Assessing the replication landscape in experimental linguistics

Word count: 9520

Kristina Kobrock
University of Osnabrück
kristina.kobrock@uos.de

Timo B. Roettger
Universitetet i Oslo
timo.b.roettger@gmail.com

Abstract Replications are an integral part of cumulative experimental science. Yet many scientific disciplines do not replicate much because novel confirmatory findings are valued over direct replications. To provide a systematic assessment of the replication landscape in experimental linguistics, the present study estimated replication rates for over 50.000 articles across 98 journals. We used automatic string matching using the Web of Science combined with in-depth manual inspections of 274 papers. The median rate of mentioning the search string "replicat*" was as low as 1.7%. Subsequent manual analyses of articles containing the search string revealed that only 4% of these contained a direct replication, i.e. a study that aims to arrive at the same scientific conclusions as an initial study by using exactly the same methodology. Less than half of these direct replications were performed by independent researchers. Thus our data suggest that only 1 in 1250 experimental linguistic articles contains an independent direct replication. We conclude that, similar to neighboring disciplines, experimental linguistics replicates very little, a state of affairs that should be reflected upon.

Keywords: replication, meta-research, journal impact factor, publishing guidelines

1 Introduction

Understanding the inner workings of human language and its cognitive underpinnings has been increasingly shaped by experimental data. With a field that builds its theories on a rapidly growing body of experimental evidence, it is of critical importance to evaluate and substantiate existing findings in the literature because evidence provided by a single study is limited (e.g., Amrhein et al. 2019). Scientists are trained to ensure the reliability and generalizability of scientific findings by conducting direct replication studies, i.e. studies that aim to arrive at the same scientific conclusions as an initial study by collecting new data and completing new analyses but using the same methodology (for a comprehensive overview of different terminological uses see Barba 2018).

Replications are an integral part of cumulative experimental science (e.g., Campbell 1969; Rosenthal 1990; Zwaan et al. 2018). Yet many scientific disciplines do not replicate a lot. Researchers from diverse fields such as psychology (Makel et al. 2012), educational science (Makel & Plucker 2014), ecology (Kelly 2019), criminology (McNeeley & Warner 2015), and economics (Mueller-Langer et al. 2019) report on very low numbers of published replications, ranging from 0.02% in ecology to 2% in criminology.

One reason for the observed lack of replication studies is the asymmetric incentive system in academia that rewards novel confirmatory findings over direct replications and null results: Replication studies are not very popular because the necessary time and resource investments are not appropriately rewarded (e.g., Koole & Lakens 2012; Nosek et al. 2012). Both successful replications (Madden et al. 1995) and repeated failures to replicate (e.g., Doyen et al. 2012) are rarely published. Even if they are, replications usually appear in less prestigious outlets than the original findings. These dynamics lead to an abundance

of positive findings in the absence of possible conflicting negative evidence (see also Fanelli 2010) and the widely held view that replications lack prestige, originality, or excitement (e.g., Lindsay & Ehrenberg 1993).

This perceived lack of prestige additionally comes with the sentiment that direct replications are unnecessary and / or uninformative. This sentiment expresses itself in two parts: Direct replications are claimed to be theoretically uninformative and conceptual replications are claimed to be sufficient to assess the robustness of a field's empirical foundation (e.g., Stroebe & Strack 2014; Crandall & Sherman 2016). However, both of these assumptions are problematic (e.g., Zwaan et al. 2018): A repeated demonstration that an effect can not be replicated is an important contribution to the field and aids it in calibrating researchers' (un)certainty in the existence of a phenomenon. Moreover, failed direct replications might uncover important moderators and boundary conditions that explain the discrepancy between an original study and a replication. Conducting a direct replication operates on the assumption that all critical elements to reproduce the original effect are understood. If the replication fails, that strong assumption has to be questioned, thus relevant auxiliary hypotheses must be reconsidered, which in turn might weaken the theory. On the other hand, successful direct replications add important data to the discourse, allowing for more precise estimation of theoretically relevant parameters, and thus help to strengthen the derivation chain between theory and predictions (Meehl 1990).

Given these arguments, we consider direct replications theoretically informative and a worthwhile endeavor. Conceptual replications on the other hand, i.e. replication attempts that have changed multiple critical design properties of the original study, are often upheld as being more valuable than direct replications because they are assumed to simultaneously address concerns about reliability of an original claim and they are able to extend the original findings.

Conceptual replications are often considered sufficient for a field to move forward under the stipulation that repeated successful conceptual replications will occur only then when the prior research identified a true effect. However, there is increasing evidence that this strong assumption is empirically not supported. Without replicating individual studies, biases caused by questionable research practices (John et al. 2012), small sample size (Button et al. 2013) and publication bias (Fanelli 2012) can lead to a set of studies that appear to form a coherent empirical foundation of an underlying theory, even if the underlying empirical claims cannot be replicated: There are now a number of widely studied theories and effects that have been supported by dozens, if not hundreds of conceptual replications, but appear to crumble in light of meta-analyses or systematic direct replication attempts (e.g., Shanks et al. 2015; Wagenmakers et al. 2016). Moreover, conceptual replications can introduce interpretational ambiguity. A failed conceptual replication can never be considered evidence against the original claim. It is always possible to attribute a failed conceptual replication to the methodological changes that were made (e.g., Pashler & Harris 2012).

In sum, direct replications are an under-appreciated tool to evaluate and cement the empirical and theoretical foundation of a field and must be considered an important complementary tool to conceptual replications.

The observed lack of replication studies across disciplines threatens the very fabric of cumulative progress in experimental science because experimental results are often taken for granted without them ever being replicated which leads to a related problem: If we don't try, we won't fail. The recent past has shown that if we try, we fail more often than we would like to: Coordinated efforts to replicate published findings have uncovered

alarmingly low rates of successful replications in fields such as psychology (Open Science Collaboration 2015), economics (Camerer et al. 2016), and social sciences (Camerer et al. 2018), a state of affairs that has been referred to as the "replication crisis" (Fidler & Wilcox 2018).

The replication crisis is not rooted in a singular cause, but pertains to a network of different practices and incentive structures, all of which conjointly lead to an increase in results that are not replicable. Researchers have identified practices that might have contributed to the wide-spread lack of replicability, including but not limited to too small sample sizes (e.g., Button et al. 2013; Vasishth et al. 2018), lack of data and materials sharing (e.g., Nosek et al. 2015), use of anti-conservative statistical methods (e.g., Yarkoni 2019), large analytical flexibility (e.g., Simmons et al. 2011), and lack of generalizability across diverse contexts and populations (Henrich et al. 2010).

These limitations are present, and maybe even exacerbated in experimental linguistic research: Access to certain linguistic populations is often limited or too cost-intensive, making it difficult to collect sufficiently large samples. Experimental linguistic research is resource-intensive because of equipment cost and complexity, elaborateness of data collection procedures, and computational requirements of data analysis and curation. This often results in studies with small sample sizes and, consequently, with low statistical power (e.g., Casillas 2021; Kirby & Sonderegger 2018). Statistical analyses in linguistics are often ignoring important assumptions (e.g., Winter & Grice 2021) and are characterized by a large number of researcher degrees of freedom (Roettger 2019). Moreover, claims about human language are often based on a small set of languages, limiting their generalizability (e.g., Levisen 2019; Majid & Levinson 2010).

In light of the large overlap in research practices between linguistics and neighboring disciplines for which low replication rates and failures of attempts to replicate have been attested, there are raising concerns about both replication rates and replicability in the field of experimental linguistics (e.g., Marsden et al. 2018; Roettger & Baer-Henney 2019; Sönning & Werner 2021). A number of failed replication attempts reported in various subfields of linguistics indicate that these concerns have to be taken seriously (e.g., Chen 2007; Morey et al. 2021; Nieuwland et al. 2018; Papesh 2015; Stack et al. 2018; Vasishth et al. 2018; Westbury 2018; Jäger et al. 2020; Nieuwland et al. 2020).

