Studies in Second Language Acquisition, **44** (2022), 1424–1448
doi:10.1017/S0272263121000541

Methods Forum

EFFECT SIZE–DRIVEN SAMPLE-SIZE PLANNING, RANDOMIZATION, AND MULTISITE USE IN L2 INSTRUCTED VOCABULARY ACQUISITION EXPERIMENTAL SAMPLES

Joseph P. Vitta*

Kyushu University

Christopher Nicklin

Rikkyo University

Stuart McLean

Momoyama Gakuin University

Abstract

In this focused methodological synthesis, the sample construction procedures of 110 second language (L2) instructed vocabulary interventions were assessed in relation to effect size–driven sample-size planning, randomization, and multisite usage. These three areas were investigated because inferential testing makes better generalizations when researchers consider them during the sample construction process. Only nine reports used effect sizes to plan or justify sample sizes in any fashion, with only one engaging in an *a priori* power procedure referencing vocabulary-centric effect sizes from previous research. Randomized assignment was observed in 56% of the reports while no report involved randomized sampling. Approximately 15% of the samples observed were constructed from multiple sites and none of these empirically investigated the effect of site clustering. Leveraging the synthesized findings, we conclude by offering suggestions for future L2 instructed vocabulary researchers to consider *a priori* effect size–driven sample planning processes, randomization, and multisite usage when constructing samples.

The research synthesis analyses in this article earned Open Materials and Open Data badges for transparent practices. The materials and data are available at https://www.iris-database.org/iris/app/home/detail?id=york%3A939467&ref=search

The authors wish to acknowledge Joy Egbert, Luke Plonsky, Ali H. Al-Hoorie, Phil Hiver, and Dayna Jost for the expert reviews regarding the report pool creation process and for their external consultations throughout the project. The authors also wish to thank the SSLA editors and the two external reviewers who provided useful and constructive feedback on multiple versions of this manuscript.

*Correspondence concerning this article should be addressed to Joseph P. Vitta, Kyushu University, Faculty of Languages and Cultures, Fukuoka, Japan. E-mail: vittajp@flc.kyushu-u.ac.jp and vittajp@gmail.com

†This article has been updated since its original publication. See https://doi.org/10.1017/S0272263121000814.

© *The Author(s), 2021. Published by Cambridge University Press.*

https://doi.org/10.1017/S0272263121000541 Published online by Cambridge University Press

INTRODUCTION

Since the 1990s and perhaps before, second language (L2) vocabulary and the continuing drive to reform quantitative practices in second language acquisition (SLA) research have been of particular interest to the field (Nicklin & Vitta, 2021). Considering the former, there has been a growing interest in vocabulary as the key to developing and accounting for L2 proficiency (McLean et al., 2020). Lewis (1993) for instance proposed an entire L2 curriculum based on the learning of multiword lexis while seminal studies such as Laufer and Nation (1995, 1999) demonstrated the primacy of vocabulary in predicting L2 outcomes. Considering the latter, L2 quantitative research in general is undergoing a methodological reform in which the field has displayed somewhat consistent and incremental improvement over time (Lindstromberg, 2016), with recent literature covering issues such as outlier treatment (Nicklin & Plonsky, 2020) and the need to check for publication bias when reporting meta-analyses (Vitta & Al-Hoorie, 2020). The reform is driven by the assumption that as a field, SLA "can only thrive and develop if our [SLA's] research demonstrates rigor" (Gass et al., 2021, p. 1).

The present study is a focused methodological synthesis conducted at the intersection between L2 instructed vocabulary experimental[1] research and the ongoing L2 methodological reform. Methodological synthesis is a type of systematic review or approach (Plonsky & Gonulal, 2015), which is an explicit process (comprising a system) whereby a set of reports meeting inclusion criteria are critically assessed to address research questions (O'Connor et al., 2008). As the name implies, a methodological synthesis concerns the methodologies observed in the reports. The present methodological synthesis was conducted to advance the generalizability and rigor of L2 instructed vocabulary research through three aspects of sample design. The first aspect considered was *sample-size determination* and involved an investigation into how researchers planned the number of participants that they recruited. Such planning references *effect sizes*, which are metrics that quantify the experimental and correlational effects observed in a given sample to facilitate generalizations about the population (Plonsky et al., 2021). Classical power is perhaps the best-known planning process and involves the determination of the minimum number of subjects required to significantly detect a predetermined (referencing relevant past studies) effect size in the sample (Brysbaert, 2019). The second aspect considered was *randomization*, which takes two distinct forms in experimental design. *Random selection* relates to the expectation that participants are chosen by chance and without bias (i.e., randomly) from the intended population (Harter, 2008), while *random assignment* concerns the allocation of participants to experimental conditions by chance (Kuehl, 2000). These two strands of randomization have historically been argued as essential features of experimental samples (Fisher, 1935). Because true random selection is impractical for most L2 researchers, particularly those investigating less common languages (Nicklin & Vitta, 2021; Vitta & Al-Hoorie, 2021), *multisite samples* have been presented as a compromise (Moranski & Ziegler, 2021), and also comprise the third aspect considered in the present study. Samples recruited from more than one location are an achievable means of attaining a sufficient number of participants to meet the requirements for L2 researchers' analyses of choice (Nicklin & Vitta, 2021), and also temper the bias that single-site samples introduce to quantitative inquiries (Morgan-Short et al., 2018).

https://doi.org/10.1017/S0272263121000541 Published online by Cambridge University Press

The case for viewing these three aspects as prerequisites for generalizable experimental sample designs is both historical and modern, with numerous prominent voices in statistical and L2 literature promoting the benefits of sample-size planning (e.g., Brysbaert, 2019; Cohen, 1988; Lakens et al., 2018), randomization (e.g., Fisher, 1935; Rogers & Révész, 2020), and multisite sampling (e.g., Moranski & Ziegler, 2021; Morgan-Short et al., 2018). However, assessment of published L2 vocabulary experimental samples in relation to these three aspects is relatively scarce. L2 vocabulary research syntheses are overwhelmingly meta-analytic and instead focus on aggregating observed effects (e.g., Lin & Lin, 2019; Zhang & Zhang, 2020). When issues such as sample size have been reviewed, such as by Elgort (2018), it was done so without overt discussion on the relationship between sample size planning and generalizations about the population. This should not be considered a flaw in Elgort's report because it was a systematic review of existing technology-mediated L2 vocabulary research. Thus, the present study aims to address this gap in the existing literature. The seed of the present methodological synthesis was rooted in a meta-analytic project promoting power-informed sample sizes for future L2 instructed vocabulary experiments (Nicklin & Vitta, 2021). During the coding process for this prior project, we observed that the majority of L2 instructed vocabulary studies included no overt sample-size planning, inconsistently reported randomization procedures, and overwhelmingly involved single-site samples. In the current study we aim to empirically quantify this prior informal observation.

LITERATURE REVIEW

Two themes in the literature underpin this current methodological synthesis. The first theme is the experimental nature of L2 instructed vocabulary research and its importance within SLA. The unpacking of this theme also highlights that vocabulary-centric methodological syntheses of sample designs appear to be lacking. The second theme relates to systematic reviews, and thus methodological syntheses, being theory-driven in relation to their focus (Siddaway et al., 2019). Accordingly, given the present study's focus, how L2 researchers and social scientists conceptualize effect size–driven sample-size planning, randomization, and multisite use is also reviewed.

L2 INSTRUCTED VOCABULARY RESEARCH: EXPERIMENTAL DESIGNS AND RESEARCH SYNTHESES

Vocabulary as a strong predictor of L2 proficiency has been established across skills, from listening and reading (e.g., McLean et al., 2020; Vafaee & Suzuki, 2020) to speaking and writing (e.g., Hulstijn et al., 2012; Vo, 2019). L2 instructed vocabulary research complements these investigations and pertains to "the best means of achieving good vocabulary learning" (Schmitt, 2008, p. 329). Nicklin and Vitta (2021) labeled the subfield as L2 instructed vocabulary acquisition (L2 IVA) to house it under instructed SLA (Loewen, 2015; Loewen & Sato, 2017) as L2 IVA inquiries are within the domain of instructed second language. Given the focus on best means, L2 IVA research tends to feature experimental designs. For instance, in a recent systematic review of technology-mediated L2 vocabulary research, Elgort (2018) observed that such designs were involved in 63 (77%) of 82 reports. These inquiries usually include a comparison of vocabulary

https://doi.org/10.1017/S0272263121000541 Published online by Cambridge University Press

learning conditions, operationalized as treatment and comparison groups, in which a set of target items are taught (Nicklin & Vitta, 2021).

L2 IVA experimental research has the potential to address theoretically grounded questions on how students learn new lexical items, where age, L1, and target items are controlled. Shintani (2013) presented a typical model of such inquiry. Her study compared task-based teaching with a teacher-led present-practice-produce (PPP) approach in the context of elementary Japanese English learners (see Willis & Willis [2007] for an overview of the debate between the two methodologies). Shintani chose a set of appropriate target adjectives and nouns according to corpus-based frequency values. The treatment group studied the words under a task-based condition featuring focus-on-form activities that were somewhat meaningful in representing language use outside of the classroom. The comparison group studied the target words under a PPP condition. Shintani's results indicated that adjectives were learned better under the treatment condition and in the discussion she explained why and how to researchers and teachers alike.

