizable parametric testing: sample size planning, randomization, and multi-site participant recruitment.

RQ2a. What are the descriptive characteristics of these reports (e.g., size of report pool, aggregate sample size)?

RQ2b. To what extent did meta-analyses in the report pool employ inferential testing? Which tests were employed?

RQ2a was addressed to provide a summative scope of the L2 vocabulary research that has been synthesized thus far in terms, for example, of how many studies L2 meta-analysts generally collate for a report. RQ2b relates to the notion that meta-analysts should conduct inferential Q-tests to determine the extent to which the studies are estimating the same effect size while accounting for within-study random error (Borenstein et al., 2009). While Q-testing is the norm for meta-analyses, Yanagisawa and Webb (2021) published a meta-analysis where each

study appeared to be a case in inferential testing but where the sample size of the individual studies was used to correct the standard errors, and thus the "spirit" behind the *Q*-test procedure appeared to be respected.

RQ3. What areas have L2 vocabulary research syntheses investigated?

As the current report presents a brief scoping review, RQ3 was posed to investigate the areas that L2 vocabulary research synthesizers have considered. This process often features selective codings for the research foci of the reports cataloged in the report pool. Zou et al. (2021) for instance provided frequencies of the digital game types (e.g., simulation and role-playing games) found in the treatment and learning conditions for L2 vocabulary acquisition studies. RQ3 in a similar vein was addressed by a selective coding categorization process where studies were grouped according to their investigative foci. This process led to the final RQ (i.e., set of sub-questions) governing the study:

RQ4a. When reviewing meta-analyses in the pool, which areas emerge as being conclusively addressed? and which areas emerge as requiring further investigations?

RQ4b. What are apparent gaps in the existing body of L2 vocabulary systematic reviews?

RQ4a and RQ4b were post-hoc and therefore exploratory and speculative in nature but are nonetheless useful for addressing areas where second ("when?" or "under what conditions?") and third ("how?") generation RQs (Zanna & Fazio, 1982) would be warranted. Few L2 reports have considered this notion and it was absent from our vocabulary-based report pool. In one of the few L2 reports to consider Zanno and Fazio's framework, Vitta and Al-Hoorie (2020) presented a suggestion that since the effectiveness of flipped classrooms was somewhat conclusive (i.e., $g = 0.58$ when corrected effect for publication bias), future researchers should pivot toward investigating *when* and *how* flipped classrooms worked in L2 settings, which are second and third generation questions, as opposed to *if* they worked, which is a first-generation question.

3 Method

3.1 Report Pool Creation

The Web of Science was searched on 09 May 2022 using the parameters listed in Table 1. The automated search selected 219 reports. One researcher screened the abstracts of these 219 selected reports and selected 34 for further investigation. A second researcher screened the same set and there were only five disagreements observed (i.e., 97.71% agreement; $\kappa = 0.91$) which was quickly resolved through discussion. At the end of the report pool creation process, 31 reports were included in the finalized report pool. The reliability was high most likely because of the small number of abstracts and the fact that the coders have experience in coding research syntheses.

Table 1. Report Pool Search Parameters

Feature	Details and explanation
Time period	01 January 1990 to 09 May 2022 – start date is beginning of Web of Science (WoS) Coverage
Indexes search	WoS Core Collection
Search terms	((meta-analysis) OR (systematic review) OR (research synthesis) OR (method* synthesis)) AND ((L2) or (second language) or (foreign language)) AND ((vocabulary) OR (lexi*)) - "topic" search: title, abstract, keywords, and *keywords plus* (a unique WoS feature) - Three sets of terms (i.e., research synthesis, L2, vocabulary terms) - L2 terms taken from Vitta and Al-Hoorie (2020) - Vocabulary terms from Nicklin and Vitta (2021)
Selection criteria	1. Report is a research synthesis report. 2. Report focuses on both L2 learners and vocabulary • Vocabulary featured alongside other constructs such as grammar (e.g., Jeon & Yamashita, 2014) retained. 3. Report is written in English.