Despite these known problems, there might be only very few published direct replications in linguistics. In their detailed assessment of replications in second language (L2) research, Marsden et al. (2018) explored 67 self-labeled L2 replication studies for a wide variety of characteristics. Their results indicate that for every 400 articles, only one replication study is published which translates into 0.25% of published studies containing a replication. Following Makel et al. (2012), we will refer to the proportion of published articles containing at least one replication as the replication rate. Moreover, the sample of Marsden et al. (2018) did not include a single direct replication study, i.e. a replication that strictly followed the design of the initial study. This is a state of affairs that is worrisome and warrants further investigation. To our knowledge, there is no systematic assessment of replication rates across experimental linguistics beyond Marsden et al. (2018). The present paper aims at filling this gap. To gauge the past and current replication landscape in experimental linguistics, track progress over time, and calibrate future policy and training initiatives, it will be useful to assess the prevalence of replications across experimental linguistics and explore their contributing factors.

The present study assesses the frequency of articles containing replications as well as the typology of replication studies that have been published in a representative sample of experimental linguistic journals from 1945 to 2020. Given the arguments presented above,

we are primarily interested in the prevalence of direct replications in the field. Our study aimed at answering two main questions: "How many published papers in experimental linguistics contain at least one direct replication?" and "Are there factors that affect the replication rates and are they either found at the journal level (e.g. journal policies, open access, journal impact factor, etc.) or at the study level (e.g. composition of authors, investigated language, etc.)?" The study consisted of two analyses: First, we assessed the frequency of articles mentioning the term replication (search string: replicat*) across 98 linguistic journals. Second, we manually categorized the type of replication studies (direct, partial, conceptual) in a subset of twenty journals. We then related their replication rates to factors like the years of publication, and the citation counts of both initial and replication study.

2 How often do journals mention the term replicat*?

The key dependent variable of the first part of this study was the rate of replication mention for journals relevant to the field of experimental linguistics.

2.1 Material and methods

The study design has been preregistered at 2021-03-08 and can be inspected at https://osf.io/a5xd7/.

In order to determine the rates of replication mention for individual journals, we drew on a method introduced by Makel et al. (2012). First, a sample of 100 journals relevant to the field of experimental linguistics was identified by making use of the search engine Web of Science (https://webofknowledge.com) (access date: 2021-03-03). We restricted the search results to journals in the web of science category "Linguistics" which had at least 100 articles published and a high ratio of articles containing the term "experiment*" in title, abstract or keywords in order to ensure that the subset contained journals that are relevant for experimental linguistics research. Among those, all articles categorized as having been published in English and between 1945-2020 were taken into account.[1]

The ratio between overall number of articles and those articles mentioning the term experiment* ranged between 6.1% and 60.3% (with a median of 11.5%) across journals. The full sample of journals can be inspected in the appendix of this article.[2]

After journal selection, we obtained the total count of articles containing the search term "replicat*" in title, abstract or keywords for each journal. Following the method presented by Makel et al. (2012), the rates of replication mention were calculated by dividing the number of articles containing the term replicat* by the total number of eligible articles for each journal. As we were only interested in experimental linguistic studies, we only considered articles containing the search term experiment* as eligible.

Rates of replication mention were then related to three journal properties: journal policies with regards to replication studies, journal impact factor and whether the journal publishes open access or not. To gain an understanding of the journal policies with regards to replication studies, we examined the journals' submission guidelines adopting

[1] The Web of Science catalog includes articles from 1945 to present. All full available years (at the date of retrieval) have been included in the analysis. The first entries for the category Linguistics date back to the year 1948 and the first hit for the search term "replicat*" was obtained for the year 1969.

[2] Two journals, namely "Language and Cognitive Processes" (since 2014: "Language, Cognition and Neuroscience") and "Literary and Linguistic Computing" (since 2015: "Digital Scholarship in the Humanities"), have been renamed. The article counts of the old and new journal names were combined under the new name. Our final sample thus included only 98 journals.

a method suggested by Martin and Clarke (2017). They grouped psychology journals into categories dependent on whether they (explicitly or implicitly) encouraged replication studies or not in their "instructions to authors" and "aims and scope" sections on the journal websites. For our analysis, we only distinguished between those journals explicitly encouraging replication studies and those that do not. We extracted journal impact factors via Journal Citation Reports (https://jcr.clarivate.com).[3] We assessed whether journals offered open access publication or not via Web of Science. We distinguished between three access categories: those journals which are listed in the Directory of Open Access Journals (DOAJ) ("DOAJ gold"), those journals that contained some open access articles ("partial") and those journals with no option to publish open access ("no") whatsoever.

We would like to stress that journal-based predictors are not static and obviously change over time. We cannot reliably capture these dynamics. Instead, we snapshoted journal policies and impact factors in the year 2019 and use this information as a (rough) proxy for our preregistered objective to relate them to replication rates. As will be discussed below, the model estimates for these predictors are characterized by large amounts of uncertainty, leaving them rather uninformative.

2.2 Results and Discussion

Out of the 52302 articles in our sample, 8437 mentioned the term experiment* in title, abstract, or keywords and were thus assumed to be articles presenting an experimental investigation. Out of these articles, 382 contained the term replicat* which results in a mention rate of 4.5% across experimental linguistic articles.

The distribution of the rate of replication mention substantially varies across journals ranging from 0 to 12.82%. The median rate of replication mention is 1.7%, a rate that is comparable to what Makel et al. (2012) have reported in their assessment of replications in psychology. Almost half of all journals (n = 42) did not mention the term in any of their articles. Figure 1 illustrates the variation across those journals that exhibited at least one mention of the term.

We statistically estimated the rate of replication mention as predicted relative to the following factors: centered journal impact factors (continuous, henceforth jif), open access type (no, partial, DOAJ gold), and replication policies (binary: either explicitly encourage or not).[4] We used Bayesian parameter estimation based on generalized linear regression models with a binomial link function.[5] The model was fitted to the proportion of replication mentions per journal using the R package brms (Bürkner 2016). We used weakly informative normal priors centered on 0 (sd = 2.5) for the intercept and Cauchy priors centered on 0 (scale = 2.5) for all population-level regression coefficients. These priors are what is referred to as regularizing (Gelman et al. 2008), thus making our model conserva-

[3] 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.

[4] We diverted from the preregistered protocol after constructive exchanges with our reviewers: We originally planned to use uncentered jif and open access as a binary covariate. Using uncentered jif would have provided an intercept representing journals with a journal impact factor of 0. Centering the variable to the mean jif of our sampled journals allows for a more intuitive interpretation of the coefficients. Second, we preregistered a dichotomization of open access policy which might obscure a more nuanced relationship between open access policy and replication rate. We thus opted for including all three levels of our open access variable in the final model. Both the preregistered and revised models are available in our repository.

[5] A possible concern of our modelling strategy might be an inflation of zeroes if there are too many journals without a single mention of the search term. A zero-inflated binomial regression can account for such an inflation. Thus, we additionally ran a zero-inflated binomial model. The resulting estimates for our parameters are highly compatible with those from the simpler binomial model. Both models are available in our repository.

Figure 1: Variation in rate of replicat* mention across those journals that exhibited at least one mention of the term. Numeric values on the right indicate the observed proportion of articles containing the string experiment* in title, abstract or keywords.

