THE JOURNAL OF ASIA TEFL
Vol. 18, No. 2, Summer 2021, 682-692
http://dx.doi.org/10.18823/asiatefl.2021.18.2.23.682

The Journal of Asia TEFL

http://journal.asiatefl.org/

e-ISSN 2466-1511 © 2004 AsiaTEFL.org. All rights reserved.

Measurement and Sampling Recommendations for L2 Flipped Learning Experiments: A Bottom-Up Methodological Synthesis

Joseph P. Vitta
Kyushu University, Fukuoka, Japan

Ali H. Al-Hoorie
Royal Commission for Jubail and Yanbu, Saudi Arabia

Introduction

With the advent of COVID-19, learning has transitioned from the classroom to online platforms (Tang et al., 2020). Interestingly, this forced migration to online teaching has coincided with the rise of the popularity of flipped learning in education in general (Låg & Sæle, 2019) and ELT and L2 in particular (Mehring, 2018). Flipped learning, in a general sense, inverts the traditional learning paradigm by presenting new content to students before and outside of the class with subsequent class time used for interacting and engaging with said content (for extensive consideration of L2 flipped learning; see Mehring, 2018). Unsurprisingly, recent research has investigated the potential of flipped classrooms against the backdrop of online teaching in response to COVID-19. Tang and colleagues (2020), for instance, investigated Chinese students' perceptions of flipped learning in online environments vis-à-vis traditional methods. It is likewise reasonable to expect L2 researchers to enhance their already substantial interest in flipped learning in response to the current migration to online teaching. What is more is that even after COVID-19 subsides, there is no reason to assume the L2 academic community's interest in flipped learning will wane as interest in it had been growing before the pandemic (see e.g., Bonyadi, 2018).

This growing interest in flipped learning and related areas such as blended learning (see Mahmud, 2018) within L2 contexts underpins this current study, a methodological synthesis. We thus present a focused systematic review of methods shortcomings in past L2 flipped learning experimental and quasi-experimental (e.g., no control group) designs ('experimental' is used hereafter as a general term). From an earlier meta-analysis (Vitta & Al-Hoorie, 2020) conducted by the researchers, several consequential and addressable methodological issues were observed that future inquiries could address to increase the rigor and trustworthiness of their findings. Given the waxing focus L2 teachers and researchers are placing on flipped learning, highlighting and observing the frequency of such issues are worthwhile as it provides a point of reference and a road map for improving future L2 flipped research. As L2 quantitative inquiries are currently undergoing a state of methodological reform (Gass, Loewen, & Plonsky, 2020), we must state clearly that we are not criticizing past inquiries, and this inquiry was conducted with the intention of improving future research into the effectiveness of flipped learning.

Selection of Methodological Issues

The methodological issues were selected via a bottom-up process during the coding of effects presented in the earlier meta-analysis (Vitta & Al-Hoorie, 2020). To be selected, an issue had to be both *consequential* and *addressable* with the researchers' observing the issue among several (as opposed to only a few) reports in the pool. Failure to report effect sizes was omitted from this review, for instance, because most reports, especially those published recently, had overtly reported them. *Consequential* implies that not addressing the issue could bias the results and/or the trust that readers would place in the results. Effect size reporting would exemplify a consequential issue given the call for researchers to report effect sizes even when nonsignificant (see Al-Hoorie & Vitta, 2019). Similarly, checking inter-rater reliability with a simple percentage of agreement instead of Cohen's kappa would be classified as inconsequential because the former metric still respects the expectation that a subsequent independent coding (McHugh, 2012) must establish reliability for nominal and/or ranked judgements. The recent call for L2 experimental designs to replace fixed-effect models (e.g., ANOVA) with mixed-effect models (see Linck & Cunnings, 2015) provides an example of a less easily *addressable* issue for the average L2 researcher currently. While use of these models are growing in the field, they do require use of specialized programs and standard research handbooks and corresponding tools (e.g., SPSS; Field, 2018) are still mainly supporting and detailing fixed-effect models and therefore we do not discuss this point further in this article. At the end of this process, five issues, subsumed under measurement and sampling, emerged as being somewhat widespread among the reports, consequential, and addressable:

