Computer Assisted
Language Instruction
Consortium

CALICO (ONLINE) ISSN 2056–9017

Article

Quantitative considerations for improving replicability in CALL and applied linguistics

Luke Plonsky

Abstract

There are a number of methodological practices commonly employed by CALL researchers that limit progress in the field. Some of these practices are particular to replication research, but most are more general and are found throughout applied linguistics. I describe in this paper two studies that are fabricated but that resemble much of what is found in the field. Each study corresponds to and contains a set of methodological issues. Following each study, I address the issues they illustrate, providing comments and suggestions for how the analyses could be modified to produce greater replicability and/or replicational value. I conclude with a summary of suggestions for quantitative reforms related to improving replication research and quantitative practices more generally in CALL and applied linguistics.

KEYWORDS: COMPUTER-ASSISTED LANGUAGE INSTRUCTION; QUANTITATIVE RESEARCH METHODS; REPLICATION; SLA

Affiliation

Northern Arizona University, Flagstaff, AZ, USA.
email: luke.plonsky@nau.edu

CALICO VOL 32.2 2015 232–244
doi : 10.1558/cj.v32i2.26857

Introduction

There is broad consensus that replicability is fundamental to advancing scientific knowledge. This is perhaps especially true in the many subdomains of applied linguistics, including computer assisted language learning (CALL), where there is an ever-pressing need to examine and establish the generalizability of technological resources across learners/populations, contexts, linguistic targets, computational tools, and so forth (see Smith & Schulze, 2013). However, engaging in replication research and, conversely, producing research that is replicable, is not a given. Indeed, there are many forces – institutional, editorial, curricular, personal, and field-specific, to name a few – that deter scholars from engaging in replication/replicable research (see Porte, 2013). Yet an additional and critical but often overlooked roadblock to replication research in the domain of CALL and throughout applied linguistics, is our handling of quantitative data.

It may come as a surprise to some researchers that I (or anyone) would suggest that the statistical techniques in a given study would have much to do with its replicability or its potential contribution as a replication. In response to such a comment I would argue that the two are actually very closely related; our data handling actually has much to do with whether and to what extent our results are able to contribute to the larger substantive context in which each study is carried out and reported. Embracing an approach that recognizes this broader context is part of the synthetic- and replicatory-minded ethic argued for in recent years by a small but vocal number of applied linguists (Rebekha Abbuhl, Alison Mackey, John Norris, Lourdes Ortega, Charlene Polio, Graeme Porte, myself). I recognize and have even demonstrated some progress in this area (e.g., Plonsky, 2014). And of course, the publication of the special issue in which this article appears and other special issues of a similar orientation in the *Journal of Second Language Writing* (2012, Issue 4), for example, provide further evidence of the improved status of replication research. By and large, however, an individualistic view of knowledge development is still prevalent and evident in the way we conduct and present quantitative analyses. A synthetic and replicatory mindset, by contrast, pushes researchers working with the same or a similar set of constructs and processes to view their work as part of a larger, collective, and connected body of research to be viewed, evaluated, and summarized in an ongoing and holistic fashion.

Further and in an effort to cast this article to the widest audience possible, I want to be clear that I do not view the mindset being described here and the statistical techniques it embodies as limited in applicability to only those studies that frame themselves as replications (or as those intending to be replicated). In fact, such studies likely already possess a heightened awareness of

the place of their work in relation to others in the same line of investigation. In my view, the points made in this paper apply to nearly all quantitative studies.

With these motivations in mind, I have identified two sets of quantitative techniques that I would like to focus on for their potential to lead to greater and more informative replication and replicability in CALL as well as elsewhere in applied linguistics. Rather than simply discuss these techniques and their conceptual foundation in abstract terms, I am going to describe two studies, each corresponding to one set of issues and techniques. These studies will appear familiar to readers not because they have read them (I made them up) but because they are typical of what is found in CALL and throughout the rest of quantitative applied linguistics research. To be clear, although these are fabricated examples, they possess potentially viable designs and data sets in the realm of CALL. The variables could be substituted with any number of variables. I will then use these descriptions to contrast the analytical problems they present with alternate techniques that lead to greater replicability and replicational value.

