Assessing replication rates in journals of experimental linguistics

Kristina Kobrock*,[a], Timo B. Roettger[b]

[a] *Department, Street, City, State, Zip*
[b] *Department, Street, City, State, Zip*

Abstract

This is the abstract. ~150 words, avoid references, optional graphical abstract, keywords (max. 6, avoid abbreviations, AE spelling)

It consists of two paragraphs.

Introduction

The replication and reproducibility of results is key to good scientific practice. Yet, various scientific disciplines are currently facing what is popularly referred to as a "reproducibility" or "replication crisis" characterized by a small amount of published replication studies and an increasing number of failed replication attempts (**fidler__reproducibility__2018?**). Researchers from fields such as psychology (Makel et al., 2012), education science (Makel and Plucker, 2014), and special education research (Makel et al., 2016) have assessed the amount of direct replications in their respective fields and report alarmingly low replication rates ranging from 0.13% in the education sciences to 1.07% in psychology publications. Coordinated efforts to replicate published findings have uncovered surprisingly low rates of successful replications ranging from 47% in psychology (Open Science Collaboration, 2015) to 61% in economics (Camerer et al., 2016) and 62% in the social sciences (Camerer et al., 2018). A number of failed replication attempts reported in various subfields of linguistics indicate that the field is not immune to these raising concerns (e.g. in language comprehension: Papesh, 2015; predictive processing: Nieuwland et al., 2018; among others: Chen, 2007; Stack et al., 2018; Westbury, 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 et al., 2018; Roettger and Baer-Henney, 2019). 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 (see also e.g. **fanelli__pressures__2010?**). In order to

*Corresponding Author
Email address: kkobrock@uni-osnabrueck.de (Kristina Kobrock)

thoroughly understand and be able to address this problem, it is important to assess the number of replication attempts and their contributing factors.

In order to evaluate the replication rate in experimental linguistics, the present study assessed the frequency and typology of replication studies that have been published in a representative sample of experimental linguistic journals from their beginnings until 2020. The study consisted of two parts: First, the frequency of self-reported replication attempts across 100 linguistic journals were assessed and the rate of replication mention was related to factors like journal impact factor, publishing policy and publication access. Second, the type of replication studies (direct, partial, conceptual) published in a subset of 20 journals was investigated and their frequency was related to factors like the year of publication, and the citation and publication year of the initial study.

Overview analysis: how often do articles 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. We intended 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?

Material and methods

The material and methods have been preregistered at DATA at the Open Science Framework and can be inspected here: https://osf.io/9ceas/.

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 has been identified by making use of the search engine "Web of Science" (https://webofknowledge.com). 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 . All English language articles from the full available range of complete years (1945-2020) were taken into account. We selected the top 100 journals according to their ratio of experimental studies. The full list of journals can be inspected here: https://osf.io/q2e9k/. The procedure described above helped us to identify journals relevant for the field of experimental linguistics. As a second step, the total number of articles containing the search term "replicat*" in title, abstract or keywords was obtained via Web of Science search for the 100 sampled journals. 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 total number of articles for each journal. As we were only interested in experimental linguistic studies, we only included articles containing the search term "experiment*" in this formula.

In order to relate the rate of replication mention to journal policies, we further examined the journals' submission guidelines adopting a 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 and Clarke, 2017, p. 3). For our analysis, we only distinguished between those journals explicitly encouraging replication studies (1) and those that do not (2-4).

Journal impact factors were 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.

Furthermore, we assessed via Web of Science whether journals published open access. We distinguished between three categories: journals which are listed in the Directory of Open Access Journals (DOAJ) ("DOAJ gold"), journals with some articles being published as open access articles ("partial") and journals with no option to publish open access ("no").

Results

Out of the 51272 articles in our sample, 8006 mentioned the term experiment* in title, abstract, or keywords. Out of these articles, 347 contained the term replic* in either abstract or title. Thus the rate of mentioning the term replic* is 0.043.

The distribution of the mention rate substantially varied across journals ranging from 0 to 0.128. 43 journals did not mention the term in any of their articles. The median mention rate across journals is 0.016 (SD = 0.033).