Measurement Recommendations
1) assessing the reliability of dependent variable measurements
2) administering pretests
Sampling Recommendations
3) conducting *a priori* power analysis
4) describing the randomized assignment procedure explicitly
5) using multi-site samples

The consequence of the selected issues becomes clear when reviewing relevant literature, especially L2 research syntheses of quantitative inquiries (e.g., Plonsky, 2013, 2014). In what follows we review each recommendation in turn against past considerations.

Reliability

As stated in the psychometric literature, reliability is a prerequisite of validity. In guidance for L2 researchers, Al-Hoorie and Vitta (2019) argued that reliability, even of trusted instruments, must always be reported. In other words, measurements that are not reliable cannot be valid. It is unsurprising that reliability has been of interest to L2 methodological syntheses with divergent findings observed. Plonsky and Gass (2011) observed 64% of reports reporting reliability while Plonsky (2014) observed 50% in reports published in the 2000s. The different scope and foci of the inquiries most likely accounts for such discrepancies.

Pretest Use

As highlighted by methodologists (e.g., Kuehl, 2000), experiments in the strictest sense do not have to include a pretest, a measurement of the outcome/dependent variable before the treatment. Recent L2 experimental guidance (Rogers & Révész, 2020), however, has conceptualized pretests as vital to establishing pre-treatment equivalency among the experimental groups. This makes sense as L2 learners will most likely come into classrooms with varying experiences learning the target language. L2 methodological syntheses have observed that experimental designs for the most part tend to use pre-tests. Farsani and Babaii (2020) found that 90% of experimental designs in graduate theses included pre-tests while Plonsky (2013) observed 67% pretest use among reports located in major L2 journals.

Power

Statistical power refers to having a large enough sample to detect a selected effect (Brysbaert, 2019). In other words, powered samples are planned around the *effect size*, i.e., quantification of the effect in the sample that the researcher intends to detect. It is related to type II error, the probability of rejecting an observed effect (size) in the sample as existing in the population when in reality it does exist there (false negative). The more power a sample has the less chance it has of rejecting such true effects, i.e. type II error. In simpler terms, having more participants equates to higher power. While post-hoc power procedures do exist, telling the reader the observed power of a finding assuming it actually exists in the population, conducting power analyses before the research referencing relevant effects (*a priori* power analysis) is considered a gold standard (Aberson, 2019). Power has thus traditionally been viewed as within the frequentist domain, but power-like analysis exists in the Bayesian realm with Kruschke (2015) describing it as “the probability of achieving the goal of a planned empirical study, if a suspected underlying state of the world is true" (p. 359). Other sample planning processes such as precision (Cumming, 2012), which determines the required sample size large enough to narrow the confidence interval for a given effect, also require the *a priori* selection of a certain effect size to execute sample planning. The overarching point is that researchers must plan their sample size referencing past effect sizes.

When such sample planning is omitted, as highlighted by Brysbaert (2019), findings become of little interpretive value vis-à-vis uncovering effects existing in the population. Underpowered studies, subsuming pilot studies, that reject small effects could have had a sample too small to detect a small but true effect existing in the population. In a similar vein, large effects in underpowered studies might be ‘flukes’ unrepresentative of the true parameter(s) within the population (see Brysbaert, 2019). Both past (Plonsky, 2013) and recent (Farsani & Babaii, 2020) L2 research syntheses have unfortunately observed that few studies (less than 1–5%) consider power in any fashion with *a priori* versus other power analyses not always clearly catalogued.

