Timo B. Roettger*

Preregistration in experimental linguistics: applications, challenges, and limitations

https://doi.org/10.1515/ling-2019-0048
Received December 30, 2019; accepted December 1, 2020; published online March 24, 2021

Abstract: The current publication system neither incentivizes publishing null results nor direct replication attempts, which biases the scientific record toward novel findings that appear to support presented hypotheses (referred to as "publication bias"). Moreover, flexibility in data collection, measurement, and analysis (referred to as "researcher degrees of freedom") can lead to overconfident beliefs in the robustness of a statistical relationship. One way to systematically decrease publication bias and researcher degrees of freedom is preregistration. A preregistration is a time-stamped document that specifies how data is to be collected, measured, and analyzed prior to data collection. While preregistration is a powerful tool to reduce bias, it comes with certain challenges and limitations which have to be evaluated for each scientific discipline individually. This paper discusses the application, challenges and limitations of preregistration for experimental linguistic research.

Keywords: confirmatory; exploratory; preregistration; publication bias; registered report; researcher degrees of freedom

1 Introduction

In recent coordinated efforts to replicate published findings, the social sciences have uncovered surprisingly low replication rates (e.g., Camerer et al. 2018; Open Science Collaboration 2015). This discovery has led to what is now referred to as the "replication crisis" in science. There are raising concerns that a similar state of affairs is true for the field of experimental linguistics because it shares with other disciplines many research practices that have been identified to decrease the replicability of published findings (e.g., Marsden et al. 2018a; Roettger and Baer-Henney 2019; Sönning and Werner this issue). Moreover, there is already mounting evidence that published experimental findings cannot be taken at face value (e.g., Chen 2007; Nieuwland et al. 2018; Papesh 2015; Stack et al. 2018; Westbury 2018, among many others). The present

***Corresponding author: Timo B. Roettger**, Department of Linguistics and Scandanavian Studies, University of Oslo, Postboks 1102 Blindern, 0317 Oslo, Norway, E-mail: timo.roettger@iln.uio.no

special issue is a welcome and timely attempt to assess the situation in linguistics and to critically discuss ways to improve linguistic research practices.

The use of the label "crisis" in the expression "replication crisis" suggests a time of intense difficulty, trouble, or even danger. It might, however, be more fruitful to think of the current situation as an opportunity. Repeated failures to replicate published findings have led to a fruitful discourse across disciplines. Researchers have identified shortcomings in how science is practiced and suggested promising ways forward to increase the transparency, reproducibility, and replicability of scientific work.[1] Even within linguistics, an increasing number of researchers have articulated their concerns about present research practices and, importantly, have offered practical advice to circumvent these problems in the future (e.g., Baayen et al. 2017; Berez-Kroeker et al. 2018; Kirby and Sonderegger 2018; Marsden et al. 2018a; Roettger 2019; Vasishth et al. 2018; Wieling et al. 2018; Winter 2011). In the same spirit, the present paper discusses a concept that marks a promising way forward in increasing the replicability of experimental linguistic research: preregistration.

A preregistration is a time-stamped document in which researchers specify prior to data collection how they plan to collect their data and/or how they plan to conduct the data analyses. In the following, I will argue that preregistration helps drawing a line between exploratory and confirmatory research. It also allows transparently tracking analytical flexibility and counteracting publication bias. Many authors have discussed the concept of preregistration across disciplines before (e.g., Nosek and Lakens 2014; Wagenmakers et al. 2012) and there are relevant discussions within the language sciences for second language research (Marsden et al. 2018b; Morgan-Short et al. 2018) and language acquisition research (Havron et al. 2020). However, I think it is worth to reiterate applications, challenges, and limitations of preregistration for experimental linguistics at large.

2 The problem: biases we live by

In the following, I will give a brief overview of relevant problems that may affect the replicability of published research and discuss how some of these problems can be

1 The terms reproducible research and replication are used ambiguously in the literature. Here I follow Claerbout and Karrenbach (1991) and refer to reproducible research as research in which "authors provide all the necessary data and the computer codes to run the analysis again, recreating the results" and a replication as a "study that arrives at the same scientific findings as another study, collecting new data (possibly with different methods) and completing new analyses." See Barba (2018) for a review of the usage of these terms.

tackled by preregistration. In what follows I will assume a particular modus of scientific inquiry in which researchers accumulate knowledge about nature and society by formulating falsifiable hypotheses and test these hypotheses on observable data for which the outcome is unknown. This confirmatory mode of scientific inquiry is particularly common in experimental subfields of linguistics. Many other subfields in linguistics are inherently observational (see Grieve this issue) and thus the proposed dichotomy between exploratory and confirmatory research as well as the concept of preregistration might not similarly apply across the language sciences. We will come back to the value of preregistration for observational studies below.