L2 vocabulary research syntheses, including L2 IVA such inquiries, have overwhelmingly been meta-analytic where the primary interest is on observed effects and not study quality (e.g., Lin & Lin, 2019; Uchihara et al., 2019; Webb et al., 2020; Yanagisawa et al., 2020; Zhang & Zhang, 2020). In the main, strong to moderate effects have been observed. For instance, in a summative review of 81 L2 IVA experiments, Nicklin and Vitta (2021) reported that treatment conditions performed 62% of a standard deviation better than comparison ones when the experiment employed independent groups (i.e., between-subject designs). Vocabulary researcher's meta-analytic focus on effect stands in contrast to L2 research syntheses in general where methodological syntheses are also somewhat common (see review in Farsani & Babaii, 2020). When vocabulary research design features have been synthesized, the focus appears to be on aspects of the treatment design with study quality often omitted. Lin and Lin (2019), for example, framed their meta-analysis of vocabulary-focused MALL (mobile application language learning; in this case mobile phone applications focusing on L2 IVA) experimental reports with a systematic review. This effort was undertaken to uncover moderator variables, such as the length of treatment and whether the target word forms were productively or receptively learned. Methodological design features reflecting study quality were omitted from the analysis. Cisco and Padrón (2012) likewise catalogued how vocabulary research reports presented strategy use instruction. Elgort (2018), in a systematic review of technology-mediated L2 vocabulary research, catalogued the average sample size of the papers in the report pool and their testing designs, but the only aspect of the report directly connected to study quality was related to enhancing vocabulary measurements' overall validity. Elgort's work in this regard was exemplary and was indeed one of the rare L2 IVA reviews investigating study quality. However, Elgort's focus was not on sample design, which is the subject of the present review. The apparent lack of focus on study quality in general and sample design study quality in particular in these reports should not be considered a flaw because such a focus is in the domain of systematic reviews and especially methodological syntheses (Plonsky & Gonulal, 2015; Siddaway et al., 2019). This lack of focus did however motivate the present study, along with the previously mentioned informal observations regarding three aspects of experimental design: sample-size planning, randomization, and multisite sampling.

https://doi.org/10.1017/S0272263121000541 Published online by Cambridge University Press

THREE RELATED ASPECTS OF EXPERIMENTAL SAMPLE DESIGN IMPACTING RIGOR AND GENERALIZABILITY

EFFECT SIZE–DRIVEN SAMPLE-SIZE PLANNING

The sample-size planning process begins with effect sizes, whether it is classical power, driven by null hypothesis significance testing (NHST), or another process such as precision (Cumming, 2012; Norouzian, 2020). In both of these approaches, researchers reference effect sizes from the relevant domain to plan the number of participants they require. Effect size metrics, as highlighted in the preceding text, quantify an observed sample's experimental and correlational effects to enable generalizations to the population. Such metrics are numerous but tend to be organized under mean difference (*d* family), strength of association (*r* family), and odds ratio categories (Plonsky et al., 2021). L2 experimental meta-analyses have usually expressed experimental effect sizes in standardized mean difference metrics, the *d* family, despite the catalogued reports sometimes utilizing different metrics (e.g., Bryfonski & McKay, 2019; Lin & Lin, 2019; Vitta & Al-Hoorie, 2020). As highlighted by Brysbaert (2019), effect size metrics convert from one to another quite easily while expressing the same underlying effect. When $d = 1.00$, a group's performance averaged one pooled standard deviation higher than another; when $d = .50$, the group was 50% of a standard deviation better.

Classical *a priori* NHST power analysis is perhaps the best-known method to plan sample sizes using effect sizes. Recent L2 methodological syntheses (e.g., Farsani & Babaii, 2020; Lindstromberg, 2016; Vitta & Al-Hoorie, 2021) have exclusively coded for classical power given NHST's prevalence among SLA quantitative research, despite arguments to move away from it (e.g., Cumming, 2012; Plonsky, 2015). As highlighted by Nicklin and Vitta (2021) and Brysbaert (2019), such power analyses should reference relevant standardized effect sizes.[2] The process entails the calculation of the sample size required to detect a certain effect size assuming predetermined alpha (α) and beta (β) thresholds. The α thresholds relates to Type I error (i.e., false positive probability) and is conventionally set at 5%, while β thresholds relate to Type II error (i.e., false-negative thresholds) and are typically set at 20% (Lakens et al., 2018). For fixed-effects testing (e.g., ANOVA and multiple regression), calculators such as *G*Power* (Faul et al. 2007) are useful for conducting *a priori* power analyses but they can be limited in their ability to model interactions. However, this issue can be addressed using simulations (see Lakens & Caldwell [2019] for details). Mixed-effects models where random effects (i.e., not a part of the experimental design) at the item and person levels likewise require *a priori* power to be addressed using simulations analyses (e.g., Brysbaert & Stevens, 2018; Green & MacLeod, 2016).

Precision has recently been introduced within the SLA arena and acts as a useful exemplar of a non-NHST *a priori*, effect size–driven sample-size planning procedure (Norouzian, 2020). This approach determines the sample size required to narrow the confidence interval (CI) for a certain effect size. Much like G*Power, Norouzian's precision calculator is turnkey whereby the researcher selects the effect size, the parametric test, and the degree to which the CI is narrowed. According to Norouzian's SLA-specific precision calculations, the trend of larger samples for small effects is present in a general sense. For example, Norouzian demonstrated that precision in regression modeling increased with sample size when effect size was held constant. Certain conditions,

https://doi.org/10.1017/S0272263121000541 Published online by Cambridge University Press

however, saw almost identical sample sizes providing adequate precision for different effect sizes (see Figure 3 in Norouzian, 2020, p. 854).

Small samples that have not been planned with relevant effect sizes are problematic in relation to both small and large effects (Brysbaert, 2019). A small sample might lead toward a false negative discovery, or Type II Error. In this situation, a researcher determines that a small effect in the sample does not exist in the population when in reality it does, but the sample was too small to significantly detect it (as argued by Brysbaert within an NHST paradigm). Fraley and Vazire (2014) demonstrated that when samples are underpowered the distribution of observed sample effects is wider than the distribution in higher-powered samples. In other words, there is a higher chance of studies observing larger and smaller effects in relation to the true population effect. This can lead to the publication of large effect "flukes," which occur when a large significant effect in a small sample is generalized to the population by the researcher when in reality it does not exist. For instance, Yang et al. (2017) reported a very large effect size ($d_s = 1.75$)[3] in favor of a sentence writing L2 IVA condition group over an essay writing group while the median between-subject *d* in SLA was estimated as 0.70 (Plonsky & Oswald, 2014). Because Yang et al.'s sample comprised only 19 and 18 students in the respective sentence and essay writing subgroups from a single Chinese university, the concern of this huge observed sample effect being a fluke relative to the population emerges. When one couples this with the observed publication bias toward large and significant effects (Fanelli, 2010), the need for L2 IVA and L2 experimental designs to plan their sample sizes with reference to relevant effect sizes before data collection begins is clear.

Given the problems that samples constructed without *a priori* size planning present, it is concerning that the process has been almost never observed when L2 methodological syntheses sought to investigate it. When reviewing L2 interaction inquiries, Plonsky and Gass (2011) observed that only three (1.72%) of 174 studies reported power in any fashion, with it being unclear if any of them employed *a priori* or post hoc procedures. Farsani and Babaii (2020) likewise observed 6 (2.10%) of 285 Iranian graduate theses reporting power without considering if any were *a priori.* In a review of quantitative reports published in *Language Teaching Research* from 1997 to 2015, Lindstromberg (2016) overtly mentioned that no study engaged in an *a priori* sample planning process. His null finding corresponded to Vitta and Al-Hoorie's (2021) observation that none of the 56 experimental reports investigating L2 flipped learning, in which new content is presented to students before class as homework (see Mehring, 2018 for more on the approach), featured *a priori* power assessments while two (3.57%) considered post hoc power considerations.

This observed, albeit limited application of post hoc (NHST) power highlights another issue with the field's failure to conduct *a priori* power analyses. Post hoc power analysis, which is classical NHST power analysis conducted after data has been collected, is routinely criticized as illogical because the power serves the purpose of determining the sample size an inquiry requires (e.g., Cumming, 2012; Field, 2018). Proponents of post hoc power emphasize its utility in assessing the power of a finding assuming that it truly exists in the population. This utility however has been noted as "pointless and potentially misleading" (Perugini et al., 2018, p. 3), with Cohen (1988) noting that sample-size planning should occur before the study is undertaken. Another option available to researchers is sensitivity analysis, which determines the minimal effect size

https://doi.org/10.1017/S0272263121000541 Published online by Cambridge University Press

that a sample can significantly detect given its size. This effect size is then the baseline that the study can accept in relation to detecting effects in the sample that generalize to the population. Some researchers have presented sensitivity analysis as a compromise between the ideal *a priori* and problematic post hoc sample-size determination analyses (Perugini et al., 2018). A sensitivity analysis probably does avoid the "illogical" flaws of its post hoc counterpart. Nevertheless, only *a priori* effect size–driven planning with relevant effect sizes located in prior research allows researchers to situate their proposed sample size in the existing body of inquiries (Brysbaert, 2019; Nicklin & Vitta, 2021).

RANDOMIZATION: SAMPLING AND ASSIGNMENT

When an experimental design features a randomized sample with random assignment of participants, it has stronger representativeness in relation to the intended population (Fisher, 1935). "Random" and "randomized" imply that chance drives either the selection or assignment process. This removes the bias that a researcher-controlled procedure might introduce, which in turn degrades the generalizability of the sample vis-à-vis the intended population.

