Submitted to Bilingualism: Language and Cognition

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

February 27, 2020

Abstract

Preregistration is an open science practice that focuses on statistically testing hypotheses specified before the data are inspected (hypothesis-driven, confirmatory research). Here, we discuss the benefits of preregistration for bilingualism research. Using a psycholinguistic example, we illustrate how preregistration serves to implement a clean distinction between hypothesis testing and data exploration. This distinction consequently helps the researcher to avoid casting post-hoc hypotheses and analyses as confirmatory ones. In keeping with the current best practices in the experimental sciences, preregistration, along with sharing data and code, should be an integral part of hypothesis-driven bilingualism research.

Keywords: Confirmatory analysis; preregistration; open science; psycholinguistics; bilingualism

Correspondence: mertzen@uni-potsdam.de. All data and code are available from https://osf.io/5ab7d/.

Introduction

An important aspect of hypothesis-driven confirmatory 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 begins. Preregistration is a relatively simple yet powerful tool for improving transparency in bilingualism research, and we suggest that, in keeping with the current best practices in the experimental sciences, researchers standardly include preregistration as an essential component of the hypothesis-driven research process. The motivation for preregistration is that it can help researchers to eliminate unconscious biases and to avoid questionable research practices such as selectively choosing and reporting one of many potential data analysis paths (*'garden-of-forking-paths'*) as well as *'HARKing'*—hypothesizing after the results are known (Kerr, 1998; Simmons, Nelson, & Simonsohn, 2011). Preregistering a study helps to clearly separate confirmatory analyses (used for hypothesis testing) from exploratory analyses (used for hypothesis generation) (de Groot, 1956/2014; Nosek, Ebersole, DeHaven, & Mellor, 2018). This distinction between confirmatory and exploratory research is often not strictly applied in bilingualism research. In this article, we point out that this distinction is crucial for robust science, and we discuss the benefits of preregistering a planned study.

While concerns about (non-)transparency and post-hoc theorizing are well-known in psychological science (Wicherts, Borsboom, Kats, & Molenaar, 2006; Simmons et al., 2011), bilingualism research is similarly weighed down by the opacity of its 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). To address these issues, two journals in the field of bilingualism, *Language Learning* and *Bilingualism: Language and Cognition*, have recently 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, our focus is on preregistration using open science platforms such as the Open Science Framework (OSF, https://osf.io/), or their dedicated registration platform Center for Open Science (https://cos.io/prereg/), and AsPredicted (https://aspredicted.org/). On these platforms, researchers have the opportunity to create a public or private, but time-stamped, non-modifiable record of a planned study *before* the data is collected, or during the process of data collection, but before it is analyzed. This type of preregistration is not subject to the often time-consuming process of peer review.

Why preregister? An example from sentence comprehension research

To illustrate the benefits of preregistration, we use an example from our own work, which attempted to replicate the findings of a previous eye-tracking reading study investigating real-time sentence comprehension (Dillon, Mishler, Sloggett, &

Phillips, 2013; Jäger, Mertzen, Van Dyke, & Vasishth, 2020).

The two studies that we use as an example examined a grammaticality illusion phenomenon in two different syntactic configurations. The example can be easily translated to bilingualism settings where, similar to our example, differential processing patterns are investigated for different syntactic constructions in L2 learners (Alemán Bañón, Fiorentino, & Gabriele, 2018; Tokowicz & MacWhinney, 2005; Foucart & Frenck-Mestre, 2012). In bilingualism, the comparison of interest may also concern 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 grammaticality illusion example investigates a phenomenon called *agreement attraction*. The subject-verb agreement dependencies shown in 1a and 1b are ungrammatical because the singular subject *The amateur bodybuilder* does not have the same number marking as the plural auxiliary *were*. Interestingly, when a plural noun (a so-called "attractor", such as *personal trainers*) intervenes between the subject and the auxiliary verb, a speedup in reading times is observed in 1a, compared to 1b where no plural noun is present (Wagers, Lau, & Phillips, 2009; Pearlmutter, Garnsey, & Bock, 1999). This result suggests that the formation of subject-verb agreement dependencies is not always constrained by grammar.

(1) a. *Subject-verb agreement; attraction*
*The amateur bodybuilder who worked with the **personal trainers** amazingly <u>were</u> competitive for the gold medal.

b. *Subject-verb agreement; no attraction*
*The amateur bodybuilder who worked with the **personal trainer** amazingly <u>were</u> competitive for the gold medal.