2.1 Exploratory and confirmatory research

Linguists who work empirically generally collect data (make observations) to understand aspects of human language, such as how it is comprehended, how it is produced, how it is acquired, or how it has evolved. In linguistics, this can involve the analysis of corpora, the analysis of crosslinguistic databases, or the analysis of experiments, and so forth. The observations are then used to formulate empirical models that capture what has been observed (e.g., Lehmann 1990). In most experimental fields, the model development is usually informed by two phases of research: exploratory and confirmatory research (e.g., Box 1976; de Groot 2014 [1956]; Nicenboim et al. 2018b; Roettger et al. 2019; Tukey 1977): Researchers *explore* patterns and relationships in their observations. Based on these observed patterns, they reason about plausible processes or mechanisms that could have given rise to these patterns in the data. They then formulate *hypotheses* as to how nature will behave in certain situations that they have not been observed yet. These hypotheses can then be tested on new data in an attempt to *confirm*[2] the empirical predictions.

Exploratory and confirmatory research are both vital components of scientific progress. Exploration has led to many breakthroughs in science. A linguistic example is the McGurk effect, i.e., perceiving a sound that lies in-between an auditorily presented component of one sound and a visually presented component of another one (McGurk and MacDonald 1976). This effect was not predicted

2 The term "confirmation" in this context refers to statements about *statistical* hypotheses. Following commonly held views in the philosophy of science (e.g., Popper 1963), we cannot *confirm* scientific hypotheses (we can only falsify them). While statistical hypotheses and empirical models can be *corroborated* by data, it must also be borne in mind that this interpretation is always conditional on the statistical model, i.e., the validity of the assumptions specified by the inferential procedure.

a priori, it was accidentally discovered (Massaro and Stork 1998) and led to predictions that since then have been confirmed on new observations (Alsius et al. 2018). Putting hypotheses under targeted scrutiny via confirmatory tests enables the accumulation of evidence in order to challenge, support, and refine scientific models.

The distinction between confirmatory and exploratory research is tremendously important. Given the complexity of the phenomena investigated in linguistics, every set of observations offers a myriad of ways to look at it and contains spurious relationships and patterns. Chance and sampling error alone will produce what might look like a meaningful pattern (e.g., Kirby and Sonderegger 2018; Nicenboim et al. 2018a; Winter 2011). If these exploratory observations are treated as they were predicted *a priori*, one might overconfidently believe in the robustness of a relationship that will not stand the test of time (e.g., Gelman and Loken 2013; Roettger 2019; Simmons et al. 2011).

2.2 To err is human

Researchers are human and humans have evolved to filter the world in *irrational* ways (e.g., Tversky and Kahneman 1974), blurring the line between exploration and confirmation (Nuzzo 2015). For example, humans see coherent patterns in randomness (Brugger 2001), they convince themselves of the validity of prior expectations ("I knew it", Nickerson 1998), and they perceive events as being plausible in hindsight ("I knew it all along", Fischhoff 1975). When hindsight bias comes into play, researchers tend to generate explanations based on observations, and, at the same time, believe that they would have anticipated this very explanation before observing the data. For example, a researcher might predict that focused constituents in a language under investigation are prosodically marked. As is, this is a vague prediction, as prosodic marking can be operationalized in different ways (e.g., Gordon and Roettger 2017). After combing through a data set of speech and measuring several plausible acoustic dimensions, the researcher discovers that the average word duration of focused words is greater than that of unfocused words. After this discovery, the researcher identifies duration, out of all acoustic dimensions that have been measured (and that could have been measured), as the one most relevant for testing the prediction. Crucially, the researcher does not mention those dimensions that *did not* show a systematic relationship (Roettger 2019). The researcher generates and tests predictions on the same data set.

