Effect-Driven Sample Sizes in Second Language Instructed Vocabulary Acquisition Research

CHRISTOPHER NICKLIN[1] AND JOSEPH P. VITTA[2]

[1]*Rikkyo University, Center for Foreign Language Education and Research, Nishi Ikebukuro 3-34-1, Toshima-ku, Japan 171-8501, Japan Email: c.nicklin@rikkyo.ac.jp*

[2]*Rikkyo University, Center for Foreign Language Education and Research, Nishi Ikebukuro 3-34-1, Toshima City, Tokyo 171–8501, Japan Email: vittajp@rikkyo.ac.jp*

The present study involved the analysis of 81 second language instructed vocabulary acquisition (L2 IVA) studies over 2 phases. In Phase I, we categorized and coded the effect sizes of the studies. Observing that the basic between- and within-subject design dichotomy lacked the sensitivity to capture the heterogeneity of observed effects, we employed a more granular approach. In both between- and within-subject designs, treatment versus comparison contrasts best represented comparisons of most interest in L2 IVA experiments, with median effect sizes (*g*) of .62 (between-subject) and .25 (counterbalanced within-subject). In Phase II, the aggregated effect sizes observed in Phase I were utilized in a priori power simulations to suggest approximate sample sizes for common L2 IVA analyses. For conservatively powered between-subject designs, the simulations suggested sample sizes ranging from 292 to 492 participants. Counterbalanced within-subject designs required 95 to 203 subjects depending on the assumed correlation between the repeated measures. The overarching implication of these simulations suggests that future L2 IVA experiments require larger samples that reference effect sizes from previous research, and we offer 3 potential solutions to the problem of obtaining larger samples.

Keywords: vocabulary; effect size; power analysis; research synthesis

SECOND LANGUAGE ACQUISITION (SLA) research is currently undergoing methodological reform, especially within its quantitative inquiries (Gass, Loewen, & Plonsky, 2020). This call for reform has often come via research syntheses such as meta-analyses or systematic reviews. Although some research syntheses have offered comprehensive reviews of second language (L2) research (e.g., Plonsky, 2013), there exists a trend of focusing on particular areas. For instance, L2 research syntheses have focused on interaction research (Plonsky & Gass, 2011), the reporting of statistics (Al–Hoorie & Vitta, 2019), and outlier treatment in L2 self-paced reading tasks (Nicklin & Plonsky, 2020) to name but a few.

The Modern Language Journal, 105, 1, (2021)
DOI: 10.1111/modl.12692
0026-7902/21/218–236 $1.50/0
© National Federation of Modern Language Teachers Associations

The current study fits into this trend and presents a review of L2 instructed vocabulary research with an eye toward reform. Specifically, (quasi)experimental[1] studies from trusted L2 journals were reviewed to extract, categorize, and aggregate effect sizes with the aim of informing sample-size determination in future research. This process began by expanding on the conventional between- and within-subject dichotomy (Plonsky & Oswald, 2014), which facilitated a more sensitive analysis of the extracted effect sizes' observed heterogeneity by differentiating among treatment, control, and comparison group contrasts. The aggregated effect sizes were then summarized according to the resulting categorization scheme. Using these effects, a set of suggested sample sizes was calculated for the most common statistical procedures utilized by L2 instructed vocabulary researchers. While there are areas within L2 teaching research other than instructed vocabulary, few areas are as topical considering the observed strong relationship with L2 proficiency

(e.g., Matthews & Cheng, 2015; McLean, Stuart, & Batty, 2020) and its popularity and emphasis among teachers (Lewis, 1993; Wilkins, 1972).

LITERATURE REVIEW

L2 instructed vocabulary acquisition (IVA) research and sample-size planning for L2 experimental designs were the two overarching themes that underpinned this study. Each is discussed in turn.

L2 Instructed Vocabulary Acquisition Research

L2 IVA is primarily concerned with how language learners acquire target lexical items and the factors that influence this process (Laufer, 2005), or "the best means of achieving good vocabulary learning" (Schmitt, 2008, p. 329). While both Laufer (2005) and Schmitt (2008) employed the 'instructed second language vocabulary learning' label, in the present study we employ 'L2 IVA.' This decision was based on the fact that instructed second language acquisition (ISLA) is a well-established domain within SLA (Loewen, 2015; Loewen & Sato, 2017) and can act as a suitable superordinate under which to subsume instructed vocabulary. L2 IVA, as a label, facilitates this hierarchal organization.

While L2 research has many subdomains on which this study could focus, L2 IVA was chosen because of its observed importance at both practitioner and researcher levels (Richards et al., 2009). Research in this area has a broad and wide-ranging influence, thus any undertaking that can improve (quasi)experimental designs within the domain is especially worthwhile. At the practitioner level, the calls for emphasizing vocabulary date back several decades. In an L2 learning handbook for teachers, Wilkins (1972) famously argued that nothing could be stated without vocabulary knowledge. Later, Lewis (1993) proposed an entire curriculum approach that focused on the acquisition of multi-word lexis, based on the belief that lexis and vocabulary were "at the heart of language" (p. 89). Nowadays, teachers have access to wordlists, such as the new general service list (NGSL; Browne, Culligan, & Phillips, 2013) and the academic word list (AWL; Coxhead, 2000), which provide empirically supported accounts of the most beneficial words that language learners should learn for varying proficiency levels, contexts, and purposes. Major publishers utilize these lists in textbooks and emphasize in their catalogues that their products have supplementary vocabulary tools (e.g., Cambridge University Press, 2019; Pearson, 2019).

Vocabulary's prominence at the practitioner level has developed in parallel with research demonstrating the strong empirical relationship between vocabulary knowledge and overall proficiency (Schmitt, 2000). In the 1990s, concrete observations were reported that established vocabulary's primacy as a predictor of L2 proficiency. Nation and Waring (1997) empirically demonstrated that receptive knowledge of the first 3,000 word families was required to read nonspecialized text in English, while Laufer and Nation (1995) observed that texts written by more proficient L2 writers used more infrequent words and had a greater number of different words being employed. Recent studies have further entrenched the case for vocabulary having a key role in L2 proficiency. For instance, a recent meta-analysis illustrated that vocabulary knowledge, in the aggregate, has a strong relationship with reading ($r = .57$) and listening ($r = .56$) proficiency (Zhang & Zhang, 2020). Structural equation modeling has likewise revealed that lexical measures (sophistication and cohesion) account for approximately 81% of Test of English as a Foreign Language (TOEFL) essay scores' variance, while syntactic measures are only indirectly involved (Kim & Crossley, 2018). Receptive knowledge of the first 3,000 word families has been shown to account for 52% of the variance of Chinese learners' International English Language Testing System (IELTS) listening scores (Matthews & Cheng, 2015). Additionally, vocabulary knowledge predicted between 38% and 61% of Test of English for International Communication (TOEIC) reading performance depending on the modality of the vocabulary testing and number of test items (McLean et al., 2020). To summarize, research has repeatedly emphasized the relationship between L2 vocabulary knowledge and proficiency across domains. Therefore, L2 IVA research is important for both L2 practitioners and researchers, and is an ideal area of focus for a study such as the present one.

Despite its importance, vocabulary acquisition, and in turn L2 IVA, has yet to emerge as a clearly defined construct within which consensus has been reached (Schmitt, 2019). This lack of consensus notwithstanding, Nation's (2013) conceptualization of vocabulary knowledge as a multidimensional construct has been one of the most consistently referenced vocabulary knowledge frameworks (e.g., Kremmel & Schmitt, 2016; McLean et al., 2020). In the present study, Nation's framework was utilized to identify what vocabulary was and what a vocabulary acquisition

FIGURE 1
Nation's Framework of Word Knowledge

Form	**Spoken**	**Receptive:** What does the word sound like? **Productive:** How is the word pronounced?
	Written	**Receptive:** What does the word look like? **Productive:** How is the word written and spelled?
	Word parts	**Receptive:** What parts are recognizable in this word? **Productive:** What word parts are needed to express the meaning?
Meaning	**Form and meaning**	**Receptive:** What meaning does this word form signal? **Productive:** What word form can be used to express this meaning?
	Concept and referents	**Receptive:** What is included in the concept? **Productive:** What items can the concept refer to?
	Associations	**Receptive:** What other words does this make us think ok? **Productive:** What other words can we use instead of this one?
Use	**Grammatical functions**	**Receptive:** In what patterns does the word occur? **Productive:** In what patterns must we use this word?
	Collocations	**Receptive:** What words or types of words occur with this one? **Productive:** What words or types of words must we use with this one?
	Constraints on use	**Receptive:** Where, when, how often would we expect to meet this word? **Productive:** Where, when, how often can we use this word?

Note. From Nation (2013, p. 49).

intervention entails. In Nation's framework, lexical items are conceptualized in terms of their forms (both single- and multi-word constructions), meanings, and use in relation to other lexical units (see Figure 1). However, Nation's framework has been criticized as being too broad and unfocused (González–Fernández & Schmitt, 2019). Despite this criticism, Nation's framework provides a somewhat accepted insight into L2 IVA that can be conceptualized in relation to target lexical forms, and is the basis for the conceptualization of vocabulary in the present study.

