The benefits of preregistration for hypothesis-driven bilingualism research

Daniela Mertzen
Department of Linguistics, University of Potsdam, Potsdam, Germany
Sol Lago
Institute for Romance Languages and Literatures, Goethe University Frankfurt, Frankfurt, Germany
Shravan Vasishth
Department of Linguistics, University of Potsdam, Potsdam, Germany

Supplementary material

For data and code accompanying this paper, visit https://osf.io/5ab7d/

***Acknowledgements**

We thank João Veríssimo, Laura de Ruiter, Cylcia Bolibaugh, and Luke Plonsky for their valuable feedback on the earlier version of this paper. This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 317633480 – SFB 1287, Projects B03 and Q (PIs: Shravan Vasishth and Ralf Engbert).

Correspondence
Email: mertzen@uni-potsdam.de

Keywords: preregistration; open science; bilingualism; psycholinguistics; confirmatory analysis; exploratory analysis

Abstract

Preregistration is an open science practice that requires the specification of research hypotheses and analysis plans *before* the data are inspected. Here, we discuss the benefits of preregistration for hypothesis-driven, confirmatory bilingualism research. Using examples from psycholinguistics and bilingualism, we illustrate how non-peer reviewed preregistrations can serve to implement a clean distinction between hypothesis testing and data exploration. This distinction helps researchers avoid casting post-hoc hypotheses and analyses as confirmatory ones. We argue that, in keeping with current best practices in the experimental sciences, preregistration, along with sharing data and code, should be an integral part of hypothesis-driven bilingualism research.

1. Introduction

An important aspect of hypothesis-driven research is *preregistration*, an open science practice that consists of the specification of research question(s), method(s) and analysis plan(s) before data collection. Preregistration is a relatively simple yet powerful tool for improving transparency in bilingualism research, and we suggest that, in keeping with current best practices in the experimental sciences, bilingualism researchers include preregistration as an essential component of hypothesis-driven research, along with other open science practices such as releasing materials, data and code alongside publications (Chambers, Feredoes, Muthukumaraswamy & Etchells, 2014; Nosek, Ebersole, DeHaven & Mellor, 2018b; Nosek & Lakens, 2014; Nosek, Ebersole, DeHaven & Mellor, 2018a; Open Science Collaboration, 2015).

There are several positions regarding the goals of preregistration. Many researchers view it as a tool specific to confirmatory research because it can help assess the falsifiability of an experimental study's predictions, control for false positive error probability in null hypothesis significance testing (NHST), and mitigate researcher biases (e.g., Lakens, 2019; Chambers, 2019; Nosek et al., 2019). Under this view, preregistration helps implement the distinction between confirmatory analyses (used for hypothesis testing) and exploratory analyses (used for hypothesis generation) (e.g., de Groot, 1956/2014; Chambers, 2019; Nosek et al., 2018b; Nosek et al., 2019; Wagenmakers, Wetzels, Borsboom, van der Maas & Kievit, 2012). More recently, preregistration has also been considered for qualitative research with the aim to make documentation of research plans more transparent (Haven & Grootel, 2019). Other research groups acknowledge the contribution of preregistration to scientific transparency, but call into question the validity of the distinction between confirmatory

and exploratory research, and the usefulness of preregistration to help implement this distinction (e.g., Devezer, Navarro, Vandekerckhove & Buzbas, 2020; Szollosi et al., 2020; Szollosi & Donkin, 2019, cf. Wagenmakers, 2019). From this point of view, a shift to the development of more explicit theories would make preregistration unnecessary.

In this paper, we take the position that preregistration is crucial to separate confirmatory from exploratory analyses. In our view, the preregistration of confirmatory hypotheses can counter questionable research practices and unconscious biases (Box 1). Consequently, it can enhance research transparency in confirmatory bilingualism (L2) research. Concerns about (non-)transparency and researcher biases are well-known in psychological science (Wicherts, Borsboom, Kats & Molenaar, 2006; Simmons, Nelson & Simonsohn, 2011). L2 research is similarly affected by a lack of clarity about pre-data collection hypotheses and analysis plan choices. This problem is compounded by the fact that L2 studies rarely release their research materials (Derrick, 2016; Marsden, Thompson & Plonsky, 2018) or their data (Larson-Hall & Plonsky, 2015; Bolibaugh, Vanek & Marsden, 2020).