Consider another example: a group of psycholinguists is convinced that a certain syntactic structure leads to processing difficulties. These processing

difficulties should be reflected in language users' reading times. The researchers run an eye-tracking experiment but do not find a significant difference in reading times. Convinced that they have overlooked something, they also measure alternative behavioral indices related to different visual fields (foveal, parafoveal, peripheral), related to different fixations of regions (e.g., first, second, third), and the duration of a fixation before exiting/entering a particular region (see von der Malsburg and Angele 2017 for a discussion of these researcher degrees of freedom in eye tracking). After several analyses of the data, a significant effect of one of these measures materializes. In hindsight, it strikes the researchers as particularly obvious that this measure was the one that shows the clearest effect and when writing up the paper, they frame it as if this linking hypothesis had been spelled out prior to data collection.

This after-the-fact reasoning is particularly tempting in linguistic research. Some languages such as English are particularly well investigated. Many other languages are either heavily underdocumented or not documented at all. It is all too easy to assume that grammatical functions, linguistic phenomena, or communication patterns that are relevant for the handful of well-investigated languages can be found in other languages, too (e.g., Ameke 2006; Bender 2011; Gil 2001; Goddard and Wierzbicka 2014; Levisen 2018; Wierzbicka 2009). For example, it has been shown that focused constituents are often phonetically longer in English (e.g., Cooper et al. 1985). Researchers' prior belief in how the next language manifests focus might be biased by their preconceptions about the languages (Majid and Levinson 2010) and cultures (Henrich et al. 2010) with which they are most familiar.

2.3 Incentivizing confirmation over exploration

Biases such as confirmation or hindsight bias are further amplified by the academic ecosystem. When it comes to publishing experimental work, exploration and confirmation are not weighted equally. Confirmatory analyses have a superior status within the academic incentive system, determining the way funding agencies assess proposals, and shaping how researchers frame their papers (Sterling 1959). In an incentive system in which high impact publications are the dominant currency to secure jobs and funding, the results of what has actually been an exploratory analysis are often presented as if they were the results of a confirmatory analysis (Simmons et al. 2011). Whether done intentionally or not, this reframing of results adds to the publishability of the proposed findings.

Within the confirmatory framework, findings that statistically support predictions are considered more valuable than null results. The lack of incentives for

publishing null results or direct replication attempts biases the scientific record toward novel positive findings. For example, Marsden et al. (2018a) investigated the prevalence of replication studies across second language research. They found a low replication rate, corresponding to only one direct replication in every 400 articles. Replication studies were on average conducted after more than six years and over a hundred citations of the original study. Thus, replications are either only performed after the original study had already impacted the field substantially or only then published if the original study was impactful.

This leads to a pervasive asymmetry with a large number of null results and direct replication attempts not entering the scientific record ("publication bias", e.g., Fanelli [2012]; Franco et al. [2014]; Sterling [1959], see also the "significance filter", Vasishth et al. [2018]). For example, Fanelli (2012) analyzed over 4,600 papers published across disciplines, estimated the frequency of papers that, having declared to have "tested" a hypothesis, reported support for it. On average, 80% of tested hypotheses were found to be confirmed based on conventional statistical standards.

What advances experimental researchers' careers and helps them obtain funding are statistically supported predictions, not null results. The prevalent expectation that the main results of a study should be predicted based on *a priori* grounds is one of the factors that have led to research practices that are inhibiting scientific progress (John et al. 2012). These questionable practices are connected to the statistical tools that are used. In most scientific papers, statistical inference is drawn by means of null hypothesis significance testing (NHST), a procedure to evaluate prediction and test hypotheses (Gigerenzer et al. 2004; Lindquist 1940). In NHST, the probability of observing a result at least as extreme as a test statistic (e.g., t-value) is computed, assuming that the null hypothesis is true (the p-value). Receiving a p-value below a certain threshold (commonly 0.05) leads to a categorical decision about the incompatibility of the data with the null hypothesis.

Within the NHST framework, any pattern that yields a p-value below 0.05 is, in practice at least, considered sufficient to reject the null hypothesis and claim that there is an effect. However, if the p-value is 0.05 and the null is actually true, there is a 5% probability that the data accidentally suggests that the null hypothesis can be refuted (a false positive, otherwise known as Type I error). If one only performs one test and follows only one way to conduct that test, then the p-value is diagnostic about its intended probability. However, the diagnostic probability of the p-value changes as soon as one performs more than one analysis (Benjamini and Hochberg 1995).