Randomized sampling is the process by which participants are randomly selected from the intended population. According to most frameworks (e.g., Harter, 2008), randomized sampling is the superordinate under which one can subsume probability sampling procedures such as stratification and clustering. Stratification, as defined by Harter, is the identification of subgroups, or strata, within the population that should be represented in the sample, with participants randomly selected from each strata. Clustering is the random selection of participants from randomly selected clusters, which are intact groups existing in the population (Trochim et al., 2016). Among L2 researchers, such sampling is rarely implemented, given the practical constraints that the field faces. In a recent review of L2 research, only 2 (8.4%) of the 282 reviewed samples employed randomized sampling, subsuming fully randomized, clustered, and/or stratified procedures (Farsani & Babaii, 2020). In a similar vein, Vitta and Al-Hoorie (2021) informally observed that none of the 56 L2 flipped learning experimental reports that were coded involved a randomized sampling process and thus the procedure went unanalyzed in the synthesis. Earlier reviews of L2 samples omitted the issue entirely with only randomized assignment being analyzed (e.g., Plonsky, 2013).

The potential of randomized sampling within L2 research is noteworthy, however. Hiver and Al-Hoorie (2020) could not replicate the conclusions and findings of You et al.'s (2016) L2 learning motivation study. One of the differences between the studies was that the replication study employed a randomized stratified sample, while You et al. employed a nonprobability snowball sample. Snowball, or chain sampling, involves participants enlisted for earlier iterations of the research assisting in participant recruitment for later iterations (see methods in You et al.). According to Trochim et al. (2016), this procedure is subject to low external validity, entailing biased results and a lack of generalizability. While Hiver and Al-Hoorie's failure to replicate cannot be fully ascribed to the sampling difference, this finding invites randomized sampling replications of prior findings from early studies that employed convenience samples.

Unlike with randomized sampling, L2 methodological syntheses have reported that L2 researchers are more prone to employ randomly assigned experimental conditions.

https://doi.org/10.1017/S0272263121000541 Published online by Cambridge University Press

Random assignment is the process of assigning participants to experimental conditions (e.g., treatment and comparison) by chance. Farsani and Babaii (2020) observed that 100 (58%) of 170 classroom experiments featured randomized assignment at either the class or participant level. This was higher than the overall 47% observed by Plonsky (2013) seven years' earlier in 606 quantitative reports published in popular L2 SSCI (Web of Science's Social Science Citation Index) journals, and also higher than the 32% observed by Lindstromberg (2016). The relatively high observed frequency of random assignment in comparison with random sampling and sample-size determination procedures allows for assessment of the association between time and random assignment. However, divergent findings have been reported. Lindstromberg observed a small correlation ($\rho = .17$) between year of publication and the use of random assignment in 76 between-subject comparison experimental designs. Plonsky (2014), however, observed that the use of random assignment in SSCI journals' reports decreased when comparing the 1990s (40%) to the 2000s (36%).

While random assignment at the participant level is the gold standard (see Rogers & Révész, 2020) and corresponds to Fisher's (1935) guidance, L2-centric reviews have made the distinction between random assignment at either the class/group or participant level. In classroom-centric SLA subfields such as L2 IVA, assessing random assignment at the class/group level is sensible as research is often conducted on intact classes. For instance, Peters (2019) randomly assigned instructed vocabulary conditions to classes and would have been coded as employing random assignment at the class level in reviews such as Plonksy and Gass's (2011). Additionally, there have been divergent findings when it comes to the balance between random assignment at the class and participant levels. Plonsky (2013) observed that in both classroom and lab conditions, random assignment was more prevalent at the participant level. Farsani and Babaii (2020), however, observed that classroom experiments presented in Iranian theses had more group-level random assignments than person level with the opposite trend observed in lab conditions. Lindstromberg's (2016) assessment of *Language Teaching Research* found that random assignment was overwhelmingly at the class level. Taken together, these syntheses paint a mixed picture with regard to random assignment at the participant and the class levels.

Despite the historical (Fisher, 1935) and current (Rogers & & Révész, 2020) case for participant-level random assignment as the gold standard, some have argued that nonrandom procedures are also suitable. Farsani and Babaii (2020), for instance, coded studies that undertook purposeful assignment. When employing this strategy, as highlighted by Farsani and Babaii, L2 researchers consider tangential aspects such as L2 proficiency during the group construction process (e.g., ensuring groups are not significantly different in IELTS scores). Such procedures, however, violate the decades-old expectation that researchers do not introduce bias into the sample with their choices (Fisher, 1935). Furthermore, purposeful assignment becomes unnecessary when linear mixed-effects models (LMMs) and multivariate models with covariates are utilized. These analyses partial out constructs (i.e., covariates and random effects) that are tangential with the experimental condition(s) in predicting the outcome L2 IVA variable (Brysbaert & Stevens, 2018).

https://doi.org/10.1017/S0272263121000541 Published online by Cambridge University Press

MULTISITE SAMPLING

Recruiting participants from the intended population across two or more institutions or locations constitutes a multisite sample. The process of recruiting from multiple locations relates to the need for randomized sampling in quantitative research (Vitta & Al-Hoorie, 2021). Moranski and Ziegler (2021) argued that single-site samples present an external validity limitation, which reduces the ability to generalize findings beyond the context in which the data was collected. The authors also connected multisite samples with power, whereby the use of the former improves the latter. Such connections help reinforce the focus of this current study. Randomized samples from meaningful populations will additionally be multisite and thus the two constructs are related. For example, if a researcher intended to construct a sample that was representative of Saudi university EFL students and randomly selected 100 of such students from a national database, the probability of all participants' being from the same university approaches zero.

A further benefit of multisite sampling is the facilitation of a "random(effect)-by site" analysis (Morgan-Short et al., 2018, p. 408). Under ideal conditions, the effect of the location should approach zero with experimental grouping variables and theory-driven covariates accounting for the dependent variable's variance. Checking for this effect of location corresponds to the call for L2 researchers to consider clustering effects within their samples and data, whereby the different locations account for influence over the model estimates and parameters (Al-Hoorie & Vitta, 2019). Although a multisite sample in L2 research would most likely be a nonprobability convenience sample conceptually, a random-by-site cluster check moves samples away from the biases that single sites introduce. As highlighted by Vitta and Al-Hoorie (2021), multisite sampling represents a compromise in relation to respecting the historical need for randomized sampling (Fisher, 1935) and recent L2 studies investing in the process (e.g., Hiver & Al-Hoorie, 2020).

Convenience samples from one location, regardless of the observed sample size, therefore have an inherent representativeness limitation in relation to generalizations beyond the setting from which they came (Moranski & Ziegler, 2021). While it is unreasonable to expect an L2 researcher to have access to national databases, multiple sites that are chosen with some randomness are perhaps within the realm of possibility. In the seminal reviews that started the L2 methodological reform (e.g., Plonsky, 2013; Plonsky & Gass, 2011), the need to consider power and thus construct robust samples was presented, but overt cataloguing of single- versus multisite samples appears to have been omitted until recently. Vitta and Al-Hoorie (2021) observed that 10 (18%) of 56 flipped learning experimental designs employed multi-site samples. Furthermore, in a review of L2 research practices, Farsani and Babaii (2020) noted that most of the observed samples were convenience samples, but the single- versus multisite issue was not assessed. From such limited data points, one cannot generalize more broadly about the state of multisite sample research in SLA. This view is somewhat reinforced by the observation that Moranski and Ziegler (2021) and Morgan-Short et al. (2018) presented position pieces with examples on the issue as opposed to dedicated methodological syntheses, a gap in the literature that the present study intends to fill by analyzing single- and multisite sampling in L2 IVA research.

https://doi.org/10.1017/S0272263121000541 Published online by Cambridge University Press

RESEARCH QUESTIONS

Situated in the literature and motivated by the apparent lack of an L2 IVA-centric review of experimental sample designs, this focused methodological synthesis addressed four general research questions (RQ1–RQ4). RQ1, RQ2, and RQ3 were grounded in the foci of related L2 sample design methodology syntheses (e.g., Farsani & Babaii, 2020; Plonsky, 2013; Vitta & Al-Hoorie, 2021). RQ4 was motivated by past assessments of how synthesized methodological features changed over time (e.g., Lindstromberg, 2016; Plonsky, 2014):

RQ1. To what extent do L2 IVA researchers engage in effect size–driven sample-size planning procedures with regard to:
  a. *A priori* procedures?
  b. Other procedures?

RQ2. To what extent do L2 IVA researchers engage in randomization with regard to:
  a. Randomized sampling, subsuming probability sampling?
  b. Random assignment, at the participant level?
  c. Random assignment, at the class/group level?

RQ3. To what extent do L2 IVA researchers:
  a. Employ multisite samples?
  b. Consider the statistical clustering effect when multisite samples are employed?

RQ4. What is the association between time and:
  a. Effect size–driven sample-size planning procedures?
  b. Randomization?
  c. Multisite use?

