Preregistration: Replication Studies in Linguistic Journals

Kristina Kobrock[1] & Timo B. Roettger[2]

[1] University of Osnabrück

[2] University of Oslo

Author Note

The authors made the following contributions. Kristina Kobrock: Conceptualization, Writing - original draft preparation, Writing - review & rditing, Data curation, Methodology, Formal analysis; Timo B. Roettger: Writing - review & editing, Methodology, Formal analysis, Supervision.

Preregistration: Replication Studies in Linguistic Journals

Introduction

Coordinated efforts to replicate published findings have uncovered surprisingly low rates of successful replications across the psychological sciences (Open Science Collaboration, 2015), economics (Camerer et al., 2016), and social sciences (Camerer et al., 2018). Experimental linguistics shares research practices that have been shown to decrease the replicability of findings. Thus, there are raising concerns about a similarly low number of replication studies conducted and published in this field (e.g. Marsden, Morgan-Short, Thompson, & Abugaber, 2018; Roettger & Baer-Henney, 2019). A number of failed replication attempts in various subfields of linguistics indicate that these concerns warrant attention (e.g. in language comprehension: Papesh, 2015; predictive processing: Nieuwland et al., 2018; among others: e.g. Chen, 2007; Stack, James, & Watson, 2018; Westbury, 2018). One driving factor for this phenomenon is an asymmetric incentive system that rewards novel confirmatory findings more than direct replications and null results. This leads to an abundance of positive findings in the absence of possible conflicting negative evidence. In order to thoroughly understand and be able to address this problem, it is important to assess the number of replication attempts and their contributing factors. Other fields such as psychology (Makel, Plucker, & Hegarty, 2012), eudcation science (Makel & Plucker, 2014), and special education research (Makel et al., 2016) have assessed the amount of direct replications in their respective field and report alarmingly low replication rates (0.13% - 1.07%).

In order to evaluate the replication rate in linguistics, the present study aims at assessing the frequency and typology of replication studies that are published in a representative sample of linguistic journals. The study consists of two parts: First, we will assess the frequency of self-reported replication attempts across 100 linguistic journals and relate the replication rate to factors related to journal policy, impact factor and publication type. Second, we will assess the type of replication studies (direct, partial, conceptual) published in a subset of 20 journals and relate their frequency to factors like the year of publication, and the citation and publication year of the original study.

Overview Analysis: Rate of Replication Mention

The key dependent variable of the first part of this study is the rate of replication mention for journals relevant to the field of experimental linguistics. In order to determine these rates for the individual journals, we will draw on a method introduced by Makel et al. (2012). We will use the search engine "Web of Science" (https://webofknowledge.com) for journal articles that contain the search term "replicat*" in title, abstract or keywords and compute the rate of replication mention.

Research Questions. We intend to answer the following research questions: How many replication studies have been published in journals representative for experimental linguistic research? How did the rate change over time and how does it relate to journal policy, impact factor, and publication type?

Sample. To obtain a representative sample of journals relevant for the field of experimental linguistics, we follow the procedure presented here: First, using the Web of Science advanced search on the "Web of Science Core Collection" database, we filter for the category "Linguistics" (WC=(Linguistics)) that lists the articles of every journal covered by Web of Science that was assigned to the subject category of Linguistics (see a list of categories here). All English language articles from the full available range of complete years (1945-2020) are taken into account. From the resulting set (159002) only those articles are selected which contain the search term "experiment*" in their title, abstract or keywords using TS="experiment*" in order to filter for experimental linguistic studies. This search results in 11093 articles. The relevant journals are selected based on the obtained article counts. From all journals that include at least one experimental linguistic study according to our criterium (418), journals with less than 100 published articles are excluded, yielding 259 remaining journals. Because we are interested in journals with a high proportion of experimental studies, we calculate the ratio of studies that contained the search term “experiment*” by the total amount of articles per journal and sort the results in descending order. Our sample constitutes the first 100 journals of that list. Counts have been obtained on the 21st February 2021. See here for more details: https://osf.io/q2e9k/

Procedure. The total number of articles containing the search term “replicat*” in title, abstract or keywords is obtained via Web of Science search for the 100 sampled journals. This number and and the total number of experimental studies described above serve as a baseline for calculating the rates of replication mention, following the method used by Makel et al. (2012). The rates of replication mention are calculated by dividing the number of articles containing the term “replicat*” by the number of articles that contain the term “experiment*” for each journal, respectively.