Experimental Designs and Sample Sizes

Experimental Designs. L2 IVA experimental designs have corresponded to the broader trends of social science research in which treatment, comparison (alternative learning method), and/or control (no learning method) conditions have been implemented in either between- (e.g., Shintani, 2011) or within-subject (e.g., Rott, 2007) designs to investigate the best ways of acquiring lexical forms. Research syntheses, such as Plonsky & Gass (2011) and Plonsky (2013), have been useful in determining how L2 researchers implement such interventions. Approximately 87% of the experimental designs in studies surveyed by Plonsky and Gass (2011) incorporated a comparison group of some kind, which was almost identical to the slightly below 86% of experimental designs assessed by Plonsky (2013). These results imply that designs involving comparison groups are clearly favored by L2 researchers. Furthermore, in their seminal investigation of effect sizes in L2 research, Plonsky and Oswald (2014) collected more than twice as many between-group ($k = 67$) than within-group contrasts ($k = 25$) from 91 meta-analyses. When synthesized, the results of these studies reveal comparison group designs as being the most frequently utilized design in L2 research.

Assessment of between-group comparison types is a potentially fruitful avenue of investigation within a research synthesis. When an experimental design utilizes a treatment, comparison, and control group, the control group predictably displays the least improvement in a posttest (e.g., Mizumoto & Takeuchi, 2009), while contrasts between the treatment and comparison groups are generally closer (e.g., Tseng, Liou, & Chu, 2020). Therefore, we should expect that contrasts involving these two groups will produce smaller effect sizes than contrasts between treatment and control groups.

Scant attention has been paid in L2 research syntheses to the existence of within-subject comparison types, such as counterbalanced designs, which in the present study subsumes switching replications (Trochim, Donnelly, & Arora, 2016) and Latin square designs (Kirk, 2012). Both of these designs involve treatment, comparison, and control conditions being counterbalanced, or rotated, across two groups in the case of the former and three or more groups in the case of the latter. Thus, the effect of treatment order is theoretically accounted for and ideally zero. Additionally, each group undergoes all conditions, resulting in a perhaps more ethical experiment (e.g., Rott, 2007). Despite the lack of attention paid to these designs, they possess two main advantages over conventional between-subject designs, consisting of (a) improved power vis-à-vis their between-subject counterparts, and (b) control for covariates and confounding variables as all subjects experience all experimental conditions (Kuehl, 2000). However, Kuehl also noted that counterbalanced designs have a major drawback in not facilitating the investigation of interactions among grouping variables. Also, although some recent meta-analyses (e.g., Lin & Lin, 2019) have conflated effects from such within-subject designs with their between-subject counterparts, researchers utilizing

counterbalanced designs must be aware that these designs are operationalizing fundamentally different hypotheses from their between-subject counterparts. Between-subject designs operationalize hypotheses posed at the group level while within-subject designs presuppose that the same subject will perform or behave differently under different conditions (Fisher, 1935).

Sample Size and Statistical Power. Another L2 issue highlighted by research syntheses is the tendency to utilize small sample sizes, which in turn affects the statistical power achieved by analyses. Statistical power is the process through which researchers determine the appropriate sample size needed for their inquiry, and relates to the probability of finding a statistically significant result if the null hypothesis of an experiment is rejected and the alternative hypothesis is therefore deemed true (Cumming, 2012). Plonsky (2013) revealed that the median group sample size in L2 research was merely 19 participants. This is problematic because samples of 19 are unlikely to provide appropriate statistical power for discovering significant effects. Small sample sizes are less likely to achieve sufficient statistical power to detect a significant effect, leading to Type II errors. The median sample size of 19 is particularly disconcerting in light of research indicating that a sample size of 90 (two groups of 45) is required to detect an effect of $d = .60$ with 80% power for a between-group comparison involving one variable with two levels (Brysbaert, 2019). Conversely, Brysbaert (2019) highlighted that underpowered studies reporting large effect sizes that are significant have an increased likelihood of being unrepresentative of the smaller true effect in the population. Therefore, readers of L2 research should be especially cautious in trusting large effects in small sample studies.

To determine the sample size required for detection of the effect under investigation, it is wise for researchers to conduct a classical a priori power analysis, using software such as G*Power (Faul et al., 2007), during the design phase of an experiment. The calculations to do so require an expected effect size, which Cohen (1988) advised to be estimated based on previous studies' data, such as Plonsky & Oswald's (2014) effect sizes for L2 research. Despite the availability of such software, Plonsky (2013) found that researchers reported power analysis in just 6 studies out of 606. Furthermore, two of those studies reported post hoc power, which is controversial. Post hoc power has been labeled as misleading because it is illogical to talk about the probability of an experimental outcome following the data collection (e.g., Cumming, 2012). Furthermore, post hoc power contributes little to an analysis because of its near-perfect association with the observed effect size when sample size is held constant (Aberson, 2019). Despite these issues, proponents of post hoc power view it as a method of understanding power when assuming that the observed effect truly exists in the population. Using this rationale, some journals request that authors report post hoc power for reasonable yet nonsignificant effect sizes (Aberson, 2019).

Despite the stated importance of a priori power analysis, it has also received criticism. Disciplines such as biometrics and epidemiology have seen power considerations countered by cost concerns and ethical arguments regarding increased harm to vulnerable populations (Bacchetti et al., 2005; Bacchetti, McCulloch, & Segal, 2008). These arguments were not against power but rather implored researchers to take a more holistic view of their inquiries' consequences. However, within L2 research these concerns become somewhat moot, because the treatments are learning conditions with relatively minimal costs. A priori power analysis has also been criticized due to its association with frequentist null hypothesis significance testing (NHST; e.g., Cohen, 1997; Cumming, 2012). However, the essence of power considerations still exists even when a reliance on p values is removed. Within the Bayesian framework, statistical power is "the probability of achieving the goal of a planned empirical study, if a suspected underlying state of the world is true" (Kruschke, 2015, p. 359), and a number of proposals exist to estimate power within a Bayesian framework (e.g., Lakens, 2014; Stefan et al., 2019), all relying on the results from previous research. Even sample-size estimation techniques such as precision (Cumming, 2012; Norouzian, 2020) require estimates of effect sizes from prior research, despite being conceptually different to classical or Bayesian power. The overarching point is that L2 IVA researchers require estimates of effect size to conduct proper experiments with informed sample sizes.

THE CURRENT STUDY

The current study's aim was to provide empirically driven sample-size estimates to aid future L2 IVA experimental research. This aim was achieved over two sequential phases. In Phase I, a meta-analytic process was undertaken to investigate the observed effects of past L2 IVA (quasi)experimental studies in relation to

FIGURE 2
Multistage Journal Selection Process

Note. SSCI = Social Sciences Citation Index; ELT = English language teaching.

different group-mean comparison types. To achieve this, a "comparison-of-most-interest" (Brooks & Johanson, 2011, p. 98) approach was assumed, whereby only comparisons of group performance that directly addressed a study's research questions or hypotheses were considered. This bivariate focus corresponded to meta-analysis literature emphasizing the need to aggregate common effect metrics across studies that utilize varying (quasi)experimental comparison types (e.g., Borenstein et al., 2009). In Phase II, the aggregated effect sizes observed in Phase I were utilized in a series of simulated power analyses to determine a set of sample-size suggestions for L2 IVA researchers. It was observed that none of the studies considered incorporated a priori power analysis referencing vocabulary-centric effects from previous research, thus the rationale of this undertaking was clear.

PHASE I: L2 INSTRUCTED VOCABULARY ACQUISITION RESEARCH EFFECT SIZES

Phase I involved acquiring a body of L2 IVA studies from five respected journals in the Social Sciences Citation Index (SSCI) in order to address the following research questions:

RQ1. What is the observed distribution of group-mean comparison types within L2 IVA (quasi)experimental design research?
RQ2. What is the observed range of effect sizes for each of the group-mean comparison types?

The SSCI focus was grounded in the observation that statistical quality has been demonstrated to vary by index with the SSCI outperforming Scopus (see Al–Hoorie & Vitta, 2019). Thus, we limited our focus to SSCI journals to increase the probability of reviewing the best-quality research. Following Zhang (2020), who similarly focused upon SSCI journals, we make no judgments about non-SSCI L2 journals and value their contributions to the field.

Methodology

Journal and Study Selection. A sample of 82 L2 IVA studies was obtained from five prominent journals in the applied linguistics field and was the result of a multistage process (see Figure 2). 'Prominence' was operationalized by a journals' SSCI index status, a decision based on the observation that such journals have higher statistical quality (Al–Hoorie & Vitta, 2019) and recent research syntheses have referenced this point in justifying their SSCI focus (e.g., Zhang, 2020). The aim of the process was to obtain a set of journals from which classroom-based vocabulary treatment effect sizes could be estimated. We did not intend to conduct a thorough meta-analysis of L2 IVA research. To that end, we required SSCI-indexed journals that published quantitative vocabulary research intended to inform classroom teaching. Stage 1 of the selection process involved reviewing Egbert's (2007) list of 35 field-specific journals. Egbert's comprehensive list of field-specific journals was one of the first to merge expert opinion with a consideration of a multitude of data points, including rejection rates and citation analysis. The list was reviewed to identify candidates that (a) were indexed in the SSCI (as of 2010), (b) had a title referencing a broad field focus (e.g., *CALL* journal excluded), and (c) included

'practitioner' or agents of 'language teaching' as their audience (as catalogued by Egbert, 2007). This analysis resulted in six candidate journals. In Stage 2, the six candidate journals were compared with two SSCI journal lists employed in quantitative L2 research reviews (Al–Hoorie & Vitta, 2019; Plonsky & Gass, 2011).[2] Stage 3 involved an external review, during which three external experts agreed that the five remaining journals could be conceptualized under a label of 'SSCI-indexed journals that are well known, somewhat trusted in relation to their quantitative papers, and intended for practice-level research.' This process resulted in the following five journals: *Language Learning*, *Language Teaching Research, The Modern Language Journal*, *System*, and *TESOL Quarterly*.