METHODOLOGY

REPORT POOL CONSTRUCTION

The report pool consisted of 110 studies retrieved from six renowned journals that were selected using a multistage systematic process. The first stage involved identifying candidates from a list of 35 field specific journals (Egbert, 2007) that were (a) SSCI indexed since 2010, (b) broad in focus (e.g., not *ReCALL*), and (c) targeted (in part) toward an audience of practitioners or agents of language teaching. SSCI indexed journals were considered important because Al-Hoorie and Vitta (2019) demonstrated that the quality of the statistical analyses is generally higher in such journals. This stage resulted in six candidate journals. The second stage involved a comparison of the six journals with two lists of SSCI journals from quantitative L2 research (Al-Hoorie &Vitta, 2019; Plonsky & Gass, 2011). Following this stage, one journal, *ELT Journal*, was removed for not being on both lists. In the final stage, three external experts confirmed that the five remaining journals could be labeled as "SSCI-indexed journals that are well-known, somewhat trusted in relation to their quantitative papers, and intended for practice-level research." The five journals were *Language Learning*, *Language Teaching Research*, *The Modern Language Journal*, *System*, and *TESOL Quarterly* and they were employed to construct the report pool for Nicklin and Vitta (2021). Subsequent expert review (two independent expert judgments) suggested that *Studies in Second Language Acquisition* should be included as its scope and prestige in the field mirrored that of *Language*

https://doi.org/10.1017/S0272263121000541 Published online by Cambridge University Press

Learning, which was included in the original report pool. To validate the inclusion, this rationale and the process of selecting the original five journals was presented to two other experts (the final external expert review) who concurred with adding *Studies in Second Language Acquisition* while also agreeing with the process and rationale for the selection of the five original journals.

From the six nominated journals, 1,053 reports, published between January 1, 2000 and December 31, 2020 were initially collected. The reports were retrieved from the EBSCO databases using the following search terms: *vocab**, *lexi**, *idiom**, *collocat**, *phrasal*, *multiword*, and *formula*, where * represented a "wild card" search term for which any character was valid. The EBSCO platform was employed for the search because individual journal platforms have inconsistent search options. EBSCO helped to systematize this process of the study (for a similar rationale for using Scopus to search journals; see Vitta & Al-Hoorie, 2020). Because the search was conducted with EBSCO host, advance online publications were omitted from the report pool. For incorporation in the final pool, the reports were required to include (a) dependent variables measuring student acquisition of a set of target vocabulary items, (b) measurements enabling the dependent variables associated with L2 IVA (e.g., a vocabulary knowledge test), and (c) vocabulary treatment and comparison conditions that could be expected to occur in the classroom. When applying criteria (b) and (c), psycholinguistic measurements such as reaction times and very short treatments (e.g., 3,000 ms) were excluded as they were outside the intended L2 IVA focus. Because the report pool was originally constructed for another study, the reports were also required to contain sufficient information for effect size calculation. However, this final criterion had no bearing on the current study's research questions. Analysis of the 1,053 reports, by the first author, resulted in a final pool of 110 that satisfied the three criteria.

The final report pool consisted of 110 reports from six journals: *Language Learning* ($k = 11$), *Language Teaching Research* ($k = 39$), *The Modern Language Journal* ($k = 9$), *System* ($k = 18$), *Studies in Second Language Learning* ($k = 19$), and *TESOL Quarterly* ($k = 14$). Supplementary Materials A contains a list of the studies comprising the report pool. A randomly selected group of 148 reports (from the 1,053) were independently coded by a second researcher to validate their inclusion in or exclusion from the final sample and confirm the criteria judgments. This process resulted in 89.80% agreement ($\kappa = .80$), which became 100% following deliberation.[4]

INSTRUMENTATION AND CODING

The present study's first three research questions were operationalized as a series of dichotomous judgments regarding each report (see Table 1). These judgments comprised the study's instrumentation. RQ1 involved judgments about the use of effect-size driven *a priori* sample-size planning procedures and procedures besides *a priori* ones. RQ2 necessitated judgments about (a) random assignment at either the participant or group level, and (b) randomized sampling. The judgment concerning the use of multisite samples in the operationalization of RQ3 required nuance. Some reports presented subsamples or sequenced experiments. In these cases, it was possible to conceptualize each report as a singular design. Accordingly, they were coded as employing multisite samples for respecting the bias that single sites introduce (Moranski & Ziegler, 2021). For

https://doi.org/10.1017/S0272263121000541 Published online by Cambridge University Press

TABLE 1. Coding Questions, Interrater Agreement, *S* index, and Notes

RQ	Description	Interrater agreement	*S* index	Coding notes
1a	Did the report feature an *a priori* effect size-driven sample-size planning process?	100%	NA	To be coded "1" (i.e., *yes*), a reference to past L2 IVA effects was required given the observed heterogeneity among L2 effect sizes (see Plonsky & Oswald, 2014).
1b	Did the report feature an effect size-driven sample-size planning process besides an *a priori* procedure?	100%	1.00	N/A
2a	Was randomized sampling, vis-à-vis the intended population, undertaken?	100%	NA	N/A
2b	Was random assignment undertaken at the participant level?	86.36%	.73	To be coded "1" (i.e., *yes*) for an overt mention of "random" or another term relating to chance.
2c	Was random assignment undertaken at the class/group level?	95.54%	.91	*See preceding note.*
3a	Was the sample constructed from the intended population across multiple educational locations (i.e., a multisite sample)?	90.90%	.82	Explicit mentioning of multiple sites was required. Successive "trials" from different locations addressing the same research question was coded as "1" (i.e., *yes*; $k = 2$).
3b	When the sample was constructed from multiple sites, were the clustering effects of the sites checked using inferential testing, by either *a priori* checking or multivariate modeling?	100%	NA	*A priori* checks or multivariate modeling coded as "1" (i.e., *yes*).

Note. RQ = Research Question

example, Hulstijn and Laufer (2001) investigated reading, reading + fill-in, and writing interventions with Dutch and Hebrew L1 learners of English. This study was coded as multisite despite the inferential tests being confined to single settings. RQ3 also involved a dichotomous, complementary judgment regarding the consideration of cluster effects between multiple sites.

Using this framework, the primary researcher analyzed each of the 110 reports from the pool and dichotomously coded them as "1" for *yes* or "0" for *no* in relation to the questions listed in Table 1. Twenty-two (20%) randomly selected reports were independently coded by two researchers (the primary and a secondary researcher). The overwhelming frequency of *no* judgments prohibited inferential assessments using Cohen's κ. As highlighted by McHugh (2012), κ statistics are driven by the chi-squared test and an abundance of *no* judgments unbalance the contingency table rendering the test inappropriate. In response to this limitation of κ, the *S Index* was employed to assess observed interrater reliability. Norouzian (2021) presented the SLA research community with a turnkey program to calculate the *S* Index metric for interrater reliability (Falotico & Quatto, 2010, 2015, as cited in Norouzian, 2021), which like κ partials out change agreement but is not sensitive to the imbalance that renders κ unusable in cases found in this current study (see Figure 3, Norouzian, 2021, p. 6). As highlighted in Table 1, all

https://doi.org/10.1017/S0272263121000541 Published online by Cambridge University Press

judgments had high (.80 to .89) to very strong (.90 to 1.00; 'NA' denotes perfect agreement with only one or zero 'yes' judgments) observed *S* Index values according to Norouzian's stated magnitude thresholds, except for random assignment at the participant level (RQ 2b; *S* Index = .73; "acceptable"). The relatively low value for the random assignment at the participant level judgment could be attributed to the ability to engage in the design process without having to overtly label it as "random assignment." Some report authors engaged in the design feature without overtly employing the "random" label (e.g., Traxler & Nakatsukasa, 2020) while others noted random assignment in past studies but omitted detailing the feature in their methodology (e.g., Choi, 2017). Power considerations, conversely, do not facilitate such variation in wording when presented in the methods.

In line with past L2 meta-analytic reports (e.g., Al-Hoorie & Vitta, 2019; Nicklin & Vitta, 2021; Vitta & Al-Hoorie 2020, 2021), the primary and secondary researchers who coded the 22 reports for the reliability check discussed disagreements to identify possible systemic issues in the coding process and to identify points of clarification. Such discussion resulted in agreement with the primary researcher's judgments and rationale behind them being accepted by the second coder while highlighting points of clarification such as double site studies where inferential tests were constrained to single sites being coded as multisite as highlighted in the preceding text.

ANALYSIS

In addressing the first three research questions, each of the 110 reports from the pool was coded according to the seven questions in Table 1, and the tallied responses (judgments) are presented visually. To address RQ4, which related to changes in sample-size planning, randomization, and multisite sampling over a 20-year period, the results for each of the dichotomous judgments (or aggregates of them) were correlated with the year of publication. To frame the correlation values further, an analysis of report publication by year was also conducted. Simple regressions with year as the dependent variable and a series of dummy variables representing each coded binary judgment as the predictors were conducted. Visual inspection of scatterplots, with standardized residuals on the y-axis and standardized predicted values on the x-axis, suggested that the homoscedasticity assumption might not have been met because cases unevenly straddled the x-axis (see Osborne & Waters [2002] for using residual scatterplot to check this assumption). This was unsurprising given the extreme frequency of "no" judgments, thus Spearman's *rho* (ρ) tests were conducted. Dellinger (2017) summarized the assumptions of Spearman's ρ as (a) paired observations, (b) ratio, interval and/or ordinal variables, and (c) a monotonic relationship (i.e., no change in direction), which are less stringent assumptions than those associated with parametric correlation testing (see Field, 2018). In presenting the case to use ρ with binary (1/0) and ranked variables, Glass (1965) highlighted the procedure's "weak" assumptions in comparison to parametric assumptions as an advantage for extreme distributions of variables (e.g., the overwhelming frequency of "no" judgments across reports in this current study). Spearman's ρ and accompanying 95% confidence intervals (CIs) were calculated using the *spearmanCI* package (de Carvalho, 2018) in *R* (R Core Team, 2020).

https://doi.org/10.1017/S0272263121000541 Published online by Cambridge University Press

Finally, the possibility of a by-journal clustering effect was assessed with Fisher's (for 2×2 contingency tables) and Fisher-Freeman-Halton (for contingency tables exceeding 2×2) Exact testing, despite the conceptual justification for viewing them as being a homogenous group (see preceding text). A cluster effect within the journals occurs if reports from one or more of the six journals comprises a subset that is more likely than the rest of the journals to include one of the three aspects of research design under investigation. Because the expected cell count was less than five in more than 20% of the cells, X^2 testing was not employed, as recommended by Field (2018). Nonsignificant associations between the journals and the dichotomous judgments operationalizing the first three research questions suggested that clustering concerns were mitigated.