There are usually many decisions researchers must make when analyzing data. For example, they have to choose how to measure a desired phenomenon and they have to decide whether any observation has to be excluded and what

predictors to include in their analysis (see Roettger [2019] for an in-depth discussion for experimental phonetics). These decisions during the analysis procedure have been referred to as "researcher degrees of freedom" (Simmons et al. [2011]; see also Gelman and Loken's [2013] garden of forking paths). If data-analytic flexibility is exploited *during* analysis, that is after observing the data, hindsight and confirmation biases can creep in and affect how researchers make their decisions. Researchers often explore many ways of analyzing the data, for all of which they have good reasons. However, the diagnostic nature of the *p*-value changes dramatically when doing so (Gelman and Loken 2013; Simmons et al. 2011). Two often-discussed instances of this problem are HARKing (Hypothesizing After Results are Known, e.g., Kerr 1998) and selective reporting (Simmons et al. 2011, also referred to as *p*-hacking). Researchers HARK when they present relationships that have been obtained after data collection as if they were hypothesized in advance, i.e., they reframe exploratory indications as confirmatory results. Researchers selectively report when they explore different analytical options until significant results are found, i.e., different possible data analytical decisions are all explored and the one data analytical path that yields the desired outcome is ultimately reported (while the others are not). The consequence of this (often unintentional) behavior is an inflation of false positives in the literature. Left undetected, false positives can lead to theoretical claims that may misguide future research (Smaldino and McElreath 2016).

In light of the outlined interaction between cognitive biases, statistical procedures, and the current incentive structure, it is important to cultivate and institutionalize a clear line between exploration and confirmation for experimental research. One tool to achieve this goal is preregistration.

3 A solution: preregistration

A preregistration is a time-stamped document in which researchers specify how they plan to collect their data and/or how they plan to conduct their confirmatory analysis (e.g., Nosek and Lakens 2014; Wagenmakers et al. 2012; see Havron et al. 2020; Marsden et al. 2018b; Morgan-Short et al. 2018 for language-related discussions).

Preregistrations can differ with regard to how detailed they are, ranging from basic descriptions of the study design to very detailed descriptions of the procedure and statistical analysis. In the most transparent version of a preregistration, all relevant materials, experimental protocols, and statistical procedures are published alongside the preregistration prior to data collection.

Preregistration draws a clear line between exploratory and confirmatory parts of a study. By doing so, it reduces researcher degrees of freedom because the conduct of a study commits to certain decisions prior to observing data. Additionally, public preregistration can help to reduce publication bias, as the number of failed attempts to reject a hypothesis can be tracked transparently.

The concept of preregistration is not new. A form of preregistration has been mandatory for clinical trials funded by the US government since 2000. Since 2005, preregistration is a precondition for publishing clinical trials in most medical journals (DeAngelis et al. 2005). In a research climate that can be characterized by an increased interest in openness, transparency and reproducibility, preregistration has become more and more common for experimental research outside of the medical field. On the Open Science Framework (osf.io), one of the most widely used preregistration platforms, there is an exponential growth of preregistrations (Nosek and Lindsay 2018), a trend that has not yet carried over to the field of experimental linguistics.

When writing a preregistration, the researcher should keep a skeptical reader in mind. The goal of the preregistration is to reassure the skeptic that all necessary decisions have been planned in advance. Ideally, the families of questions in Figure 1 should be addressed in sufficiently specific ways (see also Wicherts et al. 2016 for a detailed list of relevant researcher degrees of freedom):

What does "sufficiently specific" mean? For example, excluding participants "because they were distracted" is not specific enough because it leaves room for interpretation. Instead, one should operationalize what is meant by being "distracted", e.g., incorrectly answering more than 40% of prespecified comprehension questions. Claiming to analyze the data with linear mixed effects models is not sufficient either as there are many moving parts that can influence the results. Instead, one should specify the model structure (including at least a description of the model formula), the inferential procedure (e.g., are hypotheses evaluated by model comparison, by null-hypothesis significance testing, by Bayesian parameter estimation), and the inferential criterion (e.g., when are statistical hypotheses claimed to be confirmed?). Ideally, one publishes the analysis script with the preregistration to leave no ambiguity as to the data analysis pipeline (see the discussion in Section 4.2 on dealing with data-contingent decisions during analysis).