Randomized Assignment

Randomized (or random) assignment is one side of the randomization process which one expects to find in experimental designs, with the other being randomized sampling or randomly drawing participants from the population (Kuehl, 2000). Random assignment refers to assigning experimental conditions randomly to participants; the gold standard for this process is at the participant level (Rogers & Révész, 2020). When ‘writing up’ experiments in reports, there is a need to explicitly detail the random assignment process given its importance to experimental designs. L2 researchers, however, are sometimes precluded from implementing participant-level random assignment as they perform research on pre-existing classes/groups (Farsani & Babaii, 2020). With classroom-orientated research areas such as flipped learning, random assignment is often only possible at the class-level (e.g., Hung, 2015). Plonsky (2013) and Farsani and Babaii (2020), when coding random assignment, therefore coded the random assignment using the person- versus class/group-level distinction. Unlike power, the field has generally adopted random assignment in its experimental design processes; Farsani and Babaii observed 58% randomized assignment use, subsuming both levels, within its report pool which corresponds with the 47% observed earlier by Plonsky (2013).

Multi-site Use

Randomized sampling, subsuming other probability sampling approaches such as stratified (see Harter, 2008), are expensive and labor intensive and perhaps impractical for L2 researchers. Recent examples of such sampling procedures (e.g., Hiver & Al-Hoorie, 2020), see sample sizes over 1000 drawn from various regions within the country and/or intended population. It is therefore unsurprising that recent L2 methods guidance has either emphasized randomized assignment (e.g., Rogers & Révész, 2020) or mentioned randomization in a general sense (e.g., Gass et al., 2020). Many L2 researchers most likely are practically constrained from constructing randomized samples.

If randomized sampling is at one end of the continuum and perhaps impractical in relation to selecting participants for a sample, then single-site convenience samples are at the other end. Imagine that one randomly selected 200 Chinese High School EFL students from a national database. The probability that all of them would be from the same region yet alone the same school is for all practical purposes zero. It is therefore unsurprising that there has been a recent call to replicate past L2 studies using multi-site samples (see Morgan-Short et al., 2018). Multi-site samples, while still probably being convenience samples conceptually, create a "random(effect)-by site" (Morgan-Short et al., 2018, p. 408) which can allow for a better generalization to meaningful populations than single-site samples assuming the sites' effect is minimal vis-à-vis the fixed experimental effects. While recent L2 syntheses such as Farsani and Babaii (2020) did investigate randomized (probability) versus convenience sampling, it appears the multi- versus single-site sampling has yet to be overtly explored within L2 research syntheses.

Present Study

In the present study, we reviewed a report pool of L2 flipped experimental (full- and quasi-experimental) designs in relation to the measurement and sampling issues that were selected as being both consequential and addressable. The research question governing the present study is:

To what extent do L2 flipped experimental designs: a) report reliability, b) use pretests, c) consider power, d) employ randomized assignment, and e) use multi-site samples?

As highlighted above, these issues have been the focus of past research syntheses (e.g., Plonsky, 2014) or been recently brought to the attention of L2 researchers (e.g., multi-site use; Morgan-Short et al., 2018). The rationale of our investigation was to highlight areas that could be improved upon by future researchers while also providing empirical data on how satisfactory L2 flipped learning experimental designs appear to be in the areas investigated.

Report Pool Creation

The report pool utilized in this study was first created for a meta-analysis of experimental designs investigating L2 flipped learning interventions. Following the conventions for a robust literature search process to mitigate selection bias, multiple databases, indexes, and search engines were employed to review over 50,000 reports with over 8000 papers manually inspected as well against our search criteria (see Vitta & Al-Hoorie, 2020). The researchers also published a call for papers to capture unpublished reports.

The process terminated in August 2019 (with no beginning time constraint) and 56 reports were selected. Of these, 45 of these were journal articles and 11 were either conference proceedings (k = 4 reports) or theses and dissertations (k = 7). Our report pool total was larger than recent research syntheses such as Låg and Sæle (2019) who located 23 L2 flipped experimental designs.