To address these issues, two journals in the field of bilingualism, *Language Learning* and *Bilingualism: Language and Cognition*, have introduced a new type of article, Registered Reports, which allows researchers to submit their hypotheses, methods, and analysis protocols for peer review prior to data collection (Marsden, Morgan-Short, Trofimovich & Ellis, 2018). Here, we discuss a different approach: non-peer reviewed preregistration using open science platforms such as the Open Science Framework (OSF, https://osf.io/) or AsPredicted (https://aspredicted.org/). On these platforms, researchers have the opportunity to create a public or private, time-

stamped, non-modifiable record of a planned study prior to data inspection, either before or during data collection. Here, we argue that non-peer reviewed preregistration can counteract the questionable research practices presented below. We first illustrate them with an example from our own work on native (L1) sentence processing. Then, we discuss correlates in the L2 literature and explain how non-peer reviewed preregistrations can improve L2 research.

Box 1. Three questionable research practices and biases

- **The garden of forking paths**

In hypothesis-driven research, there are many possible data analysis paths, and one of several potential paths can be selectively chosen and reported (Gelman & Loken, 2013, 2014). For example, one could choose a particular measure, region of interest or time-window that was not originally selected for analysis, or delete outliers based on an arbitrary criterion. Such multiple analysis paths cumulatively create so many researcher degrees of freedom that one can describe them using a decision tree. This bias is often an unconscious one (Gelman & Loken, 2013, pp. 9-10):

> It's not that the researchers performed hundreds of different comparisons and picked ones that were statistically significant. Rather, they start with a somewhat-formed idea in their mind of what comparison to perform, and they refine that idea in light of the data. (...) they are using their scientific common sense to formulate their hypotheses in a reasonable way, given the data they have. The mistake is in thinking that, if the particular path that was chosen yields statistical significance, that this is strong evidence in favor of the hypothesis.

- **Multiple testing**

For purely statistical reasons, if one conducts enough statistical tests, some test will eventually come out significant. For example, in psycholinguistic eye-tracking reading research, one can easily end up conducting dozens of statistical tests to evaluate a single hypothesis. Simulations in von der Malsburg and Angele (2017) demonstrate that multiple analyses in eye-tracking dramatically inflate Type I error, leading to a large proportion of false positive rejections of the null hypothesis.

- **Post-hoc hypothesizing**

When data is analyzed without having explicitly stated the predictions, one may easily convince oneself that an unforeseen result was expected all along, and subsequently report this unexpected finding as a confirmatory one. This bias is commonly referred to as 'hypothesizing after the results are known' (*HARKing*) (Simmons et al., 2011; Kerr, 1998). This can skew the scientific record with less well-grounded theories, cherry-picked *after the fact* (Chambers, 2019).

2. Possible pitfalls of hypothesis-driven research: An example from L1 sentence processing

We briefly introduce our study, which attempted to replicate the findings of an eye-tracking reading study that compared the processing of two different syntactic dependencies (Dillon, Mishler, Sloggett & Phillips, 2013; Jäger, Mertzen, Van Dyke & Vasishth, 2020). This example can be easily translated to bilingualism settings where, similar to our example, processing patterns are investigated for different syntactic constructions, but for different speaker groups, such as native vs. non-native speakers (Felser & Cunnings, 2012; Grüter, Lew-Williams & Fernald, 2012), or successive vs. simultaneous learners (Lemmerth & Hopp, 2019; Sabourin & Vīnerte, 2015).

Our example concerns a phenomenon called *agreement attraction*. For subject-verb agreement dependencies, previous work has shown that a processing disruption elicited by an ungrammatical plural verb can be weakened if a plural noun (an "attractor") intervenes between the subject and the verb (as in 1a vs. 1b; Wagers, Lau & Phillips, 2009; Pearlmutter, Garnsey & Bock, 1999; Dillon et al., 2013). Dillon and colleagues used a within-subjects design to examine whether the attraction effect extended to ungrammatical antecedent-reflexive dependencies, where an attractor matched the reflexive in number (1c vs. 1d).