Note: In line with Al-Hoorie et al. (2022), the automated WoS search was conducted via a "topic" search to limit the number of false positive reports that would be selected at this stage of the report pool creation process.

Figure 1 details the process of the report pool process according to PRISMA (Page et al., 2021) standards. It is noteworthy that the WoS focus resulted in no duplicated reports for omission and in general a much more straightforward process. In line with past review projects in L2 research (e.g., Al-Hoorie et al., 2022; Hiver et al., 2021), the WoS restriction was used to isolate what many consider to be the best the field has to offer in terms of quality (see argument of Zhang, 2020, citing Al-Hoorie & Vitta, 2019). This restriction was further justified by the stated RQs that were posed to inform future research and practices.
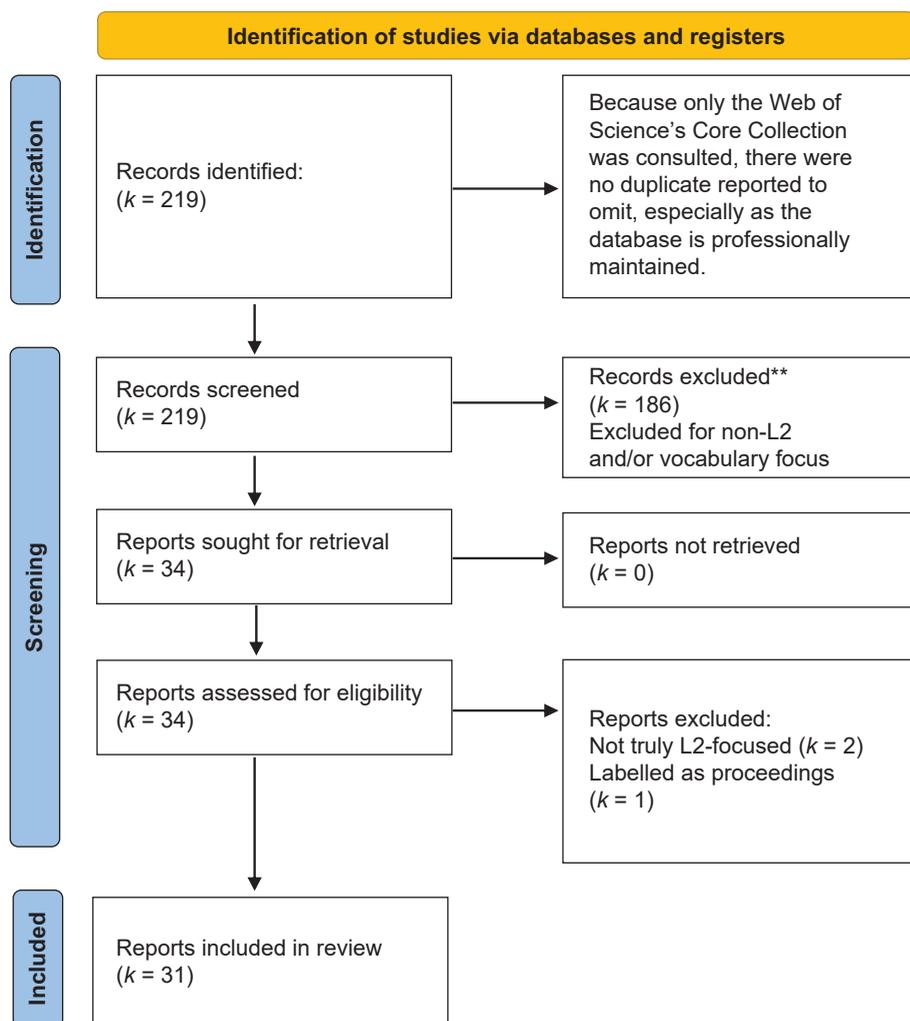
3.2 Coding

The 31 selected reports were coded by two researchers in tandem following the protocols stated in Al-Hoorie et al. (2022) as there were numerous categories requiring judgments. The tandem coding reduced the potential for individual variation of judgments and the same researchers had a high degree of reliability observed when independently coding the 219 abstracts during the screening process. The codes can be found in the shared data sheet on: https://www.iris-database.org/iris/app/home/detail?id=york%3a940835&ref=search, where each judgment is subsumed under its governing RQ.