Operationalization, Coding and Results

The five areas of inquiry driving this investigation and reflected in the governing research question were operationalized into dichotomous judgments (items) located in Table 1. A 'yes' response meant that the report was satisfactory; a 'no' meant that it was not. For some judgments, further explanation on the yes/no delineation has been present as 'additional notes.' With the exception of power, one item or judgment operationalized each issue investigated. Since other forms of power analysis besides *a priori*, which is the gold standard (Brysbaert, 2019), have been considered useful by some (Aberson, 2019), Item 3a was included to capture such iterations of power analysis.

For each report, judgments were made for the items found in Table 1. After this initial coding, a second researcher coded 25% of the report pool (*k* = 14). Because of the overwhelming instances of 'no' judgements for some items, inferential testing of judgements, e.g., Cohen's kappa, was not possible (McHugh, 2012). Instead, raw percentage of initial agreement was employed and ranged from 71.42% to 100% across the six items as highlighted in Table 1 with an average agreement of 88.09% (*SD* = 11.67%) observed. Differences were resolved via discussion.

TABLE 1
Judgments of Selected Measurement and Sampling Design Issues

Item	Additional Notes	Inter-rater Reliability
Item 1. Did the study report the reliability of the L2 outcome variable(s)' measurements?	Coded 'yes' if a standardized test was employed as outcome variable or a published paper of said exam was used.[†] Coded 'no' if qualitative validation process, e.g., peer/colleague review, of test measuring classroom outcomes such as targeted vocabulary items. Marked 'no' if reliability was reported for some but not all L2 outcomes.	78.57%
Item 2. Was a pretest employed to establish pre-treatment equivalency among the groups/experimental conditions in relation to the outcome variable?	Coded 'no' if pre-treatment measurement was not measuring the same L2 outcome as the posttest. Coded 'no' if pretest was not fully reported. Coded 'yes' if measurement was near start of treatment and demonstrated equivalency.	85.71%
Item 3a. Was power considered in the study in any fashion?[§]	Included post-hoc, compromise, etc.	100%
Item 3b. Was *a priori* power analysis conducted referencing relevant effect sizes?	Had to be reported as an *a priori* process referencing relevant L2-centric effect sizes.	100%
Item 4. Did the study engage in random assignment at either the participant or class/group level?	Coded 'yes' only if there was *explicit* description/labeling of the random assignment process. Text had to explicitly state chance was involved in assigning conditions. Some reports conflated group- and participant-level assignment by mislabeling what was clearly group-level assignment as participant-level.	71.42% (before *explicit* clarification)
Item 5. Was the sample multi-site?	Coded 'no' when report was ambiguous.	92.85%

Note 1. It may be unreasonable to expect all classroom researchers to obtain reliability data for such test data. Past papers, likewise, have been externally validated but in the strictest sense, future research should check instrument reliability.
Note 2. No study employed Bayesian analyses or mentioned sampling planning procedures such as precision (Cumming, 2012). Accordingly, power relates to NHST-centric analysis here.

As highlighted in Figure 1, the investigated methodological issues had a varying range of frequency of occurrence. In sum, L2 flipped experimental reports appeared to be more satisfactory in the measurement issues than in sampling. Considering the former, about 70% of the reports (Item 2: *k* = 39) employed a pretest and half of the reports (Item 1: *k* = 28) reported reliability according to the standards presented in Table 1. Regarding sampling, only about 36% (Item 4: *k* = 20) and about 18% (Item 5: *k* = 10) of the reports featured randomized assignment and multi-site use, respectively. No report (Item 3b) employed the gold standard of *a priori* power analysis.

