Replication Research in Applied Linguistics: A Primer Including Statistical Considerations

Caroline Handley, Asia University

Abstract

This paper provides a brief overview of replication research. It explains some important reasons for conducting such research in the field of applied linguistics and, more specifically, foreign language teaching and learning, focusing on statistical issues in quantitative data analysis. An outline of how to conduct replication research is also provided, including advice for analysing results and writing reports. The paper concludes with a description of some of the publishing opportunities available to researchers in this field.

Background and Issues

What is Replication Research?

Replication research, as the name implies, is the exact, approximate, or conceptual repetition of prior published research (e.g., Porte, 2012). In approximate or close replication, the aim is to follow the original experimental[1] methods as far as possible, although within the field of applied linguistics this is often difficult due to factors such as subjective judgements and a lack of information in the original research paper (Mackey, 2012, pp. 25-27). Recently, online databases have been established to share original research materials, making replication more feasible (e.g., Open Science Framework (https://osf.io/); IRIS database (www.iris-database.org)).

In conceptual replication, the research design is altered in some systematic way (e.g., a reading task might be adapted into a listening task). However, in all types of replication the distinct feature is that the researcher(s) are focused on the original study, and elucidating its findings, rather than conducting novel research (Porte, 2012, pp. 5-7). Replication should not be confused with multiple studies being conducted to explore the same phenomenon in different contexts or with research carried out to extend previous findings (e.g., Markee, 2017). Although quantitative research and statistical analyses are the focus of this paper, it is not intended to imply that qualitative research should be exempted from replication. Markee (2017) has argued for replication of qualitative second language acquisition studies, although he suggests adapting the terminology to "comparative re-production research" to reflect the different aim of understanding rather than explaining pedagogical practices and interactions.

Why do Replication Research?

Publication of replication studies is rare, yet is being increasingly encouraged within the social sciences (e.g., McBee & Matthews, 2014). The initial impetus was probably due to the "crisis" in psychology (e.g., Kruglanski, 2001). Although replications are now being published within the field of psychology, they are neither abundant or free from controversy (Brandt et al., 2014). However, psychologists do appear to support publication of replication studies (Fuchs, Jenny, & Fiedler, 2012). A recent initiative to replicate 100 psychology experiments conducted

[1] Within applied linguistics a quasi-experimental design is often adopted (e.g., classroom research), rather than a "true" experimental design, however, for simplicity the term 'experiment' is used for both.

in 2008 (Open Science Collaboration, 2015) found the mean effect size of the replications was half that of the original studies and only 36% of replications produced statistically significant results, compared to 97% of the original studies. The methodological practices of psychology have guided research within applied linguistics (e.g., Gass, 1993), implying replication issues in psychology might also be shared. Some authors within applied linguistics are calling for more replication research; for example, Norris and Ortega (2000) state there is the need for "replication as a central undertaking of primary research in cumulative scientific endeavour" (p. 491).

As described in more detail below, replication studies, along with meta-analyses, should not be perceived as a critique of earlier studies, but ways to gain a deeper understanding of the phenomena of interest and enable greater confidence in the knowledge of any scientific field. Within the social sciences low power is a frequent issue, which increases the risk of Type I errors (rejecting the null hypothesis when it is true); replication reduces this risk (Hüffmeier, Mazei, & Schultze, 2016).

Replication does not only further collective knowledge, it can also provide rich educational potential for individual researchers. Replicating a previously published experiment can be a valuable experience for novice researchers, enabling the development of greater research expertise. Studies that have undergone peer review and been published should meet the standards of the field (Marsden, Mackey, & Plonsky, 2016, p. 17), providing good models to follow. Paul Meara established a distance PhD program in applied linguistics utilising replication, reasoning that following a tested model constrains the study to a realistic scope and aim, and enables attention to be focused on implementation and interpretation (Fitzpatrick, 2012).