4 Results and Discussion

4.1 RQ1: Scope of L2 Vocabulary Research Syntheses

Of the 31 coded reports, there were 24 meta-analytic research syntheses, 6 systematic reviews, and 1 coded as being both (see Table 2 for details). Besides

Identification of studies via databases and registers

Identification		
	Records identified: (*k* = 219)	Because only the Web of Science's Core Collection was consulted, there were no duplicate reported to omit, especially as the database is professionally maintained.

Screening		
	Records screened (*k* = 219)	Records excluded** (*k* = 186) Excluded for non-L2 and/or vocabulary focus
	Reports sought for retrieval (*k* = 34)	Reports not retrieved (*k* = 0)
	Reports assessed for eligibility (*k* = 34)	Reports excluded: Not truly L2-focused (*k* = 2) Labelled as proceedings (*k* = 1)

Included	
	Reports included in review (*k* = 31)

Source: Adapted from Page et al. (2021).
Note: Studies/reports contrast removed because all reports were L2 vocabulary-based syntheses.

Figure 1. Report Pool Creation Flow Chart.

this trend, there were also strong observed preferences for experimental and methodological foci among reviewed meta-analyses, and systematic reviews respectively. As argued by Nicklin and Vitta (2021), L2 instructed vocabulary research lends itself to experimental designs. The focus on methods in systematic reviews could be attributed to the ongoing methodological review in L2 research as these reports were published after the seminal work of Plonsky and Gass (2011), which acted as a call for reform that was perhaps the first to garner wide-spread attention.

The two systematic reviews coded as thematic reviews interestingly focused upon the inferential models that vocabulary researchers presented. Choi and

Table 2. Features of the Research Syntheses in the Report Pool

RQ	Type	k
1a	Meta-analysis	25[a]
	Systematic review	7[b]
1b	Experimental meta-analysis	21
	Correlational meta-analysis	4
1c	Methodological synthesis	4[c]
	Other systematic review	3[d]

[a]Including 24 bona fide meta-analyses and Nicklin and Vitta (2021), who aggregated effect sizes without *Q* testing and overtly stated the aggregated effects for power simulations as opposed to a wide-scoping inquiry seeking to make generalizable claims about the population.

[b]Lin and Lin (2019) claimed a systematic review and meta-analysis design but the report was coded here as meta-analysis only because the systematic review was not executed according to the protocols of O'Connor et al. (2008). Nicklin and Vitta (2021) coded as both meta-analysis and systematic review.

[c]Elgort (2018) and Yang et al. (2021) coded as methodological syntheses as their reviews synthesized areas to answer research questions mostly focused on methods.

[d]Two thematic reviews and one broad overview of digital gaming and vocabulary research that could have been labelled a scoping review despite the authors' label of systematic review.

Zhang (2021) reviewed multivariate models where grammar and vocabulary predicted reading, concluding that there was insufficient evidence to say which was the stronger predictor. Because the authors reviewed effect sizes thematically, the report was not a meta-analysis nor a methodological synthesis as the methods of the report pool's studies were not in focus. In a similar vein, Squires et al. (2020) thematically reviewed the models and findings of studies investigating cognate influences and multilingual children's vocabulary without aggregating data in a meta-analytic manner.

4.2 RQ2: Details of Aggregated Studies in L2 Vocabulary Research Syntheses

RQ2a. The 31 reviewed reports featured 1,198 synthesized studies with 85,154 reported participants (see Table 3 for further details). To put these figures into context, Plonsky and Oswald (2014) executed a broad search of L2 quantitative research and synthesized 394 primary studies and 94 meta-analyses where the total number of participants comprised 604,000. While the data points in Table 3 are smaller than Plonsky and Oswald's aggregated data, these values are still reasonable considering the strict vocabulary focus and Web of Science restriction of this current review.