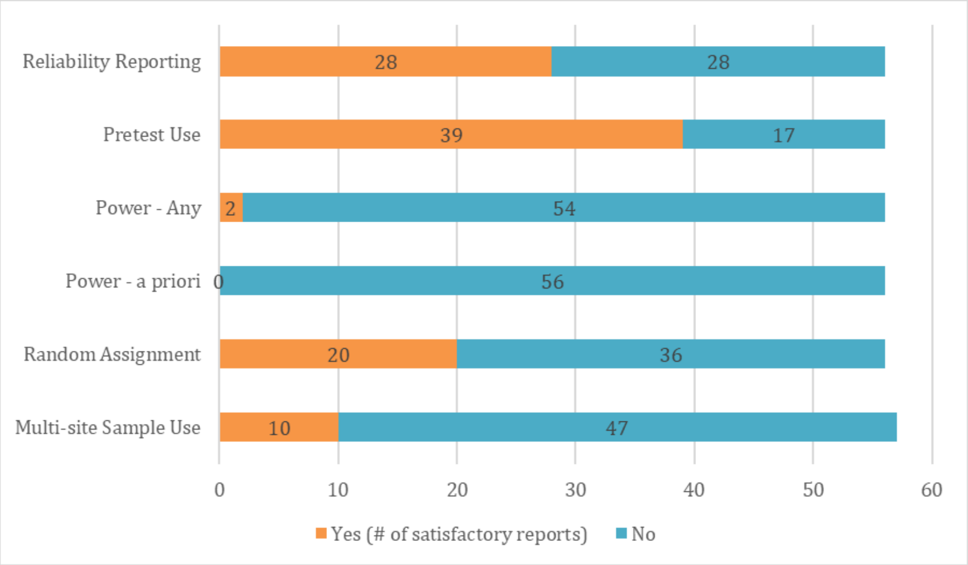


Figure 1. Frequency of Measurement and Sampling Issues in L2 Flipped Experimental Reports.

Discussion

The implications of the current study's findings are two-fold. First, L2 flipped experimental designs report reliability and use pretests with a similar frequency observed in other L2 research syntheses. Second, there is room to improve sampling designs in future L2 flipped experiments. We discuss each in turn.

Reliability and Pretest Use

The frequency of reliability checking (50%, *k* = 28) was within the range of observations observed in past methodological syntheses. Consider, for instance, that Plonsky and Gass (2011) observed 64% of reports reported reliability while Plonsky (2014) observed that 50% of reports published in the 2000s had also done so. Reports in the pool can therefore be referenced for appropriate practice. When multiple-choice tests or other close-ended instruments are employed to facilitate the outcome variable, a reliability metric such as Cronbach's alpha or KR-21 is needed to demonstrate the observed internal consistency of the measurement. Al-Harbi and Alshumaimeri (2016) provided such information when reporting the reliability of their grammar measurement. When the measurement is derived from rater (human) judgements, inter-rater reliability is the standard to establish the reliability of the judgments. Bonyadi (2018) provides a positive example of this process in validating judgments of oral performance. When there is approximate balance among the possible judgements, inferential testing and corresponding metrics such as Cohen's kappa (κ) can be used to correct for the contribution of chance to the observed percentage of agreement.

As with reliability, 39 reports (around 70%) included a pretest which corresponded to observations such as Plonsky's (2013) observed 67%. Pretest use is somewhat straightforward compared to reliability, but future research should be mindful of how pretests can affect inferential testing. A handful of reports employed ANCOVA (see Field, 2018) where the outcome (dependent) variable was either a gain score (posttest–pretest) or the posttest and the covariate was the pretest. Future research should be mindful however that such designs (pretest as predictor/covariate) have been observed to create issues in relation to model estimates (Lord's paradox; see Allison, 1990, p. 96) with gain scores alone as dependent variable perhaps yielding more accurate results in fixed-effect models.