(1) a. *Subject-verb agreement; attraction*

*The amateur bodybuilder who worked with the **personal trainers** amazingly were competitive for the gold medal.

b. *Subject-verb agreement; no attraction*

*The amateur bodybuilder who worked with the **personal trainer** amazingly were competitive for the gold medal.

c. *Reflexive; attraction*

*The amateur bodybuilder who worked with **the personal trainers** amazingly injured themselves on the lightest weights.

d. *Reflexive; no attraction*

*The amateur bodybuilder who worked with **the personal trainer** amazingly injured themselves on the lightest weights.

Building on work by Sturt (2003), they argued that, unlike subject-verb agreement configurations, the processing of antecedent-reflexive dependencies should be syntactically constrained (Chomsky, 1981). If so, attraction effects were expected in subject-verb dependencies but not in antecedent-reflexive dependencies, yielding an interaction between dependency type and attraction.

Dillon et al. (2013) analyzed multiple reading measures and observed the predicted interaction only in total reading time. This result was taken as support for the hypothesis that subject-verb agreement and reflexives show different susceptibility to agreement attraction, and thus are differentially constrained by syntactic principles. In our large-sample replication study (Jäger et al., 2020), the goal was to replicate the statistically significant interaction in total reading time from the original study. Our confirmatory analysis of total reading time showed no effect, while the exploratory analyses of first-pass regressions and regression-path durations did (Table 1).

	Dillon et al. results (N = 40)			**Jäger et al. replication (N = 181)**		
Dependent measure	Estimate	Standard Error	z/t value	Estimate	Standard Error	z/t value
First fixation duration	1.74	5.54	0.31	-2.54	4.22	-0.60
First pass reading time	0.46	16.16	0.03	-3.14	7.74	-0.41
First pass regressions	-0.07	0.19	-0.35	**-0.20**	**0.09**	**-2.30**
Regression-path duration	-5.07	30.74	-0.16	**-44.02**	**21.09**	**-2.09**
Re-reading time	-64.86	39.9	-1.63	5.28	13.11	0.40
Total reading time	**-54.72**	**25.02**	**-2.19**	2.19	14.24	0.15

Table 1. *Comparison of the findings by Dillon et al. (2013) and Jäger et al. (2020). The table shows the interaction effect of Dependency type × Attraction, computed using generalized linear mixed models (effects on first-pass regressions were estimated using a logit link function). The interaction effect was expected to have a negative sign. Significant effects at a 0.05 α-level are shown in bold. Note that the published analyses in Jäger et al. (2020) differ from the ones we present here due to different model assumptions made in the present paper for expository purposes.*

The study by Dillon and colleagues and our attempted replication serve to illustrate the potential issues of the garden of forking paths, multiple testing and posthoc theorizing. First, even for a confirmatory replication study, where one analyzes the same region and reading measure that showed the interaction in the original study, garden of forking paths scenarios arise if an analysis path is not defined prior to data inspection. For example, different decisions regarding statistical tests and outlier treatment could still be made after data inspection.

Second, for the analyses of the Dillon et al. study and our replication study, six statistical tests were conducted. Testing six eye-tracking measures increases the Type I error probability from 5% to 26.5% (i.e., $1 - 0.95^6 = 0.265$) (Bonferroni, 1936). It is possible to correct for multiple testing. For example, a Bonferroni correction would

require an adjusted Type I error of 0.05⁄6 for the six statistical tests we conducted, which implies that the absolute critical t-/z-value would be 2.64. If this criterion were used, there would be no significant effects in either the original study or the replication attempt (see observed z/t-values in Table 1). A better solution to the multiple testing problem may be to avoid it altogether by having precise predictions about the dependent measure(s), and focus on (Bayesian) estimation of effects rather than NHST (e.g., Norouzian, 2020; Gelman & Carlin, 2014; Gelman et al., 2014; Kruschke, 2014).