Example Study 1

Example Study 1 posed the following research question: Do game-based activities lead to greater L2 pragmatics development compared to traditional, classroom instruction? The participants, 30 learners of Spanish as a foreign language, were divided into two groups (Game and Traditional). Those in the Game group utilized an avatar-type interface to complete a series of brief tasks designed to elicit the participants' use of requests. The Traditional group completed similar role-play tasks but in a classroom environment. Learners' pragmatic competence was measured by means of two equivalent forms of a 30-item written discourse completion task on requests, given before and after the treatment.

Once the data were collected, the researcher ran an independent samples *t*-test to compare the two groups' scores on the pretest. No statistically significant difference was found between them (see Table 1; $t = 1.37$, $p > 0.05$). Then, in order to answer the research question, the researcher compared the two groups' posttest scores, again, using a *t*-test. This time, however, the difference was statistically significant: $t = 2.32$, $p < 0.05$. The author concluded that the game-based learning environment is superior to traditional instruction in terms of fostering pragmatic development and argued in the discussion for corresponding changes to be made to foreign language curricula.

On the surface, there is nothing immediately objectionable here; we have all read dozens of studies and attended just as many, probably more, conference presentations just like this. There are, however, several serious problems with the author's approach which diminish the project's value both as a stand-alone

study and, more importantly, from a replicatory and synthetic point of view. I will now discuss these problems in turn.

Table 1: Example Study 1: Pre- and posttest results

	Pretest	Posttest
Condition	*M* (*SD*)	*M* (*SD*)
Traditional (*n* = 15)	7 (2)	14 (3)
Game (*n* = 15)	8 (2)	17 (4)

1. The absence of evidence is not evidence for absence

More specifically, the lack of a statistically significant difference on the pretest does not mean that the two groups are actually equal (Godfroid & Spino, in press). Although the author's approach is quite common, it exemplifies what Cumming (2012) refers to as the 'slippery slope of nonsignificance' (p. 31); just because the difference between the groups is not statistically significant, we cannot assume that there is no difference. Or, in statistical terms, $p > 0.05 \neq d = 0$, where d refers to a standardized mean difference effect size. Using this index of practical significance, we see that the difference between the groups, although it appears very small, is actually not very close to zero at all: $d = 0.5$. (one half of a standard deviation unit). Though not large by most scales, a difference of this magnitude is certainly one that should be taken into account when interpreting posttest results. Unfortunately, differences between groups like this one, much more often than not, are dismissed on account of the lack of a statistical difference on pre-test scores. As discussed below this unperceived difference also has implications for the rest of the study's results.

The larger issue at hand here, though, is that of our field's extremely heavy reliance on null hypothesis significance testing (NHST; see Norris, in press; Plonsky, in press). Given a large enough and more powerful (statistically speaking) sample, this same difference in pretest scores would indeed be statistically significant. With this understanding and the view to group differences shown through the *d* effect size, the author would likely want to consider pretest performance as a covariate for posttest comparisons.

2. The world is not black or white, statistically significant or non-significant

Closely tied to the previous points is another based on the researcher's use of statistical significance. The approach embodied by NHST prompts researchers to boil their results down to a dichotomy. Such a perspective, also found frequently in the yes/no wording of research questions, is a shameful waste of data and drastically unproductive from a replicatory and synthetic point

of view. In the case of Example Study 1, it is the researcher's black and white thinking, informed by the inherently problematic *p*-value, that caused the author to dismiss entirely the substantial difference in pre-test scores. And it is with this same kind of thinking that s/he might claim unreserved superiority of game-based over traditional pragmatics instruction. The magnitude of the difference (i.e., the effect size, *d*) between the groups' posttest scores is $d = 0.85$. It is important to note that this effect size indicates a fairly precise and substantial difference between conditions that is not at all indicated by the marker of '$p < 0.05$'. However, if we then subtract from this value the previously unaccounted for effect size for the pre-treatment differences ($d = 0.5$), the posttest effect size drops to a much smaller $d = 0.35$.