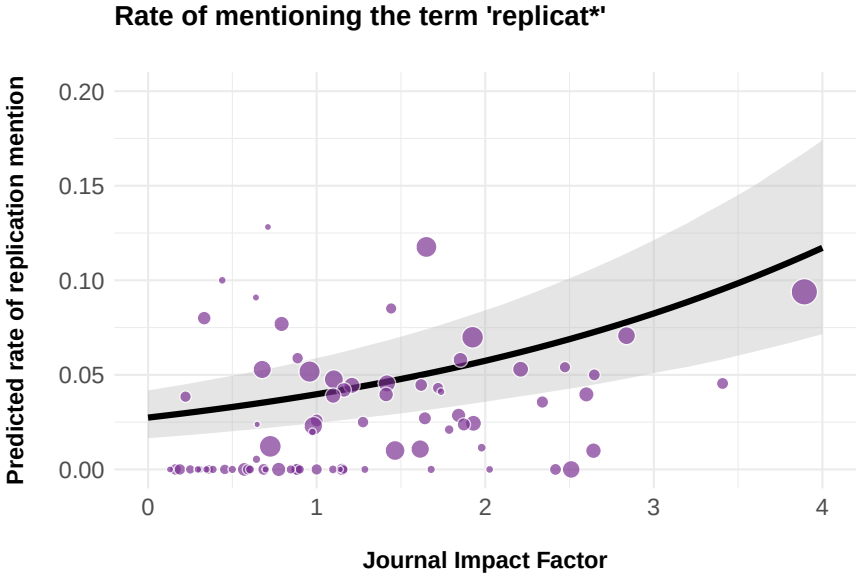
Following preregistered protocol, we estimated the mention rate as predicted relative to the following factors: journal impact factors (continuous), open access (binary: open access journal or not), and replication policies (binary: either explicitly encourage or not). We used Bayesian parameter estimation based on generalized linear regression models with a binomial link function. The model will be 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 zero (scale = 2.5) for all population-level regression coefficients. These priors are what is referred to as regularizing (Gelman et al., 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).

The model estimates a replication mention rate of 0.03 [0.02, 0.04] at a JIF of 0 and estimates a rather robust increase of the rate with each unit of JIF (log odds = 0.39 [0.29, 0.49]). FigureX illustrates this relationship.

(ref:plot_mention_jif) Rate of mentioning the term 'replic*' across sampled journals plotted against journal impact factor. Each point represents one journal. Point size indicates the ratio of papers categorized as experimental (i.e. larger points indicate journals with more experimental articles. Line and shading indicate model predictions and 95% credible intervals.