In total, 809 unique studies were retrieved from the five journals under the following search parameters:

1. Published between January 2000 and March 2020.
2. Selected from the EBSCO search platform with the search terms *vocab**, *lexi**, *idiom**, *collocat**, *phrasal*, *multi-word*, and *formula* (where the asterisk denotes a 'wildcard' search term, which maximizes the results).

For inclusion, a study's hypotheses or research questions were required to (a) involve dependent variables relating to Nation's (2013) framework (see Figure 1), or be connected to it through the arguments of others (e.g., Tseng & Schmitt, 2008), (b) involve a treatment that students would encounter in a classroom context (thus excluding psycholinguistic studies), (c) involve a discrete set of target lexical items, and (d) be reported adequately enough to allow for an estimation of effect. After applying these criteria to the selected studies, 727 studies were rejected by the second researcher for failing to fulfill the selection criteria, leaving an initial sample of 82 studies for analysis (see Online Supporting Information A). To validate these judgments, the first researcher independently coded 148 studies and 89.80% initial agreement ($\kappa = .80$) was observed with 100% agreement reached after discussion.

Extracting Group-Mean Comparisons. Each of the 82 sample studies was then reviewed to identify the group-mean comparisons—that is, "comparison-of-most-interest" (Brooks & Johanson, 2011, p. 98)—that specifically addressed each study's research questions. This process involved a mapping of observed comparisons to research questions and hypotheses, and only research questions and hypotheses meeting the inclusion criteria described in the previous section were considered for data extraction. For instance, comparisons that reflected an extensive reading treatment were excluded for not being exclusively focused on target-form acquisition, while comparisons involving a lexical diversity-related dependent variable were excluded for measuring forms outside of the set target. 'True' control group comparisons, involving control groups that were never exposed to the target vocabulary, were only extracted when the researcher used them to specifically address a research question. To calibrate judgments, the researchers extracted group-mean comparisons from 12 studies in tandem employing a collaborative and iterative process and discussing and refining inclusion and exclusion criteria. After this calibration, the first researcher processed the remaining 70 studies. The second researcher checked 17 randomly selected studies (20.73% of the study sample), and only 1 study saw any disagreement (94.11% agreement) on the effects extracted in relation to stated research questions or hypotheses. After discussion, agreement was reached. In total, 462 group-mean comparisons were extracted.

Categorization of Group-Mean Comparison Types. Each mean comparison ($n = 462$) was then coded as being: (a) between-group, treatment–control posttest comparison (B-TCtrl), (b) between-group, treatment–comparison (group) posttest contrast (B-TCom), (c) within-group, pretest to posttest comparison (W-PP), (d) within-group, one sample, observed difference from 0 (W-O), (e) within-group, treatment–comparison (condition) contrast (W-TCom), or (f) within-group, counterbalanced groups (W-CB), in which all subjects experience all conditions but in subgroups where order of exposure is varied (see Table 1). Given the novelty of this categorization framework, both researchers independently coded all 462 group-mean comparisons and 90.91% initial agreement ($\kappa = .86$) was observed, with 100% agreement achieved after further discussion (see Online Supporting Information B for coding scheme). B-TCtrl was specifically for 'true' controls, in which the participants were not exposed to the target words and/or a learning condition (e.g., Shintani, 2011). The 'one sample' category (W-O) was created because one study (Lee & Muncie, 2006) stated a research question pointing toward a pretest–posttest comparison but then failed to adequately report the inferential test. However, as this group only contained a single study, it was removed from

TABLE 1
Group-Mean Comparison Types

Comparison	Code	Description of Type	Example
Between group	B-TCtrl	Treatment group–'true' control group comparisons	A treatment group reads and practices target collocations, while a true control group is not exposed to the collocations beyond the pretest (e.g., Webb & Kagimoto, 2009)
	B-TCom	Treatment group–comparison group comparisons	A treatment group studies target vocabulary with strategies, while a comparison group studies target vocabulary without strategies (e.g., Mizumoto & Takeuchi, 2009)
Within group	W-PP	Pretest–posttest comparisons	The results of a treatment group's pretest and posttest results are compared (e.g., Mizumoto & Takeuchi, 2009)
	W-TCom	Treatment group with one or more comparison conditions	An experimental group undergoes a retrieval (treatment) condition and then a context interference (comparison) condition and the proportion of correct answers produced under each condition are compared (e.g., van den Broek et al., 2018)
	W-CB	Counterbalanced groups	All participants experience all conditions, and presentation order is counterbalanced to account for the effects of treatment order (e.g., Rott, 2007)

the sample, leaving 81 studies and 453 mean comparisons for further analysis.

Effect Size Coding. When possible, effect sizes were estimated for the 453 mean comparisons from *M* (mean), *SD* (standard deviation), and *n* (sample size). For between-group effects, Cohen's d_s (Cohen, 1988) was initially calculated utilizing pooled *SD*s, which provide the most accurate estimate of the population variance (Lakens, 2013). For within-group effects, an effect size specifically for within-group comparisons, d_{av}, was calculated. Hedges's (1981) *g* correction was applied to all resulting effect sizes to account for small sample sizes (see Online Supporting Information C for details regarding effect size calculations). When studies failed to report these data, test statistics were employed instead. This extraction process was calibrated with the same 12 studies from the group-mean comparison extraction process. With the exception of one miscoded *n*, perfect agreement was observed, and we point to the calibration process as an explanation for this.

At the end of this process, 453 *g* values (*g*s)[3] were coded and as in Plonsky & Oswald (2014), we included more than one effect per study to capture the full heterogeneity of our sample—but we acknowledge the clustering trade-off of this design choice. Due to skewed distributions, median and interquartile range (IQR) for each group-mean comparison type was reported instead of *M* and *SD*. All negative *g* values were converted to absolute values because magnitude, not direction, was of interest.

Hedges's *Q* test and the accompanying I^2 effect size (Hedges & Olkin, 1985) were considered to estimate how much of the variance in the Hedges's *g* effect sizes was true. We avoided this approach for three reasons: First, our study was not a comprehensive meta-analysis, and only five journals were investigated. Second, the full heterogeneity of effects was of interest and descriptive analysis facilitated this (see Plonsky & Oswald, 2014). Third, effects were clustered by paper and therefore not independent, which is an assumption of the *Q* test and I^2.

Results

The results addressing the first RQ are displayed in Table 2, and indicate that the most common group-mean comparison type to address L2 IVA hypotheses and research questions in the study sample was B-TCom ($n = 237$), which accounted for more than half of the comparisons of

TABLE 2
Descriptive Statistics for Number of Effect Sizes and Comparison Types by Journal

			Group-Mean Comparison Type				
Journal	*k*	*n*	B-TCtrl	B-TCom	W-PP	W-TCom	W-CB
Language Learning	10	98	16	28	17	12	25
LTR	35	184	1	89	31	17	46
MLJ	8	31	0	27	1	0	3
System	16	80	10	56	3	8	3
TESOL Quarterly	12	60	2	37	15	0	6
Total	81	453	29	237	67	37	83

Note. *k* = number of studies; *n* = number of effect sizes; B-TCtrl = between-group treatment–control; B-TCom = between-group treatment–comparison; W-PP = within-group pretest–posttest; W-TCom = within-group treatment–control; W-CB = within-group counterbalanced groups; *LTR* = *Language Teaching Research*; *MLJ* = *The Modern Language Journal*.

TABLE 3
Minimum, Median, and Maximum Total Sample Sizes by Comparison Type

Type	*n*	Minimum	Median	Maximum
B-TCtrl	29	36	161	161
B-TCom	237	24	99	729
W-PP	67	17	54	178
W-TCom	37	20	41	65
W-CB	83	20	54	778

Note. B-TCtrl = between-group treatment–control; B-TCom = between-group treatment–comparison; W-PP = within-group pretest–posttest; W-TCom = within-group treatment–control; W-CB = within-group counterbalanced groups. B-TCtrl median and maximum are identical because 15 of the 29 effect sizes came from a single study. See Methods and General Discussion for justification of clustering trade-off.