There are several websites that offer services to preregister studies: two of the most discussed platforms are AsPredicted (AsPredicted.org) and the preregistration forms on the Open Science Framework (osf.io). These platforms afford time-logged reports and either make them publicly available or grant anonymous access only to a specific group of people (such as reviewers and editors during the peer-review process). AsPredicted.org is a rather slim version of what is necessary for a preregistration. One author on the research team simply answers nine questions about the

Figure 1: Questions that should be answered in a preregistration.

planned project, and a PDF file of the pre-registration is generated. The Open Science Framework template is more detailed and asks for specifics about the study design and data analysis. The preregistration can conveniently be associated with an OSF repository and linked to materials, data, analyses scripts, and preprints.

A particularly promising version of preregistration is a peer-reviewed *Registered Report* (Nosek and Lakens 2014; Nosek et al. 2018). Registered Reports include the theoretical rationale and research question(s) of the study as well as a detailed methodological description that aims to answer those questions. In other words, a Registered Report is a full-fledged manuscript that *does not present results*. These

reports are assessed by peer reviewers, who offer critical feedback on how well the proposed method addresses the research question. This critical feedback helps the authors to refine their methodological design and potentially identify critical flaws. Upon sufficient revision, the study plan might get accepted *in-principle*, irrespective of whether the results confirm the researchers' predictions or not.

The idea behind Registered Reports is by no means new either. Similar proposals have been made as early as 1966 by Robert Rosenthal (1966) (cited by Chambers 2017). The first journal to implement this article format was *Cortex* almost 50 years later, with an increasing number of journals following suit. As of time of writing, there are already 286 scientific journals (and counting) that have adopted Registered Reports including linguistic journals like *Bilingualism: Language and Cognition*, *Biolinguistics*, *Cognitive Linguistics*, *Discourse Processes*, *Language Learning*, and *Language & Speech*.

The Registered Report workflow effectively counteracts the dynamics that lead to publication bias and has already been shown to produce a more realistic amount of null results than regular publication routes. For example, Allen and Mehler (2019) showed that out of 113 analyzed Registered Reports that explicitly declared to have tested a hypothesis, only 40% found confirmatory evidence. Similarly, Scheel et al. (Scheel et al. 2021) found 44% confirmed hypothesis in a sample of 71 Registered Reports. While certainly more of these analyses are needed to come to a firmer conclusion about the extent to which Registered Reports reduce publication bias, these numbers stand in stark contrast to those presented by Fanelli (2012) who showed an average of 80% confirmed findings in published papers across disciplines.

4 Applications, challenges, and limitations

At first sight, there are many challenges that come with preregistering linguistic studies (see Nosek et al. 2018 for a general discussion; see also Marsden et al. 2018a). In this section, I will discuss some illustrative examples from experimental linguistics (psycholinguistics and phonetics) that differ in their data collection procedure, the accessibility of data, the time of analysis (before or after data collection), the type of analysis, and so forth.

4.1 Using pre-existing data

Consider the following scenario: A group of researchers (henceforth research group A) is interested in whether the predictability of a word affects its

pronunciation. They plan to use an already existing data set: the HCRC Map Task Corpus (Anderson et al. 1991). They want to assess the predictability of words and they plan to extract fundamental frequency (f_0)[3] as the relevant acoustic dimension. The HCRC Corpus had already been collected when they formulated their research hypothesis. One may object, therefore, that preregistrations cannot be applied in such studies.

While testing hypotheses on pre-existing data is not ideal, preregistration of the analyses can still be performed. Ideally, of course, researchers want to limit researcher degrees of freedom *prior* to having seen the data. However, a large amount of recent advancements in our understanding of language resulted from secondary data analyses based on already existing corpora. Nevertheless, researchers can (and should) preregister analyses after having seen pilot data, parts of the study, or even whole corpora. When researchers generate a hypothesis which they intend to confirm with preexisting data, they can preregister analysis plans and commit to how evidence will be interpreted before analyzing the data. For example, research group A can preregister the exact way they are going to measure predictability (e.g., Do they consider monogram, bigram, trigram probabilities? Are these indices treated as separate predictors or combined into one predictability index? If they are combined, how?), they can preregister possible control factors that might affect the dependent variable (e.g., dimensions that are known to affect f_0, such as speaker sex and illocutionary force), they can *a priori* define which data will be excluded (e.g., Are they looking at content words only? Are they only looking at words that fall into a certain range of lexical frequency?), etc.