Reporting and interpreting these effect sizes and their corresponding confidence intervals is absolutely critical for both replication studies and for studies such as this one that may be replicated in some form in the future (see Norris *et al.*, in press; Plonsky & Oswald, 2014). In the former, in order to fully explain and contextualize the findings, the researcher may have to take the additional step to calculate the effect size(s) from the original study. In addition, and on a more conceptual level, whether or not the researcher views their study as a replication per se, it is most likely building on a larger body of work in a given domain and should be considered as such. In these cases, full and precise (i.e., not dichotomous) reporting will help future studies integrate their findings into an existing body of research. Due consideration should also be given to the effects of studies in that body of research. In the case of Example Study 1, this might mean examining the results of studies comparing the effects of game-based instruction targeting other linguistic features or of studies targeting other speech acts, for example. As an added bonus, a closer look at similar studies and their effect sizes can also be used to conduct an a priori power analysis, which will help the researcher determine how many participants they need to reliably detect a truly statistical relationship or effect (see Plonsky, in press).

In addition to these points, the unreliability of *p* values is especially poignant in the context of replication research. That is, *p* values vary as a function of sample size such that a given effect size can be statistically significant for one *N* but not statistically significant for another, smaller *N* (see Plonsky, in press). It is perhaps for this reason more than any other that effect sizes such as *d* are imperative in the realm of replication, where researchers are interested in observing true patterns of effects across studies.

Finally, these same practices are also helpful from the synthetic or meta-analytic perspective, where it is often difficult to extract effect sizes from all candidate studies (see Larson-Hall & Plonsky, in press), which is fundamental to a study's contribution to cumulative knowledge in a given domain (e.g., Norris & Ortega, 2006; Oswald & Plonsky, 2010).

3. Beware of overgeneralizations

The findings from Example Study 1 are used to imply that game-based pragmatics instruction in general is superior to that of traditional instruction. There are multiple problems with this interpretation. First, pragmatic competence is of course much broader than requests, the target feature of the study. And in fact, there are likely many types of request structures that were not taught in the treatment that may or may not be as amenable to either of the two treatments. Second, the author is too liberal in his/her claims regarding the superiority of game-based learning over traditional instruction, particularly in light of the comments and re-analysis presented above. And third, we cannot assume that these findings would hold across all learner types, L1s, L2s, proficiency levels, ages, and so forth. Whether or not and to what extent they would hold are indeed empirical questions that merit replication research.

It is in this very function – indicating the generalizability of a study's findings across different contexts, learners, materials, targeted features, and so forth – that I find the greatest benefit of replication research (Gass & Valmori, in press; Plonsky, 2012). Here again, from both a synthetic and replicatory standpoint, the author of Example Study 1 should recognize the substantive limitation of the study and, rather than attempt to generalize beyond what was tested, encourage others to conduct (or conduct him/herself) targeted replications to assess the external validity of these findings across other demographic, instructional, and linguistic conditions. By doing so, s/he would also greatly facilitate research synthesis and/or meta-analysis once sufficient findings had accumulated within the domain.

Example Study 2

The primary goal of Example Study 2 was to gain a better understanding of reading comprehension generally as well as specifically with respect to its relationship to the use of one particular type of technological tool. Toward this end, the following research question was put forth: Is the use of text-based mobile apps related to reading comprehension in English as a second language (ESL)? The data for this study were collected using two instruments, which were administered to a sample of 90 students in a university-based intensive English program in the US: (a) a 100-item test of reading comprehension; and (b) a questionnaire which asked participants to rate on a scale of 1 (never) to 6 (very often) their use of text-based mobile apps in English and which collected demographic information such as their first language (L1) and age.

In order to address the research question, the author divided the sample into two groups based on their median scores for frequency of use of

text-based mobile apps (see results in Table 2). The Low and High groups' scores on the reading comprehension test were then compared using an independent samples *t*-test. The test indicated that the difference in reading comprehension in favor of the High use group was statistically significant: $t = 3.91$ ($p < 0.05$). The author's target journal also requires effect sizes, so along with the results of the *t*-test, s/he dutifully reported $d = 1.45$, labeling it as a 'large' difference.