TABLE 4
Effect Size (*g*) Median with 25th and 75th Percentiles and Interquartile Range by Group-Mean Comparison Type

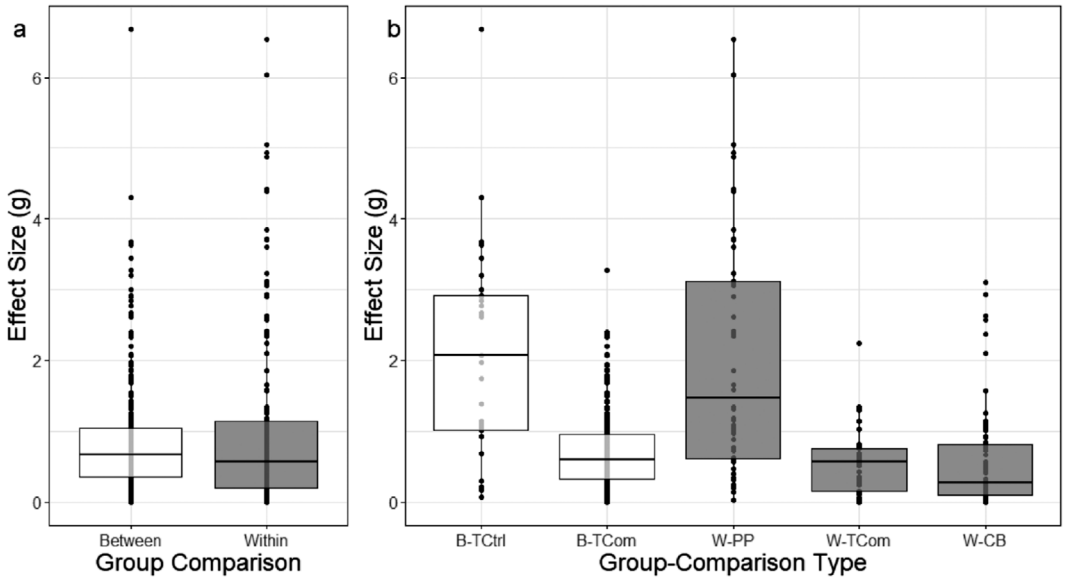
Comparison	Type	*n*	25%	Median (*g*)	75%	IQR
Between groups	General	266	.35	.67	1.04	.70
	B-TCtrl	29	1.01	2.07	2.91	1.91
	B-TCom	237	.33	.62	.97	.64
Within groups	General	187	.21	.66	1.65	1.44
	W-PP	67	.77	1.84	3.26	2.49
	W-TCom	83	.15	.57	.75	.60
	W-CB	37	.09	.25	.79	.69

Note. IQR = interquartile range; B-TCtrl = between-group treatment–control; B-TCom = between-group treatment–comparison; W-PP = within-group pretest–posttest; W-TCom = within-group treatment–control; W-CB = within-group counterbalanced groups. Values have been rounded, therefore 25% + IQR does not always equal 75%.

most interest. This was followed by W-CB ($n = 83$), W-PP ($n = 67$), W-TCom ($n = 37$), and B-TCtrl ($n = 29$). Table 3 displays the minimum, median, and maximum sample sizes employed by L2 IVA researchers for each comparison type.

In order to address the second RQ, regarding the observed range of effect sizes for each of the group-mean comparison types, the median *g*s and accompanying 25th and 75th percentiles for each group-mean comparison type were calculated (see Table 4). Due to the heavily skewed effect size distributions rendering parametric tests inappropriate, decisions regarding significant differences between groups were based upon

FIGURE 3
Boxplots of Effect Sizes (*g*) for (a) Between- and Within-Group Comparisons and (b) Group-Mean Comparison Types



nonoverlap between boxplots (Streiner, 2018). The boxplots in Figure 3a compare the general between- and within-group effect sizes, while Figure 3b displays the same set of effect sizes distributed between the group-mean comparison types utilized by L2 IVA researchers. In both Figures 3a and 3b, between- and within-group comparisons are unshaded and shaded, respectively. Figure 3a suggests that the general between- (median = .67; IQR = .70) and within-group (median = .66; IQR = 1.44) effect sizes observed in L2 IVA research are almost identical, with the within-group having a larger spread. However, once the effect sizes are organized by group-mean comparison type, clusters inside these two general groups become visible. The lack of overlap between the B-TCtrl (median = 2.07; IQR = 1.91) and B-TCom (median = .62; IQR = .64) boxplots in Figure 3b indicates that a significant difference exists between the effect sizes observed for these two group-mean comparison types. A similar pattern is visible for within-group effect sizes, whereby the W-PP (median = 1.84; IQR = 2.49) effect sizes are significantly larger than those of the W-TCom (median = .57; IQR = .60) and W-CB (median = .25; IQR = .69) groups, as attested to by the lack of overlap. The boxplots illustrate that the distribution of the effect sizes from W-TCom and W-CB designs are similar, despite the median effect of the W-TCom effect sizes being twice as large as the W-CB median.

Discussion

The first RQ pertained to the observed distribution of group-mean comparison types within L2 IVA (quasi)experimental designs. The results indicated that a simple between- and within-group distinction lacked the sensitivity required to account for the contrasting effect sizes associated with the varying designs, particularly for within-group designs. With regard to between-group designs, the majority of the 266 effect sizes resulted from B-TCom comparisons (n = 237; 89.10%), along with 29 (10.90%) B-TCtrl effect sizes (see Table 2). With regard to the 187 effect sizes from within-group designs, 67 (35.83%) were from W-PP comparisons, 83 (44.39%) were from W-CB comparisons, and 37 (19.79%) were W-TCom designs. The existence of these distinct comparison types clustered within the general within-group division is potentially problematic for determining sample sizes. For example, if a researcher powered a W-CB design with the general within-subject value (g_{av} = .66), the resulting sample size would be underpowered because the median W-CB value is lower (g_{av} = .25; see Table 4). The more fine-grained categorization scheme of this current study appears to be more suitable for these estimates.

The observed range of effect sizes for each of the five group-mean comparison types guided the second research question. In the case of

between-group comparisons, the majority were B-TCom and the distribution of their effect sizes was similar to Plonsky and Oswald's (2014) between-group thresholds for L2 research. Whereas the distribution of Plonsky and Oswald's 25th percentile, median, and 75th percentile meta-analytic *d*s was .38, .62, and 1.19 (IQR = .81), the distribution of B-TCom g_ss in the present study was .33, .62, and .97 (IQR = .64). This indicates that between-group L2 IVA research effect sizes are predominantly comparable to L2 research in general. However, the effect sizes observed in B-TCtrl comparisons (median $g_s = 2.07$; IQR = 1.91) formed a significantly different group, as revealed by a lack of overlap between the boxplots (see Figure 3b). The largest B-TCtrl effect size was observed in response to a research question asking "Does participation in input-based tasks enable young L2 learners to acquire new vocabulary?" (Shintani, 2011, p. 140), resulting in a remarkably large effect size ($g_s = 6.68$).[4]

The excessively large effect sizes observed for B-TCtrl comparisons suggest that comparisons between true control groups and treatments should not be the focus of research questions and sample-size determinations. Such effect sizes have led some meta-analysts to exclude true control comparisons from their analyses. For instance, Bryfonski and McKay (2019) coded task-based versus other learning condition contrasts in their meta-analysis of task-based interventions while effects involving true controls appeared to be omitted. This is not to posit that control groups are without use. When utilized wisely, control groups allow researchers to establish baselines, detect test reactivity, and discover confounding variables (Kuehl, 2000). Also, without control groups it becomes difficult to distinguish whether improvement between times of testing was the result of the treatment, test effects, or participants adapting to the test demands. Nevertheless, control group comparisons are perhaps unsuitable for specifically addressing L2 IVA hypotheses because they do not concern the substantial issue of which learning condition is better for learning vocabulary. Control groups that have not learned the target forms will most likely underperform compared to treatment and comparison groups. Such comparisons have and will result in oversized effect sizes that potentially bias the results of meta-analyses and power analyses.

Of the three within-subject group-mean comparison types, Figure 3b illustrates that the W-PP group effect sizes were significantly larger than both the W-TCom and W-CB groups, as attested to by lack of overlap between the boxplots. This suggests a difference between the effect sizes garnered from pretest–posttest comparisons and those obtained from treatment–comparison condition comparisons, regardless of the design of the latter (e.g., one group undergoing two learning conditions in counterbalanced designs). Furthermore, the general within-group effect size estimates (see Figure 3a) lacked the sensitivity to detect this discrepancy.

As with the B-TCtrl comparison types, remarkably large effect sizes were observed for certain W-PP comparisons and stemmed from research questions in which the answer was clear beforehand. For instance, in an otherwise exemplary study, Webb and Kagimoto (2009) exposed a group of learners to a set of collocations that they barely knew according to pretest results (M = 4.25 [out of 56]; $SD = 2.55$). The learners underwent a treatment and, unsurprisingly, considerable improvement was observed in the immediate posttest. This research design resulted in one of the largest observed effect sizes in L2 IVA research ($g_{av} = 4.93$). Small samples powered to such effect sizes would have little face validity. As with between-subject treatment–control comparisons, we are not suggesting L2 IVA researchers should avoid pretest–posttest comparisons, but that samples should not be powered to the tremendous effect sizes that result from straightforwardly predictable research questions. Furthermore, as discussed regarding B-TCtrl, W-PP comparisons also do not address meaningful L2 IVA hypotheses of relative learning method effectiveness.

Phase I was undertaken as a meta-analytic process to guide Phase II's power simulations. The results in Phase I pointed to treatment–comparison contrasts, B-TCom and W-CB, as being the most suitable in relation to representing L2 IVA comparisons of most interest. Although they have a use, B-TCtrl and W-PP comparisons resulted in large effects that did not address the substantive questions of which L2 IVA learning conditions worked better. While W-TCom comparisons did facilitate such substantive enquiries, these designs lack the counterbalancing to account for the effects of order of condition exposure (Kuehl, 2000). Thus, B-TCom and W-CB were the comparison types utilized in Phase II of the current study.