RESULTS

In the *a priori* clustering check, nonsignificant results for Fisher-Freeman-Halton Exact tests were observed in six of the seven judgments, $.258 \leq p_s \leq 1$. A significant effect, $p < .001$, was observed between journal and multisite usage, where two journals, *Modern Language Journal* (MLJ) and *Language Learning* (LL), employed more multisite samples (9 [45.00%] of 20 total reports) than the other four journals (7 [7.78%] of 90 reports). Further Exact tests revealed nonsignificant differences within the two sets of journals: Set 1 (MLJ & LL), $p = .406$ (Fisher's Exact test); Set 2 (other four journals), $p = .360$ (Fisher-Freeman-Halton Exact test). Because of the externally confirmed label capturing all six journals and the belief that there was no reason to assume any journal would favor multisite samples more or less than the others, we acknowledge the finding but still treat the six journals as a homogenous group.

Figure 1 displays the results of the dichotomous coding questions that were detailed in Table 1. Regarding the first research question, merely nine (8.18%)[5] reports considered power in any fashion, with only one (0.91%) report engaging in *a priori* sample-size determination procedures referencing past L2 IVA experimental effects (RQ1a), and eight

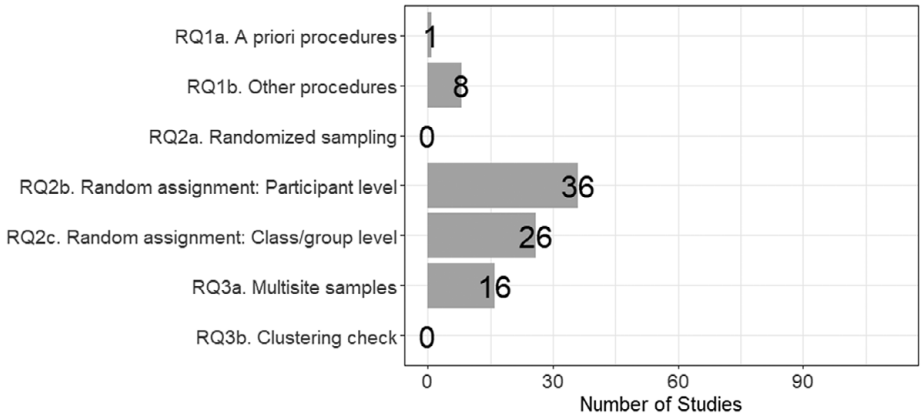


FIGURE 1. Sample-size determination, randomization, and multisite usage in 110 L2 IVA research reports. *Note*: Although the maximum count is 36, the bar chart displays up to 110 to emphasize how infrequently the procedures were utilized.

https://doi.org/10.1017/S0272263121000541 Published online by Cambridge University Press

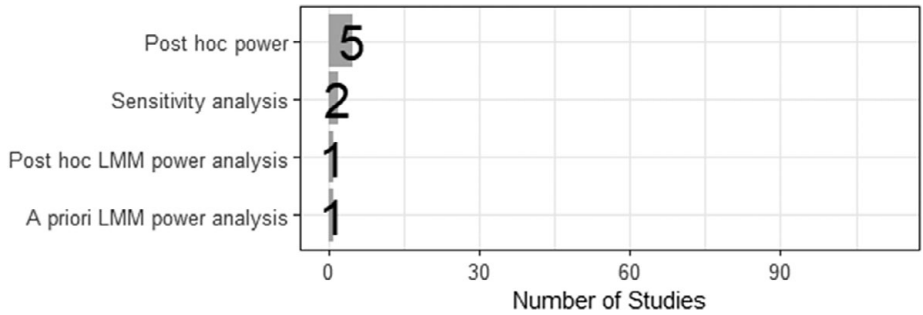


FIGURE 2. Sample-size determination procedures utilized by L2 IVA researchers across 110 reports. *Note*: Although the maximum count is five, the bar chart displays up to 110 to emphasize how infrequently the procedures were utilized.

(7.27%) performing other procedures. A follow-up analysis was conducted to investigate which procedures L2 IVA researchers utilized to determine sample sizes. The nine reports in which sample-size planning was utilized were reviewed and coded according to the following coding scheme, which was developed as the process progressed: 1 = *Post hoc power*; 2 = *Sensitivity analysis*; 3 = *Post hoc LMM power simulation*; and 4 = *A priori LMM power simulation.* The results in Figure 2 show that post hoc power was the most common procedure, accounting for five (4.55%) reports, while two (1.82%) reports involved sensitivity analysis, one (0.91%) included a post hoc LMM power analysis, and one (0.91%) contained an *a priori* LMM power analysis. With respect to RQ2, no report engaged in randomized sampling procedures (RQ2a). However, 62 (56.36%) reports featured random assignment, with participant-level assignment being more frequent (RQ2b; $k = 36$ [32.72%]) than class/group level (RQ2c; $k = 26$ [23.63%]). Finally, the results for RQ3 show that while 16 (14.55%) of the reports included samples constructed from multiple sites (RQ3a), none presented an empirical clustering check (RQ3b).

Spearman's ρ was utilized to answer RQ4 regarding the associations between time and (a) effect size–driven sample-size planning procedures, (b) random assignment, and (c) multisite sample use. Because there was only one report coded "1" (i.e., *yes*) for *a priori* planning (RQ1a), it was aggregated with the eight reports featuring other procedures (RQ1b) to represent any sample-size planning procedure. Random assignment at the group and participant levels was conflated given the observation that they are both suitable for L2 experimental designs (Rogers & Révész, 2020; Vitta & Al-Hoorie, 2021) and the desire to present the results as parsimoniously as possible. The assumptions of ρ were met given the ordinal and ranked nature of the variables and a visual inspection of scatterplots revealed no monotonicity issues (see Supplementary Materials B). Figure 3 shows that the number of L2 IVA experimental reports published in the six journals increased with time. However, there were no practical associations between time and effect size–driven sample-size planning $\rho = .18$ [–.03; .38] and random assignment, $\rho = .06$ [–.13; .25], $n = 110$, with CIs inclusion of zero implying nonsignificant results (Cumming, 2012). A small[6] negative relationship was observed between time and multisite sample usage, $\rho = –.30$ [–.48, –.12], $n = 110$, with zero outside the CI rendering the relationship significant. In other words, L2 IVA experimental reports employed slightly less multisite samples as time went on. This negative relationship should be read with

https://doi.org/10.1017/S0272263121000541 Published online by Cambridge University Press

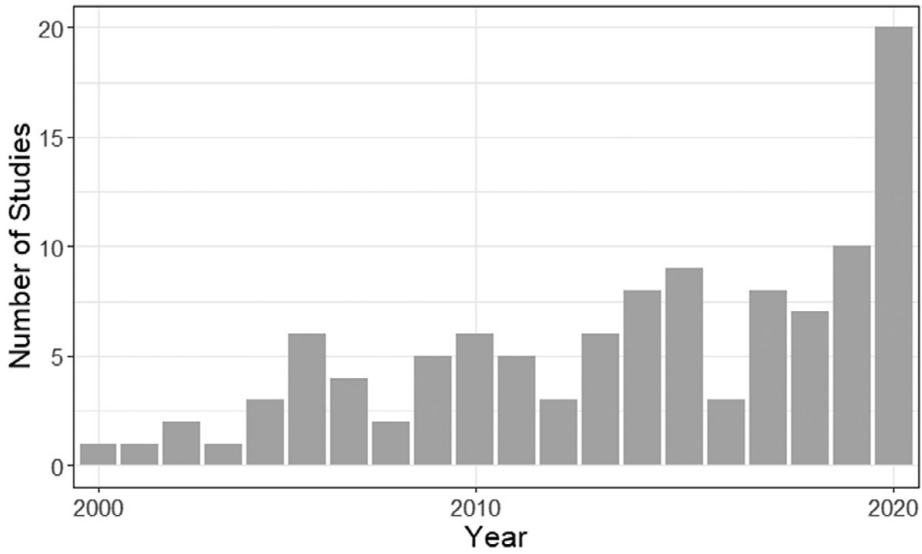


FIGURE 3. Number of L2 IVA reports published by year.

caution, however, given that only 16 reports were coded as "1" (i.e., *yes*) and that the upper boundary value ($\rho = -.12$) of the CI interval implied no practical association.

DISCUSSION

The results of the current study can be summarized thus to address the four research questions:

(a) Few reports featured samples that were planned with effect size–driven sample-size planning procedures (RQ1).
(b) Random sampling was entirely omitted while random assignment was inconsistently employed (RQ2).
(c) A small minority of the reports featured multisite samples (RQ3).
(d) Only multisite sample use displayed a practical and significant association with time where the direction was negative (RQ4).

For each of the three sample design areas investigated, the findings are further analyzed by situating them among relevant observations of past L2 syntheses and then guidance is given in the area using positive examples from the report pool. Finally, the current study's limitations are presented to constrain the analysis of findings.