Dillon and colleagues used a within-subjects design to examine whether similar attraction effects were observed for reflexive-antecedent dependencies (as in 2). Building on work by Sturt (2003), they argued that, unlike subject-verb agreement configurations, the processing of reflexive-antecedent dependencies should be syntactically constrained by Principle A of the binding theory (Chomsky, 1981), which rules out the plural attractor noun as an antecedent for the reflexive. Therefore, no difference in processing time is expected at the reflexive in condition 2a compared to 2b, i.e., no agreement attraction should be observed. Due to the predicted speedup in 1a vs. 1b (subject-verb agreement conditions) but not in 2a vs. 2b (reflexive conditions), an interaction between dependency type and attraction effect was expected.

(2) a. *Reflexive; attraction*
*The amateur bodybuilder who worked with **the personal trainers** amazingly injured <u>themselves</u> on the lightest weights.

b. *Reflexive; no attraction*
*The amateur bodybuilder who worked with **the personal trainer** amazingly injured <u>themselves</u> on the lightest weights.

In an eyetracking study, it is possible to test for the interaction in a range of dependent measures, some of which are usually assumed to reflect 'early' and some 'late' cognitive processes (Clifton, Staub, & Rayner, 2007; Vasishth, von der Malsburg, & Engelmann, 2012). Dillon et al. (2013) conducted statistical tests for multiple dependent measures and observed evidence for the predicted interaction only in total reading time (Table 1). This significant interaction seems to confirm the hypothesis that subject-verb agreement and reflexives show different patterns for the number agreement manipulation.

	Dillon et al. results (N = 40)			**Jäger et al. replication (N = 181)**		
Dependent measure	Estimate	Std. Error	z/t value	Estimate	Std. 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	39.90	-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 a summary of the difference in the attraction effect in the subject-verb agreement vs. reflexive conditions, computed using generalized linear mixed models (effects on first-pass regressions were estimated using a logit link function). 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. An interaction effect with a negative sign was expected. The so-called early reading measures examined are first-fixation duration, first-pass reading time, first-pass regressions, and regression-path duration; late measures are re-reading time, and total reading time. The rows with significant effects at a 0.05 α-level are shown in bold.

In our large-sample replication study (Jäger et al., 2020), the goal was to replicate the significant interaction found in total reading time in the original Dillon et al. study. Here, by 'replicating' we mean reproducing the claimed statistically significant interaction with a negative sign in total reading time. 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).

The original study and the replication study both seem to show *some* evidence of an interaction, i.e., an attraction effect for subject-verb agreement but not for reflexives; however, this interaction occurs in different measures across the two studies. Importantly, 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 (de Groot, 1956/2014). Therefore, the evidence

for different processing profiles in agreement vs. reflexives is inconclusive at best. When we have equivocal results like these, what steps could be taken to come closer to answering our research question? We could conduct a confirmatory study. Here, all the usual researcher degrees of freedom (dependent measures, critical regions, etc.) should be fixed prior to collecting the data (Chambers, 2019); subsequently, hypothesis testing can be carried out. It is for such confirmatory hypothesis tests that preregistration is designed. Next, we discuss why confirmatory analyses and an exploratory analyses should be treated differently.

Why exploratory analysis is not the same as confirmatory hypothesis testing

The garden-of-forking-paths problem. Why is it problematic if we conduct an exploratory analysis, and then conclude that the predicted interaction has been found, if we obtain a statistically significant effect? The first important issue is the so-called *garden-of-forking-paths* problem (Gelman & Loken, 2013, 2014). There are many analysis paths that could have been chosen: one could have deleted some participants' data using some criterion (e.g., deleting extreme values that lie 2.5 or 3 standard deviations away from the mean), or chosen a different region of interest than originally planned. Such multiple analysis paths cumulatively create so many researcher degrees of freedom that one can describe them using a binary decision tree, hence the term garden-of-forking-paths. We follow Gelman and Loken (2013)'s terminology here to emphasize the point that the garden-of-forking-paths issue is often an unconscious bias (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 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.

If we decide to report one of the many possible analysis paths, at least one of these paths will often yield a desired pattern that looks like a signal but is really just noise.

Multiple testing. The second important issue, which is closely related to the garden-of-forking-paths, is the multiple comparisons problem: for purely statistical reasons, if one conducts enough statistical tests, *some* test will eventually come out significant. In psycholinguistics, one can easily end up conducting dozens or sometimes even hundreds of statistical tests to evaluate a single hypothesis. Simulations in von der Malsburg and Angele (2017) demonstrated that multiple analyses in eyetracking dramatically inflate Type I error, leading to a large proportion of false positive rejections

of the null hypothesis. For the analyses of the Dillon et al. study and the Jäger et al. replication study shown in Table 1, six statistical tests were conducted. Testing six eye-tracking measures increases the false positive probability from 5% to 26.5% (i.e., $1-0.95^6=0.265$). 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. Any observed t-/z-values would have to reach this threshold to be considered significant. None of the observed t-values in Table 1 for either the Dillon et al. (2013) or the Jäger et al. (2020) study reach the 2.64-threshold. Thus, if we were to use the corrected Type I error, there would be no significant effects in either the original study or the replication attempt. A better solution to the multiple testing problem than the Bonferroni correction would be to avoid multiple testing altogether and to have a precise prediction about the dependent measure(s) and the critical sentence region to be analyzed.