In order to relate the rate of replication mention to journal policies, we further examine the journals’ submission guidelines adopting the procedure used by Martin and Clarke (2017). They grouped psychology journals into four classes determined by what was stated in the “instructions to authors” and “aims and scope” sections on the websites of the respective journals: (1) Journals which stated that they accepted replications; (2) Journals which did not state they accepted replications but did not discourage replications either; (3) Journals which implicitly discouraged replications through the use of emphasis on the scientific originality of submissions, (4) Journals which actively discouraged replications by stating explicitly that they did not accept replications for publication (Martin & Clarke, 2017, p. 3).

Journal impact factors are extracted via Journal Citation Reports (https://jcr.clarivate.com). The 2019 journal impact factors are calculated by dividing the citations in 2019 to items published in 2017 and 2018 by the total number of citable items in 2017 and 2018. The open access category of journals is assessed via Web of Science. We distinguish between three categories: journals which are listed on the Directory of Open

Access Journals (DOAJ) (“DOAJ gold”), journals with some articles being published as open access articles (“partial”) and journals with no openly accessible articles (“no”).

Data Analysis. We will use Bayesian parameter estimation based on generalized linear regression models with a binomial link function in order to estimate the rate of replication mention relative to the following predictors: journal impact factors (continuous), open access (binary: open access journal or not), and replication policies (binary: either explicitly encourage or not). The model will be fitted to the proportion of replication mentions per journal using the R package brms (Bürkner, 2016). We will use (weakly) informative normal priors centered on -2.1973 (corresponding to a 10% base rate, sd = 2.5) for the intercept since we expect very low base rates (e.g. Makel et al., 2012). We will use weakly informative Cauchy priors centered on zero (scale = 2.5) for all population-level regression coefficients. These priors are what is referred to as regularizing (Gelman, Jakulin, Pittau, Su, & others, 2008), i.e. our prior assumption is agnostic as to whether the predictors affect the dependent variable, thus making our model conservative with regards to the predictors under investigation. Four sampling chains with 2000 iterations each will be run for each model, with a warm-up period of 1000 iterations. For relevant predictor levels and contrasts between predictor levels, we will report the posterior probability for the rate of replication mention. We summarize these distributions by reporting the posterior mean and the 95% credible intervals (calculated as the highest posterior density interval).

Detailed Analysis: Types and Contributing Factors

Methods

The second part of the analysis aims at obtaining a better understanding of the underlying mechanisms of replication attempts published in the field of experimental linguistics. Because the term “replication” is commonly used in ambiguous ways, the articles that contain the search term “replicat*” require further analysis to determine whether the articles in question indeed report a replication study or use the term in a different way.

Research Questions. We are interested in which kinds of replication studies are published and which factors contribute to their publication. We aim at investigating what types of replication studies are prevalent in the field. We are further interested in the relationship of direct replications and whether the paper was published as open access or not, the number of citations of the initial study and the years between publication of the initial study and the replication attempt.

Sample. From the superset of 100 journals obtained above, the first 20 journals (i.e. those journals with the highest proportion of experimental studies) are selected for a more detailed analysis. We exclude those journals for which less than 2 hits (TS=(replicat*)) can be obtained. This method yields a total number of 274 articles (see here for a list of article counts per journal: https://osf.io/f3yp8/). Because of the skewed distribution of our sample (114 hits for Journal of Memory and Language, and less than 40 for all other journals), we randomly select 50 out of the 114 articles for the Journal of Memory and Language to achieve a more balanced distribution of papers across journals by drawing from a uniform distribution in R without replacement (see here for details).

Procedure. The sampling procedure above results in 210 possible self-labeled replication studies. In a first step, we will identify whether the article indeed presents a replication study or not. By reading title and abstract of the paper a first intuition of what the article is about can be obtained. The main task is to assess whether the authors claim that their underlying aim was to replicate or reproduce findings or methods of another study (henceforth initial study). A search for occurrences of the search term "replicat" in the text and an assessment of the paragraph before the Methods section as well as the first paragraph of the Discussion section (following the procedure specified by Makel et al. (2016)) helps to further identify the intention communicated by the authors regarding their use of the term replication. If the authors communicate that (one of) the underlying aim(s) was to replicate an original study, this article can be treated as a replication. It then qualifies for further analysis after the coding scheme that can be viewed here: https://osf.io/ct2xj/.