Methodological issues in L2 research

Should research in applied linguistics follow the same trends as in psychology, described above, the accumulated knowledge of the field may not be very certain. An indication that this notion warrants concern has been provided by meta-analyses of L2 pedagogical studies and direct investigation of the methodology of peer-reviewed journal articles (reviewed below). Meta-analysis is related to replication in that both are conducted to deepen understanding of previous primary research. The difference is that meta-analysis has been developed to combine

the results of multiple studies to provide a better estimation of the size of the effect of interest (e.g., Cumming, 2012). Two meta-analyses of L2 research, which have concomitantly focused on methodological features of the target studies, are discussed here, in terms of the issues raised regarding statistical reporting. Next, recent methodological studies are summarised. These two secondary research sources illustrate the need for replication (and meta-analysis) to better understand the knowledge base of L2 research.

Norris and Ortega (2000) conducted a meta-analysis of research published in the 1980s and 1990s into the effectiveness of second language (L2) instruction to evaluate the inconsistent results in this area. They found that not only was incomplete reporting common, but also analysis in the majority of the papers was a comparison of group means, reporting purely whether a statistically significant difference was found. Only one out of 77 studies[2] that met their inclusion criteria reported effect sizes, despite this being recommended practice (e.g., American Psychological Association, 2009[3]). As the aim in meta-analysis is to discover a more accurate estimate of the effect size of the phenomenon of interest, Norris and Ortega (2000) calculated this from the means and standard deviations reported in each experiment. However, as many studies did not provide even this basic information, they often had to employ alternative statistical methods, based on *t* (*t* test) or *F* (ANOVA) values. Notwithstanding these efforts, they could only calculate effect value for 45 (58%) of the studies (p. 444). In addition, design issues sometimes made it difficult to define the contrasts between groups. Their research analysis showed that:

> both descriptive and inferential statistical techniques and data were reported inconsistently among the study reports, with the consequence that the analytic strategies, tools, and outcomes were often not sufficiently clear to enable readers to understand what was actually observed in the primary research. (p. 458)

These concerns about the quality of quantitative analysis in L2 research are compounded as although large effect sizes were found for explicit instruction, there was also great variation in

[2] 66 of the studies were published between 1990 and 1998, the final publication year included.
[3] The most recent edition available is cited here, but this recommendation has been in effect since the 4th edition, published in 1994.

the size of the effect between studies (as calculated by 95% confidence intervals (CI)). In other words, L2 instruction is effective, particularly explicit instruction, but replication is needed to ascertain how effective each type of instruction is, and in what contexts.

In a more recent meta-analysis of L2 strategy instruction, Plonsky (2011) found that of the 61 studies that met the criteria for inclusion only eight reported an effect size statistic and none reported Cohen's *d* value. He also found that only 25 of the studies reported estimates of measurement reliability, but that those studies produced larger effect sizes. Studies which included a pre-test also yielded greater effect sizes than those that did not. In other words, it is possible that the design and analysis of an experiment could be a factor in the results obtained. Plonsky (2011) recommended that L2 strategy instruction research would benefit from better statistical reporting and more transparent methodology in published papers, to enable more exact replication, comparison and interpretation. Elsewhere he has also highlighted issues with the increasing reliance on ANOVA tests within the field of second language acquisition, which has led to researchers turning continuous variables into discrete categories and ignoring interrelationships between independent variables, thus losing information (Plonsky & Oswald, 2017).

In a direct investigation of methodological and analytical issues in language learning research, comparing 606 articles published in two journals in the 1990s and 2000s, Plonsky (2014) found that design factors, such as pre- and delayed post-test, and the reporting of descriptive and inferential statistics are improving. However, many issues remain in the field, such as an over-reliance on NHST and low power. Similarly, in an analysis of three decades of interaction research in second language acquisition (174 quantitative studies published between 1981 and 2009), Plonsky and Gass (2013) found that although statistical tests were increasingly used, there was "little evidence for an increase in statistical sophistication" (p. 344). However, they also noted indications of improved reporting of analyses, such as an increase over time in the studies reporting standard deviation and effect size, despite the persistence of incomplete reporting.