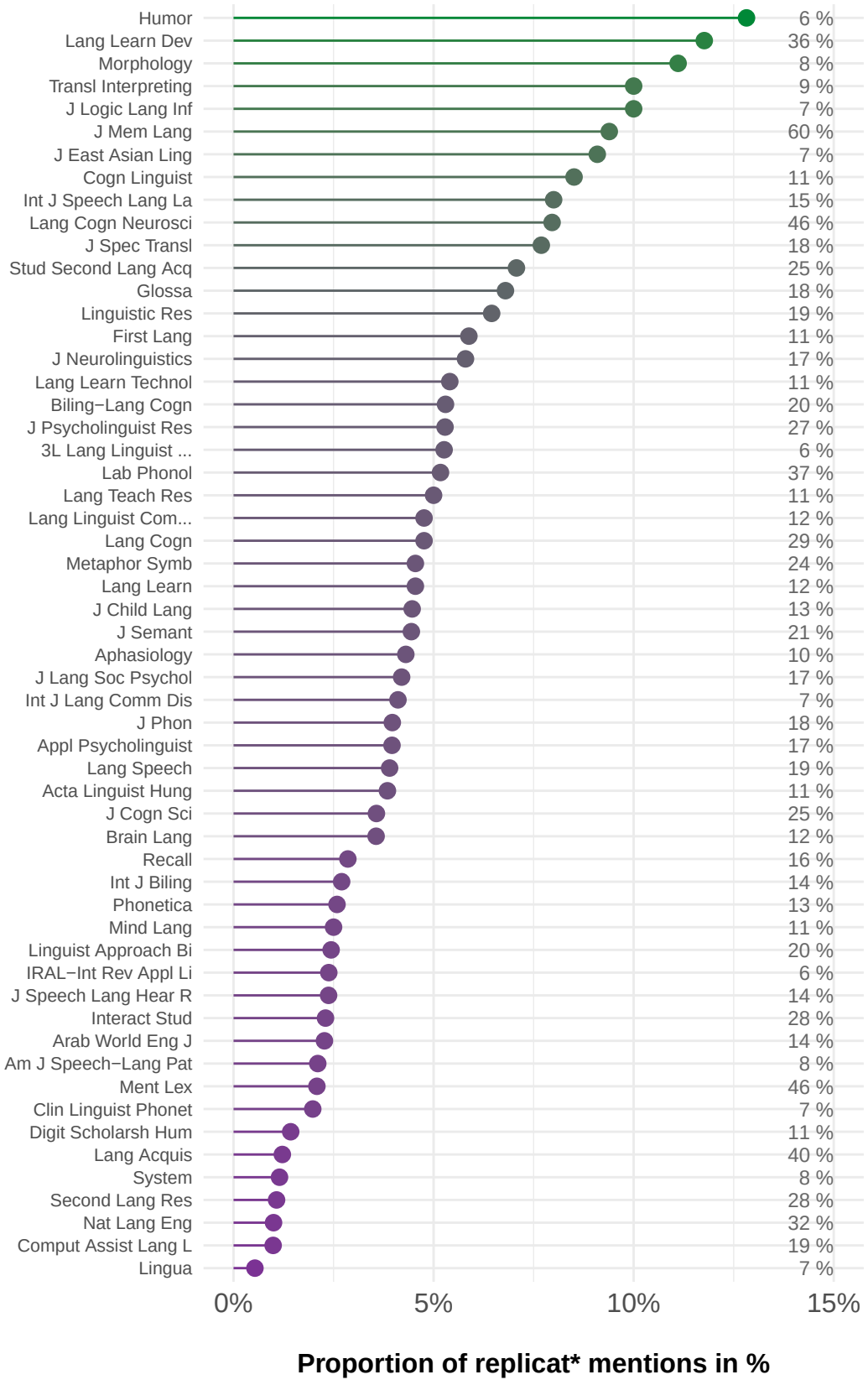
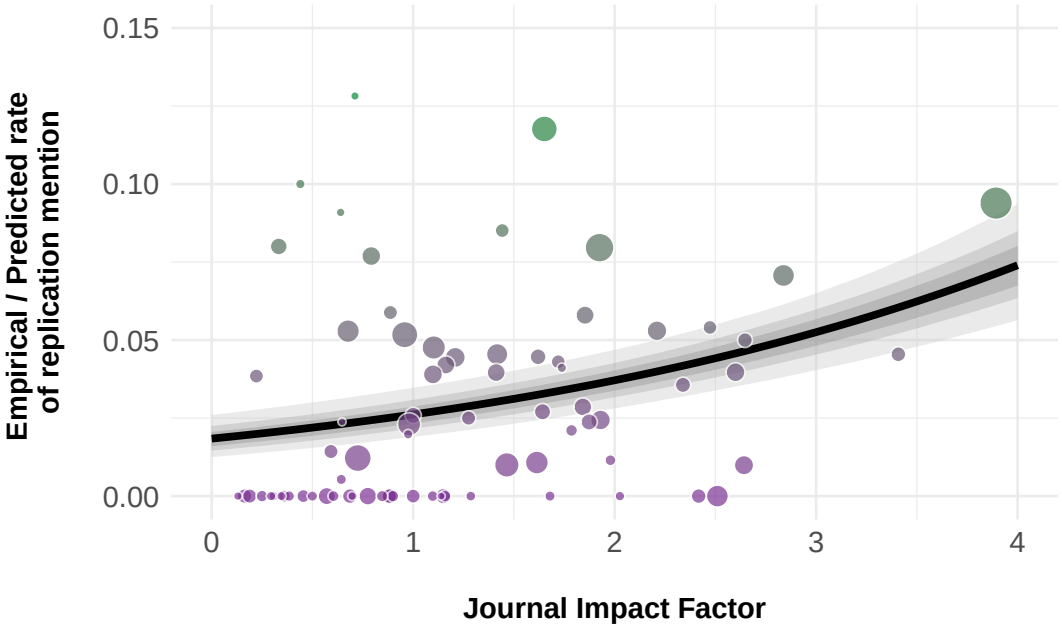


Figure 2: Estimated and empirical rate of mentioning the term 'replicat*' across sampled journals plotted against their journal impact factor. Each point represents one journal. Point size indicates the proportion of papers categorized as experimental (i.e. larger points indicate journals with more experimental articles). Line and shading indicate model predictions for journals with partial open access and 50/75/95% credible intervals.



tive with regards to the predictors under investigation. Four sampling chains with 2000 iterations each have been run for each model, with a warm-up period of 1000 iterations. For relevant predictor levels and contrasts between predictor levels, we 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).

The model estimates the proportion of replication mentions as 3.7% [2.8, 4.7] for the average journal impact factor of our sample and the most common open access category "partial". The model estimates that the mention rate increases with each integer unit of jif (log odds = 0.36 [0.27, 0.46]). Figure 2 illustrates this relationship.

Further explorations, however, indicate that jif is correlated with the number of experimental studies reported in a journal (Spearman correlation = 0.42).[6] Given the observed correlation, it remains unclear if the term replicat* is really used more often in high impact journals or simply more common in journals that generally publish more experimental studies (which tend to have higher jifs).

The model estimates the impact of whether the journal allows for open access publishing or not and whether replications are explicitly encouraged or not both as positive, i.e. the term replication is mentioned more often in both open access journals and in journals that explicitly encourage direct replications. Furthermore, the model suggests an ordinal

[6] This exploratory analysis was not preregistered.

relationship between open access categories and replication mention, characterized by higher rates in DOAJ gold open access journals than in partial open access journals which have higher rates than journals without any open access. However, due to the small number of journals that explicitly encourage direct replications (2 out of 98), and the relatively small number of open access journals (11 out of 98), the uncertainty around all these estimates is substantial (rates increase for DOAJ gold open access: 1.6% [-1.5, 6.1]; rates decrease for no open access: -1.9% [-4.4, 7]; and rates increase for encouraging replication policies: 0.7% [-1.1, 2.8]). We thus won't discuss these results further.

3 How many articles containing the term replicat* are actual replications?

The second part of the study had two aims: First, the term replication is commonly used in ambiguous ways, so articles containing the search term were further analyzed to determine whether they indeed reported a replication study or whether they used the term in a different way. Second, we further investigated what types of replication studies are published and whether replications are becoming more frequent over time. Our target estimand is the proportion of experimental articles containing at least one replication.

3.1 Material and methods

From the superset of 98 journals obtained above, the 19 journals[7] with the highest proportion of experimental studies were selected for a more detailed analysis while excluding journals for which less than 2 hits (TS=replicat*) could be obtained (see at https://osf.io/f3yp8/ for a list of article counts per journal). The sampling procedure above resulted in 274 possible self-labeled replication studies with publication years ranging between 1989 and 2020. We included the full set of articles in our sample for manual coding.

We identified whether the article in question indeed contained a replication study or not. Parts of the papers that were examined were title and abstract of the paper, text before and after occurrences of the search term replicat*, the paragraph before the Methods section as well as the first paragraph of the Discussion section (following and adapting the procedure specified by Makel et al. 2016). If the authors explicitly claimed that (one of) their research aim(s) was to replicate the result or methods of an initial study, this article was treated as a replication and was submitted to further analysis according to the preregistered coding scheme which can be inspected at https://osf.io/ct2xj/.