Assuming that the authors did not make any drastic changes to the initial study *without* reporting them, number and type of changes made by the replication study are extracted. The replication studies are classified according to three types: direct replication (0 changes), partial replication (1 change) and conceptual replication (2 or more changes), following Marsden et al. (2018). We note the nature of the change as one of the following categories (yes/no): experimental paradigm, sample, materials/experimental set-up, dependent variable, independent variable, and control. Coding the articles also involves examining the factors open access of that article (yes/no), years between initial study and replication attempt, author overlap of initial study and replication attempt (yes/no), citation counts of both studies and the language under investigation. The information on whether the article is open access as well as citation counts and years of publication for both studies can be obtained from Web of Science. An author overlap is attested when one of the authors is a (co-)author on both articles.

Data Analysis. We will use Bayesian parameter estimation based on generalized linear regression models with a logit link function in order to estimate the rate of direct replications relative to the following predictors: year of publication (continuous), open access (binary: open access article or not), time lag between publication of initial study and replication attempt (continuous) and number of citations of initial study (continuous). The model will be fitted to whether the replication mention was a direct replication or not using the R package brms (Bürkner, 2016). The model further includes random intercepts for individual journals to account for varying rates of direct replications across journals. We will use (weakly) informative normal priors centered on -2.1973 (corresponding to 10% base rate, sd = 2.5) for the intercept (since we expect very low base rates) and weakly informative Cauchy priors centered on zero (scale = 2.5) for all population-level regression coefficients. Four sampling chains with 2000 iterations each will be run for each model, with a warm-up period of 1000 iterations. For relevant predictor levels and contrasts between predictor levels, we will report the posterior probability for the rate of direct replication. We summarize these distributions by reporting the posterior mean and the 95% credible intervals (calculated as the highest posterior density interval).

References

Bürkner, P.-C. (2016). Brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., … Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280), 1433–1436. https://doi.org/10.1126/science.aaf0918

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johanesson, M., … Wu, H. (2018). Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature*, *2*, 637–644. https://doi.org/10.1038/s41562-018-0399-z

Chen, J.-Y. (2007). Do Chinese and English speakers think about time differently? Failure of replicating Boroditsky (2001). *Cognition*, *104*(2), 427–436.

Gelman, A., Jakulin, A., Pittau, M. G., Su, Y.-S., & others. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, *2*(4), 1360–1383.

Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty: Replication in the education sciences. *Educational Researcher*, *43*(6), 304–316.

Makel, M. C., Plucker, J. A., Freeman, J., Lombardi, A., Simonsen, B., & Coyne, M. (2016). Replication of special education research: Necessary but far too rare. *Remedial and Special Education*, *37*(4), 205–212.

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science*, *7*(6), 537–542.

Marsden, E., Morgan-Short, K., Thompson, S., & Abugaber, D. (2018). Replication in Second Language Research: Narrative and Systematic Reviews and

Recommendations for the Field. *Language Learning*, *68*(2), 321–391. https://doi.org/10/gc3h3b

Martin, G. N., & Clarke, R. M. (2017). Are psychology journals anti-replication? A snapshot of editorial practices. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.00523

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., … Huettig, F. (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *eLife*, *7*, e33468. https://doi.org/10.7554/eLife.33468.001

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251). https://doi.org/10.1126/science.aac4716

Papesh, M. H. (2015). Just out of reach: On the reliability of the action-sentence compatibility effect. *Journal of Experimental Psychology: General*, *144*(6), e116–e141. https://doi.org/10.1037/xge0000125

Roettger, T. B., & Baer-Henney, D. (2019). Toward a replication culture: Speech production research in the classroom. *Phonological Data and Analysis*, *1*(4), 1–23.

Stack, C. M. H., James, A. N., & Watson, D. G. (2018). A failure to replicate rapid syntactic adaptation in comprehension. *Memory & Cognition*, *46*(6), 864–877. https://doi.org/10.3758/s13421-018-0808-6

Westbury, C. (2018). Implicit sound symbolism effect in lexical access, revisited: A requiem for the interference task paradigm. *Journal of Articles in Support of the Null Hypothesis*, *15*(1), 1–12.