It was unsurprising that the systematic reviews omitted effect sizes, but it was surprising that 4 of the 25 meta-analytic reports omitted summary statistics for the total aggregated number of participants, while 9 omitted summaries for aggregated effect sizes. This highlights a potential quality control issue because *Q*-testing and inferential alternatives require this information for their calculations (Borenstein et al., 2009; Yanagisawa & Webb, 2021).

Table 3. Descriptive Details of L2 Vocabulary Research Syntheses' Aggregated Studies and their Components

	Meta-analyses (k = 24)	Systematic reviews (k = 6)	Nicklin and Vitta (2021; MA & SR)
Number of studies	791	326	81
Number of participants	78,811	6,343	NA
Number of effect sizes	3,656	NA	453

Note: MA = meta-analysis; SR = systematic review.

RQ2b. Of the 25 meta-analyses, 17 featured Q-testing as the inferential assessment testing the significance of the reported aggregated effect size(s) and/or moderator analyses. Of the eight meta-analyses not using Q-testing, 5 featured parametric testing where procedures were stated to treat each effect size as a "case." Three (3) reports had no inferential assessment with only Nicklin and Vitta (2021) offering a justification as to why. In the case of Nicklin and Vitta, Q-testing was avoided to capture the full heterogeneity of studies' effect sizes by using a simpler approach to ascertain effect sizes for a series of power analyses.

As a post-hoc to the analysis of RQ2, we decided to ascertain the relationship between the frequency of the publication of L2 vocabulary research syntheses and time. As Figure 2 illustrates, the number of reports has increased with time, which was confirmed via Spearman's *rho* analysis, n = 10, ρ = 0.74 [0.21, 0.93]. *Rho* was used to quantify the trend illustrated in Figure 2 as opposed to making a generalization beyond the sample. Nevertheless, this trend demonstrates that the interest in synthesizing L2 vocabulary research is growing, at least in published Web of Science reports, and thus the time is ripe to consider the best directions going forward.

4.3 RQ3: Categorization of Topics

In the bottom-up coding addressing this question, it became apparent that only some areas had been investigated with enough frequency to warrant the construction of a category. In other words, a grouping category was constructed and presented only if more than three reports fit into it. The rationale for this approach is twofold. Firstly, the categories listed represent areas where several synthesis efforts have taken place and we can consider how robust the findings are in the aggregate because they are not beholden to one or two author teams. Secondly, areas not represented in these categories should be viewed by future designers of L2 vocabulary research syntheses as being potentially worthy of further exploration. Having stated these points, the three categories constructed were (1) technology use and vocabulary subsuming CALL, MALL, and digital games, and excluding glossing (k = 9, of which one was exclusively MALL), (2) vocabulary as a predictor of reading (k = 4, of which one also considered listening), and (3) glossing of target vocabulary items (k = 5). Although glossing research syntheses did mostly focus upon technology (e.g., Ramezanali et al., 2021), the theoretical underpinning of the synthesized research was distinct and as four glossing
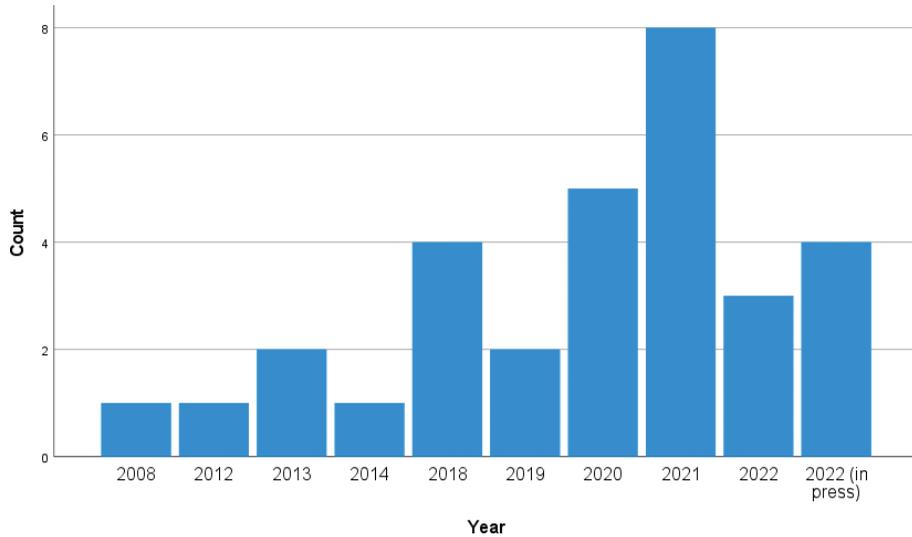
Figure 2. Frequency of L2 Vocabulary Research Syntheses' Publication over Time.