Reforming and Improving Sampling Issues

Power considerations were almost completely ignored within our report pool and this corresponded to past L2 methodological syntheses (e.g., Plonsky, 2013). As highlighted by Brysbaert (2019), *a priori* power is the gold standard for effect size planning and requires a pre-determined effect to begin the process. In the larger study preceding this report, the researchers (Vitta & Al-Hoorie, 2020) observed a (corrected for publication bias) aggregated effect of g = 0.58 where flipped interventions were more effective than non-flipped comparisons. This value corresponded to Plonsky and Oswald's (2014) median between-subject effect of d = 0.70 in group comparisons in L2 research. Calculating power from g = 0.58 for a 3-group design (e.g., treatment, comparison, control) shows that researchers need 120 participants with 40 per group (calculated with G*Power; Faul, Erdfelder, Lang, & Buchner, 2007; see Appendix A). Post-hoc comparisons would have to be executed using tests such as Tukey or Games-Howell to maintain significance with this sample size. A two-group design requires two groups of 48 participants (N = 96). Should researchers wish to be conservative in their power calculations, the aggregated effect from an overarching review of education research (d = 0.40; see Hattie, 2009) could be employed as opposed to g = 0.58.

Random assignment and multi-site use can work in tandem to enhance the generalizability facilitated by samples found in future L2 flipped experimental designs. As highlighted above, random assignment is observed in about a third of the report pool which was somewhat lower than observations in past research syntheses, e.g., 58% observed by Farsani and Babaii (2020). Nevertheless, 20 reports did implement this feature and these can act as positive examples. Hung (2015) for instance provides a clear rationale for and description of random assignment at the class level. To be able to better utilize multi-site samples, researchers could collaborate with each other to construct samples with multiple locations reflective of the intended population. Overt descriptions of such collaboration was missing from the report pool, but consider as an example vocabulary within an Asian TEFL context, Japan, which like flipped interventions are often investigated at the classroom level. McLean, Kramer, and Beglar (2015) presented a multi-authored report where the co-authors facilitated different sites from which to construct their sample. Where one can improve on this approach would be the addition of a degree of randomness. Assume that six researchers provide access to six Vietnamese universities. After *a priori* power calculations, the researchers could select the required sites (from the six available universities) to meet the power-determined sample size threshold. Alternatively, participants could be drawn randomly from the six sites to attain an appropriately powered sample.

Conclusion

In this focused methodological synthesis, 56 L2 flipped experimental designs were analyzed in relation to their reporting of reliability, pretest use, power considerations, use of random assignment and multi-site samples. Taken together, the designs seemed adequate in handling reliability and including pretests, but there is room for improvement in future L2 flipped research. Power considerations and multi-site samples were largely missing, and so future flipped research could consider these issues in order to enhance the generalizability of findings.

Data Availability: Data sheet and report pool bibliography, available upon request to corresponding author.

Acknowledgements

The authors acknowledge Dr. Jeffery Mehring and Christopher Nicklin for their assistance with this project.

The Authors

Joseph P. Vitta (corresponding author) is an Associate Professor at Kyushu University, Fukuoka, Japan.

Kyushu University – Faculty of Languages and Cultures
744 Motooka, Nishi-ku, Fukuoka-shi, Fukuoka-ken, 819-0395
Tel: +81 092-802-2125
Email: vittajp@flc.kyushu-u.ac.jp
ORCID: 0000-0002-5711-969X

Ali H. Al-Hoorie is an assistant professor at the Jubail English Language and Preparatory Year Institute, Royal Commission for Jubail and Yanbu, Saudi Arabia.

Jubail English Language and Preparatory Year Institute
Royal Commission for Jubail and Yanbu
Jubail Industrial City 31961
Saudi Arabia
Email: hoorie_a@jic.edu.sa
ORCID: 0000-0003-3810-5978

References

Aberson, C. L. (2019). *Applied power analysis for the behavioral sciences*. Routledge.

Al-Harbi, S. S., & Alshumaimeri, Y. A. (2016). The flipped classroom impact in grammar class on EFL Saudi secondary school students' performances and attitudes. *English Language Teaching, 9*(10), 60–80.