When extracting number and types of changes made to the initial study, we assumed that the authors of a replication study did not make any drastic changes *without* reporting them. Following Marsden et al. (2018), replication studies were classified according to the number of changes made into three categories: direct replication (0 changes), partial replication (1 change) and conceptual replication (2 or more changes). We noted the nature of methodological changes as one of the following categories: experimental paradigm, sample, materials/experimental set-up, dependent variable, independent variable, and control. Table 1 shows examples for the five categories that were used for identifying to which of the three types of replications an article belonged. All of the changes that have been identified by the manual coding procedure are changes that have been reported by the authors of the replication study. Most of these changes have been made by the authors in order to

[7] Due to "Language and Cognitive Processes" being renamed to "Language Cognition and Neuroscience", we did not reach the preregistered target sample of 20.

achieve specific goals: Either they aimed at showing that an effect extends to another language, that it is robust across different experimental paradigms or subject groups or how different kinds of measurements, manipulations and controls affect the observed results. As such, we did not consider slight changes in the stimulus materials like the correction of typos but only changes that were identified by the authors as expected to change the results or improve the study in a significant way. We also recorded the language under investigation. The information on whether the article was published open access as well as citation counts and years of publication for both studies were obtained from Web of Science. An author overlap was attested when at least one author was a (co-)author on both studies. During the coding procedure of the articles, we encountered edge cases that we did not anticipate in our preregistration: When several self-labeled replication studies were mentioned in one article, we chose the first mentioned study for our analysis. If there were one independent, but also one or more inner-paper replications, i.e. experiments that first replicated an independent initial study and then replicated results from a study in the same article, we selected the independent replication for analysis. Note that since our target estimand is the rate of published articles that contain at least one replication, this choice does not artificially reduce the replication rate.

Table 1: Types of changes that determined the type of replication study with examples.

type of change	examples
experimental paradigm	explicit change in experimental paradigm, e.g. artificial grammar learning paradigm □ oddball paradigm
sample	explicit change in population under investigation for the purpose of generalizability, e.g. children □ adults, English □ French, monolinguals □ bilinguals
materials / set-up	explicit change in material or experimental set-up, e.g. change in materials due to a different language, general changes to the stimulus material or presentation in order to improve the study (except for small changes like typos)
dependent variable	explicit change in operationalization/measurement of dependent variable(s) due to theory change or a different measurement technique, e.g. response times □ ERP component
independent variable	explicit change in operationalization/measurement of manipulated variable(s), e.g. the inclusion or omission of specific manipulation conditions

type of change	examples
control	explicit change in control variable(s), e.g. adding or excluding a specific control variable

3.2 Results and Discussion

Out of the 274 articles in the subsample, 262 (95.6%) indeed presented experimental linguistics research. The remaining 12 (4.4%) were not experimental in nature, but rather comments, reviews or computational studies. Out of the 262 experimental studies, 151 were self-claimed replications according to our criteria. The remaining 111 mentions were articles that mentioned the term in other contexts or studies that did not specify the concrete aim of replicating an initial study's design or results. Moreover, many papers used the term "replicated" in a broad sense that roughly translates into "finding a similar result", thus not qualifying as a replication study as defined above. Out of the replication studies, we categorized 86 (57%) as conceptual, 56 (37.1%) as partial, and only 11 (7.3%) as direct replications.

Looking closer at direct replications, 5 studies were independent studies, i.e. there was no overlap between authors of the initial study and the replication study. Out of these independent direct replication studies, 3 were self-labeled as successful replications. In other words, our sample included only two failed, independent, direct replication attempts. These low rates indicate that replication attempts, and especially direct replication attempts, are rather rare in the experimental linguistics literature - an observation that is in line with replication rates estimated for other research fields (Makel et al. 2012; Makel & Plucker 2014; Mueller-Langer et al. 2019).

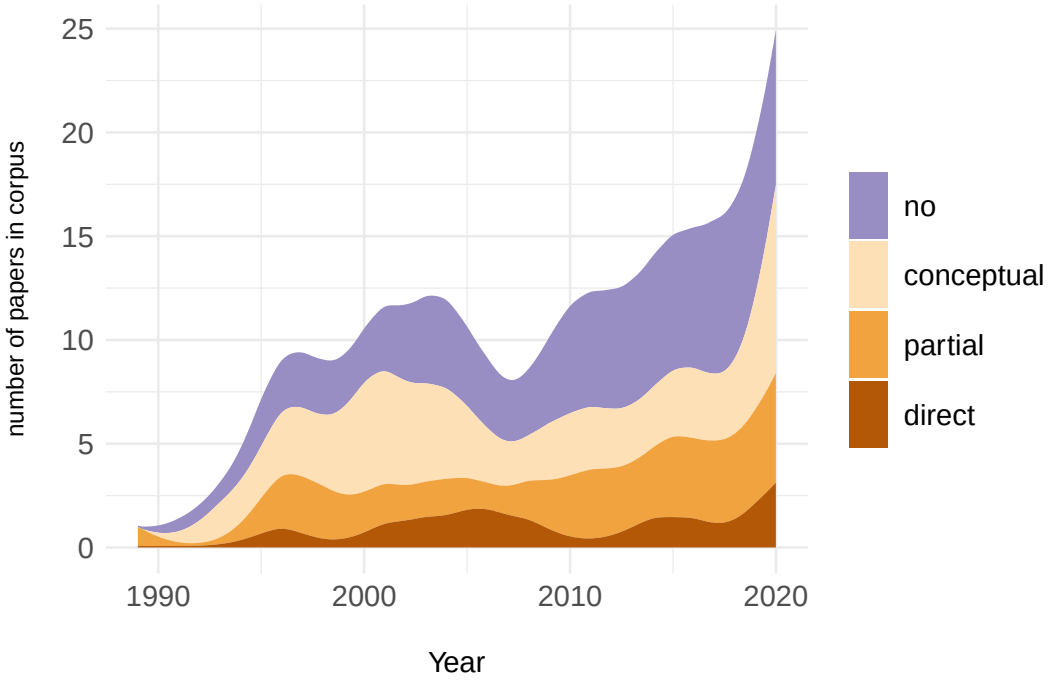
Figure 3 illustrates the development of replication studies throughout publication years. While the overall number of studies increased over the years, the proportion of direct replications remained stable at best. However, it seems as if there is an increasing number of partial and conceptual replications that was published within the last few years.[8] This increase could represent a shift towards replication practices as a direct consequence of renewed attention to the concept of replications caused by the replication crisis.

One possible reason for the fact that (direct) replication rates are not increasing for the field according to our analysis could be that experimental linguistics predominantly replicates experimental findings across languages, making the studies by definition only partial/conceptual replications. However, only 19.9% of replications targeted a different language than the initial study. The majority of replication efforts were conducted within the same language as the initial study. In fact, 67.5% of all replication studies in our sample had one variety of English as the main language of investigation either in the replication or in the corresponding initial study.

The median number of years between an initial and a replication study is 7 years. Initial studies were on average 50.1 times cited before a replication was published which corresponds to an average yearly citation rate of 7.2 citations. This average citation rate is well above the impact factor of core linguistic journals (median journal impact factor in superset: 1.1). Replication studies were on average only 21 times cited which corresponds

[8] Given the small number of direct replications in our sample, both a descriptive assessment and an inferential assessment as preregistered are very uninformative. The reader is directed to the supplementary materials if they are interested in the model outputs of the preregistered analysis.

Figure 3: Development of amount of replication studies published over time. Bandwith of kernel density estimation is set to 0.8.



to an average yearly citation rate (calculated up to the time of analysis) of 0.5 citations. These results are in line with Marsden et al.'s (2018) assessment of second language research. They found that replication studies were on average conducted after more than six years and after over a hundred citations of the original study. They concluded that replications are either only performed or only published after the original study had already substantially impacted the field. Our findings are in line with this interpretation for experimental linguistics. The observed smaller number of citations of replication studies compared to corresponding initial studies is also in line with the lack of perceived value of replication studies reported in other fields (e.g., Koole & Lakens 2012; Nosek et al. 2012).