Third, suppose that the effect that was expected a priori at the critical auxiliary verb or the reflexive had been found further downstream in the sentence or even before the critical region. Without specifying the critical region in advance, one could easily have found a post-hoc theory for the effect showing up in another region and reported this as if it had been predicted all along.

Finally, both the original and the replication study show some evidence of the effect of interest. However, the effect occurs in different measures across the two studies. Because of the exploratory nature of the first-pass regression and regression-path duration results in the replication attempt, we cannot treat these hypothesis tests as confirmatory ones. Exploratory analyses per se are an important part of doing science, but they should be presented as such (e.g., Bishop, 2020; de Groot, 1956/2014; Nosek et al., 2018b).

3. Problematic research practices in L2 research

The issues above can also arise in L2 research. Two common examples of forks in the analysis path are outlier treatment and the selection of interest regions in reading studies. For example, a synthesis of methodological decisions in L2 self-paced reading (SPR) research showed a variety of outlier removal criteria across 64 studies, such as standard

deviations around the mean, reading time cutoffs, or both (Marsden, Thompson & Plonsky, 2018; see Nicklin & Plonsky, 2020, for discussion of outlier treatment). Moreover, L2 reading studies on the same grammatical phenomena can vary substantially in their selection of interest regions. For a subset of the L2 SPR studies on local ambiguity processing synthesized in Marsden, Thompson and Plonsky (2018), some studies reported statistical analyses for the ambiguous sentence region, and other studies for some, or all, of the subsequent regions. In addition, the critical regions varied between studies, consisting of a single word or several words combined.

A closely related problem to the selective reporting of interest regions is conducting statistical tests for many different regions, and/or eye-tracking measures. Godfroid (2020) reported that an average of 3.4 eye-tracking measures per study are analyzed in the L2 eye-tracking literature, further inflating Type I error probability. The Type I error issue might be particularly prevalent in L2 studies because many of them use frequentist NHST and only report binary decisions about the presence or absence of an effect without also reporting effect estimates (Marsden, Thompson & Plonsky, 2018). One unfortunate consequence is that other researchers cannot gain knowledge about the magnitude of an effect across studies, or conduct meta-analyses due to the lack of information from previous studies (Plonsky, 2013; Larson-Hall & Plonsky, 2015; Plonsky & Oswald, 2014; Al-Hoorie & Vitta, 2019; for an introduction to meta-analyses in bilingualism research, see Plonsky & Oswald, 2015; Plonsky, Sudina & Hu, 2020).

Finally, as in our example on L1 processing, post-hoc hypothesizing, i.e., changing a hypothesis to match the findings, may reduce the reproducibility of L2 research (Marsden, Morgan-Short, Trofimovich & Ellis, 2018; Marsden, Morgan-Short, Thompson & Abugaber, 2018; Chambers, 2019). Possibly partly due to the issues raised

above, and low statistical power (Cohen, 1962, 1988; Brysbaert, 2020), inconsistent findings also occur in L2 research. Some examples include the role of crosslinguistic influence in syntactic processing (Dussias, Dietrich & Villegas, 2015; Lago, Mosca & Stutter Garcia, 2020), the existence of a bilingual advantage in attentional systems (Bialystok, 2017; Paap, Anders-Jefferson, Mason, Alvarado & Zimiga, 2018), and the role of morphological decomposition in inflected vs. derived forms during word recognition in native vs. non-native speakers (Clahsen & Veríssimo, 2016; Feldman & Kroll, 2019). Next, we discuss how a non-peer reviewed preregistration can be implemented to improve L2 research.

4. Non-peer reviewed preregistration in psycholinguistic research

For preregistration to counter questionable research practices and biases, it is not sufficient to a priori specify the dependent measure(s), because many researcher degrees of freedom remain. A complete preregistration requires a full description of the research questions and hypotheses, study design, methods, speaker group selection criteria, data collection procedure, participant sample size or stopping rule, outcome variable(s), as well as an analysis plan including statistical models, information on data exclusion and statistical inference criteria. This does not only ensure greater transparency, but it can also keep in check one's biases because analysis decisions are made public prior to data analysis, preventing selective reporting of effects. For example, assume that for a planned study we preregister no outlier exclusions, but later find an effect only when removing certain data points. This could be reported as an exploratory finding. Without preregistration, it may be tempting to report the most 'interesting' result as confirmatory, preventing other researchers from evaluating the findings in light of the analysis choices.