PHASE II: POWER ANALYSES

The RQs for Phase II of the present study were primarily concerned with uncovering appropriate sample sizes for future L2 IVA research. In order to answer the RQs, a series of a priori power simulations was conducted incorporating the

aggregated effect sizes from Phase I. Before the power analyses were conducted, it was deemed necessary to determine which (quasi)experimental designs L2 IVA researchers favored. Therefore, the first RQ for this phase was posed to uncover the (quasi)experimental design preferences of L2 IVA researchers:

RQ3. What is the observed distribution of (quasi)experimental testing designs within L2 IVA research?

This question acted as an a priori analysis to drive the subsequent RQs. In accordance with Phase I observations, the aggregated effect sizes from B-TCom and W-CB designs were utilized to answer the proceeding questions:

RQ4. What sample sizes are required to achieve appropriate power in between-subject testing designs favored by L2 IVA researchers?

RQ5. What sample sizes are required to achieve appropriate power in within-subject testing designs favored by L2 IVA researchers?

For RQ4, the 25th percentile ($g_s = .33$; hereafter *small*) and median ($g_s = .62$; hereafter *medium*) B-TCom effect sizes were employed in a priori power analysis simulations, while the median W-CB effect size ($g_{av} = .25$) was utilized in the calculations for RQ5. The 25th percentile W-CB effect size was omitted from consideration because it did not seem credible that L2 IVA researchers would deliberately set out to investigate a treatment that they predicted would result in an effect size so small ($g_{av} = .09$).

Methodology

In order to address RQ3, the testing designs observed in the sample of 81 L2 IVA (quasi)experimental designs were coded and counted. The design categories employed by Plonsky (2013) were modified to reflect L2 IVA design choices while striving for parsimony. In the end, studies' designs were coded as: 1 = analysis of covariance (ANCOVA) family, 2 = analysis of variance (ANOVA) family, 3 = multiple parametric bivariate tests (e.g., independent *t* tests), 4 = multiple nonparametric tests (e.g., Mann–Whitney U), 5 = linear and generalized mixed-effect (multilevel) modeling, and 6 = other/nondefined. Multivariate ANCOVA (MANCOVA) and multivariate ANOVA (MANOVA) were subsumed under 1 and 2, respectively, as when a study did implement such a design (e.g., Rott, 2007), the comparison of most interest was located in a subsequent one-way ANCOVA or ANOVA. In other words, the summative effect sizes derived from Roy's largest root did not address L2 IVA hypotheses. Factorial ANOVA designs (e.g., Kim, 2008) were likewise conflated with Category 2 as they could easily be reconceptualized as one-way designs, and researchers never considered the summative effect of the complete model to address the L2 IVA hypotheses. Both researchers independently coded the studies and 93.90% agreement ($\kappa = .90$) was observed.

The results addressing RQ3 (see Results section) revealed L2 IVA researchers' preference for ANOVA and ANCOVA testing designs. Therefore, ANOVA and ANCOVA drove the power analyses addressing RQ4 and RQ5. Following the conventions of educational and broader social science methods, the a priori power simulations assumed Type I (α) and Type II (β) error thresholds of 5% and 20% (i.e., 80% power, $1 - \beta$), respectively (Field, 2018). Because ANOVA and ANCOVA testing designs often require post hoc comparisons, these were also considered in the a priori power simulations. To achieve this in the power simulations, the *g*s obtained in Phase I required conversion into partial eta-squared[5] (η_p^2; see Online Supporting Information C for details). For the between-subject power calculations, the observed small ($g_s = .33$) and medium ($g_s = .62$) B-TCom g_ss were converted, resulting in η_p^2 effect sizes of .03 and .09, respectively. The within-group g_zs (.21, .25, and .31) were also transformed resulting in η_p^2 effect sizes of .04, .05, and .09, respectively.

Once all the relevant effect sizes were calculated, samples sizes were calculated with G*Power (Faul et al., 2007). Two testing designs were investigated to determine sample sizes for between-group comparisons: one-way omnibus ANOVA and ANCOVA and one-way omnibus ANOVA and ANCOVA with post hoc comparisons. One-way AN(C)OVA power simulations return the same number of participants and thus they were conflated in the simulations (Faul et al., 2007). G*Power provided drop-down menus and boxes to insert the parameters of the calculations for all simulations. For AN(C)OVA with post hoc tests, a one-tailed *t* test setting was used because the preceding omnibus determined the direction of the different contrasts (Brysbaert, 2019). Also, a conservative Bonferroni adjustment ($\alpha/3$) accounted for three comparisons (e.g., treatment vs. comparison, treatment vs. control, comparison vs. control). This decision was based on the fact

that ANOVA analyses involving post hoc tests that hold sample size constant from the omnibus analysis between experimental groups (e.g., Tukey or Games–Howell) spread the Type I error of the omnibus ANOVA across the three comparisons. This maintains Type I error control (significance) while losing power and increasing Type II error (Gravetter & Wallnau, 2015). The G*Power output provided sample sizes for two groups, and following Brysbaert (2019), a third group was added of equal size.

To ascertain sample sizes for within-group comparisons, a priori power simulations were conducted with the three η_p^2 effect sizes (.04, .05, and .09) derived from the g_zs in three separate analyses, which approximated weak, medium, and strong correlations between the repeated measures.[6] For each of the three effect sizes, G*Power was utilized and appropriate sample sizes were determined for three-group counterbalanced repeated measures ANOVA (RM-ANOVA; three measurements: treatment, comparison, control with three counterbalanced groups) and RM-ANOVA with post hoc comparisons. The standard G*Power menu for RM-ANOVA requires the correlation coefficient for the repeated measures. However, because this was accounted for in the η_p^2 effect sizes the 'as in SPSS' G*Power option was selected, which removes the requirement to enter the correlation into the calculation for a second time. As with the simulations presented in Brysbaert (2019), the calculations assumed the default nonsphericity correction of 1. Figures illustrating the parameters utilized in all G*Power calculations are provided in Online Supporting Information D.

Results

With regard to the first RQ for Phase II (RQ3), Table 5 illustrates that despite L2 researchers' increasing interest in mixed-effect modeling (e.g., Cunnings & Finlayson, 2015), the majority of L2 IVA researchers utilized either ANOVA ($k = 51$; 62.96%), multiple t tests ($k = 10$; 12.34%), or ANCOVA ($k = 9$; 11.11%). L2 IVA researchers' clear preference for ANOVA is typical of L2 research in general (e.g., Plonsky, 2013) and because more than 73% of the sample studies utilized ANOVA or ANCOVA, the focus of Phase II's power simulations on these designs was justified.

Once ANOVA and ANCOVA were identified as L2 IVA researchers' experimental testing designs of choice, a priori power analyses utilizing the between-group effect sizes from Phase I were conducted in order to address the Phase II second RQ (RQ4). Table 6 displays the suggested sample sizes to detect medium and small effect sizes for B-TCom AN(C)OVA, and AN(C)OVA with post hoc test designs. For researchers wishing to employ a simple design involving a posttreatment contrast of treatment and comparison groups, samples of 292 and 84 are required to achieve power, assuming small and medium B-TCom effects, respectively. When a control group or a second comparison is added, 357 or 105 participants are required (one-way AN[C]OVA). The number of subjects per group decreases paradoxically because the omnibus test requires the significance to only exist somewhere among the groups (see Brysbaert, 2019). When post hoc comparisons are conducted to isolate the contrasts that meaningfully address the research questions, the number of required participants increases to 492, or 164 per group.

TABLE 5
Distribution of (Quasi)Experimental Design Across 81 L2 Instructed Vocabulary Acquisition Studies

Testing Design	*k*	%
ANCOVA family	9	11.11
ANOVA family	51	62.96
Multiple bivariate parametric tests	10	12.34
Multiple nonparametric tests	3	3.70
Linear and generalized mixed-effect modeling	5	6.17
Other	3	3.70

Note. ANCOVA = analysis of covariance; ANOVA = analysis of variance.

The results in Table 7 relate to the third RQ in Phase II (RQ5), which was posed to ascertain appropriately powered sample sizes for within-subject L2 IVA (quasi)experimental designs. The correlation coefficients (r) in the table represent small, medium, and large repeated measure correlations extracted from L2 IVA studies (see Online Supporting Information E). These coefficients were then utilized with the observed medium effect for W-CB ($g_{av} = .25$) to calculate g_z, which was necessary to calculate η_p^2, which in turn was required for the G*Power analysis (see Online Supporting Information C for the effect size calculation equations). Table 7 shows that when $r = .50$, effect sizes g_{av} and g_z are equal (.25). When the correlation is between zero and approaching .50, g_z becomes less than g_{av}, with the difference narrowing as the correlation approaches .50. When the observed correlation is greater than .50, g_z becomes greater than g_{av}, with the difference

TABLE 6
Between-Subject (80%) Power Simulations and Suggested Sample Sizes in Total (N) and by Group (n)

Effect Size	g_s	η_p^2	Design	α	N	n
Small	.33	.03	t test	.05	292	146
			One-way omnibus AN(C)OVA	.05	357	119
			AN(C)OVA + post hoc tests	.02	492	164
Medium	.62	.09	t test	.05	84	42
			One-way omnibus AN(C)OVA	.05	105	35
			AN(C)OVA + post hoc tests	.02	144	48

Note. ANCOVA = analysis of covariance; ANOVA = analysis of variance. See Online Supporting Information D for exact values. Exact value of corrected α (+ post hoc tests) is .0167 (.05/3). Effect sizes assumed for post hoc tests' power calculations were stated g_s values.