Table 2: Example Study 2: Reading comprehension scores for low and high app use groups

	M (*SD*)
Low Use Group ($n = 45$)	32 (29)
High Use Group ($n = 45$)	70 (23)

Two additional analyses were carried out. The author was familiar with research showing that L1–L2 distance moderated reading comprehension (e.g., Jeon & Yamashita, 2014). To examine this relationship statistically, s/he ran a one-way ANOVA with reading comprehension scores again as the dependent variable and learner L1 as the independent or grouping variable (see descriptive statistics in Table 3). The difference between these mean scores, however, was not statistically significant ($F = 1.94$, $p > 0.05$), and this part of the analyses was therefore not included in the manuscript.

Table 3: Example Study 2: Reading comprehension test scores across L1 groups

	M (*SD*)
L1 Chinese ($n = 30$)	53 (29)
L1 Spanish ($n = 30$)	68 (24)
L1 Arabic ($n = 30$)	61 (33)

In addition, the author, who was also an instructor in the program where the study was conducted, had heard many of the older students – yet few of the younger ones – complain of difficulty reading in English. In order to check whether reading comprehension in English might be related to learner age, the researcher ran a correlation between reading comprehension and age. The correlation of $r = -0.56$ was statistically significant ($p < 0.05$), providing empirical confirmation of an inverse relationship between age and reading comprehension.

Like Example Study 1, Example Study 2 reads like a typical study, yet it is marred by conceptual and statistical flaws that greatly limit its potential to contribute to L2 theory, future research (e.g., replications), and practice.

1. Preserving variance and multivariate thinking (or, why *t*-tests and ANOVAs are overrated and overused)

If the last three or four decades of research have taught us anything, it is that L2 learning, teaching and use are complex, multivariate, and graded phenomena. I am all in favor of simple statistics, and I am aware that additional sophistication can reduce the interpretability of results. However, in the case of Example Study 2 and many others like it, the multiple analyses could have been carried out in a way that is more efficient, more eloquent, and that is more true to the data and the constructs/relationships in question (see below).

On a related note, our field's reliance on analyses that examine group differences has become so strong that we force our continuous data into the square peg of *t*-tests and ANOVAs. In Example Study 2, the author was interested first of all in the relationship between reading scores and use of text-based mobile apps. Both were measured as continuous variables, yet the author chopped the latter into two groups to allow for a comparison of a means-type analysis. By doing so the researcher traded precious variance for what might appear to be a clearer (yes/no) result compared to one that requires more interpretation but that is more informative: a correlation of $r = 0.62$ ($r^2 = 0.38$, an effect size indicating the percentage of shared variance between the two variables). (Note: In a replication study, conducting the same analyses as the original study will facilitate comparability. However, if the original study's analyses are mistaken as in Example Study 2, I would suggest conducting the analyses appropriately and requesting the original data set so you can re-run them on that study appropriately and then compare the two).

We see this all the time in applied linguistics: Researchers force continuously-scaled independent variables (e.g., motivation) into categorical ones (low, high), without any theoretical justification for doing so, just to make them more amendable to an ANOVA-type analysis. We insist on looking for differences rather than asking about the extent of the relationship between variables. I view this practice as related to the dichotomous view of data/analyses referred to above and embodied by NHST. I also suspect that researchers are more comfortable with analyses that compare means in part because they appear to give an unambiguous result: different or not different. Correlations, on the other hand, express the relationship between two variables as a matter of degree, which is often more informative but requires more interpretation on the part of the researcher.

The other two analyses in study 2 were largely innocuous. The second ANOVA involved a truly categorical independent variable (L1), and a comparison of means was warranted. And the third analysis, the correlation between reading comprehension and age, was also fine.