Post-hoc hypothesizing. To conduct a confirmatory analysis of a precise prediction can not only counteract researcher degrees of freedom, but also hypothesizing after the results are known (*HARKing*) (Chambers, 2019; Simmons et al., 2011). As an example, suppose that the effect that was expected at the critical auxiliary verb or the reflexive for the Dillon et al. (2013) and Jäger et al. (2020) studies had been found further downstream in the sentence or even before the critical region. Without specifying the critical region a priori, one could easily have found an explanation for the effect showing up in another region and reported this as if it had been predicted all along.

How to clearly separate confirmatory and exploratory research

Preregistration of confirmatory hypotheses after a pilot experiment. If we lack prior knowledge of a particular psycholinguistic phenomenon, we could pilot an experiment, conduct as many exploratory analyses as desired to identify which reading measure(s) might show the predicted effect, and generate hypotheses from this.

This is where preregistration comes in. After the exploration stage, we preregister our prediction of the expected effect *for a particular reading measure at a particular region*. We then re-run our study with the same items and a new (larger) participant sample, conducting the statistical analysis predefined in our time-stamped preregistration document. This will be our confirmatory analysis used for statistical inference. Of course, we are also free to conduct exploratory analyses for the new data set, but this exploration should be clearly identified as such. Recent implementations of this distinction between exploratory vs. confirmatory stage in psycholinguistics are Nicenboim, Vasishth, Engelmann, and Suckow (2018), Nicenboim, Vasishth, and Rösler (2019). We want to reiterate that exploratory analyses per se are an important part of doing science. In fact, Bem (2004), for example, actively encourages us to go on so-called "fishing expeditions", but such statements can easily be misunderstood. A

recent paper by Bishop (2020) agrees with Bem (2004) but with the following caveat:

> The problem comes when exploratory research is repackaged as if it were hypothesis testing, with the hypothesis invented after observing the data (HARKing), and the paper embellished with p-values that are bound to be misleading because they were p-hacked from numerous possible values, rather than derived from testing an a priori hypothesis.

Preregistration of confirmatory hypotheses based on previous research. If we already know something about the effect of interest from the previous literature, the dependent measure and analysis region previously reported to be significant could serve as the basis for a preregistration of a confirmatory replication study. This is not always straightforward. For example, Dillon et al. (2013) tested for an effect in multiple reading measures, and found the predicted interaction in total reading times, while our replication study only observed this interaction during the exploratory analysis stage in first-pass regressions and regression-path durations. Now, let's assume that, based on linguistic theory, we believe that the interaction should be found in the early stages of reading. If we want to get closer to answering the question of whether the first-pass regressions and regression-path duration results truly reflect evidence of agreement attraction in subject-verb dependencies but not in reflexives, the only way to test this is by conducting a planned study with a sufficiently large participant sample size where the analysis plan is laid out 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.

In a preregistration, we can specify the expected interaction at the critical sentence region, in *first-pass regressions* based on our previous outcome and based on linguistic theory. If planning to carry out such a confirmatory test, it is not sufficient to specify the dependent measure and the to-be-tested region. Too many researcher degrees of freedom remain. A complete preregistration requires a full description of the research questions and hypotheses, study design, methods, data collection procedure, participant sample size, outcome variable(s), an analysis plan including statistical models and information on data exclusion and statistical inference criteria. Preregistration templates are available on OSF and AsPredicted for novel as well as for replication studies. If one prefers to create a Registered Report-type preregistration (i.e., in manuscript format), it is possible to upload a preregistration manuscript (pre-data collection or, at least, pre-data analysis). Importantly, 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, but rather, one needs to create a time-stamped version. This time-stamped, non-editable version can be made public either immediately or it can be embargoed until, for example, the associated paper is submitted or published. If the registration is withdrawn at any stage after creating a "frozen" version of it, some meta data (the title, the authors, a short

description, reason for withdrawing the preregistration) will remain publicly available (see http://bit.ly/withdrawprereg). A new version of the preregistration can be made available at any point before data is analyzed. If opting for a manuscript-style registration, we have found it helpful to use the existing templates as a checklist to ensure that no important questions have been left unanswered in the manuscript.

We have previously made an attempt at such a manuscript-style preregistration (Vasishth, Mertzen, Jäger, and Gelman (2018): https://osf.io/eyphj/). The frozen, non-editable version can be inspected at https://osf.io/dgewb.