reports were observed in the report pool, a separate category was constructed. These categories, as a final note, should be considered with caution as the claim of homogeneity of the reports subsumed under them is tenuous.

The remaining 17 research syntheses in the report pool represented a divergent range of interests, but that is not to say that the topics were scattershot. Instead, there were cases where the same researchers appeared to offer multiple complementary reports. Uchihara et al. (2019), for instance, published a meta-analysis on the effects of repetition on incidental vocabulary learning and 2 years later, Yangisawa and Webb (2021) meta-analyzed research on the task involvement load's predictiveness on vocabulary learning. In between these reports was another meta-analysis by Webb et al. (2020) on intentional learning and vocabulary gains. In a similar vein, Vitta et al. (2021) and Nicklin and Vitta (2021) published research syntheses on the quality of L2 vocabulary experimental sampling practices along with sample size recommendations.

4.4 RQ4: Future Directions

RQ4a. In the three categories constructed to address RQ3, some meta-analyses presented aggregated effect sizes to support three empirical trends (see shared data for studies and summary of findings):

- Technology enhances the acquisition of vocabulary.
- Glossing enhances the acquisition of vocabulary.
- Vocabulary predicts reading proficiency (and listening proficiency [see Zhang & Zhang, 2020]).

These trends can be viewed as addressing first-generation RQs (Zanna & Fazio, 1982), which merely ask "if" a treatment is effective. For example: *Is technology effective in supporting the acquisition of vocabulary?* Future researchers can extend on these trends to ask second- (e.g., "when?" and "under what conditions?") and third-generation RQs (e.g., "how" a given treatment works) under Zanna and Fazio's framework.

What such research projects would look like are modeled using examples from the report pool. Consider glossing as an example with Abraham's (2008) finding that glossing had strong yet nonsignificant effects for more proficient learners. A second-generation RQ-governed project could address the conditions under which this finding holds true. To illustrate the potential of a third-generation RQ consider Zhang and Zhang (2020), who presented a moderator analysis detailing that vocabulary knowledge when measured via meaning recall had a significantly stronger association with reading than when such vocabulary knowledge was measured via form recall[1] or meaning recognition. To explain the "how" of this finding, researchers could design mixed-methods studies where qualitative data helps to explain what meaning recall is testing to give it its significantly stronger association.

RQ4b. Because there were only seven systematic reviews, it is difficult to make generalizations. However, one report did present findings challenging one of the core principles of vocabulary research: its primacy in the language learning process. Choi and Zhang (2021) reviewed multivariate models where vocabulary and grammar predicted reading and reported that it was inconclusive as to which construct provided the stronger association. This contradicts the assumption that vocabulary is the prime competency of receptive skills and even productive skills (see review in Nicklin & Vitta, 2021). On one hand, only 17 reports were synthesized and the claim was not assessed via inferential meta-analytic testing. On the other hand, Choi and Zhang (2021) pointed to second-generation RQs being needed in highlighting that factors such as proficiency, context, and other covariates would affect the observed effect sizes in these models, and thus this report is really a call to improve research methods when investigating how grammar and vocabulary predict reading. In doing so, this report complements the general drive to improve SLA quantitative research as the field matures (see Gass et al., 2021).