Al-Hoorie, A. H., & Vitta, J. P. (2019). The seven sins of L2 research: A review of 30 journals' statistical quality and their CiteScore, SJR, SNIP, JCR Impact Factors. *Language Teaching Research, 23*(6), 727–744. https://doi.org/10.1177/136216881876719

Allison, P. D. (1990). Change scores as dependent variables in regression analysis. *Sociological Methodology, 20,* 93–114.

Bonyadi, A. (2018). The effects of flipped instruction on Iranian EFL students' oral interpretation performance. *The Journal of Asia TEFL, 15*(4), 1146–1155.

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), 1–38. https://doi.org/10.5334/joc.72

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

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*(4), 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*(2), 175–191. https://doi.org/10.3758/BF03193146

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

Gass, S., Loewen, S., & Plonsky, L. (2020). Coming of age: The past, present, and future of quantitative SLA research. *Language Teaching,* 1-14. https://doi.org/10.1017/s0261444819000430.

Harter, R. (2008). Random sampling. In P. Lavrakas (Ed.), *Encyclopedia of survey research methods* (pp. 683–684). Sage.

Hattie, J. A. C. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*.: Routledge.

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*(1), 48–102 https://doi.org/10.1111/lang.12371

Hung, H.-T. (2015). Flipping the classroom for English language learners to foster active learning. *Computer Assisted Language Learning, 28*(1), 81–96.

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

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

Låg, T., & Sæle, R. G. (2019). Does the flipped classroom improve student learning and satisfaction? A systematic review and meta-analysis. *AERA Open, 5*(3), 1-17. https://doi.org/10.1177/23328584 19870489

Linck, J. A., & Cunnings, I. (2015). The utility and application of mixed-effects models in second language research. *Language Learning, 65*(S1), 185–207. https://doi.org/10.1111/lang.12117

Mahmud, M. M. (2018). Technology and language–What works and what does not: A meta-analysis of blended learning research. *The Journal of Asia TEFL*, *15*(2), 365–382. https://doi.org/10.18823/asiatefl.2018.15.2.7.365

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica, 22*(3), 276–282. https://doi.org/10.11613/bm.2012.031

McLean, S., Kramer, B., & Beglar, D. (2015). The creation and validation of a listening vocabulary levels test. *Language Teaching Research, 19*(6), 741–760. https://doi.org/10.1177/1362168814567889

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

Morgan-Short, K., Marsden, E., Heil, J., Issa II, B. I., Leow, R. P., Mikhaylova, A., Mikołajczak, S., Moreno, N., Slabakova, R., & Szudarski, P. (2018), Multisite replication in second language acquisition research: Attention to form during listening and reading comprehension. *Language Learning, 68*(2), 392–437. https://doi.org/10.1111/lang.12292

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*(4), 655–687. https://doi.org/10.1017/S0272263113000399

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

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

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

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.

Tang, T., Abuhmaid, A. M., Olaimat, M., Oudat, D. M., Aldhaeebi, M., & Bamanger, E. (2020). Efficiency of flipped classroom with online-based teaching under COVID-19. *Interactive Learning Environments*, 1–12. https://doi.org/10.1080/10494820.2020.1817761