3.3 Case study of Journal of Memory and Language

The Journal of Memory and Language (JML) accounts for the largest part of articles in our sample (114 out of 274) and is the journal with the highest impact factor (3.9). We conducted a subset analysis of articles published in JML because we were interested in whether our results were affected by this skewed sample.[9] We find that 70 (61.9%) of the 113 experimental JML papers contain replication studies. Of these, 35 (50%) are conceptual, 30 (42.9%) are partial, and 5 (7.1%) are direct replication studies which is in line with the results for the whole sample. Only 3 of the studies published in JML were independent direct replication studies (one of which was successful). We conclude that

[9] Originally, this subset-analysis was planned because in an earlier version of this paper we sampled 50 from the 114 articles published in JML. Following a reviewer's suggestion, we later submitted the full set of JML articles to manual coding. But we keep this analysis to show that our results apply to the whole field and are not mainly influenced by one journal.

we have little reason to believe that the large proportion of JML articles in our sample substantially affected our overall results and are confident that our results apply to the field rather than to one journal.

4 General discussion

The current study aimed at providing a comprehensive survey of published replications in experimental linguistic research. By analyzing the publication history of over 50000 articles across 98 journals that publish experimental linguistic research, our study found that 4.5% of experimental linguistic publications used the term replicat* in title, abstract or keywords. A more thorough analysis of 274 sampled experimental articles containing the term replicat* revealed that only around half of the hits represented actual replication studies, reducing the effective replication rate to 2.5%. This rate is slightly higher than reports of comparable investigations in psychology (1.6%, Makel et al. 2012), educational science (0.1%, Makel & Plucker 2014), and economics (0.1%, Mueller-Langer et al. 2019). The higher rate might be due to a methodological choice, however. Due to large plurality of methods in linguistics, we calculated the replication rate based on only those articles that contained the term experiment* (as opposed to all articles in the sample), reducing the denominator substantially.

A closer look at the nature of replication studies revealed that the majority of replication studies were studies that diverted from the initial study by at least one design choice. Only 7.3% were direct replications, i.e. studies that directly repeated an initial study without self-reported changes to the design, and only five of these were replications conducted by an independent team of researchers. Taking together replicat* mention rate and actual replication rate, 0.08% of experimental studies are independent direct replications in the field of linguistics. In other words, only 1 in 1250 experimental linguistic articles contains an independent direct replication. This clearly indicates that replication attempts, and especially independent direct replication attempts, are still very rare in the experimental linguistics literature.

Before interpreting the results and offering possible ways forward, we need to discuss two important caveats to our study. First, if research articles were not framed as experimental, then they were not included in the analysis. Similarly, if experimental articles were not framed as replications, then they were not categorized as such. These are clear limitations to our search strategy and might lead to an underestimation of the true replication rate. Assuming the false negative rate is not zero, the reported replication rates might change after correction. To circumvent this methodological problem, a large sample of articles would have to undergo manual coding which is not feasible for a large-scale assessment. Future research using alternative assessment methods (possibly machine learning techniques) or more in-depth investigation of either subfields (e.g., Marsden et al. 2018) or specific journals might result in different replication rates. However, the existence of replication studies that are not referred to as such might also reflect a more general problem: If studies are not framed as replications by using the term replication, readers' ability to connect research to its intellectual precedents is severely limited.

Second, our assessment of replication types relied on two assumptions. On the one hand, we assume that the authors disclosed changes to the initial study in a transparent way. On the other hand, we assume that if changes are disclosed, we were able to extract and interpret these changes accurately. Neither of these assumptions must hold, thus any rates that are generated here are necessarily only a rough proxy of the true replication rate. Nevertheless, given that our findings seem to align well with evidence from other fields

as well as an in-depth analysis of a subfield of linguistics (Marsden et al. 2018), we are confident that our conclusion holds.

Although the present study is the first systematic assessment of replication rates in linguistics, our conclusions are hardly surprising. Academic incentive systems do not reward replication studies. Neither journals nor funders encourage them. For example, Martin and Clarke's (2017) survey results suggest that in 2015 only 3% of psychology journals explicitly state that they will consider publishing replications. Similarly, out of the 98 journals in our sample, only 2 encouraged direct replications. And even if one manages to publish a replication, replication studies are characterized by much lower yearly citation counts compared to corresponding initial studies, leading to a lack of perceived prestige (e.g., Koole & Lakens 2012; Nosek et al. 2012; Marsden et al. 2018). Direct replications simply do not seem worth their costs.

In order to overcome the asymmetry between the cost of direct replication studies and the presently low academic payoff for it, we must re-evaluate the value of direct replications. Funding agencies, journals, but also editors and reviewers, need to start valuing direct replication attempts as much as they value novel findings. For example, we could either dedicate existing journal space to direct replications (e.g. as its own article type) or create new journals that are specifically dedicated to replication studies. Journals could help normalizing replication studies by calls for special issues dedicated to replications of influential findings like e.g. the recent call by the Journal of Memory and Language.[10] Another alternative is the Pottery Barn rule, implemented by for example Royal Society Open Science: Once the journal has published a study, it commits to publishing all direct replications of this study.[11]

At the same time, we should attempt to find more resource-efficient ways to both identify replication targets and conduct replication studies. We believe, most people would agree that not every study needs direct replication. Take for example the McGurk effect, i.e. perceiving a sound that lies in-between an auditory presented component of one sound and a visually presented component of another one (McGurk & MacDonald 1976). This phenomenon is probably replicated in dozens of linguistic classrooms every semester across the globe. On the other hand it might be a good idea to evaluate more critically whether a given study is worth replicating. Resources can be saved if studies with poor experimental design, unsuitable measurement approach or inept model specifications are ruled out from direct replication attempts (Yarkoni 2019). Finding convenient yet effective tools to identify worthwhile replication targets is an active meta-scientific field (e.g., Coles et al. 2018; Isager et al. 2021a; Hardwicke et al. 2018) and feasible algorithms are currently developed and tested (Isager et al. 2021b). When it comes to more accessible ways to conduct replication studies, several authors have suggested involving our students more rigorously (e.g., de Leeuw et al. 2019; Frank & Saxe 2012; Grahe et al. 2012; Roettger & Baer-Henney 2019), possibly creating a rich learning experience for our students while at the same time reducing the resource costs of replication studies. Alternatively, resources can be pooled across multi-lab replication efforts, effectively reducing the costs for individual researchers and labs (e.g., Frank et al. 2017; Nieuwland et al. 2018; Open Science Collaboration 2015). The StudySwap platform, for example, allows researchers to identify independent labs for conducting a replication attempt of one's study, thus helping researchers to assess the independent replicability of their findings prior to publication (Chartier et al. 2018).

[10] https://www.journals.elsevier.com/journal-of-memory-and-language/call-for-papers/replicating-influential-findings

[11] https://royalsociety.org/blog/2018/10/reproducibility-meets-accountability/

We are confident that the field of linguistics can function as a role model for neighboring fields. Although major meta-scientific discourses are held in other fields, linguistics has demonstrated quick uptake of methodological reforms time and time again. A point in case is the swift uptake of Registered Reports[12], a new article form in which a study proposal is reviewed before the research is undertaken. While the uptake across disciplines is slow, linguistics has already at least 12 high-impact journal outlets that offer Registered Reports. Moreover, an increasing number of reproducibility initiatives founded in the field during the last few years give hope that the field is continuing to evaluate their past, current, and future practices and successfully face the challenges ahead. This paper was an attempt to contribute to this development. We hope our assessment allows future efforts to track progress over time and calibrate policies across experimental linguistics.

Appendix

Table 2: The full sample of journals sorted by their ratio of experimental linguistics articles.