A further concern in L2 research is studies often involve a small number of participants. Although Norris and Ortega (2000) found evidence of publication bias, in that almost all the published studies in their meta-analysis reported at least one statistically significant result, small sample sizes often do not have enough statistical power to obtain a significant result, even when

there actually is an important effect in the target population (Cohen, 1992). For example, Cohen (1992) calculated that to detect a medium effect size between two independent groups from a population at the alpha (α) probability level of .05 (the standard *p* value accepted as statistically significant in most research in applied linguistics) each group would need $N = 64$. In other words, if a study is designed to detect a difference in test scores between two groups of students (control and treatment) and a medium effect size is anticipated based on previous research, then to reliably reject a false null hypothesis a minimum of 128 participants is required. This contrasts starkly with the average group size of approximately 20 participants in L2 research (Plonsky, 2013; Plonsky & Gass, 2011). Therefore, it is unsurprising that methodological reviews of quantitative L2 research have concluded that studies are typically underpowered, undermining the reliability of the results obtained and conclusions derived (e.g., Plonsky, 2013; Plonsky & Gass, 2011).

Given the typically small number of participants in any one study, replication and meta-analysis provide the means to combine population samples and power, reducing the possibility of not rejecting the null hypothesis when it is false (Type II error; e.g., Cumming, 2012). Replication can also reduce random variation, which is particularly problematic with small sample sizes, reducing the possibility of rejecting the null hypothesis when it is true (Type I error; e.g., Cumming, 2008). In both instances replication decreases the chance that the difference in means between two or more groups (the *p* value obtained) is unduly influenced by outliers (e.g., Lindstromberg, 2016).

Finally, L2 research published within Japan is consistent with international-level research regarding methodological issues. In a review of the articles published throughout the first 30 years of *JALT Journal*, Stapleton and Collett (2010) found that of papers published between 2001 and 2008 in which statistical significance was reported only 14% also reported effect size.

Conducting Replication Research

Selecting a study

The starting point for a replication study should be the identification of a robust reason why an experiment (or series of experiments) should be repeated. Typically, this means that the original results should be theoretically important and relevant. In addition, the study may not yet

have been extended across diverse contexts, may show evidence of a design flaw, or might contain conflicting results, either within the paper or between it and other published research (e.g., Mackey, 2012). Replication could also be warranted if the original research was conducted with a small number of participants and it is possible to recruit a much larger number, increasing the statistical power (e.g., Cohen, 1992).

Recently transparency in research is being encouraged and supported within the field of applied linguistics, with organisations and websites being created to promote and enable replication. Possibly the most relevant of these at present is the IRIS database (www.iris-database.org), a repository of materials and tools that have been used in L2 research. It is also now possible to pre-register your replication on the Open Science Framework (https://osf.io/).

Doing the research

When carrying out a replication, the most important thing to remember is to remain focused on the original study. Its methodology must be scrutinised and followed as closely as possible, and any differences should be reported and explained. Of course, the actual participants will be different, and the implications of this for the ability of your research to replicate the original study should be discussed in your paper.

It is important to ensure the number of participants in your study (or sample size) is large enough to provide the statistical power necessary to confirm the original effect (e.g., Cohen, 1992; Maxwell, Kelley, & Rausch, 2008), otherwise failure to replicate could be a design flaw in your experiment. As explained above, power can be understood as the ability to reject the null hypothesis when it is false. In this regard, if the analysis involves Null Hypothesis Significance Testing (NHST) it is vital to understand that statistical significance is directly related to sample size (participant numbers) as well as effect size (e.g., Cohen, 1992; Cumming, 2012).