In addition, if our published preregistration committed to a predicted effect for a particular region and measure, based on theory or previous findings, we can no longer convince ourselves that a surprising result was originally predicted and restate the hypotheses post-hoc.

One may argue that if one has strong theoretical predictions, preregistration is redundant because the analysis choices are predetermined by the theory. However, Silberzahn et al. (2018) convincingly illustrated that different analysis choices can be made even under highly constraining conditions. Their study recruited 29 research groups in the psychological sciences to answer *the same research question for one particular dataset*. Of the 29 groups, 20 observed a significant and nine a non-significant result. Strikingly, the range of effect estimates reported by the different research groups allowed for different conclusions.

Although we take the view that preregistration without peer review can be an effective way to reduce unconscious biases in one's work, the lack of peer review means that the preregistration of a study can be as thorough or as vague as the researcher deems appropriate. Vaguely specified research plans still allow for many possible analysis paths, and selective reporting of effects. Consequently, it is up to the scientific community to make non-peer reviewed preregistration a success or a failure: only a thoroughly implemented preregistration and a precisely followed research plan can reduce unconscious biases and help to separate confirmatory hypothesis tests from exploratory ones.

4.1 Selecting dependent measures for a preregistration

If one wants to preregister a study, but lacks prior knowledge of a particular phenomenon, an experiment could be piloted and exploratory analyses conducted to identify which measure(s) show the predicted effect. One could then generate hypotheses from this and test them in a confirmatory study (e.g., Nicenboim, Vasishth, Engelmann & Suckow, 2018; Nicenboim, Vasishth & Rösler, 2020). If, on the other hand, there are previous findings on a phenomenon, these could serve as the basis for a preregistration. However, when the literature shows equivocal results as discussed above, what steps could be taken to consolidate the support in favor of or against a theory? This is not straightforward. For example, in the Dillon et al. (2013) study and our replication study, the effect of interest was observed in different reading measures. If, based on linguistic theory, we believe that the effect of interest should be found in earlier reading measures (first-pass regression and regression-path duration as in our replication study), the only way to test this is by conducting a replication study. This replication should aim for a sufficiently large participant sample and a sufficiently precise effect estimate, and specify the dependent measure(s) and critical region(s) in advance. Otherwise, in a future study we may find some other dependent measure showing the effect, which may again tempt us to draw a bullseye around the arrow that happened to land where it did.

4.2 How to get started with a non-peer reviewed preregistration

Preregistration templates are available on OSF and AsPredicted for novel studies as well as for replication studies (e.g., https://bit.ly/OSFtemplates; https://bit.ly/AsPredtemplate). If one prefers to create a Registered Report-type preregistration (i.e., in manuscript format), it is possible to upload a preregistration manuscript on OSF. It is not enough to

upload this document to the project's public repository, because the preregistration could be removed or replaced at any point. Rather, one needs to create a time-stamped, non-editable version which can be made public either immediately or it can be embargoed until, for example, the associated paper is submitted or published. If the preregistration is withdrawn at any stage after creating a "frozen" version of it, some meta data (title, authors, description, reason for withdrawing preregistration) will remain publicly available. A new version of the preregistration can be made available before the data are inspected. We have previously made attempts at such manuscript-style preregistrations, e.g., for Vasishth, Mertzen, Jäger and Gelman (2018) (see https://osf.io/dgewb for the non-editable preregistration).

5. Conclusion

We have used examples from L1 sentence processing and the L2 literature to illustrate some of the problems that can arise during the research process. We then discussed how preregistration allows researchers to better separate confirmatory and exploratory analyses, which can help them counter questionable research practices and unconscious biases. Our view is that, if done thoroughly, non-peer reviewed preregistration would greatly benefit the bilingualism community. We suggest that the hypothesis-driven L2 research process should standardly include preregistration, in addition to the release of materials, data and code upon publication to increase research transparency and reproducibility.