In addition to the controversial finding of Choi and Zhang (2021), the lack of quality-focused methodological syntheses is something vocabulary researchers can address in the future. Of the five reports coded as methodological syntheses, only two directly addressed issues of quality while making recommendations. These pertained to the quality of L2 experimental samples (Nicklin & Vitta, 2021; Vitta et al., 2021) but the spectrum of quality in L2 quantitative research is much broader than that (see Al-Hoorie & Vitta, 2019; Plonsky & Gass, 2011), and this is an area for future exploration. The quality of qualitative research can also be assessed but from the clear imbalance favoring meta-analyses it is evident that L2 vocabulary research, at least what has been synthesized, tends to be quantitative. This is an important point as the methods reform in the field is still relatively new (Gass et al., 2021), and recommendations for improvements based upon empirical observations have been made. For example, just 1 of 110 L2 experimental vocabulary reports in top journals considered power in an a priori manner (Vitta et al., 2021).

5 Conclusion

In this brief scoping review of L2 vocabulary syntheses ($k = 31$), it was observed that vocabulary-based research synthesizers prefer to conduct meta-analyses of experimental effects, as opposed to systematic reviews or meta-analyses of correlational effects. In line with the growing popularity of research syntheses across academia in general, the amount of these reports with a vocabulary focus published per year has increased with time. It was also observed that just under half of these reports could be subsumed under three general categories: technology and vocabulary use, glossing effect on vocabulary acquisition, and vocabulary as a predictor of reading. While trends did appear in the other half of the reports, they were too infrequent to categorize. The findings of this study lead to the call for future L2 vocabulary research to consider second- and third-generation RQs (Zanna & Fazio, 1982), such as "when/under what conditions" and "how" certain phenomena in vocabulary research exist, while highlighting gaps to be addressed by more advanced and robust research approaches.

Authors' Endnote

In his discussion of this paper, Dr. Kyle noted that he could only mostly replicate the observed 219 hits in the WoS with the stated search string on his end (207–208 hits). After further correspondence with Dr. Kyle, the authors were able to reconfirm the 219 hits (correcting for index date updates) on the WoS (from the second author's access to the index) in this paper, but it does seem that location and institutional access might be causing the slight discord observed. Both Dr. Kyle and the authors agree that this is an intriguing area to be further explored.

Note:

1. The authors claimed a significant difference despite marginal CI overlap between form and meaning recall. We have given the authors the benefit of the doubt in this regard.

References

Abraham, L. B. (2008). Computer-mediated glosses in second language reading comprehension and vocabulary learning: A meta-analysis. *Computer Assisted Language Learning, 21*(3), 199–226. https://doi.org/10.1080/09588220802090246

Ahn, E. J., & Kang, H. (2018). Introduction to systematic review and meta-analysis. *Korean Journal of Anesthesiology, 71*(2), 103–112. https://doi.org/10.4097/kjae.2018.71.2.103

Al-Hoorie, A. H., Quint Oga-Baldwin, W. L., 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*. Advance online publication. 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*(6), 727–744. https://doi.org/10.1177/1362168818767191

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis.* Wiley. https://doi.org/10.1002/9780470743386

Choi, Y., & Zhang, D. B. (2021). The relative role of vocabulary and grammatical knowledge in L2 reading comprehension: A systematic review of literature. *International Review of Applied Linguistics in Language Teaching, 59*(1), 1–30. https://doi.org/10.1515/iral-2017-0033

Elgort, I. (2018). Technology-mediated second language vocabulary development: A review of trends in research methodology. *CALICO Journal, 35*(1), 1–29. https://doi.org/10.1558/cj.34554

Gass, S., Loewen, S., & Plonsky, L. (2021). Coming of age: The past, present, and future of quantitative SLA research. *Language Testing, 54*(2), 245–258. https://doi.org/10.1017/S0261444819000430

Hiver, P., Al-Hoorie, A. H., Vitta, J. P., & Wu, J. (2021). Engagement in language learning: A systematic review of 20 years of research methods and definitions. *Language Teaching Research.* Advance online publication. https://doi.org/10.1177/13621688211001289

Jeon, E. H., & Yamashita, J. (2014). L2 reading comprehension and its correlates: A meta-analysis. *Language Learning, 64*(1), 160–212. https://doi.org/10.1111/lang.12034

Lin, J. J., & Lin, H. (2019). Mobile-assisted ESL/EFL vocabulary learning: A systematic review and meta-analysis. *Computer Assisted Language Learning, 32*(8), 878–919. https://doi.org/10.1080/09588221.2018.1541359