In some instances, it may be appropriate to extend the original research to achieve better design principles. For example, if the original study reported a treatment effect at post-test, but did not include a pre-test or delayed post-test (an issue in some L2 studies, e.g., Plonsky, 2014), these additional steps would enable further, more reliable, interpretation of treatment effects.

Analysing the data

This section deals with what should be done to analyse your data, space constraints do not permit discussion of how to perform the analyses. However, there is an abundance of textbooks on research methods in applied linguistics that can be consulted. It should be noted that some of the statistical tests discussed below cannot be performed using the SPSS package (IBM, 2013). However, all can be carried out using the free software R (R Core Team, 2016) or R Studio (RStudio Team, 2015) and the various packages that have been created for these platforms. R and R Studio provide more complete analytical tools than SPSS at the expense of a steeper learning curve. Collett (2017) provides a short introduction to statistics with R, although for those wishing to use this software a more complete, book-length guide would be required. For applied linguistics researchers new to R an appropriate introductory text would be Levshina (2015) or Larson-Hall (2015).

The first steps in analyzing your data should be to replicate the original analyses. This involves preparing the same descriptions of the data as provided in the original paper; any tables and figures, as well as descriptive and inferential statistics. However, additional analyses may also be required. For example, in applied linguistics research papers NHST is common (e.g., Norris and Ortega, 2000; Plonsky, 2013), and often only statistically significant (and non-significant) results are reported. However, failure to replicate a statistically significant result does not automatically entail the contradiction of the original study (e.g. Kline, 2004). If the results follow the same trend, the replication supports the original, and the non-significance should be interpreted as the result of random sampling variation or insufficient power (see Tversky and Kahneman, 1971). One of the reasons for not relying solely on NHST is that *p* values are unreliable, as they are likely to vary greatly over replications (e.g. Cohen, 1994; Cumming, 2008; 2012).

Therefore, in reporting a replication, estimation techniques should also be used, involving the calculation of a point estimate and 95%[4] confidence intervals (CI), as well as effect sizes (e.g., American Psychological Association, 1994; Cumming, 2008; 2012). If the original study compares two or more groups (i.e., control and treatment group or pre- and post-tests within a group), provided the group means and standard deviations are reported it is relatively simple to

[4] This is the most commonly used interval; there may be justification for using 90% or 99% intervals, depending on your study.

calculate the effect size of the original study, if this was not reported. This provides a much better comparison between studies than *p* values, enabling interpretation of the extent of the impact of the interventions, instead of a dichotomous interpretation of significance or not (e.g., Cumming, 2012). Recommendations for L2 pedagogy should be based on how much of an effect there is, not whether the effect is statistically unlikely to be due to chance variability (in other words, the results are improbable if the null hypothesis were true).

An additional suggestion for replications that are consistent with the original results is to combine the findings of the original and replication into a mini meta-analysis (e.g., Cumming, 2012). This combines the power of both studies and, as consistency entails overlap in confidence intervals (CI), will also decrease the length of the CIs, reducing the range of the estimated population mean around the point estimate of the sample mean (i.e., increasing the level of precision regarding the actual value of the population mean). This increased accuracy means that the effect can be reported with greater security.

Writing the replication paper

A basic overview on writing a replication paper is given here, with the aim of indicating the most important considerations. For more detailed information the reader is referred to Brown (2012), on which the following recommendations are based. Replication papers should follow the standard style in the field, including an introduction, method, results, discussion, and conclusion. A summary of the original study should be included, highlighting the most pertinent aspects for the replication. The rationale for the replication should be explained and reference made to any relevant literature published since the year prior to the original study. It is vital that the methodology is presented as transparently as possible, so readers can scrutinise the effect this may have had on your ability to reproduce similar findings to those of the original experiment. The same applies to the analyses, as described above. The discussion should compare your findings to the original, whereas the conclusion should include the limitations and implications of the replication. In addition, in keeping with the issues highlighted with reporting results above, the discussion should focus on interpreting the substantive significance of your results (i.e., the inferences from the effect size for L2 teaching or learning, with reference to prior literature), rather than the statistical significance (Kline, 2004).