TABLE 7
Within-Subject (80%) Power Simulations and Suggested Total Sample Sizes for the Median W-CB Effect Size Observed in Phase I ($g_{av} = .25$)

r	g_z	η_p^2	Design	α	N
.31	.21	.04	t test	.05	180
			RM-ANOVA (counterbalanced)	.05	114
			RM-ANOVA + post hoc tests	.02	203
.50	.25	.06	t test	.05	128
			RM-ANOVA (counterbalanced)	.05	84
			RM-ANOVA + post hoc tests	.02	144
.68	.31	.09	t test	.05	84
			RM-ANOVA (counterbalanced)	.05	57
			RM-ANOVA + post hoc tests	.02	95

Note. W-CB = within-group counterbalanced groups comparison; RM-ANOVA = repeated measures analysis of variance. See Online Supporting Information D for exact values. Exact value of corrected α (+ post hoc tests) is .0167 (.05/3). Effect sizes assumed for post hoc tests' power calculations were stated g_z values.

widening as the value approaches 1.00. Thus, Table 7 reflects the variation on sample-size requirements resulting from the correlation between the repeated measurements, holding g_{av} constant at .25.

Based upon the effect size aggregated from within-group counterbalanced designs, the minimum sample sizes required to achieve 80% power in L2 IVA research are 180, 128, or 84 for a pairwise t test between two conditions, depending on the strength of the correlation between the repeated measures. When researchers use three or more conditions, the counterbalanced RM-ANOVA is efficacious because it determines the direction of differences among the conditions and thus one-way t tests can be employed at the post hoc level, which increases power while holding sample size constant (see Brysbaert, 2019). However, for a counterbalanced RM-ANOVA with post hoc tests to determine which group comparison was responsible for an effect, the sample-size numbers increase to 203, 144, or 95, depending on the strength of the correlation between the repeated measures.

Discussion

The sample sizes suggested in Table 6 for appropriate statistical power in L2 IVA between-subject (quasi)experimental designs varied from 292 to 492 participants to detect a relatively small effect, and from 84 to 144 to detect a relatively medium effect. These results are comparable to Brysbaert's (2019) default frequentist analysis sample sizes for psychology, which suggested 200 to 435 participants when $d = .40$, and 90 to 195 when $d = .60$ (see 'I = II > III' values in Table 7 in Brysbaert, 2019). In juxtaposing Brysbaert's suggestions with ours, we must highlight that our approach was slightly different. In our calculations, assumed effects were a comparison (or comparisons) of most interest (competing L2 IVA interventions) within the model(s). Although these requirements

are daunting, the initial shock induced upon discovering the large sample sizes suggested by power analysis is "far less than the pain of actually running dozens of subjects and finding highly uncertain estimates" (Kruschke, 2015, p. 395). While it might be tempting for future researchers to power according to the medium effect size, $g_s = .62$, in order to justify smaller sample sizes, caution should be exercised. Many studies in the sample were simple designs that omitted covariates and confounds. Also, the median effect sizes for L2 research observed by Plonsky and Oswald (2014) in between-, $d = .70$, and within-subject, $d = 1.00$, designs are larger than the median found in psychology and social sciences, $d = .40$ (Brysbaert, 2019). This is most likely due to the relative immaturity of L2 methods in general (Plonsky & Oswald, 2014), which are improving and currently under a state of reform. It is therefore prudent, when possible, to attempt to power according to the small effect size found in Table 6.

In comparison, the effect sizes suggested for within-group designs in Table 7 are more achievable. A robust counterbalanced design involving a treatment, comparison, and control condition necessitates between 95 and 203 participants, depending on the correlation between the repeated measurements. A simpler two-condition design requires 84 to 180 subjects. These figures are larger than those proposed by Brysbaert (2019), who suggested between 52 ($d = .40$) and 24 ($d = .60$) participants for a simple within-subject t test comparing two repeated measures and between 75 and 35 participants (same effect parameters) for comparing three within-subject repeated measurements (see 'I = II > III' values in Table 7 in Brysbaert, 2019). However, this discrepancy resulted from the effect sizes in the present study (see Table 7) being smaller than those utilized by Brysbaert while, as highlighted earlier, our power analyses focused on the effect of the comparison of most interest. The recommended samples sizes in Table 7 also demonstrate the great extent to which correlation between repeated measures affects required sample sizes for adequate power. Because these three r coefficients were estimated from correlations calculated from just three studies (see Online Supporting Information E), they might not be representative of all L2 IVA research. Until more research investigates these correlations, or more researchers share their data so that these values can be calculated, it is prudent for a priori planning to assume a correlation coefficient of either .31 or .50. Assuming an r of .68 would result in an underpowered sample if the observed post hoc correlation is less. As a final note, RM-ANOVA alone cannot address L2 IVA hypotheses when there are three or more conditions and its purpose is to facilitate one-way post hoc comparisons, which conserves power.

GENERAL DISCUSSION

The findings of the present study indicate that future L2 IVA researchers should attempt to construct larger samples in order to achieve appropriate power. We discuss this here and suggest collaboration among researchers across institutions and alternative statistical analyses as solutions for this problem. Finally, the pedagogical implications and limitations of the study are presented.

Phase II, which referenced meta-analyzed effects uncovered in Phase I, suggests that future L2 IVA research requires larger participant numbers than those in Phase I's study sample. The median total sample size observed in B-TCom and W-CB comparisons was 99 and 54, respectively (see Table 3). While two or three groups of 50 participants would meet most of the suggested sample-size thresholds assuming a medium effect ($g = .62$), they would not satisfy the small effect size thresholds. The observed W-CB median of 54 participants is smaller than all suggestions in Table 7. Future sample sizes should thus be planned using effect-size-driven a priori power analyses.

It is important to remember that none of the sample studies conducted a priori power analysis referencing effect sizes from previous studies. Caution should be exercised when interpreting results from such samples because they might be too small to detect small population effects, while observed large effects might not actually exist (Brysbaert, 2019). In a recent study, Tseng et al. (2020) employed a mixed ANOVA design and observed a nonsignificant difference between paired (teacher- and individual-centered) and teacher-centered virtual environment experimental (treatment vs. comparison) groups' performance on the posttest ($d = .49$). With a subsample of 48 (two groups of 24), it is entirely possible that this effect is real within the population but the study was underpowered to detect the difference. Conversely, Rassaei (2020) observed huge effects between a dynamic glossing mobile-assisted language learning (MALL) vocabulary treatment group and a nondynamic glossing comparison group in relation to receptive ($d = 2.87$) and productive ($d = 2.20$) vocabulary acquisition. Because each group only had 13 subjects from the same school, caution should be exercised if assuming such a large effect exists in the broader population. Both of these examples point

to why it is advisable for researchers to consider effect sizes from previous studies and to power their samples accordingly.

However, caution is also warranted when power analysis is informed by effect sizes determined from previous research. As mentioned previously, the median sample for L2 research has been estimated as merely 19 participants (Plonsky, 2013). Brysbaert's (2020) simulation of within-group comparisons demonstrated that sample sizes consisting of fewer than 30 participants rarely detected the correct effect size with accuracy. The results were even worse for between-group comparisons and designs involving interactions. Furthermore, if an effect size of $d = .60$ is necessary for a significant result, and if only significant results are published, the effect sizes accrued from previous research will tend toward $d = .60$, regardless of the true population effect.[7] When these points are considered, it is possible that effect sizes obtained from previous research, including those presented in the present study, are somewhat influenced by sampling error. Therefore, the sample sizes indicated from power analyses conducted with empirically derived effect sizes should be the minimum sought.

It would be unreasonable to call for larger sample sizes without offering suggestions for how the issue could be resolved. The first solution that we propose is collaboration between researchers. Previous L2 vocabulary research displays evidence of such collaboration. For instance, the relationship between vocabulary size measurements and TOEIC reading scores was investigated by Stoeckel et al. (2019), and involved 200 English-as-a-foreign-language participants from four universities across Japan. Also, 214 participants from the same population spread across two universities were recruited by McLean, Kramer, and Beglar (2015) in order to validate a listening vocabulary levels test. Such collaboration between researchers and institutions not only allows researchers to conduct experiments with larger samples but also involves samples from multiple testing sites. This offers an advantage because single-site samples are only truly representative of the education location from which they come, with broader generalization being merely argumentative (see Morgan–Short et al., 2018; Vitta & Woollock, 2019). As highlighted previously, there are single-site studies in our sample that contain either reasonable yet nonsignificant effects (e.g., Tseng et al., 2020) or very large effects (e.g., Rassaei, 2020) from small samples in single settings. Replication of these studies incorporating larger samples that have been appropriately powered could be facilitated by collaboration among researchers across institutions. In presenting this solution, we concede that when the target language is not commonly taught, such collaboration might be more challenging but the benefits of the effort are clear.