The problem here was the lack of consideration of these variables in a unified analytical approach. Conducting our analyses in a more holistic, multivariate way has at least three major benefits. First, it limits the number of analyses needed, thereby preserving experiment-wise power and leading to fewer false positives. Second, when we limit ourselves to conducting only bivariate analyses, we are unable to detect potential relationships between or among the independent variables. For instance, in Example Study 2, use of text-based apps and age were both correlated with reading proficiency but they may also be correlated with each other, which would indicate that there may be shared variance among all three variables. And third, related to this point, advancing theory in any area of empirical inquiry requires modeling the relationships between multiple variables simultaneously. In statistical terms, this step often takes the form of the amount of variance in the dependent variable that can be accounted for by the predictor variables, both individually and in unison. In the case of Example Study 2, the three unique analyses carried out (ANOVA, ANOVA, correlation) can be fruitfully consolidated into a single main analysis: multiple regression.

A hierarchical multiple regression with use of text-based apps in the first model (think: covariate in ANCOVA) and the remaining variables in the second model provides the following results. All together, the three independent variables (or 'predictors', as they are usually called in multiple regression) are able to explain a substantial 42% of the variance in reading comprehension scores. The vast majority of this amount (38%) was accounted for by the variable the author was most interested in, text-based app usage. (Note that, by no accident, this is the exact same value we found when we squared the correlation between text-based app use and reading proficiency scores.) The other two variables, L1 background and age, only explained an additional 4% of the variance in reading comprehension scores. This is not entirely surprising given the lack of a statistically significant difference in reading scores across different L1 groups. However, given the fairly strong negative correlation found between age and reading scores ($r = -0.56$), we might have expected this variable to explain a greater portion of the variance in our criterion variable. In other words, how do we make sense of the fact that the results of the multiple regression analysis appear to contradict those of the (bivariate) correlation analysis (age*reading)?

As I mentioned earlier, multiple regression is able to account for multiple relationships simultaneously. Although both age and use of text-based apps are associated with reading comprehension (see Table 4), the two predictors are also strongly correlated with each other (see Larson-Hall, 2010, Figure 7.2, or simply picture a three-way Venn diagram). The multiple regression analysis shows that the variance in reading scores that is explained by text-based app use is largely unique to that variable and not shared with age or L1. Multiple

regression is especially helpful in cases like this, which are quite common, in that it helps us to identify overlapping relationships between independent/predictor variables and the dependent variable that would otherwise (using bivariate analyses) be interpreted as unique.

Table 4: Example Study 2: Correlations (*r*) between continuous variables

Variable	Reading	Age
Text-based app use	0.62	–0.82
Reading	-	–0.56

2. Non-statistically significant results can be important and should not be omitted

There was one other serious error in Example Study 2 that should be addressed. Upon finding that the difference in reading comprehension scores for the three L1 groups was not statistically significant, the author buried the result rather than reporting it. This practice, that of suppressing results with $p > 0.05$, is as common in our field as it is unfortunate. If a researcher has a theoretical, practical, and/or empirical motivation for conducting a particular analysis, the results of that analysis should be reported regardless of the outcome (see Norris, Plonsky, Ross, & Schoonen, in press). The practice of reporting only statistically significant findings and omitting those with $p > 0.05$: (a) restricts the refinement of theory; (b) fails to inform our colleagues who might replicate our work and who might examine the same set of variables/relationships; and (c) produces an inflated or biased view of the relationship or effect in question at the meta-analytic level (see e.g., Oswald & Plonsky, 2010).

Imagine a scenario where theory in a given area predicts that X technological tool has a positive effect on L2 learning. However, the true population effect of X is actually $d = 0$; any deviations from that value are the result of sampling and measurement error. If 20 studies examine the effect of X, all using an alpha level of 0.05, we would expect just one, more or less, to find a statistically significant effect. And if only this one study reports the finding, the entirety of the available empirical evidence on the effect of X will confirm something that is not true. This scenario is an exaggeration – but only slight one – of what can and does happen (Plonsky, 2013). CALL researchers in particular, some of whom may have a vested interest in promoting the utility of technology for L2 learning and teaching, must be particularly vigilant and careful of such practices.