\begin{figure}



{

}

\caption{(ref:plot_mention_jif)} \end{figure}

Given the small number of journals that explicitly encourage direct replications (2 out of 98),

```
##                             Journals Ratio of experimental studies
## 1       JOURNAL OF MEMORY AND LANGUAGE                         60.34
## 2                        MENTAL LEXICON                         45.71
## 3                  LANGUAGE ACQUISITION                         39.61
## 4  LANGUAGE COGNITION AND NEUROSCIENCE                         38.81
## 5                  LABORATORY PHONOLOGY                         37.42
## 6    LANGUAGE LEARNING AND DEVELOPMENT                         36.17
## 7         NATURAL LANGUAGE ENGINEERING                         32.05
## 8    LECTURE NOTES IN COMPUTER SCIENCE                         30.67
## 9                LANGUAGE AND COGNITION                         29.17
## 10                  INTERACTION STUDIES                         27.88
```

```
##    Rate of replication mention
## 1                         9.39
## 2                         2.08
## 3                         1.22
## 4                         6.99
## 5                         5.17
## 6                        11.76
## 7                         1.00
## 8                         0.00
## 9                         4.76
## 10                        2.30
```

```
##                                       Journals Rate of replication mention
## 1 HUMOR INTERNATIONAL JOURNAL OF HUMOR RESEARCH                      12.82
```

```
##                                                                                                journals
## 1                                          HUMOR INTERNATIONAL JOURNAL OF HUMOR RESEARCH
## 2                                                    LANGUAGE LEARNING AND DEVELOPMENT
## 3                                                                                MORPHOLOGY
## 4 TRANSLATION & INTERPRETING THE INTERNATIONAL JOURNAL OF TRANSLATION AND INTERPRETING
## 5                                                JOURNAL OF LOGIC LANGUAGE AND INFORMATION
## 6                                                    JOURNAL OF MEMORY AND LANGUAGE
##   policy
## 1      3
## 2      3
## 3      3
## 4      3
## 5      2
## 6      3
```

```
##                                                                                                webpag
## 1                                                http://www.humorstudies.org/JournalCenter.ht
## 2 https://www.tandfonline.com/action/authorSubmission?show=instructions&journalCode=hlld2
## 3                                                                https://www.springer.com/journal/1152
## 4                http://trans-int.org/index.php/transint/about/submissions#authorGuideline
## 5                                      https://www.springer.com/journal/10849/aims-and-scop
## 6                          https://www.journals.elsevier.com/journal-of-memory-and-language
##   binary_policy             jif openaccess
## 1             0           0.711    partial
## 2             0           1.651    partial
## 3             0 not retrievable    partial
## 4             0 not retrievable  DOAJ gold
## 5             0           0.440    partial
## 6             0           3.893    partial
```

Discussion

- too little replication attempts in experimental linguistics

- journals guidelines generally don't encourage replication studies
- ...

Detailed analysis: types and contributing factors

The second part of the study aimed 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 contained the search term "replicat*" required further analysis to determine whether the articles in question indeed reported a replication study or used the term in a different way.

We were interested in which kinds of replication studies are published and which factors contribute to their publication. We aimed at investigating what types of replication studies are prevalent in the field. We were 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.

Material and methods

The material and methods have been preregistered on Open Science Framework and can be inspected here: https://osf.io/9ceas/.

From the superset of 100 journals obtained above, the first 20 journals (i.e. those journals 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 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 selected (see here for details) 50 out of the 114 articles for the Journal of Memory and Language to achieve a more balanced distribution of papers across journals. The sampling procedure above resulted in 210 possible self-labeled replication studies.

In a first step, we identified whether the article in question indeed presented a replication study or not. The relevant parts of the papers were title and abstract of the paper, sentences around occurrences of the search term "replicat" as well as the paragraph before the Methods section and the first paragraph of the Discussion section (following the procedure specified by Makel et al. (2016)). If the authors explicitly claimed that (one of) their research aim(s) was to replicate or reproduce findings or methods of an initial study, this article was treated as a replication. It then qualified for further analysis after the coding scheme that can be viewed here: 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. The replication studies were 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 noted 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. We also noted 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 one of the authors was a (co-)author on both articles.