The second potential solution for underpowered L2 IVA experimental designs is the utilization of linear mixed-effects models (LMMs). LMMs are a form of regression that allows variance from both fixed and random effects to be modeled. Of particular benefit for vocabulary treatments is the fact the variance exhibited by both the participants and the items can be modeled (Baayen, 2008; Linck & Cunnings, 2015). Although LMM analysis is the gold standard in fields such as psycholinguistics, it is still emergent in L2 research (Siyanova–Chanturia & Spina, 2020). For instance, LMM analysis was utilized in merely five (6.17%) of the studies in the present sample. Claims have been made that LMMs offer an advantage over ANOVA with regard to statistical power and precision when operationalizing experimental designs (Singmann & Kellen, 2019). However, we must stress that LMMs do not constitute a 'silver bullet' solution for all underpowered and undersampled L2 IVA research designs. For instance, the complexity in the 'maximal models' promoted by Barr et al. (2013) have been shown to result in a loss of power (Matuschek et al., 2017). In other words, LMMs provide a clear power advantage under some conditions but perhaps not all. Despite this, they undoubtedly offer potential for certain research designs. For example, Brysbaert and Stevens (2018) proposed 40 participants and 40 items per condition as sufficient for psycholinguistic research. Because psycholinguistic and L2 IVA research are substantially different, power analysis packages such as SIMR (Green & MacLeod, 2016) should be utilized to investigate the requirements for L2 IVA (quasi)experimental designs.

Finally, L2 IVA could potentially account for the biases inherent in small sample sizes through Bayesian analysis, which is being promoted by L2 researchers (e.g., Norouzian, de Miranda, & Plonsky, 2019; Ross & Mackey, 2015). A Bayesian approach considers a parameter as a variable with a multitude of possible values, unlike frequentist analyses, which assume a true population parameter (Lee & Wagenmakers, 2013). The uncertainty surrounding a parameter of interest is governed by probability, and is dependent upon prior information, such as data from previous studies. In Bayesian analyses, observed data are used to update the prior information, which then becomes the posterior information and can be used as prior information for future data.

Importantly, unlike frequentist approaches, Bayesian analyses are not governed by the central limit theorem and thus small to moderate sample sizes are potentially sufficient for reasonable results (Miočević, Levy, & van den Schoot, 2020).

It must be emphasized that the Bayesian advantage with respect to smaller sample sizes is contingent on prior information. For realization of these benefits, Bayesian analysis requires power analysis, which necessitates parameter values from a theoretically informed distribution (prospective power) or a posterior distribution (retrospective power). Hundreds, if not thousands, of simulations are then performed using random samples of data generated from these distributions, and power is defined by the proportion of times the goal of the analysis is accomplished (Kruschke, 2015). It is also vital to emphasize that this method is distinct from Bayesian analysis utilizing uniformed priors, such as the default values provided in JASP (JASP Team, 2020). Previous research has suggested that Bayesian analyses using such uniformed priors require approximately double the number of participants for appropriate power when compared with their frequentist counterparts (Brysbaert, 2019). Furthermore, the utilization of Bayesian analysis to account for smaller sample sizes should only be considered as a last resort, and collecting more data from larger samples should always be prioritized (van de Schoot & Miočević, 2020).

The pedagogical implications of the present study relate to theory development. Empirical research is essential for developing the pedagogical theory that fuels frontline language teachers' practice. If L2 IVA research continues to produce methodological solutions to pedagogical questions based upon underpowered (quasi)experimental research designs, it is not clear whether these solutions will replicate between teaching contexts, or even be truly beneficial for language learners. By employing appropriately powered sample sizes determined by the methodology and/or recommendations proposed here, empirical L2 IVA research will have a greater chance of guiding successful language teaching pedagogy.

Despite our best intentions, the present study is not without limitations, specifically relating to generalizability, publication bias, and clustering. First, the present study focused on L2 IVA research and our approach's generalizability to other L2 subdomains is unclear. Our between-group effect sizes, however, were similar to those found for L2 research in general (Plonsky & Oswald, 2014). Thus, it is very possible that some of this study's findings, such as the different range of effect sizes observed for different between- and within-group designs, will generalize to other subdomains. Second, our sample's derivation from five journals means that the ranges of effect sizes might be inflated due to publication bias (Correll et al., 2020). However, even if the effect sizes are inflated, the sample-size recommendations that they produced are smaller than sample-size recommendations that would be produced by smaller effect sizes. Thus, our conclusion that L2 IVA sample sizes are too small still stands, and the sample sizes suggested previously become the bare minimum to be considered with the respective (quasi)experimental designs. Finally, by following Plonsky & Oswald (2014) and analyzing each comparison embedded with studies, this study was limited by a clustering issue, visible in the identical values for median and maximum B-TCtrl sample size in Table 3. However, this decision allowed us to capture the full heterogeneity within our sample. Furthermore, the variable most affected, B-TCtrl, was not included in Phase II of the study.

CONCLUSION

The present study involved a two-phase process to reform sample-size planning within L2 IVA experimental research. In Phase I, effect sizes from 81 L2 IVA studies were aggregated and analyzed by group-mean comparison type. The conventional within- and between-subject design categorization scheme was observed as lacking sensitivity, and the scheme was expanded to include treatment comparisons, control comparisons, pretest–posttest comparisons, and counterbalanced designs. The results of Phase I indicated that treatment-versus-comparison performance in both between- and within-subject designs (coded as B-TCom and W-CB) captured the L2 IVA main effects most accurately. In Phase II, the median effect sizes from these comparisons were then utilized in a priori power simulations, and a set of effect-driven sample sizes were proposed for future research. The required sample sizes were larger than the majority of sample sizes employed in previous L2 IVA research, indicating that future researchers should strive to recruit larger samples. To this end, L2 IVA researchers are advised to collaborate with researchers from similar teaching contexts in order to increase sample sizes and to consider that mixed-effect models and Bayesian analysis under certain conditions may present a power advantage vis-à-vis frequentist ANOVA (fixed-effect) designs.

ACKNOWLEDGMENTS

The authors would like to acknowledge Joy Ebgert, Luke Plonsky, and Ali H. Al–Hoorie for their review of the journal selection process. The authors also thank Marc Brysbaert, Garrett DeOrio, Daniël Lakens, Erin Maer, Stuart McLean, and Reza Norouzian for their valuable assistance with this project.

NOTES

[1] (Quasi)experimental conflates quasi-experimental and experimental designs, and is used to conserve space.

[2] One of the six candidates, *ELT Journal*, was not included in both lists and was removed. From this observation came the tacit implication that the journal might not be viewed as having a strong focus on quantitative inquiries, and this was confirmed by the researchers via a review of its published aims and scope.

[3] *g*s subsumes g_s and g_{av} hereafter.

[4] Input-based treatment ($M = 21.9$ [sic], $SD = 2.8$ [sic], $n = 13$); true control group ($M = 4.7$ [sic], $SD = 2.1$ [sic], $n = 12$).

[5] Although η^2 and η_p^2 are different, and in certain parts of this study η^2 is more appropriate, we have employed η_p^2 throughout for three reasons. First, it is the effect size referred to in the equations. Second, all instances of η^2 in this study are equal to η_p^2 because there is nothing to partial out. Finally, we employed η_p^2 throughout for the sake of consistency.

[6] See Online Supporting Information E. Because of underreporting of results, only three studies were utilized to estimate the correlation values.

[7] We would like to thank Marc Brysbaert for highlighting this to us.

Open Research Badges

This article has earned an Open Data badge for making publicly available the components of the research methodology needed to reproduce the reported procedure and analysis. All materials are available at http://www.iris-database.org.

REFERENCES

Aberson, C. L. (2019). *Applied power analysis for the behavioral sciences*. New York: Routledge, Taylor & Francis Group.

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.

Baayen, R. H. (2008). *Analyzing linguistic data*. Cambridge: Cambridge University Press.

Bacchetti, P., McCulloch, C., & Segal, M. (2008). Simple, defensible sample sizes based on cost efficiency. *Biometrics*, *64*, 577–585.

Bacchetti, P., Wolf, L., Segal, M., & McCulloch, C. (2005). Ethics and sample size. *American Journal of Epidemiology*, *161*, 105–110.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255–278.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Hoboken, NJ: Wiley.

Brooks, G. P., & Johanson, G. A. (2011). Sample size considerations for multiple comparison procedures in ANOVA. *Journal of Modern Applied Statistical Methods*, *10*, 97–109.

Browne, C., Culligan, B., & Phillips, J. (2013). The new general service list. Accessed 7 December 2020 at http://www.newgeneralservicelist.org

Bryfonski, L., & McKay, T. H. (2019). TBLT implementation and evaluation: A meta-analysis. *Language Teaching Research*, *23*, 603–632.

Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition*, *2*, 1–38.

Brysbaert, M. (2020). Power considerations in bilingualism research: Time to step up our game. *Bilingualism: Language and Cognition*. Advance online publication. https://doi.org/10.1017/S1366728920000437

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition*, *1*, 9.

Cambridge University Press. (2019). *2019 ELT Cambridge University Press international catalogue*. Cambridge: Cambridge University Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York: Routledge Academic.

Cohen, J. (1997). The Earth is round ($p < .05$). In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 21–35). Mahwah, NJ: Lawrence Erlbaum.

Correll, J., Mellinger, C., McClelland, G. H., & Judd, C. M. (2020). Avoid Cohen's 'small', 'medium', and 'large' for power analysis. *Trends in Cognitive Sciences*, *24*, 200–207.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*, 213–238.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York: Routledge.

Cunnings, I., & Finlayson, I. (2015). Mixed effects modeling and longitudinal data analysis. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 159–181). New York: Routledge.

Egbert, J. (2007). Quality analysis of journals in TESOL and Applied Linguistics. *TESOL Quarterly*, *41*, 157–171.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191.

Field, A. (2018). *Discovering statistics using IBM SPSS statistics*. London: SAGE.