Journal	no. articles	no. exp. articles	ratio of exp. articles in %
Journal Of Memory And Language	2,012	1,214	60.34
Mental Lexicon	105	48	45.71
Language Acquisition	207	82	39.61
Language Cognition And Neuroscience	1,373	628	45.74
Laboratory Phonology	155	58	37.42
Language Learning And Development	141	51	36.17
Natural Language Engineering	312	100	32.05
Lecture Notes In Computer Science	150	46	30.67
Language And Cognition	144	42	29.17
Interaction Studies	312	87	27.88

[12] http://cos.io/rr

Journal	no. articles	no. exp. articles	ratio of exp. articles in %
Second Language Research	338	93	27.51
Journal Of Psycholinguistic Research	1,691	454	26.85
Studies In Second Language Acquisition	389	99	25.45
Computational Linguistics	521	130	24.95
Journal Of Cognitive Science	114	28	24.56
Metaphor And Symbol	278	66	23.74
Lecture Notes In Artificial Intelligence	113	26	23.01
Journal Of Semantics	218	45	20.64
Linguistic Approaches To Bilingualism	204	41	20.10
Bilingualism Language And Cognition	753	151	20.05
Computer Assisted Language Learning	531	101	19.02
Linguistic Research	166	31	18.67
Language And Speech	1,521	282	18.54
Journal Of Specialised Translation	141	26	18.44
Glossa A Journal Of General Linguistics	561	103	18.36
Journal Of Phonetics	1,389	252	18.14

Journal	no. articles	no. exp. articles	ratio of exp. articles in %
Journal Of Neurolinguistics	806	138	17.12
Applied Psycholinguistics	1,202	202	16.81
Journal Of Language And Social Psychology	711	119	16.74
Recall	214	35	16.36
Phonology	190	31	16.32
Interpreting	131	20	15.27
Eurasian Journal Of Applied Linguistics	115	17	14.78
International Journal Of Speech Language And The Law	171	25	14.62
Journal Of Language And Education	145	21	14.48
Linguistics Vanguard	146	21	14.38
Arab World English Journal	952	132	13.87
Journal Of Speech Language And Hearing Research	3,389	463	13.66
International Journal Of Bilingualism	542	74	13.65
Phonetica	862	116	13.46
Journal Of Child Language	1,711	224	13.09
Procesamiento Del Lenguaje Natural	107	14	13.08
Applied Linguistics Research Journal	177	23	12.99

Journal	no. articles	no. exp. articles	ratio of exp. articles in %
Natural Language Semantics	145	18	12.41
Journal Of Quantitative Linguistics	258	32	12.40
Brain And Language	3,680	449	12.20
Language And Linguistics Compass	178	21	11.80
Language Learning	1,314	154	11.72
Corpus Linguistics And Linguistic Theory	156	18	11.54
Review Of Cognitive Linguistics	182	21	11.54
Language Teaching Research	524	60	11.45
Interpreter And Translator Trainer	231	26	11.26
Poznan Studies In Contemporary Linguistics	322	36	11.18
Mind Language	728	80	10.99
First Language	312	34	10.90
Pragmatics Cognition	193	21	10.88
Acta Linguistica Hungarica	243	26	10.70
Syntax A Journal Of Theoretical Experimental And Interdisciplinary Research	150	16	10.67
Cognitive Linguistics	443	47	10.61

Journal	no. articles	no. exp. articles	ratio of exp. articles in %
Journal Of Research In Applied Linguistics	283	30	10.60
Language Learning Technology	352	37	10.51
Aphasiology	1,999	209	10.46
Digital Scholarship In The Humanities	636	70	11.01
Probus	157	15	9.55
Innovation In Language Learning And Teaching	168	16	9.52
International Journal Of English Linguistics	786	71	9.03
Translation Interpreting The International Journal Of Translation And Interpreting	114	10	8.77
Across Languages And Cultures	164	14	8.54
Morphology	106	9	8.49
American Journal Of Speech Language Pathology	1,132	95	8.39
Revue Roumaine De Linguistique Romanian Review Of Linguistics	205	17	8.29
Intercultural Pragmatics	245	20	8.16
Child Language Teaching Therapy	249	20	8.03

Journal	no. articles	no. exp. articles	ratio of exp. articles in %
Language Awareness	262	21	8.02
Gesture	143	11	7.69
Journal Of The International Phonetic Association	221	17	7.69
System	1,131	87	7.69
Metaphor And Symbolic Activity	134	10	7.46
Iberica	203	15	7.39
Lingua	2,551	187	7.33
Annual Review Of Applied Linguistics	151	11	7.28
Linguistica Antverpiensia New Series Themes In Translation Studies	138	10	7.25
Terminology	127	9	7.09
Annual Review Of Linguistics	101	7	6.93
Journal Of Logic Language And Information	146	10	6.85
Journal Of French Language Studies	117	8	6.84
Clinical Linguistics Phonetics	1,480	101	6.82
Language And Linguistics	281	19	6.76
International Journal Of Language Communication Disorders	1,080	73	6.76
Nordic Journal Of Linguistics	150	10	6.67

Journal	no. articles	no. exp. articles	ratio of exp. articles in %
Journal Of East Asian Linguistics	338	22	6.51
Language And Literature	246	16	6.50
3l Language Linguistics Literature The Southeast Asian Journal Of English Language Studies	293	19	6.48
Babel Revue Internationale De La Traduction International Journal Of Translation	264	17	6.44
Humor International Journal Of Humor Research	607	39	6.43
International Journal Of Corpus Linguistics	239	15	6.28
Iral International Review Of Applied Linguistics In Language Teaching	671	42	6.26
International Journal Of Applied Linguistics	163	10	6.13

Abbreviations

DOAJ = Directory of Open Access Journals,
JIF = journal impact factor,
L2 = second language,
JML = Journal of Memory and Language

Data availability

All data and analyses are available online at https://osf.io/9ceas/.

Acknowledgements

We would like to thank Brian Dillon and two anonymous reviewers for their insightful comments and suggestions. All remaining errors are our own.

Funding information

KK is supported by TODO.

Competing interests

The authors have no competing interests to declare.

Authors' contributions

KK supervised the project and was responsible for its administration. KK and TR conceptualized the project, decided on the methodology and analyzed the data. KK took the lead on data curation and writing. TR provided visualizations and reviewed and edited the text.

References

Amrhein, Valentin & Trafimow, David & Greenland, Sander. 2019. Inferential statistics as descriptive statistics: There is no replication crisis if we don't expect replication. *The American Statistician* 73(sup1). 262–270. Publisher: Taylor & Francis.

Barba, Lorena A. 2018. Terminologies for Reproducible Research https://doi.org/https://doi.org/10.48550/arXiv.1802.03311. http://arxiv.org/abs/1802.03311. ArXiv: 1802.03311.

Button, Katherine S & Ioannidis, John PA & Mokrysz, Claire & Nosek, Brian A & Flint, Jonathan & Robinson, Emma SJ & Munafo, Marcus R. 2013. Power failure: why small sample size undermines the reliability of neuroscience. *Nature reviews neuroscience* 14(5). 365–376. https://doi.org/https://doi.org/10.1038/nrn3475. Publisher: Nature Publishing Group.

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

Camerer, Colin F. & Dreber, Anna & Forsell, Eskil & Ho, Teck-Hua & Huber, Jürgen & Johannesson, Magnus & Kirchler, Michael & Almenberg, Johan & Altmejd, Adam & Chan, Taizan & Heikensten, Emma & Holzmeister, Felix & Imai, Taisuke & Isaksson, Siri & Nave, Gideon & Pfeiffer, Thomas & Razen, Michael & Wu, Hang. 2016. Evaluating replicability of laboratory experiments in economics. *Science* 351(6280). 1433–1436. https://doi.org/https://doi.org/10.1126/science.aaf0918.

Camerer, Colin F. & Dreber, Anna & Holzmeister, Felix & Ho, Teck-Hua & Huber, Jürgen & Johanesson, Magnus & Kirchler, Michael & Nave, Gideon & Nosek, Brian A. & Pfeiffer, Thomas & Altmejd, Adam & Buttrick, Nick & Chan, Taizan & Chen, Yiling & Forsell, Eskil & Gampa, Anup & Heikensten, Emma & Hummer, Lily & Imai, Taisuke & Isaksson, Siri & Manfredi, Dylan & Rose, Julia & Wagenmakers, Eric-Jan & Wu, Hang. 2018. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature* 2. 637–644. https://doi.org/https://doi.org/10.1038/s41562-018-0399-z.