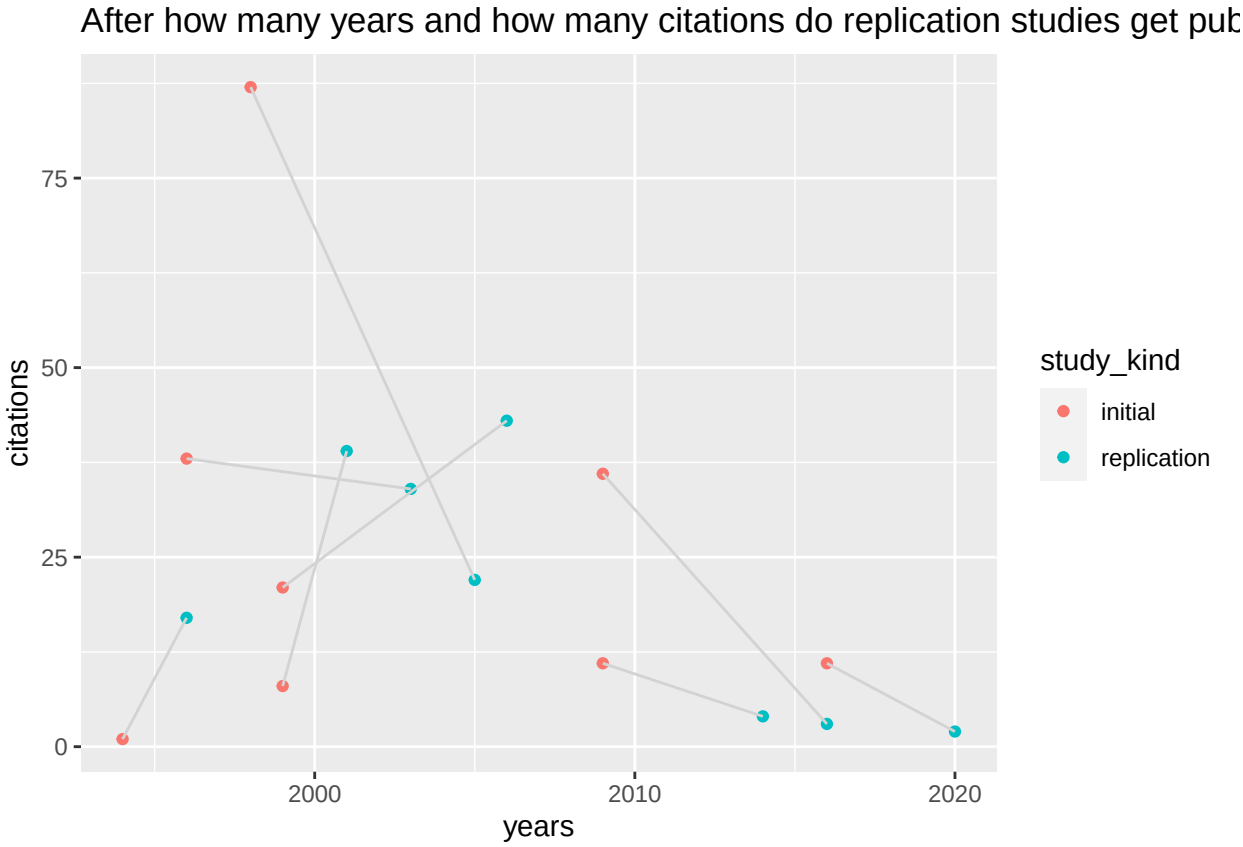
Results

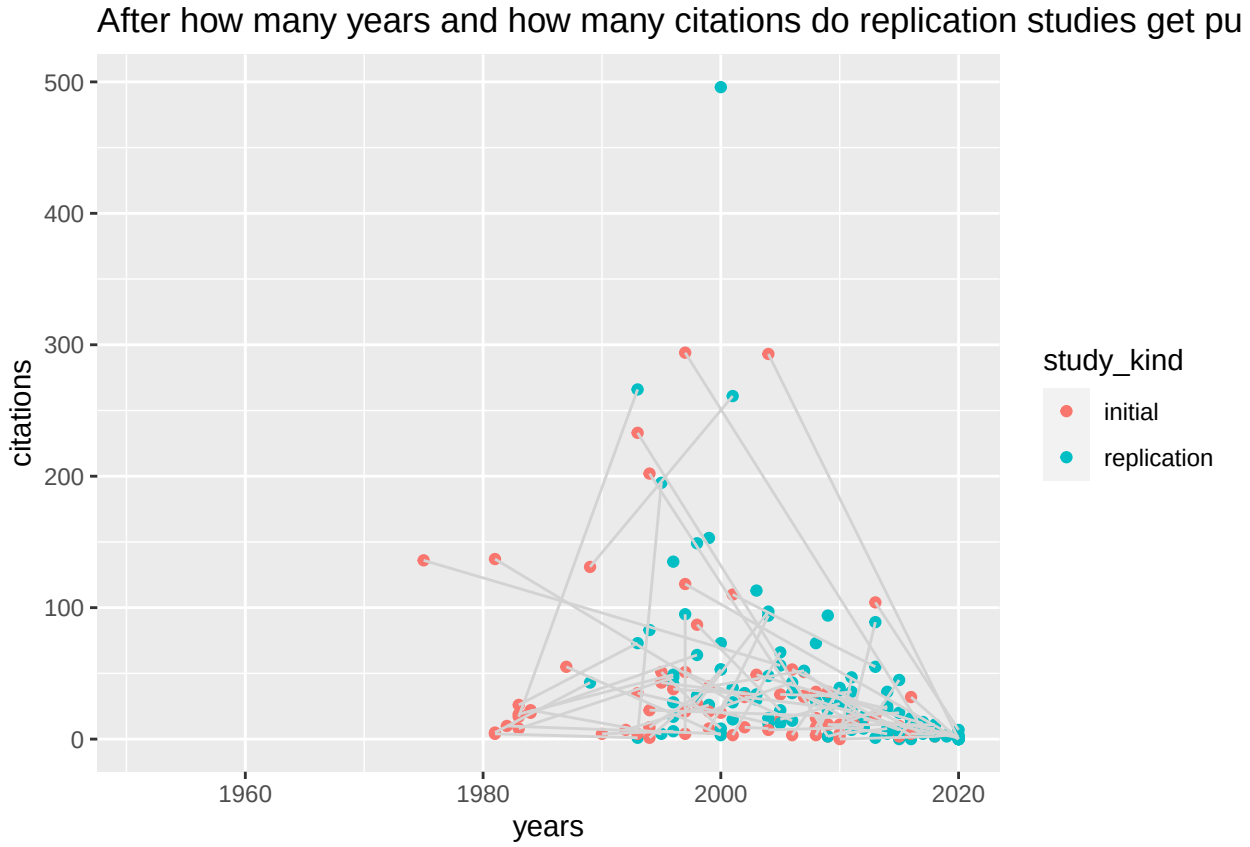
```
##   mean_years
## 1   8.810127
```

```
##   median_years
## 1            7
```

```
##   mean_cit_init
## 1      41.08108
```

```
##   median_cit_init
## 1            19.5
```


After how many years and how many citations do replication studies get pub
citations
75
50
25
0
2000
2010
2020
years
study_kind
initial
replication




```
## type_replication
## conceptual     direct    partial
##       0.44       0.09       0.47

##                  auth_overlap
## type_replication    0    1
##       conceptual 0.18 0.26
##       direct     0.06 0.03
##       partial    0.24 0.24

## [1] 0.04
```

Discussion

General discussion

- compare rate of replication mention to previous studies in different fields –> broader picture

Caveats

This procedure is necessarily only a rough proxy of relevant experimental linguistic articles published in the field and several articles might thus have been overlooked and not been included in the analysis.

Appendices

identified as A, B, etc.

References

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

Camerer, C.F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., Wu, H., 2016. Evaluating replicability of laboratory experiments in economics. Science 351, 1433–1436. doi:10.1126/science.aaf0918

Camerer, C.F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johanesson, M., Kirchler, M., Nave, G., Nosek, B.A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., Isaksson, S., Manfredi, D., Rose, J., Wagenmakers, E.-J., Wu, H., 2018. Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. Nature 2, 637–644. doi: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, 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, 1360–1383.

Makel, M.C., Plucker, J.A., 2014. Facts are more important than novelty: Replication in the education sciences. Educational Researcher 43, 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, 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, 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, 321–391. doi:gc3h3b

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

Nieuwland, M.S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., Zu Wolfsthurn, S.V.G., Bartolozzi, F., Kogan, V., Ito, A., Mézière, D., Barr, D.J., Rousselet, G.A., Ferguson, H.J., Bush-Moreno, S., Fu, X., Tuomainen, J., Kulakova, E., Husband, M.E., Donaldson, D.I., Kohút, Z., Rueschemeyer, S.-A., Huettig, F., 2018. Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. eLife 7, e33468. doi:10.7554/eLife.33468.001

Open Science Collaboration, 2015. Estimating the reproducibility of psychological science. Science 349. doi: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, e116–e141. doi: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, 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, 864–877. doi: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, 1–12.