Competing interests: The authors declare none.

References

Bialystok, E. (2017). The bilingual adaptation: How minds accommodate experience. *Psychological Bulletin, 143*(3), 233–262. doi:10.1037/bul0000099

Bishop, D. V. (2020). The psychology of experimental psychologists: Overcoming cognitive constraints to improve research: The 47th Sir Frederic Bartlett Lecture. *Quarterly Journal of Experimental Psychology*, *73*(1), 1–19. doi:10.1177/1747021819886519

Bolibaugh, C., Vanek, N., & Marsden, E. (2020). *Towards a credibility revolution in bilingualism research: Open data and materials as stepping stones*. Submitted for publication.

Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze,* 8, 3-62.

Brysbaert, M. (2020). Power considerations in bilingualism research: Time to step up our game. *Bilingualism: Language and Cognition*, 1–6. doi:10.1017/S1366728920000437

Chambers, C. D. (2019). *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice*. Princeton, NJ: Princeton University Press.

Chambers, C. D., Feredoes, E., Muthukumaraswamy, S. D., & Etchells, P. J. (2014). Instead of "playing the game" it is time to change the rules: Registered reports at AIMS Neuroscience and beyond. *AIMS Neuroscience*, *1*, 4–17. doi:10.3934/Neuroscience2014.1.4

Chomsky, N. (1981). *Lectures on government and binding*. Dordrecht: Foris.

Clahsen, H., & Veríssimo, J. (2016). Investigating grammatical processing in bilinguals: The case of morphological priming. *Linguistic Approaches to Bilingualism*, *6*(5), 685–698. doi:https://doi.org/10.1075/lab.15039.cla

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology*, *65*(3), 145.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

de Groot, A. (1956/2014). The meaning of "significance" for different types of research [translated and annotated by Eric-Jan Wagenmakers, Denny Borsboom, Josine Verhagen, Rogier Kievit, Marjan Bakker, Angelique Cramer, Dora Matzke, Don Mellenbergh, & Han L. J. van der Maas]. *Acta Psychologica*, *148*, 188–194. doi:https://doi.org/10.1016/j.actpsy.2014.02.001

Derrick, D. J. (2016). Instrument reporting practices in second language research. *TESOL Quarterly*, *50*(1), 132–153. doi:10.1002/tesq.217

Devezer, B., Navarro, D. J., Vandekerckhove, J., & Buzbas, E. O. (2020). The case for formal methodology in scientific reform. *bioRxiv*. doi:10.1101/2020.04.26.048306

Dillon, B., Mishler, A., Sloggett, S., & Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: Experimental and modeling evidence. *Journal of Memory and Language*, *69*(2), 85–103. doi:https://doi.org/10.1016/j.jml.2013.04.003

Dussias, P., Dietrich, A. J., & Villegas, Á. (2015). Cross-language interactions during bilingual sentence processing. In J. W. Schwieter (Ed.), *The Cambridge Handbook of Bilingual Processing* (pp. 349–366). Cambridge Handbooks in Language and Linguistics. Cambridge University Press. doi:10.1017/CBO9781107447257.016

Feldman, L., & Kroll, J. (2019). Learning and Using Morphology and Morphosyntax in a Second Language. *Oxford Research Encyclopedia of Linguistics.* Oxford University Press. Retrieved from https://oxfordre.com/linguistics/view/10.1093/acrefore/9780199384655.001.0001/acrefore-9780199384655-e-604. doi: https://doi.org/10.1093/acrefore/9780199384655.013.604

Felser, C., & Cunnings, I. (2012). Processing reflexives in a second language: The timing of structural and discourse-level constraints. *Applied Psycholinguistics*, *33*(3), 571–603. doi:10.1017/S0142716411000488

Gelman, A., & Carlin, J. B. (2014). Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641–651.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (Third). Boca Raton, FL: Chapman and Hall/CRC.

Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no 'fishing expedition' or 'p-hacking' and the research hypothesis was posited ahead of time. Retrieved from http: //www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, *102*. doi:460.10.1511/2014.111.460

Godfroid, A. (2020). *Eye Tracking in Second Language Acquisition and Bilingualism: A Research Synthesis and Methodological Guide*. New York, NY: Routledge. doi:10.4324/9781315775616

Grüter, T., Lew-Williams, C., & Fernald, A. (2012). Grammatical gender in L2: A production or a real-time processing problem? *Second Language Research*, *28*(2), 191–215. doi:10.1177/0267658312437990

Haven, T. L., & Grootel, D. L. V. (2019). Preregistering qualitative research. *Accountability in Research*, *26*(3), 229–244. doi:10.1080/08989621.2019.1580147

Al-Hoorie, A. H., & Vitta, J. P. (2019). The seven sins of L2 research: A review of 30 journals' statistical quality and their CiteScore, SJR, SNIP, JCR Impact Factors. *Language Teaching Research*, *23*(6), 727–744. doi:10.1177/1362168818767191

Jäger, L. A., Mertzen, D., Van Dyke, J. A., & Vasishth, S. (2020). Interference patterns in subject-verb agreement and reflexives revisited: A large-sample study. *Journal of Memory and Language*, *111*. doi:https://doi.org/10.1016/j.jml.2019.104063

Kerr, N. L. (1998). Harking: Hypothesizing After the Results are Known. *Personality and Social Psychology Review*, *2*(3), 196–217. doi:10.1207/s15327957pspr0203_4

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. London: Academic Press.

Lago, S., Mosca, M., & Stutter Garcia, A. (2020). The role of crosslinguistic influence in multilingual processing: Lexicon versus syntax. *Language Learning*. doi:10.1111/lang.12412

Lakens, D. (2019). The value of preregistration for psychological science: A conceptual analysis. PsyArXiv. doi:10.31234/osf.io/jbh4w

Larson-Hall, J., & Plonsky, L. (2015). Reporting and interpreting quantitative research findings: What gets reported and recommendations for the field. *Language Learning*, *65*(S1), 127–159. doi:10.1111/lang.12115

Lemmerth, N., & Hopp, H. (2019). Gender processing in simultaneous and successive bilingual children: Cross-linguistic lexical and syntactic influences. *Language Acquisition*, *26*(1), 21–45. doi:10.1080/10489223.2017.1391815

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. doi:10.1111/lang.12286

Marsden, E., Morgan-Short, K., Trofimovich, P., & Ellis, N. C. (2018). Introducing Registered Reports at Language Learning: Promoting transparency, replication, and a synthetic ethic in the language sciences. *Language Learning*, *68*(2), 309– 320. doi:10.1111/lang.12284

Marsden, E., Thompson, S., & Plonsky, L. (2018). A methodological synthesis of self-paced reading in second language research. *Applied Psycholinguistics*, *39*(5), 861–904. doi:10.1017/S0142716418000036

Nicenboim, B., Vasishth, S., Engelmann, F., & Suckow, K. (2018). Exploratory and confirmatory analyses in sentence processing: A case study of number interference in German. *Cognitive Science*, *42*. doi:10.1111/cogs.12589

Nicenboim, B., Vasishth, S., & Rösler, F. (2020). Are words pre-activated probabilistically during sentence comprehension? Evidence from new data and a Bayesian random-effects meta-analysis using publicly available data. *Neuropsychologia*, *142*, 107427. doi:https://doi.org/10.1016/j.neuropsychologia.2020.107427

Nicklin, C., & Plonsky, L. (2020). Outliers in L2 Research in Applied Linguistics: A Synthesis and Data Re-Analysis. *Annual Review of Applied Linguistics*, *40*, 26–55. doi:10.1017/S0267190520000057

Norouzian, R. (2020). Sample size planning in quantitative L2 research: A pragmatic approach. *Studies in Second Language Acquisition*, 1–22. doi:10.1017/S0272263120000017