Campbell, Donald T. 1969. Reforms as experiments. *American Psychologist* 24(4). 409.
Casillas, Joseph V. 2021. Interlingual interactions elicit performance mismatches not "compromise" categories in early bilinguals: Evidence from meta-analysis and coronal stops. *Languages* 6(1). 9. https://doi.org/https://doi.org/10.3390/languages6010009. Publisher: Multidisciplinary Digital Publishing Institute.
Chartier, Christopher R. & Riegelman, Amy & McCarthy, Randy J. 2018. StudySwap: A Platform for Interlab Replication, Collaboration, and Resource Exchange. *Advances in Methods and Practices in Psychological Science* 1(4). 574–579. https://doi.org/https://doi.org/10.1177/2515245918808767.
Chen, Jenn-Yeu. 2007. Do Chinese and English speakers think about time differently? Failure of replicating Boroditsky (2001). *Cognition* 104(2). 427–436. https://doi.org/https://doi.org/10.1016/j.cognition.2006.09.012.
Coles, Nicholas A & Tiokhin, Leonid & Scheel, Anne M & Isager, Peder M & Lakens, Daniël. 2018. The costs and benefits of replication studies. *Behavioral and Brain Sciences* 41. https://doi.org/https://doi.org/10.1017/S0140525X18000596. Publisher: Cambridge University Press.
Crandall, Christian S & Sherman, Jeffrey W. 2016. On the scientific superiority of conceptual replications for scientific progress. *Journal of Experimental Social Psychology* 66. 93–99. https://doi.org/https://doi.org/10.1016/j.jesp.2015.10.002. Publisher: Elsevier.
de Leeuw, Joshua R & Andrews, Jan & Livingston, Ken & Franke, Michael & Hartshorne, Josh & Hawkins, Robert & Wagge, Jordan. 2019. Using replication studies to teach research methods in cognitive science. *Perspectives on Psychological Science* 7(6). 600–604.
Doyen, Stéphane & Klein, Olivier & Pichon, Cora-Lise & Cleeremans, Axel. 2012. Behavioral priming: it's all in the mind, but whose mind? *PloS one* 7(1). e29081. https://doi.org/https://doi.org/10.1371/journal.pone.0029081.
Fanelli, Daniele. 2010. Do pressures to publish increase scientists' bias? An empirical support from US States Data. *PLoS ONE* 5(4). e10271. https://doi.org/https://doi.org/10.1371/journal.pone.0010271.
Fanelli, Daniele. 2012. Negative results are disappearing from most disciplines and countries. *Scientometrics* 90(3). 891–904. https://doi.org/https://doi.org/10.1007/s11192-011-0494-7. Publisher: Akadémiai Kiadó, co-published with Springer Science+ Business Media BV ….
Fidler, Fiona & Wilcox, John. 2018. *Reproducibility of Scientific Results*. https://plato.stanford.edu/entries/scientific-reproducibility/.
Frank, Michael C & Bergelson, Elika & Bergmann, Christina & Cristia, Alejandrina & Floccia, Caroline & Gervain, Judit & Hamlin, J Kiley & Hannon, Erin E & Kline, Melissa & Levelt, Claartje & others. 2017. A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy* 22(4). 421–435. https://doi.org/https://doi.org/10.1111/infa.12182. Publisher: Wiley Online Library.
Frank, Michael C & Saxe, Rebecca. 2012. Teaching replication. *Perspectives on Psychological Science* 7(6). 600–604. https://doi.org/https://doi.org/10.1177/1745691612460686. Publisher: Sage Publications Sage CA: Los Angeles, CA.
Gelman, Andrew & Jakulin, Aleks & Pittau, Maria Grazia & Su, Yu-Sung & others. 2008. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics* 2(4). 1360–1383. https://doi.org/https://doi.org/10.1214/08-AOAS191.

Grahe, Jon E & Reifman, Alan & Hermann, Anthony D & Walker, Marie & Oleson, Kathryn C & Nario-Redmond, Michelle & Wiebe, Richard P. 2012. Harnessing the undiscovered resource of student research projects. *Perspectives on Psychological Science* 7(6). 605–607. https://doi.org/https://doi.org/10.1177/1745691612459057. Publisher: Sage Publications Sage CA: Los Angeles, CA.

Hardwicke, Tom E & Tessler, Michael Henry & Peloquin, Benjamin N & Frank, Michael C. 2018. A Bayesian decision-making framework for replication. *Behavioral and Brain Sciences* 41. https://doi.org/https://doi.org/10.1017/S0140525X18000675. Publisher: Cambridge University Press.

Henrich, Joseph & Heine, Steven J & Norenzayan, Ara. 2010. The weirdest people in the world? *Behavioral and brain sciences* 33(2-3). 61–83. https://doi.org/https://doi.org/10.1017/S0140525X0999152X. Publisher: Cambridge University Press.

Isager, Peder M & van Aert, Robbie & Bahník, Štěpán & Brandt, Mark J & DeSoto, K Andrew & Giner-Sorolla, Roger & Krueger, Joachim I & Perugini, Marco & Ropovik, Ivan & van't Veer, Anna E & Vranka, Marek & Lakens, Daniel. 2021a. Deciding what to replicate: A decision model for replication study selection under resource and knowledge constraints. *Psychological Methods* https://doi.org/https://doi.org/10.1037/met0000438. Publisher: American Psychological Association.

Isager, Peder M & van't Veer, Anna & Lakens, Daniel. 2021b. Replication value as a function of citation impact and sample size https://doi.org/https://doi.org/10.31222/osf.io/knjea. Publisher: MetaArXiv.

John, Leslie K & Loewenstein, George & Prelec, Drazen. 2012. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science* 23(5). 524–532. https://doi.org/https://doi.org/10.1177/0956797611430953. Publisher: Sage Publications Sage CA: Los Angeles, CA.

Jäger, Lena A. & Mertzen, Daniela & van Dyke, Julie A. & Vasishth, Shravan. 2020. Interference patterns in subject-verb agreement and reflexives revisited: A large-sample study. *Journal of Memory and Language* 111. https://doi.org/https://doi.org/10.1016/j.jml.2019.104063.

Kelly, Clint D. 2019. Rate and success of study replication in ecology and evolution. *PeerJ* 7. e7654. https://doi.org/https://doi.org/10.7717/peerj.7654. Publisher: PeerJ Inc.

Kirby, James & Sonderegger, Morgan. 2018. Mixed-effects design analysis for experimental phonetics. *Journal of Phonetics* 70. 70–85. https://doi.org/https://doi.org/10.1016/j.wocn.2018.05.005. Publisher: Elsevier.

Koole, Sander L. & Lakens, Daniël. 2012. Rewarding replications: A sure and simple way to improve psychological science. *Perspectives on Psychological Science* 7(6). 608–614. https://doi.org/https://doi.org/10.1016/10.1177/1745691612462586.

Levisen, Carsten. 2019. Biases we live by: Anglocentrism in linguistics and cognitive sciences. *Language Sciences* 76. 101173. https://doi.org/https://doi.org/10.1016/j.langsci.2018.05.010. Publisher: Elsevier.

Lindsay, R Murray & Ehrenberg, Andrew SC. 1993. The design of replicated studies. *The American Statistician* 47(3). 217–228. https://doi.org/http://dx.doi.org/10.1080/00031305.1993.10475983. Publisher: Taylor & Francis.

Madden, Charles S. & Easley, Richard W. & Dunn, Mark G. 1995. How journal editors view replication research. *Journal of Advertising* 24(4). 77–87. https://doi.org/https://doi.org/10.1080/00913367.1995.10673490.

Majid, Asifa & Levinson, Stephen C. 2010. The language of perception across cultures. In *Talk presented at the XXth Congress of European Chemoreception Research Organization, Symposium on "Senses in language and culture"*. Avignon, France. https://doi.org/https://doi.org/10.1093/chemse/bjq126.

Makel, Matthew C & Plucker, Jonathan A. 2014. Facts are more important than novelty: Replication in the education sciences. *Educational Researcher* 43(6). 304–316. https://doi.org/https://doi.org/10.3102/0013189X14545513.