A challenge for preregistering preexisting data analyses is how much the analyst knows about the data set. The researchers might have read the seminal paper by Aylett and Turk (2004) who analyzed the same data set, the HCRC Map Task Corpus, to answer a related research question. Aylett and Turk looked at the relationship between word duration and the predictability of the word. In situations like this, it is important to record who has observed the data before the analysis and what observations and summary reports are publicly available and potentially known to the authors. If the authors are blind to already published investigations on a data set, the authors could still test 'novel' predictions. However, if the data set has already been queried with respect to the specific research question, the authors may wish to apply a different analysis, in which case they could pre-register the rationale and analysis plan of said analysis.

Regardless of prior knowledge about the data, possible biases in statistical inference can still be minimized by being transparent about what was known prior

3 Fundamental frequency is the lowest frequency of a periodic waveform. In the context of speech, fundamental frequency closely corresponds to what we perceive as pitch.

to analysis and preregistering the analysis plan. This procedure still reduces researcher degrees of freedom (see Weston et al. 2019 for a collection of resources for secondary data analysis including a preregistration template).

4.2 Changing the preregistration

Deviations from a data collection and analysis plan are common, especially in research that deals with less accessible populations, clinical populations, or populations spanning certain age groups. Researchers also often lack relevant knowledge about the sample, the data collection or analysis. Consider the following scenario: A team of researchers (henceforth research group B) is interested in processing consequences of focus in German children. They hypothesize that focused words are accessed more quickly than words that are not in focus. In their planned experiment, three- to seven-year-olds react to sentences with target words being either in focus or not in a visual world paradigm. The children's eye movements are recorded during comprehension. The researchers planned to test 70 children but 12 of the 70 children fell asleep during the experiment, a state of affairs that renders their data useless. The sleepiness of children was not anticipated and therefore not mentioned as a data exclusion criterion in the preregistration.

In this scenario, it is possible to change the preregistration and document these changes alongside the reasons as to why and when changes were made. This procedure still provides substantially lower risk of cognitive biases impacting the conclusions compared to a situation without any preregistration. It also makes these changes to the analysis transparent and detectable.

Another important challenge when preregistering a study is specifying appropriate statistical models in advance. Preregistering data analyses necessitates knowledge about the nature of the data. For example, research group B might preregister an analysis assuming that the residuals of the model are normally distributed. After collecting their data, they realize that the data has heavy right tails, calling for a log-transformation or a statistical model without the assumption that residuals are normally distributed. The preregistered analysis is not appropriate. One solution to this challenge is to define data analytical procedures in advance that allow them to evaluate distributional aspects of the data and potential data transformations irrespective of the research question. Alternatively, one could preregister a decision tree. This may be particularly useful for people using linear mixed-effects models, which are known to occasionally fail to converge. Convergence failures indicate that the iterative optimization procedure fails to reach a stable solution, which renders the model results uninterpretable. In

order to remedy such convergence issues, a common strategy is to remove complex random effect terms incrementally from the model, a strategy which often comes with other risks such as inflated Type-I error rates (Barr et al. 2013; Matuschek et al. 2017). Since one cannot anticipate whether a model will converge or not, a plan of how to reduce model complexity can be preregistered in advance.

On a related point, research group A working on the HCRC corpus (see 4.1 above) may realize that they did not anticipate an important set of covariates in their statistical model. Words in phrase-final position often exhibit a rising f_0 contour. The last word in an utterance is also often the most predictable, so phrase positions might be confounding predictability. Ideally, then, the analysis should control for phrase position. In large multivariate data sets, it is important to not only consider a large number of possible covariates but also to consider the many ways how to handle possible collinearity between these covariates (Tomaschek et al. 2018). All of these decisions influence the final results (Roettger 2019; Simmons et al. 2011) and should be anticipated as much as possible. One helpful way to anticipate data-analytical decisions is to examine data from publicly accessible studies (an underappreciated benefit of making data publicly available). Another way to deal with these complex decisions is splitting the data set into two parts (a process called cross-validation, see Stone 1974). One may use the first part to evaluate the feasibility of an analysis and preregister a confirmatory analysis for the second part (Fafchamps and Labonne 2017).