Nosek, B. A., Beck, E. D., Campbell, L., Flake, J. K., Hardwicke, T. E., Mellor, D. T., ... Vazire, S. (2019). Preregistration is hard, and worthwhile. *Trends in Cognitive Sciences*, *23*(10), 815–818. doi:https://doi.org/10.1016/j.tics.2019.07.009

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018a). Reply to Ledgerwood: Predictions without analysis plans are inert. *Proceedings of the National Academy of Sciences*, *115*(45), E10518–E10518. doi:10.1073/pnas.1816418115

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018b). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600–2606. doi:10.1073/pnas.1708274114

Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology*, *45*, 137–141. doi:http://dx.doi.org/10.1027/1864-9335/a000192

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Paap, K. R., Anders-Jefferson, R., Mason, L., Alvarado, K., & Zimiga, B. (2018). Bilingual advantages in inhibition or selective attention: More challenges. *Frontiers in Psychology*, *9*, 1409. doi:10.3389/fpsyg.2018.01409

Pearlmutter, N. J., Garnsey, S. M., & Bock, K. (1999). Agreement processes in sentence comprehension. *Journal of Memory and Language*, *41*(3), 427–456. doi:https://doi.org/10.1006/jmla.1999.2653

Plonsky, L. (2013). Study quality in SLA: An Assessment of Designs, Analyses, and Reporting Practices in Quantitative L2 Research. *Studies in Second Language Acquisition*, *35*(4), 655–687. doi:10.1017/S0272263113000399

Plonsky, L., & Oswald, F. L. (2014). How Big Is 'Big'? Interpreting Effect Sizes in L2 Research. *Language Learning*, *64*(4), 878–912. doi:10.1111/lang.12079

Plonsky, L., & Oswald, F. (2015). Meta-analyzing second language research. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 106–128). Routledge: New York, NY, USA.

Plonsky, L., Sudina, E., & Hu, Y. (2020). *Applying meta-analysis to research on bilingualism: An introduction*. In press.

Sabourin, L., & Vīnerte, S. (2015). The bilingual advantage in the stroop task: Simultaneous vs. early bilinguals. *Bilingualism: Language and Cognition*, *18*(2), 350–355. doi:10.1017/S1366728914000704

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., ... Nosek, B. A. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337–356. doi:10.1177/2515245917747646

Simmons, J., Nelson, L., & Simonsohn, U. (2011). False-positive psychology. *Psychological Science*, *22*(11), 1359–1366. doi:https://doi.org/10.1177/0956797611417632

Sturt, P. (2003). The time-course of the application of binding constraints in reference resolution. *Journal of Memory and Language, 48,* 542–562.

Szollosi, A., & Donkin, C. (2019). *Arrested theory development: The misguided distinction between exploratory and confirmatory research*. doi:10.31234/osf.io/suzej

Szollosi, A., Kellen, D., Navarro, D. J., Shiffrin, R., van Rooij, I., Van Zandt, T., & Donkin, C. (2020). Is preregistration worthwhile? *Trends in Cognitive Sciences*, *24*(2), 94–95. doi:https://doi.org/10.1016/j.tics.2019.11.009

Vasishth, S., Mertzen, D., Jäger, L. A., & Gelman, A. (2018). The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language*, *103,* 151–175. doi:https://doi.org/10.1016/j.jml.2018.07.004

von der Malsburg, T., & Angele, B. (2017). False positives and other statistical errors in standard analyses of eye movements in reading. *Journal of Memory and Language*, *94*, 119–133. doi:https://doi.org/10.1016/j.jml.2016.10.003

Wagenmakers, E.-J. (2019). A breakdown of "preregistration is redundant, at best". https://www.bayesianspectacles.org/a-breakdown-of-preregistration-is-redundant-at-best/. Accessed: 2020-08-15.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*(6), 632–638. doi:10.1177/1745691612463078

Wagers, M. W., Lau, E. F., & Phillips, C. (2009). Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language*, *61*(2), 206–237. doi:https://doi.org/10.1016/j.jml.2009.04.002

Wicherts, J., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *The American Psychologist, 61,* 726–728. doi:10.1037/0003-066X.61.7.726