EFFECT SIZE–DRIVEN SAMPLE-SIZE PLANNING IN L2 IVA EXPERIMENTS

Only one (0.09%) of 110 reports presented an *a priori* sample-size planning procedure, which corresponded to the null observations of Vitta and Al-Hoorie (2021), Lindstromberg (2016), and the observation of Plonsky (2013) where only 4 of 606 reports might have engaged in a power procedure that was not a post hoc process. The observation that

https://doi.org/10.1017/S0272263121000541 Published online by Cambridge University Press

nine (8.18%) of 110 reports (RQ1a + RQ1b) employed any effect size–driven procedure was substantially higher than what has recently been presented to the field (e.g., 2.10% [Farsani & Babaii, 2020]; 3.57% [Vitta & Al-Hoorie, 2021]) and seminal syntheses such as Plonsky and Gass (2011; 1.72%). The virtually null association ($\rho = .18$) between effect size–driven sample-size planning and time reported in this current study can probably be ascribed to such low frequency of *yes* observations. As a final note, no report analyzed here presented a planning procedure outside of NHST-driven power analyses. This corresponds to past observations (e.g., Nicklin & Vitta, 2021; Vitta & Al-Hoorie, 2021) while highlighting the opportunity to engage with other procedures such as precision (Norouzian, 2020).

Only one report in the pool involved *a priori* sample-size planning. Elgort et al. (2020) investigated their participants' declarative and nondeclarative knowledge of 90 target vocabulary items using meaning generation and self-paced reading tasks, respectively. Because the authors' data analysis was conducted with LMMs, they conducted a power simulation using *R* along with the *simr* package (Green & MacLeod, 2016). Simr enables simulated data to be constructed based upon parameters from previous research, such as random effect variance and variance-covariance matrices. Researchers select the number of items and participants to be modeled, which can cover a specified range, and the number of simulations to be performed, which can number hundreds or even thousands. The package reports power based upon the number of simulations that resulted in a significant result, along with 95% CI. If, for instance, the package reports that 89% [81, 95] of simulations with 45 participants were significant, but only 82% [74, 87] were significant with 40 participants, then a researcher is justified to conduct their experiment with 45 participants. Elgort et al. conducted a simr power analysis with the results from a previous meaning generation task (Elgort, 2017), which were collected from an L2 sample with comparable proficiency and backgrounds to Elgort et al.'s (2020) participants. Future L2 IVA researchers are advised to adhere to this procedure when planning sample sizes for experiments involving LMM analyses. Furthermore, researchers should report the results of LMMs in full, including coefficients for random effects and correlations between random intercepts and slopes, which are all essential for future power simulations based upon those models.

When researchers use simpler fixed-effect testing designs, they can employ tools such as G*Power or Norouzian's (2020) R functions for calculating precision to plan the number of participants required beforehand. Unlike simulations, these calculations are somewhat straightforward. Because no reports in this current study reported such procedures, Nicklin and Vitta (2021) is employed here to model the process with fixed-effect designs. This prior study, using G*Power, demonstrated that between-subject experimental designs including treatment, comparison, and true control groups required 492 or 144 participants (split equally into three groups) to meet the respective thresholds. These figures assume the observed small ($g_s = .33$) or medium ($g_s = .62$) L2 IVA effect sizes, where the omnibus one-way ANOVA is followed by post hoc one-way *t*-tests, a testing procedure also modeled by Brysbaert (2019). Smaller samples could meet power threshold holds using post hoc comparison tests such as *Tukey*. However, these tests are somewhat problematic because they increase the chance of Type II error (Gravetter & Wallnau, 2015). Because the reports catalogued by Nicklin and Vitta mostly omitted covariate controls, the researchers suggested powering between-subject designs to the small effect size. In contrast, within-subject counterbalanced designs offered a power

https://doi.org/10.1017/S0272263121000541 Published online by Cambridge University Press

advantage whereby 95 to 203 participants are required depending on the correlation between the repeated measures. Higher and lower correlation coefficients respectively inflate or deflate the observed meta-analytic effect size suggested in the study ($g_{av} = .25$). Norouzian's precision tool operates in a similar manner, but its use is still emerging in SLA and none of the 110 reports synthesized here utilized the procedure. As a final note, the eight reports catalogued in the current study that involved procedures other than *a priori* ones are by no means flawed but we have chosen not to review them given the recent emphasis on planning sample sizes *before* the study is undertaken (Brysbaert; 2019; Brysbaert & Stevens, 2018; Nicklin & Vitta, 2021; Vitta & Al-Hoorie, 2021).

RANDOMIZATION

No catalogued L2 IVA report involved randomized sampling. This corresponds to past L2 syntheses such as Plonsky (2013, 2014) and Vitta and Al-Hoorie (2021) that have omitted the area altogether from their analyses. It is also noteworthy that recent guidance for L2 experimental designs have emphasized random assignment in lieu of randomized sampling (e.g., Rogers & Révész, 2020). Regarding the other plane of randomization, the observed use of random assignment (56.36%) was approximately the same as the recently observed 58% by Farsani and Babaii (2020), but higher than what most L2 methodological syntheses have observed (e.g., 47% [Plonsky, 2013]; 36% [Vitta & Al-Hoorie, 2021]; 32% [Lindstromberg, 2016]). The near zero association between its use in the report pool and time ($\rho = .06$) points to L2 IVA researchers being mindful of the practice for some time. The preference for random assignment at the participant level corresponded to Plonsky (2013) who observed the same, while contradicting Farsain and Babaii (2020) who observed the opposite. The observed split in the present study was slightly in favor of participant-level assignment over class/group (58% to 42%) and is thus not overwhelming in terms of redefining the general trend observed in SLA research. More investigations into the area are required.

Regarding randomization guidance for future L2 IVA experimental research, it is probably unreasonable to assume that L2 IVA experimental researchers will have access to the resources required for randomized sampling. For those intending to do so, Hiver and Al-Hoorie's (2020, p. 71) process is useful. The authors referenced Korean government data (Korean Statistical Information Service, 2016 as cited in Hiver & Al-Hoorie) to construct strata in the sample representing population demographic, regional, and socioeconomic differences and then randomly selected participants from each of these strata. Randomized assignment, however, is within the practical scope of L2 IVA experimental design. Sometimes L2 researchers are forced by contextual constraints to use intact classes, but as highlighted by Rogers and Révész (2020) random assignment at the class level is an acceptable alternative. Busse et al. (2020) provides a useful example of the utility of this procedure. The study explicitly stated that the two intact groups were randomly assigned to either the experimental condition of "multilingual approach with affective-experiential activities" (p. 371) or the comparison condition comprising regular teaching. With such explicit labeling of random assignment, the reader can assume that the researchers worked to eliminate injecting bias into the design in relation to the assignment of learning conditions. In counterbalanced within-subject designs where learners learn target words under different conditions and subgroups experience the

https://doi.org/10.1017/S0272263121000541 Published online by Cambridge University Press

conditions in different orders of exposure, there is also room for such explicit labeling. When executing such a design, Folse (2006) featured "a random assignment of practice condition to each (counterbalanced) group" (p. 278) and thus met an expectation of experimental sample design dating back to Fisher (1935). The overarching point is that L2 IVA experimental samples, regardless of their variation across designs, can feature random assignment.

MULTISITE SAMPLE USE

A randomized sample from a meaningful population having all participants from a single school or university has a probability of virtually zero (Nicklin & Vitta, 2021; Vitta & Al-Hoorie, 2021). Such single-site convenience samples accordingly present bias and external validity limitations that are difficult to overcome no matter the sample size (Morgan-Short et al., 2018; Moranski & Ziegler, 2021). Multisite sample use appears underresearched in past L2 methodological syntheses but the observed frequency of multisite sample use in this current study (16 [14.54%] of 110 L2 IVA experimental reports) corresponds to Vitta and Al-Hoorie (2021), who reported that 18% of 56 L2 flipped learning experiments featured such samples. These low frequencies highlight the need for L2 experimental designs to employ multisite samples. Furthermore, more L2-focused methodological syntheses investigating the issue are required. The observation that no multisite sample report checked the statistical clustering effect (RQ3b) among different sites highlights a tangential area of improvement. As highlighted by Morgan-Short et al., these different sites constitute a random effect and when its influence is marginal we can feel more confident about the generalizations that the sample can facilitate about the population.

The observed small, yet trustworthy (CI boundaries outside of zero; Cumming, 2012) negative association between time and multisite use in L2 IVA experimental research was a surprising finding as L2 research is undergoing a reform with observed improvement over time (e.g., Lindstromberg, 2016). From the report pool (see Supplementary Materials A), one might notice that several researchers have published numerous reports over time. L2 IVA, given its classroom focus, could have naturally gravitated toward single-site samples to which these researchers had access. This is by no means a criticism of these inquiries and indeed L2 IVA is squarely within the instructed SLA arena where this classroom focus is justified (Loewen & Sato, 2017). Vocabulary represents an area of interest to both teachers and researchers (Nicklin & Vitta, 2021) and thus teachers who decided to conduct L2 IVA experiments would use their classrooms and schools in conducting their research. In both cases, the observed growth in L2 IVA research over time visible in Figure 3 could also contribute to the negative association where more single-site studies increased with the growth of L2 IVA publications.

What becomes clear in light of these findings is that L2 IVA experimental designs have room for improvement in terms of the frequency and manner of multisite sample use. To help realize such improvement, we present four points of guidance for constructing future L2 IVA experimental samples. To provide such guidance, reports from our pool featuring multisite samples are referenced in relation to how future L2 IVA designs could improve upon them. The first point of guidance is that the multisite sample planning process should begin with an *a priori* effect size-driven sample-size planning process. None of the

https://doi.org/10.1017/S0272263121000541 Published online by Cambridge University Press

16 multisite reports catalogued in this current synthesis engaged in any such process and this can be improved upon by future designs. Nicklin and Vitta (2021) provided L2 IVA-specific effect sizes and G*Power *a priori* power calculations for several common statistical analyses adopted by L2 IVA researchers that can be used as a starting point for sample-size planning. For those wishing to use precision, these L2 IVA effect sizes could be entered into Norouzian's (2020) R programs for calculating precision.