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:10.1177/1362168820981403

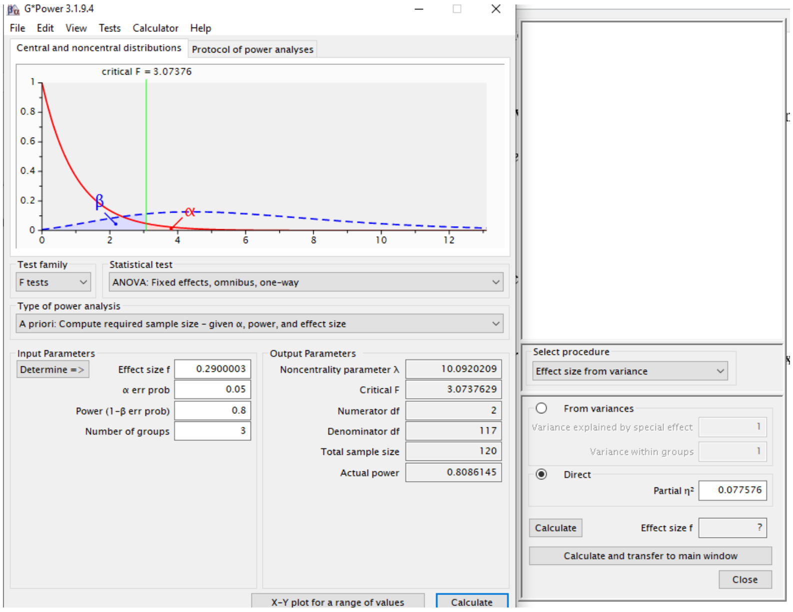
(Received March 10, 2021; Revised May 20, 2021; Accepted June 18, 2021)

Appendix A

Power Calculations

g = .58 equates to eta-squared = .077567 where eta-squared = d^2 / (d^2 + 4) – extrapolated from formula in Brysbaert (2019)

Three group one-way ANOVA


G*Power 3.1.9.4
File
Edit
View
Tests
Calculator
Help
Central and noncentral distributions
Protocol of power analyses
critical F = 3.07376
1
0.8
0.6
0.4
0.2
0
0
2
4
6
8
10
12
β
α
Determine =>
Calculate
Calculate and transfer to main window
Close
X-Y plot for a range of values
Calculate

Test family: F tests

Statistical test: ANOVA: Fixed effects, omnibus, one-way

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters

Effect size f 0.2900003
α err prob 0.05
Power (1-β err prob) 0.8
Number of groups 3

Output Parameters

Noncentrality parameter λ 10.0920209
Critical F 3.0737629
Numerator df 2
Denominator df 117
Total sample size 120
Actual power 0.8086145

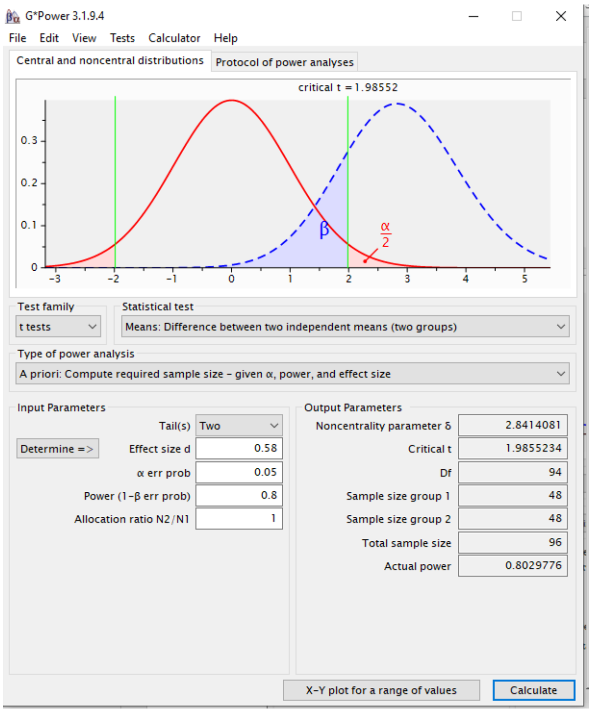
Select procedure: Effect size from variance

○ From variances
Variance explained by special effect 1
Variance within groups 1

● Direct
Partial η² 0.077576

Effect size f ?

Two-group t-test



Test family: t tests

Statistical test: Means: Difference between two independent means (two groups)

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters

Tail(s) Two
Effect size d 0.58
α err prob 0.05
Power (1-β err prob) 0.8
Allocation ratio N2/N1 1

Output Parameters

Noncentrality parameter δ 2.8414081
Critical t 1.9855234
Df 94
Sample size group 1 48
Sample size group 2 48
Total sample size 96
Actual power 0.8029776