Differences between preregistration on open science platforms and Registered Reports

Although preregistration without peer review is an effective way to reduce unconscious biases in one's work, it has some potential disadvantages when compared to Registered Reports. Registered Reports offered by journals have the assurance of peer review and of guaranteed publication, once a preregistered manuscript has been accepted. By contrast, for preregistrations without peer review, publication is not guaranteed. The lack of peer review means that the preregistration of a study can be as thorough or as vague as the researcher deems appropriate. A vaguely specified research question and analysis plan can still allow for a plethora of statistical testing possibilities. Consequently, it is up to the scientific community to make 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.

Best practices for open science

To ensure truly transparent research, in addition to using preregistration, it is essential to release all research materials, data and code along with the study publication. Releasing materials is a good first step toward allowing other researchers to attempt to replicate a study (Open Science Collaboration, 2015).

Publication of the data and analysis code is important so that the steps leading to a result can be fully retraced. Furthermore, it allows other researchers to conduct exploratory analyses which may help to formulate predictions for future (preregistered) studies, and to synthesize the available evidence on a particular topic (Mahowald, James, Futrell, & Gibson, 2016).

Conclusion

We have used a practical example to illustrate the usefulness of preregistration in cases where previous studies provide inconclusive evidence about the existence of a processing effect. While our example focused on native processing, bilingual processing research shows similar cases of inconclusive findings, such as the existence of a bilingual advantage in attentional systems (Bialystok, 2017; Paap, Anders-Jefferson,

Mason, Alvarado, & Zimiga, 2018), the role of crosslinguistic influence in syntactic processing (Dussias, Dietrich, & Villegas, 2015), and the decomposition of inflected forms during word recognition in native vs. non-native speakers (Clahsen & Veríssimo, 2016; Feldman & Kroll, 2019). Existing results in these domains could be used to preregister precise research questions including information about speaker groups, dependent measures and analysis plan(s) for expected effects. This process can counteract unconscious biases and can prevent the reporting of an exploratory result as a confirmatory one. We suggest that the hypothesis-driven research process should standardly include preregistrations, along with other open science practices such as releasing materials, data and code alongside publications to increase research transparency and reproducibility in bilingualism research.

Acknowledgements

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 317633480 – SFB 1287, Project B03 (PIs: Shravan Vasishth and Ralf Engbert).

References

Alemán Bañón, J., Fiorentino, R., & Gabriele, A. (2018). Using event-related potentials to track morphosyntactic development in second language learners: The processing of number and gender agreement in Spanish. *PLOS ONE*, *13*(7), 1–35. doi:10.1371/journal.pone.0200791

Bem, D. (2004). Writing the empirical journal article. In J. M. Darley, M. P. Zanna, & H. L. R. III (Eds.), *The compleat academic: A career guide* (pp. 185–219). American Psychological Association.

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

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

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

Clifton, C., Staub, A., & Rayner, K. (2007). Eye Movements in Reading Words and Sentences. In R. V. Gompel, M. Fisher, W. Murray, & R. L. Hill (Eds.), *Eye movements: A window on mind and brain* (Chap. 15). Elsevier.

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, and 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

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* (349?366). Cambridge Handbooks in Language and Linguistics. Cambridge University Press. doi:10.1017/CBO9781107447257.016

Feldman, L. B. & Kroll, J. F. (2019). Learning and using morphology and morphosyntax in a second language. Oxford University Press. Retrieved from https://oxfordre.com/linguistics/view/10.1093/acrefore/9780199384655.001.0001/acrefore-9780199384655-e-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

Foucart, A. & Frenck-Mestre, C. (2012). Can late L2 learners acquire new grammatical features? Evidence from ERPs and eye-tracking. *Journal of Memory and Language*, *66*(1), 226–248. doi:https://doi.org/10.1016/j.jml.2011.07.007

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

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

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

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

Mahowald, K., James, A., Futrell, R., & Gibson, E. (2016). A meta-analysis of syntactic priming in language production. *Journal of Memory and Language*, *91*, 5–27. doi:10.1016/j.jml.2016.03.009

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. (2019). *Are words pre-activated probabilistically during sentence comprehension? Evidence from new data and a Bayesian random-effects meta-analysis using publicly available data*. Submitted. Retrieved from https://psyarxiv.com/2atrh/

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

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

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

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.

Tokowicz, N. & MacWhinney, B. (2005). Implicit and explicit measures of sensitivity to violations in second language grammar: An event-related potential investigation. *Studies in Second Language Acquisition*, *27*(2), 173–204. doi:10.1017/S0272263105050102

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

Vasishth, S., von der Malsburg, T., & Engelmann, F. (2012). What eye movements can tell us about sentence comprehension. *Wiley Interdisciplinary Reviews: Cognitive Science*, 125–134. doi:https://doi.org/10.1002/wcs.1209

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

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