After estimating the number of participants required, the second point highlights the need for researchers to collaborate with colleagues to gain access to multiple sites from the intended population that are representative of differences and diversity within it. The fact that 12 of the 16 multisite reports in our pool had multiple authors indirectly substantiates the need for such collaboration. As a reference point regarding representativeness, Kim (2008) made L2 IVA inferences about the American ESL population from data collected from a midwestern American university and southeastern one. Data from the Eastern seaboard and West Coast locations would add to the robustness of the generalizations that future replications of the study could have (Morgan-Short et al., 2018). However, it is important to note that multisite samples are still convenience samples at the conceptual level, as argued in the preceding text, and represent only a compromise between the ideal randomized sample and flawed single-site samples (Vitta & Al-Hoorie, 2021).

To better realize this compromise, the third guidance point pertains to how future designs could incorporate a degree of randomization to these convenience samples. We exemplify this using Webb and Kagimoto (2009), who wisely drew their sample from two universities in Fukuoka and investigated verb-noun collocations such as *raise questions.* A replication of the study could involve collaborators to enable access to sites across Japan. After utilizing *a priori* effect size–driven sample-size planning, the replication researchers could recruit access to 150% to 200% of the participants required to meet the predetermined threshold. Sites could then be randomly excluded until 120% of the participants remain. This random exclusion works to somewhat mitigate the bias that a convenience sample of multiple locations would have.

To check for the clustering effect of multiple sites, which constitutes the final guidance point, there are two options that most L2 IVA researchers could execute. The first option is an *a priori* check. In this current study, such an assessment revealed that for all coded sample design judgments, except multisite sample use, the effects of which journal published the report were nonsignificant. When using multisite samples in L2 IVA research, a nonsignificant effect of the location grouping variable allows for a circumspect claim that clustering concerns are mitigated and thus generalizability to the population are enhanced. The second option involves analysis with LMMs where the sites can be incorporated as a random effect (Morgan-Short et al., 2018) and the contribution of the fixed (experimental) and random effects can be ascertained separately. The lower the effect of the sites, the stronger the evidence is that the data from the multisite sample is representative of the intended population.

LIMITATIONS

The analysis of the findings of this current study, a focused methodological synthesis, should be viewed in conjunction with its unavoidable limitations, of which three are most consequential. First, only six journals have been analyzed and thus publication and

https://doi.org/10.1017/S0272263121000541 Published online by Cambridge University Press

selection bias are concerns. The journal selection process, however, was systematized to represent perhaps the highest quality L2 IVA experimental research as SSCI journals have been observed to present higher quality quantitative research than journals not in the index (Al-Hoorie & Vitta, 2019). Even with this SSCI-focus, the observed frequency of reports satisfying the three aspects of experimental design that were considered was still low overall. The next limitation was the focused nature of this synthesis. There are other important aspects to sample design such as under- and over-investigated populations. Andringa and Godfroid (2020) highlighted for instance that Western, educated, industrialized, rich, and democratic (WEIRD; Henrich et al. [2010], as cited in Andringa & Godfroid, 2020]) samples dominate L2 quantitative research and thus such populations are overinvestigated. The areas presented in the present study are important, but not everything that should guide future L2 IVA experimental sample construction. Finally, the low number of reports satisfying the three aspects under consideration constrained the statistical analysis presented in this report. This was especially the case in the correlational analysis between time and coded sample design aspects addressing RQ4.

CONCLUSION

In this focused methodological synthesis, we analyzed 110 L2 IVA experimental reports and synthesized three sample design features: effect size–driven sample-size planning, randomization, and multisite use. From relevant literature (e.g., Cohen, 1988; Fisher, 1935; Morgan-Short et al., 2018), there is a strong argument that the three areas investigated in this current study work together to support the generalizability of inferential testing conducted on data from the samples. With the exception of randomized assignment, few reports satisfied the required criteria of these areas and thus this report acts as a call for sample design improvement in future L2 IVA experimental research. The low frequency of reports coded as meeting expectations in these areas corresponded to past syntheses (e.g., Farsani & Babaii, 2020; Lindstromberg, 2016; Plonsky, 2013; Vitta & Al-Hoorie, 2021). Unlike some past L2 research syntheses, where observed quantitative design quality, in the main, improved with time (e.g., Lindstromberg, 2016), time did not associate with most of the methodological features synthesized here. The only significant association was a small negative effect between time and multisite use which was surprising given the ongoing state of L2 methodological reform (Gass et al., 2021). As highlighted in the discussion, however, the interpretation of this negative correlation needs to be made with caution. Finally, guidance on how future L2 IVA experimental research could improve in the three areas synthesized was provided referencing results and examples from the report pool.

SUPPLEMENTARY MATERIALS

To view supplementary material for this article, please visit http://doi.org/10.1017/S0272263121000541.

COMPETING INTEREST

None.

https://doi.org/10.1017/S0272263121000541 Published online by Cambridge University Press

NOTES

[1]Experimental is used as a general term subsuming full experimental designs and quasi-experimental designs where part of the design (e.g., true control group) was omitted (Kuehl, 2000).

[2]Lakens et al. (2018) suggested that unstandardized effect sizes can be used for power analyses, but this view is not widespread among relevant literature, such as Faul et al. (2007).

[3]In line with L2 experimental research syntheses, all experimental effect sizes presented in this report have been transformed to *d* (for equations; see Brysbaert, 2019 and Brysbaert & Stevens, 2018). See Lakens (2013) for explanation and use of different *d* family metrics.

[4]The purpose of this discussion was to identify systemic flaws in the primary coder's decision process while identifying areas to explain to readers in the methods section. We thank the SSLA reviewers for highlighting the need to present this.

[5]Stated percentages are out of 110 reports.

[6]Effect sizes are interpreted according to Plonsky and Oswald's (2014) thresholds for L2 research.

REFERENCES

Al-Hoorie, A. H., & Vitta, J. P. (2019). The seven sins of L2 research: A review of 30 journals' statistical quality and their CiteScore, SJR, SNIP, JCR impact factors. *Language Teaching Research*, *23*, 727–744. https://doi.org/10.1177/1362168818767191.

Andringa, S., & Godfroid, A. (2020). Sampling bias and the problem of generalizability in applied linguistics. *Annual Review of Applied Linguistics*, *40*, 134–142. https://doi.org/10.1017/S0267190520000033.

Bryfonski, L., & Mckay, T. H. (2019). TBLT implementation and evaluation: A meta-analysis. *Language Teaching Research*, *23*, 603–632. https://doi.org/10.1177/1362168817744389.

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. https://doi.org/10.5334/joc.72.

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition*, *1*, 9. https://doi.org/10.5334/joc.10.

Busse, V., Cenoz, J., Dalmann, N., & Rogge, F. (2020). Addressing linguistic diversity in the language classroom in a resource-oriented way: An intervention study with primary school children. *Language Learning*, *70*, 382–419. https://doi.org/10.1111/lang.12382.

Choi, S. (2017). Processing and learning of enhanced English collocations: An eye movement study. *Language Teaching Research*, *21*, 403–426. https://doi.org/10.1177/1362168816653271.

Cisco, B. K., & Padrón, Y. (2012). Investigating vocabulary and reading strategies with middle grades English language learners: A research synthesis. *Research in Middle Level Education*, *36*, 1–23. https://doi.org/10.1080/19404476.2012.11462097.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Routledge Academic. https://doi.org/10.4324/9780203771587.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge. https://doi.org/10.4324/9780203807002.

de Carvalho, M. (2018). spearmanCI: Jackknife Euclidean/empirical likelihood inference for Spearman Rho. R package version 1.0. https://cran.rproject.org/web/packages/spearmanCI/spearmanCI.pdf.

Dellinger, J. (2017). Correlation, Spearman. In M. Allen (Ed.), *The SAGE Encyclopedia of Communication Research Methods* (pp. 274–275). Sage.

Egbert, J. (2007). Quality analysis of journals in TESOL and applied linguistics. *TESOL Quarterly*, *41*, 157–171. https://doi.org/10.1002/j.1545-7249.2007.tb00044.x.

Elgort, I. (2017). Incorrect inferences and contextual word learning in English as a second language. *Journal of the European Second Language Association*, *1*, 1–11. http://doi.org/10.22599/jesla.3.

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

Elgort, I., Beliaeva, N., & Boers, F. (2020). Contextual word learning in the first and second language: Definition placement and inference error effects on declarative and nondeclarative knowledge. *Studies in Second Language Acquisition*, *42*, 7–32. https://doi.org/10.1017/S0272263119000561.

Fanelli, D. (2010). Do pressures to publish increase scientists' bias? An empirical support from US States data. *PLoS ONE*, *5*, e10271. https://doi.org/10.1371/journal.pone.0010271.

https://doi.org/10.1017/S0272263121000541 Published online by Cambridge University Press

Farsani, M. A., & Babaii, E. (2020). Applied linguistics research in three decades: A methodological synthesis of graduate theses in an EFL context. *Quality & Quantity*, *54*, 1257–1283. https://doi.org/10.1007/s11135-020-00984-w.

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. https://doi.org/10.3758/BF03193146.

Field, A. (2018). *Discovering statistics using IBM SPSS statistics*. Sage Publications.