It is important to remember that the rationale for replication is that no single study is conclusive, so in discussing your results they may be said to support, limit, or contradict the original, but cannot be claimed to prove or disprove it, nor any underlying theory (e.g., Brandt et al., 2014). Similarly, although in conducting the replication areas may be found in which the original could be critiqued, it should never be referred to disrespectfully (Brown, 2012).

Publishing Replication Research

With the increasing recognition of the importance of replication research the opportunities for publishing such papers are also expanding. There are now a few international peer-reviewed SLA journals that include a special section for the publication of replication research, such as *Language Teaching* and *Studies in Second Language Acquisition*. Moreover, these sections are not typically being filled each issue (Marsden, Mackey, & Plonsky, 2016). In addition, prizes are being introduced to incentivise replication research, such as the IRIS Replication Award, which is an annual award for any second language research utilising materials maintained in its database.

Although the promotion of replication research may be only slowly gaining momentum, within the broader framework of the social sciences there have recently been some significant developments, such as the world's first grants fund specifically for replication studies (Baker, 2016). In 2017 the Social Sciences Replication Project (http://www.socialsciencesreplicationproject.com/) was launched, with the aim of replicating 21 experiments published since 2010, including three studies involving learning and/or testing. In addition, Elsevier, a major global academic publishing house, produced four virtual special issues on replication to facilitate such research (De Weerd-Wilson & Gunn, 2017). Despite the common perception among social science journal editors that replication is uncreative (Easley, Madden, & Gray, 2013), it offers great potential for gaining an international-level, peer-reviewed research publication.

Conclusion

This brief introduction to replication is intended to encourage such research and summarise the main considerations. References have been provided throughout to useful sources of more detailed information, and in particular the reader is recommended to refer to *Replication Research in Applied Linguistics*, edited by Graeme Porte (2012). As Cohen (1994) stated, "For generalization, psychologists must finally rely, as has been done in all the older sciences, on replication" (p. 997). Replication studies should also be valued within applied linguistics, as they contribute towards consolidating knowledge and improving the validity and transparency of research and its interpretation. Only through replication will researchers be able to reduce the noise of sampling variation and uncover the actual patterns that they aim to understand.

References

American Psychological Association. (2009). *Publication manual of the American Psychological Association* (6th Ed.). Washington, DC: American Psychological Association.

Baker, M. (2016). Dutch agency launches first grants programme dedicated to replication. *Nature*. Retrieved from www.nature.com/news/dutch-agency-launches-first-grants-programme-dedicated-to-replication-1.20287

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., Grange, J. A., Perugini, M., Spies, J. R., & van 't Veer, A. (2014). *The replication recipe: What makes for a convincing replication? Journal of Experimental Social Psychology, 50*, 217-224.

Brown, J. D. (2012). Writing up a replication report. In G. Porte, *Replication research in applied linguistics* (pp. 173-197). Cambridge, England: Cambridge University Press.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-9.

Cohen, J. (1994). The Earth is round *($p < .05$)*. *American Psychologist, 49*(12), 997-1003.

Collett, P. (2017). Moving towards better quantitative data analysis in FLL research. *JALT postconference publication: Transformation in language education*. Retrieved from http://jalt-publications.org/node/4/issues/2017-08_2016.1

Cumming, G. (2008). Replication and *p* intervals: *p* values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science, 3*(4), 286-300.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis. Routledge* (Kindle version). Retrieved from Amazon.co.uk.

De Weerd-Wilson, D., & Gunn, W. (2017). How Elsevier is breaking down barriers to reproducibility. Retrieved from https://www.elsevier.com/connect/how-elsevier-is-breaking-down-barriers-to-reproducibility

Easley, R. W., Madden, C. S., & Gray, V. (2013). A tale of two cultures: Revisiting journal editors´ views of replication research. *Journal of Business Research, 66*: 1457–1459.