Munn, Z., Peters, M. D. J., Stern, C., Tufanaru, C., McArthur, A., & Aromataris, E. (2018). Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Medical Research Methodology, 18*(1), 143. https://doi.org/10.1186/s12874-018-0611-x

Nicklin, C., & Vitta, J. P. (2021). Effect-driven sample sizes in second language instructed vocabulary acquisition research. *The Modern Language Journal, 105*(1), 218–236. https://doi.org/10.1111/modl.12692

O'Connor, D., Green, S., & Higgins, J. P. (2008). Defining the review question and developing criteria for including studies. In J. P. Higgins & S. Green (Eds.), *Cochrane handbook for systematic reviews of interventions* (pp. 81–94). Wiley-Blackwell.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., & Moher, D. (2021). The Prisma 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, n71. https://doi.org/10.1136/bmj.n71

Plonsky, L., & Gass, S. (2011). Quantitative research methods, study quality, and outcomes: The case of interaction research. *Language Learning, 61*(2), 325–366. https://doi.org/10.1111/j.1467-9922.2011.00640.x

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning, 64*(4), 878–912. https://doi.org/10.1111/lang.12079

Raeisi-Vanani, A., Plonsky, L., Wang, W., Lee, K., & Peng, P. (2022). Applying meta-analytic structural equation modeling to second language research: An introduction. *Research Methods in Applied Linguistics, 1*(3), 100018. https://doi.org/10.1016/j.rmal.2022.100018

Ramezanali, N., Uchihara, T., & Faez, F. (2021). Efficacy of multimodal glossing on second language vocabulary learning: A meta-analysis. *TESOL Quarterly, 55*(1), 105–133. https://doi.org/10.1002/tesq.579

Squires, L. R., Ohlfest, S. J., Santoro, K. E., & Roberts, J. L. (2020). Factors influencing cognate performance for young multilingual children's vocabulary: A research synthesis. *American Journal of Speech-Language Pathology, 29*(4), 2170–2188. https://doi.org/10.1044/2020_AJSLP-19-00167

Uchihara, T., Webb, S., & Yangisawa, A. (2019). The effects of repetition on incidental vocabulary learning: A meta-analysis of correlational studies. *Language Learning, 69*(3), 559–599. https://doi.org/10.1111/lang.12343

Vitta, J. P., & Al-Hoorie, A. H. (2020). The flipped classroom in second language learning: A meta-analysis. *Language Teaching Research.* Advance online publication. https://doi.org/10.1177/1362168820981403

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

Webb, S., Yangisawa, A., & Uchihara, T. (2020). How effective are intentional vocabulary-learning activities? A meta-analysis. *The Modern Language Journal, 104*(4), 715–738. https://doi.org/10.1111/modl.12671

Yangisawa, A., & Webb, S. (2021). To what extent does the involvement load hypothesis predict incidental language L2 vocabulary learning? A meta-analysis. *Language Learning, 71*(2), 487–536. https://doi.org/10.1111/lang.12444

Yang, X., Kuo, L. J., Eslami, Z. R., & Moody, S. M. (2021). Theoretical trends of research on technology and L2 vocabulary learning: A systematic review. *Journal of Computers in Education*, *8*(4), 465–483. https://doi.org/10.1007/s40692-021-00187-8

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

Zhang, S., & Zhang, X. (2020). The relationship between vocabulary knowledge and L2 reading/listening comprehension: A meta-analysis. *Language Teaching Research, 26*(4), 696–725. https://doi.org/10.1177/1362168820913998.

Zou, D., Huang, Y., & Xie, H. (2019). Digital game-based vocabulary learning: Where are we and where are we going? *Computer Assisted Language Learning, 34*(5–6), 751–777. https://doi.org/10.1080/09588221.2019.1640745

Zhang, X. (2020). A bibliometric analysis of second language acquisition between 1997 and 2018. *Studies in Second Language Acquisition, 42*(1), 199–222. https://doi.org/10.1017/S0272263119000573