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

Folse, K. S. (2006). The effect of type of written exercise on L2 vocabulary retention. *TESOL Quarterly*, *40*, 273–293. https://doi.org/10.2307/40264523.

Fraley, R. C., & Vazire, S. (2014). The N-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PloS One*, *9*, e109019. https://doi.org/10.1371/journal.pone.0109019.

Gass, S., Loewen, S., & Plonsky, L. (2021). Coming of age: The past, present, and future of quantitative SLA research. *Language Teaching*, *54*, 245–258. https://doi.org/10.1017/s0261444819000430.

Glass, G. V. (1965). A ranking variable analogue of biserial correlation: Implications for short-cut item analysis. *Journal of Educational Measurement*, *2*, 91–95. https://doi.org/10.1111/j.1745-3984.1965.tb00396.x.

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

Green, P., & MacLeod (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*, 493–498. https://doi.org/10.1111/2041-210X.12504.

Harter, R. (2008). Random sampling. In P. Lavrakas (Ed.), *Encyclopedia of Survey Research Methods* (pp. 683–684). SAGE Publications. https://doi.org/10.4135/9781412963947.n440.

Hiver, P., & Al-Hoorie, A. H. (2020). Reexamining the role of vision in second language motivation: A preregistered conceptual replication of You, Dörnyei, and Csizér (2016). *Language Learning*, *70*, 48–102. https://doi.org/10.1111/lang.12371.

Hulstijn, J. H., & Laufer, B. (2001). Some empirical evidence for the involvement load hypothesis in vocabulary acquisition. *Language Learning*, *51*, 539–558. https://doi.org/10.1111/0023-8333.00164.

Hulstijn, J. H., Schoonen, R., de Jong, N. H., Steinel, M. P., & Florijn, A. (2012). Linguistic competences of learners of Dutch as a second language at the B1 and B2 levels of speaking proficiency of the Common European Framework of Reference for Languages (CEFR). *Language Testing*, *29*, 203–221. https://doi.org/10.1177/0265532211419826.

Kim, Y. (2008). The role of task-induced involvement and learner proficiency in L2 vocabulary acquisition. *Language Learning*, *58*, 285–325. https://doi.org/10.1111/j.1467-9922.2008.00442.x.

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

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*, 1–12. https://doi.org/10.3389/fpsyg.2013.00863.

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M.A. J., Argamon, S. E., … Zwaan, R. A. (2018). Justify your alpha. *Nature Human Behaviour*, *2*, 168–171. https://doi.org/10.1038/s41562-018-0311-x.

Lakens, D., & Caldwell, A. R. (2019). Simulation-based power-analysis for factorial ANOVA designs. *PsyArXiv*. https://doi.org/10.31234/osf.io/baxsf.

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

Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing*, *16*, 33–51. https://doi.org/10.1177/026553229901600103.

Lewis, M. (1993). *The lexical approach*. 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. https://doi.org/10.1080/09588221.2018.1541359.

Lindstromberg, S. (2016). Inferential statistics in Language Teaching Research: A review and ways forward. *Language Teaching Research*, *20*, 741–768. https://doi.org/10.1177/1362168816649979.

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

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, *22*, 276–282. https://doi.org/10.11613/bm.2012.031.

https://doi.org/10.1017/S0272263121000541 Published online by Cambridge University Press

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. https://doi.org/10.1177/0265532219898380.

Mehring, J. (2018). The flipped classroom. In J. Mehring and A. Leis (Eds.), *Innovations in flipping the language classroom: Theories and practices* (pp. 1–10). Springer.

Moranski, K., & Ziegler, N. (2021). A case for multisite second language acquisition research: Challenges, risks, and rewards. *Language Learning*, *71*, 204–242. https://doi.org/10.1111/lang.12434.

Morgan-Short, K., Marsden, E., Heil, J., Issa II, B.I., Leow, R.P., Mikhaylova, A., Mikołajczak, S., Moreno, N., Slabakova, R. and Szudarski, P. (2018), Multisite replication in second language acquisition research: Attention to form during listening and reading comprehension. *Language Learning*, *68*, 392–437. https://doi.org/10.1111/lang.12292.

Nicklin, C., & Plonsky, L. (2020). Outliers in L2 research in applied linguistics: A synthesis and data re-analysis. *Annual Review of Applied Linguistics*, *40*, 25–55. https://doi.org/10.1017/S0267190520000057.

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

Norouzian, R. (2020). Sample size planning in quantitative L2 research: A pragmatic approach. *Studies in Second Language Acquisition*, *41*, 849–870. https://doi.org/10.1017/S0272263120000017.

Norouzian, R. (2021). Interrater reliability in second language meta-analyses: The case of categorical moderators. *Studies in Second Language Acquisition*. Advance online publication. https://doi.org/10.1017/s0272263121000061.

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.

Osborne, J. W, & Waters, E. (2002). Four assumptions of multiple regression that researchers should always test. *Practical Assessment, Research, & Evaluation*, *8*, 1–5. https://doi.org/10.7275/r222-hv23.

Perugini, M., Gallucci, M., & Costantini, G. (2018). A practical primer to power analysis for simple experimental designs. *International Review of Social Psychology*, *31*, 1–20. http://doi.org/10.5334/irsp.181.

Peters, E. (2019). The effect of imagery and on-screen text on foreign language vocabulary learning from audiovisual input. *TESOL Quarterly*, *53*, 1008–1032. https://doi.org/10.1002/tesq.531.

Plonsky, L. (2013). Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research. *Studies in Second Language Acquisition*, *35*, 655–687. https://doi.org/10.1017/S0272263113000399.

Plonsky, L. (2014). Study quality in quantitative L2 research (1990–2010): A methodological synthesis and call for reform. *Modern Language Journal*, *98*, 450–470. https://doi:10.1111/j.1540-4781.2014.12058.x.

Plonsky, L. (2015). Statistical power, p values, descriptive statistics, and effect sizes: A "back-to-basics" approach to advancing quantitative methods in L2 research. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 23–45). Routledge. https://doi.org/10.4324/9781315870908.

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

Plonsky, L., and Gonulal, T. (2015). Methodological synthesis in quantitative L2 research: A review of reviews and a case study of exploratory factor analysis. *Language Learning*, *65*, 9–36. https://doi.org/10.1111/lang.12111.

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

Plonsky, L., Sudina, E., & Hu, Y. (2021). Applying meta-analysis to research on bilingualism: An introduction. *Bilingualism: Language and Cognition*. Advance online publication. https://doi.org/10.1017/S1366728920000760.

R Core Team (2020). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. https://www.R-project.org/.

Rogers, J., & Révész, A. (2020). Experimental and quasi-experimental designs. In J. McKinley & H. Rose (Eds.), *The Routledge handbook of research methods in applied linguistics* (pp. 133–143). Routledge.

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

Shintani, N. (2013). The effect of focus on form and focus on forms instruction on the acquisition of productive knowledge of L2 vocabulary by young beginning-level learners. *TESOL Quarterly*, *47*, 36–62. https://doi.org/10.1002/tesq.54.

https://doi.org/10.1017/S0272263121000541 Published online by Cambridge University Press

Siddaway, A. P., Wood, A. M., & Hedges, L. V. (2019). How to do a systematic review: A best practice guide for conducting and reporting narrative reviews, meta-analyses, and meta-syntheses. *Annual Review of Psychology*, *70*, 747–770. https://doi.org/10.1146/annurev-psych-010418-102803.

Traxler, R. E., & Nakatsukasa, K. (2020). The effectiveness of voice-on and voice-off instruction on ASL vocabulary acquisition. *Language Teaching Research*, *24*, 273–286. https://doi.org/10.1177/1362168818791601.

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

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

Vafaee, P., & Suzuki, Y. (2020). The relative significance of syntactic knowledge and vocabulary knowledge in second language listening ability. *Studies in Second Language Acquisition*, *42*, 383–410. https://doi.org/10.1017/S0272263119000676.

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., & Al-Hoorie, A. H. (2021). Measurement and sampling recommendations for L2 flipped learning experiments: A bottom-up methodological synthesis. *The Journal of Asia TEFL*, *18*, 682–692. https://doi.org/10.18823/asiatefl.2021.18.2.23.682.

Vo, S. (2019). Use of lexical features in non-native academic writing. *Journal of Second Language Writing*, *44*, 1–12. https://doi.org/10.1016/j.jslw.2018.11.002.

Webb, S., & Kagimoto, E. (2009). The effects of vocabulary learning on collocation and meaning. *TESOL Quarterly*, *43*, 55–77. https://doi.org/10.1002/j.1545-7249.2009.tb00227.x.

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

Willis, D., & Willis, J. (2007). *Doing task-based teaching*. Oxford University Press.

Yang, Y., Shintani, N., Li, S., & Zhang, Y. (2017). The effectiveness of post-reading word-focused activities and their associations with working memory. *System*, *70*, 38–49. https://doi.org/10.1016/j.system.2017.09.012.

Yanagisawa, A., Webb, S., & Uchihara, T. (2020). How do different forms of glossing contribute to L2 vocabulary learning from reading? A meta-regression analysis. *Studies in Second Language Acquisition*, *42*, 411–438. https://doi.org/10.1017/so272263119000688.

You, C., Dörnyei, Z., & Csizér, K. (2016). Motivation, vision, and gender: A survey of learners of English in China. *Language Learning*, *66*, 94–123. https://doi.org/10.1111/lang.12140.

Zhang, 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.

https://doi.org/10.1017/S0272263121000541 Published online by Cambridge University Press