Conclusion

There is great potential in replication research as a tool to advance L2 theory and inform L2 pedagogy. This potential can only be reached, however, if

replication studies and the original studies that inform them are based on sound methodological and analytical practices. I have illustrated here a number of weaknesses in these areas that are both common and detrimental to progress in CALL and elsewhere in applied linguistics. Rather than dwell on these problems, however, I would like to look forward and to see the field move toward improved practice. Toward this end I will close with a concise summary of suggestions provided in the paper:

- The field's reliance on null hypothesis significance testing is doing far more harm than good. *p* values are unreliable and uninformative. Quantitative researchers should avoid them, focusing instead on descriptive statistics including effect sizes and confidence intervals.
- Related to the previous point, researchers should stop thinking in terms of dichotomous, direction-only results. This point has implications throughout the research process, from posing and wording research questions to choosing analyses to interpreting data.
- A lack of a statistically significant difference should not be interpreted as indicating no difference at all, particularly in the case of pretest data. Here again the graded approach embodied by effect sizes is of great use. Furthermore all results, whether or not they are statistically significant, should be reported along with their corresponding descriptive statistics (e.g., for means-based comparisons: mean, standard deviations, and confidence intervals).
- 'Replication' is best conceived of broadly such that all studies examining a similar set of variables are treated as replications, whether or not the authors refer to them as such. Doing so will help researchers think about and approach their studies from a synthetic point of view and as part of a larger empirical trajectory.
- Researchers must be careful not to overgeneralize results. The limitations of a study's external validity must be recognized and stated along with direction for carrying out replication studies examining the generalizability of a given set of findings.
- Models and studies of language learning and use are inherently multivariate. More of our analyses ought to be multivariate as well.
- There are very few instances when a continuously measured variable can be justifiably converted into a categorical one. Whenever possible, it is preferable to preserve the variance in continuous variables.
- Multiple regression provides a very useful, powerful, and informative alternative to a combination of *t*-tests/ANOVAs and correlations.

About the author

Luke Plonsky is Assistant Professor, Applied Linguistics at Northern Arizona University, Flagstaff, AZ, USA.

References

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

Gass, S., & Valmori, L. (in press). Replication in interaction and working memory research: Révész (2012) and Goo (2012). *Language Teaching*. http://dx.doi.org/10.1017/S0261444815000038

Godfroid, A., & Spino, L. (in press). Reconceptualizing reactivity research: Absence of evidence is not evidence of absence. *Language Learning*.

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

Larson-Hall, J. (2010). *A guide to doing statistics in second language research using SPSS*. New York: Routledge.

Larson-Hall, J., & Plonsky, L. (in press). Reporting and interpreting quantitative research findings: What gets reported and recommendations for the field. In J. M. Norris, S. Ross, & R. Schoonen (Eds), *Improving and extending quantitative reasoning in second language research*. Malden, MA: Wiley.

Norris, J. M. (in press). Statistical significance testing in second language research: Basic problems and suggestions for reform. In J. M. Norris, S. Ross, & R. Schoonen (Eds.), *Improving and extending quantitative reasoning in second language research*. Malden, MA: Wiley.

Norris, J. M., & Ortega, L. (2006). The value and practice of research synthesis for language learning and teaching. In J. M. Norris & L. Ortega (Eds), *Synthesizing research on language learning and teaching,* 3–50. Philadelphia, PA: John Benjamins.

Norris, J. M., Plonsky, L., Ross, S. J., & Schoonen, R. (in press). Guidelines for reporting quantitative methods and results in primary research. *Language Learning*.

Oswald, F. L., & Plonsky, L. (2010). Meta-analysis in second language research: Choices and challenges. *Annual Review of Applied Linguistics,* 30, 85–110. http://dx.doi.org/10.1017/S0267190510000115

Plonsky, L. (2012). Replication, meta-analysis, and generalizability. In G. Porte (Ed.), *Replication research in applied linguistics,* 116–132. New York: Cambridge University Press.

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. http://dx.doi.org/10.1017/S0272263113000399

Plonsky, L. (in press). 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*. New York: Routledge.

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

Porte, G. (2013). Who needs replication research? *CALICO Journal,* 30, 10–15. http://dx.doi.org/10.11139/cj.30.1.10-15

Smith, B., & Schulze, M. (2013). Thirty years of the CALICO Journal – replicate, replicate, replicate. *CALICO Journal,* 30 (1), i–iv. http://dx.doi.org/10.11139/cj.30.1.i-iv