Fisher, R. A. (1935). *The design of experiments*. Edinburgh, UK: Oliver and Boyd.

Gass, S., Loewen, S., & Plonsky, L. (2020). Coming of age: The past, present, and future of quantitative SLA research. *Language Teaching*. Advance online publication. https://doi.org/10.1017/s0261444819000430

González–Fernández, B., & Schmitt, N. (2019). Word knowledge: Exploring the relationships and order of acquisition of vocabulary knowledge components. *Applied Linguistics*, *41*, 481–505.

Gravetter, F. J., & Wallnau, L. B. (2015). *Statistics for the behavioral sciences*. Boston, MA: Cengage Learning.

Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*, 493–498.

Hedges, L. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*, 107–128.

Hedges, L., & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego, CA: Academic Press.

JASP Team. (2020). JASP (Version 0.14) [Computer software]. Accessed 17 December 2020 at https://jasp-stats.org/

Kim, M., & Crossley, S. A. (2018). Modeling second language writing quality: A structural equation investigation of lexical, syntactic, and cohesive features in source-based and independent writing. *Assessing Writing*, *37*, 39–56.

Kim, Y. (2008). The role of task-induced involvement and learner proficiency in L2 vocabulary acquisition. *Language Learning*, *58*, 285–325.

Kirk, R. E. (2012). *Experimental design: Procedures for the behavioral sciences* (4th ed.). Thousand Oaks, CA: SAGE.

Kremmel, B., & Schmitt, N. (2016). Interpreting vocabulary test scores: What do various item formats tell us about learners' ability to employ words? *Language Assessment Quarterly*, *13*, 377–392.

Kruschke, J. K. (2015). *Doing Bayesian data analysis: A tutorial with R and BUGS* (2nd ed.). Burlington, MA: Academic Press.

Kuehl, R. O. (2000). *Design of experiments: Statistical principles in research design and analysis*. Pacific Grove, CA: Duxbury.

Lakens, D. (2013). Calculating and reporting effects sizes to facilitate cumulative science: A practical primer for *t*-tests and ANOVAs. *Frontiers in Psychology*, *4*, 1–12.

Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, *44*, 701–710.

Laufer, B. (2005). Focus on form in second language vocabulary learning. In S. H. Foster–Cohen, M. Garcia–Mayo, & J. Cenoz (Eds.), *EuroSLA yearbook: Volume 5* (pp. 223–250). Philadelphia, PA: John Benjamins.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, *16*, 307–322.

Lee, M. D., & Wagenmakers, E.–J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge: Cambridge University Press.

Lee, S. H., & Muncie, J. (2006). From receptive to productive: Improving ESL learners' use of vocabulary in a postreading composition task. *TESOL Quarterly*, *40*, 295–320.

Lewis, M. (1993). *The lexical approach*. Hove, UK: Longman Teaching Publications.

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

Linck, J. A., & Cunnings, I. (2015). The utility and application of mixed-effects models in second language research. *Language Learning*, *65*, 185–207.

Loewen, S. (2015). *Instructed second language acquisition*. New York: Routledge.

Loewen, S., & Sato, M. (2017). Instructed second language acquisition (ISLA): An overview. In S. Loewen & M. Sato (Eds.), *The Routledge handbook of instructed second language acquisition* (pp. 1–12). New York: Routledge.

Matthews, J., & Cheng, J. (2015). Recognition of high frequency words from speech as a predictor of L2 listening comprehension. *System*, *52*, 1–13.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315.

McLean, S., Kramer, B., & Beglar, D. (2015). The creation and validation of a listening vocabulary levels test. *Language Teaching Research*, *19*, 741–760.

McLean, S., Stewart, J., & Batty, A. O. (2020). Predicting L2 reading proficiency with modalities of vocabulary knowledge: A bootstrapping approach. *Language Testing*, *37*, 389–411.

Miočević, M., Levy, R., & van den Schoot, R. (2020). Introduction to Bayesian statistics. In R. van den Schoot & M. Miočević (Eds.), *Small sample solutions: A guide for applied researchers and practitioners* (pp. 3–12). New York: Routledge.

Mizumoto, A., & Takeuchi, O. (2009). Examining the effectiveness of explicit instruction of vocabulary learning strategies with Japanese EFL university students. *Language Teaching Research*, *13*, 425–449.

Morgan–Short, K., Marsden, E., Heil, J., Issa, B. I., II, Leow, R. P., Mikhaylova, A., … Szudarski, P. (2018), Multisite replication in second language acquisition research: Attention to form during listening and reading comprehension. *Language Learning*, *68*, 392–437.

Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge. Cambridge University Press.

Nation, I. S. P., & Waring, R. (1997). Vocabulary size, text coverage and word lists. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 6–19). Cambridge: Cambridge University Press.

Nicklin, C., & Plonsky, L. (2020). Outliers in L2 research: A synthesis and data re-analysis from self-paced reading. *Annual Review of Applied Linguistics*, *40*, 25–55.

Norouzian, R. (2020). Sample size planning in quantitative L2 research: A pragmatic approach. *Studies in Second Language Acquisition*, *42*, 849–870.

Norouzian, R., de Miranda, M., & Plonsky, L. (2019). A Bayesian approach to measuring evidence in L2 research: An empirical investigation. *Modern Language Journal*, *103*, 248–261.

Pearson. (2019). *Pearson ELT catalogue*. New York: Pearson.

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.

Plonsky, L., & Gass, S. (2011). Quantitative research methods, study quality, and outcomes: The case of interaction research. *Language Learning*, *61*, 325–366.

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*, 878–912.

Rassaei, E. (2020). Effects of mobile-mediated dynamic and nondynamic glosses on L2 vocabulary learning: A sociocultural perspective. *Modern Language Journal*, *104*, 284–303.

Richards, B., Daller, M., Malvern, D., Meara, P., Milton, J., & Treffers–Daller, J. (2009). *Vocabulary studies in first and second language acquisition: The interface between theory and application*. Basingstoke, UK: Palgrave Macmillan.

Ross, S. J., & Mackey, B. (2015). Bayesian approaches to imputation, hypothesis testing, and parameter estimation. *Language Learning*, *65*, 208–227.

Rott, S. (2007). The effect of frequency of input-enhancements on word learning and text comprehension. *Language Learning*, *57*, 165–199.

Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge: Cambridge University Press.

Schmitt, N. (2008). Instructed second language vocabulary learning. *Language Teaching Research*, *12*, 329–363.

Schmitt, N. (2019). Understanding vocabulary acquisition, instruction, and assessment: A research agenda. *Language Teaching*, *52*, 261–274.

Shintani, N. (2011). A comparative study of the effects of input-based and production-based instruction on vocabulary acquisition by young EFL learners. *Language Teaching Research*, *15*, 137–158.

Singmann, H., & Kellen, D. (2019). An introduction to mixed models for experimental psychology. In D. H. Spieler & E. Schumacher (Eds.), *New methods in cognitive psychology* (pp. 4–31). East Sussex, UK: Psychology Press.

Siyanova–Chanturia, A., & Spina, S. (2020). Multi-word expressions in second language writing: A large-scale longitudinal learner corpus study. *Language Learning*, *70*, 420–463.

Stefan, A. M., Gronau, Q. F., Schönbrodt, F. D., & Wagenmakers, E.–J. (2019). A tutorial on Bayes factor design analysis using an informed prior. *Behavior Research Methods*, *51*, 1042–1058.

Stoeckel, T., Stewart, J., McLean, S., Ishii, T., Kramer, B., & Matsumoto, Y. (2019). The relationship of four variants of the Vocabulary Size Test to a criterion measure of meaning recall vocabulary knowledge. *System*, *87*, 102161.

Streiner, D. L. (2018). Statistics commentary series: Commentary # 14—Boxplots. *Journal of Clinical Psychopharmacology*, *38*, 5–6.

Trochim, W. M., Donnelly, J. P., & Arora, K. (2016). *Research methods: The essential knowledge base*. Boston, MA: Cengage Learning.

Tseng, W.–T., Liou, H.–J., & Chu, H.–C. (2020). Vocabulary learning in virtual environments: Learner autonomy and collaboration. *System*, *88*, 102190.

Tseng, W.–T., & Schmitt, N. (2008). Toward a model of motivated vocabulary learning: A structural equation modeling approach. *Language Learning*, *58*, 357–400.

van de Schoot, R., & Miočević, M. (2020). Introduction. In R. van de Schoot & M. Miočević (Eds.), *Small sample solutions: A guide for applied researchers and practitioners* (pp. viii–x). New York: Routledge.

van den Broek, G. S. E., Takashima, A., Segers, E., & Verhoeven, L. (2018). Contextual richness and word learning: Context enhances comprehension but retrieval enhances retention. *Language Learning*, *68*, 546–585.

Vitta, J. P., & Woollock, A. R. (2019). Improving Korean university EFL program instruction through language learning strategy research. *Asian EFL Journal*, *19*, 113–138.

Webb, S., & Kagimoto, E. (2009). The effects of vocabulary learning on collocation and meaning. *TESOL Quarterly*, *43*, 55–77.

Wilkins, D. A. (1972). *Linguistics in language teaching*. Cambridge, MA: MIT Press.

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

Zhang, X. (2020). A bibliometric analysis of second language acquisition between 1997 and 2018. *Studies in Second Language Acquisition*, *42*, 199–222.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.