Fitzpatrick, T. (2012). Conducting replication studies: Lessons from a graduate program. In G. Porte, *Replication research in applied linguistics* (pp. 151-170). Cambridge, England: Cambridge University Press.

Fuchs, H., Jenny, M., & Fiedler, S. (2012). Psychologists are open to change, yet wary of rules. *Perspectives on Psychological Science, 7,* 639-642.

Gass, S. (1993). Editorial: Second language acquisition: Cross-disciplinary perspectives. *Second Language Research, 9*, 95–98.

Hüffmeier, J., Mazei, J., & Schultze, T. (2016). Reconceptualizing replication as a sequence of different studies: A replication typology. *Journal of Experimental Social Psychology, 66*, 81–92.

IBM Corp. (2013). *IBM SPSS statistics for windows, version 22.0* [computer software]. Armonk, NY: IBM Corp.

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association.

Kruglanski, A. W. (2001). That "vision thing": The state of theory in social and personality psychology at the edge of the new millennium. *Journal of Personality and Social Psychology*, *80*, 871-875.

Larson-Hall, J. (2015). *A guide to doing statistics in second language research using R*. Routledge. Retrieved from http://cw.routledge.com/textbooks/9780805861853/R/full-version.pdf

Levshina, N. (2015). *How to do linguistics with R: Data exploration and statistical analysis*. Amsterdam, Netherlands: John Benjamins.

Lindstromberg, S. (2016). Inferential statistics in *Language Teaching Research*: A review and ways forward. *Language Teaching Research, 20*(6), 741-768.

Mackey, A. (2012). Why (or why not), when and how to replicate research. In G. Porte, *Replication research in applied linguistics* (pp. 21-46). Cambridge, England: Cambridge University Press.

Markee, N. (2017). Are replication studies possible in qualitative second/foreign language classroom research? A call for comparative re-production research. *Language Teaching, 50*(3), 367-383.

Marsden, E., Mackey A., & Plonsky, L. (2016). The IRIS Repository: Advancing research practice and methodology. In A. Mackey & E. Marsden (Eds.), *Advancing methodology and practice: The IRIS Repository of Instruments for Research into Second Languages* (pp. 1-21). New York, NY: Routledge.

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology, 59*, 537–563.

McBee, M. T., & Matthews, M. S. (2014). Welcoming quality in non-significance and replication work, but moving beyond the *p*-value: Announcing new editorial policies for quantitative research in *JOAA. Journal of Advanced Academics, 25*(2), 73-87.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251). Retrieved from http://science.sciencemag.org/content/349/6251/aac4716.full?ijkey=1xgFoCnpLswpk&keytype=ref&siteid=sci

Porte, G. (2012). *Replication research in applied linguistics*. Cambridge, England: Cambridge University Press.

Plonsky, L. (2011). The effectiveness of second language strategy instruction: A meta-analysis. *Language Learning, 61*, 993–1038.

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

Plonsky, L. (2014). Study quality in quantitative L2 research (1990–2010): A methodological synthesis and call for reform. *The Modern Language Journal, 98*(1), 450-470.

Plonsky, L., & Gass, S. (2011). Quantitative research methods, study quality, and outcomes: The case of interaction research. *Language Learning, 61*(2), 325-366.

Plonsky, L., & Oswald, F. L. (2017). Multiple regression as a flexible alternative to ANOVA in L2 research. *Studies in Second Language Acquisition*, *39*(3), 579-592.

R Core Team. (2016). *R: A language and environment for statistical computing* [computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

RStudio Team. (2015). *RStudio: Integrated development for R* [computer software]. Boston, MA: RStudio, Inc. Retrieved from http://www.rstudio.com/

Stapleton, P., & Collett, P. (2010). JALT Journal turns 30: A retrospective look at the first three decades. *JALT Journal, 32*(1), 75-90.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin, 76*(2), 105-110.