Makel, Matthew C. & Plucker, Jonathan A. & Freeman, Jennifer & Lombardi, Allison & Simonsen, Brandi & Coyne, Michael. 2016. Replication of Special Education Research: Necessary but Far Too Rare. *Remedial and Special Education* 37(4). 205–212. https://doi.org/https://doi.org/10.1177/0741932516646083.

Makel, Matthew C. & Plucker, Jonathan A. & Hegarty, Boyd. 2012. Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science* 7(6). 537–542. https://doi.org/https://doi.org/10.1177/1745691612460688.

Marsden, Emma & Morgan-Short, Kara & Thompson, Sophie & Abugaber, David. 2018. Replication in Second Language Research: Narrative and Systematic Reviews and Recommendations for the Field. *Language Learning* 68(2). 321–391. https://doi.org/https://doi.org/10.1111/lang.12286.

Martin, G. N. & Clarke, Richard M. 2017. Are Psychology Journals Anti-replication? A Snapshot of Editorial Practices. *Frontiers in Psychology* 8. https://doi.org/https://doi.org/10.3389/fpsyg.2017.00523.

McGurk, Harry & MacDonald, John. 1976. Hearing lips and seeing voices. *Nature* 264(5588). 746–748. https://doi.org/https://doi.org/10.1038/264746a0. Publisher: Nature Publishing Group.

McNeeley, Susan & Warner, Jessica J. 2015. Replication in criminology: A necessary practice. *European journal of criminology* 12(5). 581–597. https://doi.org/https://doi.org/10.1177/1477370815578197. Publisher: Sage Publications Sage UK: London, England.

Meehl, Paul E. 1990. Why summaries of research on psychological theories are often uninterpretable. *Psychological reports* 66(1). 195–244. https://doi.org/https://doi.org/10.2466/pr0.1990.66.1.195. Publisher: SAGE Publications Sage CA: Los Angeles, CA.

Morey, Richard D & Kaschak, Michael P & Díez-Álamo, Antonio M & Glenberg, Arthur M & Zwaan, Rolf A & Lakens, Daniël & Ibáñez, Agustín & García, Adolfo & Gianelli, Claudia & Jones, John L & others. 2021. A pre-registered, multi-lab non-replication of the action-sentence compatibility effect (ACE). *Psychonomic Bulletin & Review* https://doi.org/https://doi.org/10.3758/s13423-021-01927-8. Publisher: Psychonomic Society.

Mueller-Langer, Frank & Fecher, Benedikt & Harhoff, Dietmar & Wagner, Gert G. 2019. Replication studies in economics—How many and which papers are chosen for replication, and why? *Research Policy* 48(1). 62–83. https://doi.org/https://doi.org/10.1016/j.respol.2018.07.019. Publisher: Elsevier.

Nieuwland, Mante S. & Arkhipova, Yana & Rodríguez-Gómez, Pablo. 2020. Anticipating words during spoken discourse comprehension: A large-scale, pre-registered replication study using brain potentials. *Cortex; a journal devoted to the study of the nervous system and behavior* 133. 1–36. https://doi.org/https://doi.org/10.1016/j.cortex.2020.09.007.

Nieuwland, Mante S. & Politzer-Ahles, Stephen & Heyselaar, Evelien & Segaert, Katrien & Darley, Emily & Kazanina, Nina & Zu Wolfsthurn, Sarah Von Grebmer & Bartolozzi, Federica & Kogan, Vita & Ito, Aine & Mézière, Diane & Barr, Dale J. & Rousselet, Guillaume A. & Ferguson, Heather J. & Bush-Moreno, Simon & Fu, Xiao & Tuomainen, Jyrki & Kulakova, Eugenia & Husband, Matthew E. & Donaldson, David I. & Kohút, Zdenko & Rueschemeyer, Shirley-Ann & Huettig, Falk. 2018. Large-scale replication study reveals a limit on probabilistic prediction in language comprehension.

eLife 7. e33468. https://doi.org/https://doi.org/10.7554/eLife.33468.001.

Nosek, Brian A & Alter, George & Banks, George C & Borsboom, Denny & Bowman, Sara D & Breckler, Steven J & Buck, Stuart & Chambers, Christopher D & Chin, Gilbert & Christensen, Garret & others. 2015. Promoting an open research culture. *Science* 348(6242). 1422–1425. https://doi.org/https://doi.org/10.1126/science.aab2374. Publisher: American Association for the Advancement of Science.

Nosek, Brian A & Spies, Jeffrey R & Motyl, Matt. 2012. Scientific utopia II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science* 7(6). 615–631. https://doi.org/https://doi.org/10.1177/1745691612459058.

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

Papesh, Megan 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/https://doi.org/10.1037/xge0000125.

Pashler, Harold & Harris, Christine R. 2012. Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science* 7(6). 531–536. https://doi.org/https://doi.org/10.1177/1745691612463401. Publisher: Sage Publications Sage CA: Los Angeles, CA.

Roettger, Timo B. 2019. Researcher degrees of freedom in phonetic research. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 10(1). https://doi.org/http://dx.doi.org/10.5334/labphon.147. Publisher: Ubiquity Press.

Roettger, Timo B & Baer-Henney, Dinah. 2019. Toward a replication culture: Speech production research in the classroom. *Phonological Data and Analysis* 1(4). 1–23. https://doi.org/http://dx.doi.org/10.31234/osf.io/q9t7c.

Rosenthal, Robert. 1990. Replication in behavioral research. *Journal of Social Behavior and Personality* 5(4). 1.

Shanks, David R & Vadillo, Miguel A & Riedel, Benjamin & Clymo, Ashley & Govind, Sinita & Hickin, Nisha & Tamman, Amanda JF & Puhlmann, Lara. 2015. Romance, risk, and replication: Can consumer choices and risk-taking be primed by mating motives? *Journal of Experimental Psychology: General* 144(6). e142. https://doi.org/https://doi.org/10.1037/xge0000116. Publisher: American Psychological Association.

Simmons, Joseph P & Nelson, Leif D & Simonsohn, Uri. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science* 22(11). 1359–1366. https://doi.org/https://doi.org/10.1177/0956797611417632. Publisher: Sage Publications Sage CA: Los Angeles, CA.

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

Stroebe, Wolfgang & Strack, Fritz. 2014. The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science* 9(1). 59–71. https://doi.org/https://doi.org/10.1177/1745691613514450. Publisher: Sage Publications Sage CA: Los Angeles, CA.

Sönning, Lukas & Werner, Valentin. 2021. The replication crisis, scientific revolutions, and linguistics. *Linguistics* 59(5). 1179–1206. https://doi.org/https://doi.org/10.1515/ling-2019-0045. Publisher: De Gruyter Mouton.

Vasishth, Shravan & Mertzen, Daniela & Jäger, Lena A. & Gelman, Andrew. 2018. The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language* 103. 151–175. https://doi.org/https://doi.org/10.1016/

j.jml.2018.07.004.

Wagenmakers, E-J & Beek, Titia & Dijkhoff, Laura & Gronau, Quentin F & Acosta, A & Adams Jr, RB & Albohn, DN & Allard, ES & Benning, Stephen D & Blouin-Hudon, E-M & others. 2016. Registered replication report: strack, martin, & stepper (1988). *Perspectives on Psychological Science* 11(6). 917–928. https://doi.org/https://doi.org/10.1177/1745691616674458. Publisher: Sage Publications Sage CA: Los Angeles, CA.

Westbury, Chris. 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. https://doi.org/https://doi.org/10.1016/j.bandl.2004.07.006.

Winter, Bodo & Grice, Martine. 2021. Independence and generalizability in linguistics. *Linguistics* 59(5). 1251–1277. https://doi.org/https://doi.org/10.1515/ling-2019-0049. Publisher: De Gruyter Mouton.

Yarkoni, Tal. 2019. The generalizability crisis. *Behavioral and Brain Sciences* 1–37. https://doi.org/https://doi.org/10.1017/s0140525x20001685. Publisher: Cambridge University Press.

Zwaan, Rolf A. & Etz, Alexander & Lucas, Richard E. & Donnellan, M. Brent. 2018. Making replication mainstream. *Behavioral and Brain Sciences* 41. E120. https://doi.org/https://doi.org/10.1017/S0140525X17001972.