Regardless of how hard one tries – certain details of the data collection or analysis sometimes cannot be fully anticipated. But as long as one is transparent about the necessary changes to a preregistration, researcher degrees of freedom are reduced. Alternatively – and orthogonal to increasing transparency – researchers could run both the preregistered analysis and the changed analysis to evaluate the robustness of a finding. This enables researchers to either show that decisions after having seen the data do not influence the results or to transparently communicate possible divergences in their results that are due to critical decisions.

4.3 Exploration beyond preregistered protocol

After following the preregistered protocol, research team B, who is interested in focus processing, may end up with a null result. They could not find any relationship between focus and comprehension times. However, they would like to explore their results further and look at different measures related to eye movements. They may think that further exploration is prohibited because they preregistered only analyses related to comprehension times. Finding themselves in a similar situation, research group B has stuck to the preregistered protocol and

reports their results at a conference, where they receive feedback and suggestions for further exploration. Like research team B, however, they may feel imprisoned by the preregistered analysis plan.

Perceiving preregistration as too rigid is a commonly articulated concern (e.g., Goldin-Meadow 2016). It is, however, unwarranted. Preregistration only constraints the confirmatory part of an analysis and does not impact exploration at all. After testing their predictions, researchers are free to explore their data sets, which is, as argued above, an essential component of the scientific discovery process (Nosek et al. 2019; Wagenmakers et al. 2012) and in fact an integral aspect of linguistic research in general (Grieve, this issue). Preregistration simply draws a line between confirmation and exploration, which we can and should clearly flag in our manuscripts (see APA style guidelines, https://apastyle.apa.org/jars/quant-table-1.pdf). That means, when exploring, researchers generate (rather than test) new hypotheses, which would then need to be substantiated by subsequent empirical investigation. This calls for a certain modesty when reporting the findings of the exploration as the generalizability of these findings needs to be considered with caution.

4.4 No time for preregistration

Some research is based on student projects or grants with quick turnarounds. Researchers sometimes need to deliver academic currency within a short time scale. This time pressure might make preregistrations not feasible. While this is a legitimate concern for Registered Reports (the peer-reviewed version of preregistration), it does not necessarily apply to preregistrations in general. Writing up the methodological plan prior to data collection and analysis might strike one as additional work, but it is arguably either just shifting the work load or saving time in the long run. On the one hand, the method section of any paper needs to be written up eventually, so preregistering a study merely shifts that part of the process to an earlier point in time. The preregistration platforms mentioned in Section 3 make this easy, so there is not even any learning curve to consider. Moreover, writing up the preregistration leads to a more critical assessment of the method in advance and might lead to elimination of critical flaws in the design. Fixing these issues at an early stage saves time and resources. When it comes to Registered Reports, the eventual findings cannot be CARKed (Criticized After Results are Known; see Nosek and Lakens 2014) and subsequently rejected by the journal contingent on whether the results corroborate or contradict the researchers' predictions or established views. Editors, reviewers, and authors thus save valuable time and resources in the long run.

4.5 No a priori predictions

Research group A might be at the beginning of a new research program and does not have concrete predictions as to what aspects of the acoustic signal to measure or how to operationalize predictability. At the beginning of a research program, researchers rarely have very concrete hypotheses about a system under investigation. In these cases, it is highly appropriate to explore available data and generate new hypotheses. The researchers might explore several acoustic parameters and different predictability indices in a first stage during the discovery process. Research group B might collect some pilot data to identify potential challenges with their preplanned statistical analysis. This is fine, and for this stage of the research project, preregistering a study is not necessarily the best workflow. However, these exploratory studies are often written up in a way that recasts them as hypothesis-testing (Kerr 1998). This is, again, not necessarily an intentional process. Cognitive biases in tandem with the academic incentive system are often hidden driving forces of such decisions. The nature of statistical procedures (and common misconceptions about them) further facilitate this process. For example, using null hypothesis significance testing, a *p*-value has only known diagnosticity of false positive rates when one tests prespecified hypotheses and corrects for the number of hypotheses tested. In exploratory analyses, the false positive rate is unknown. Preregistration can still be a valuable tool in these exploratory settings as it constrains researcher degrees of freedom at the analysis stage and makes them transparent. It is also conceivable to first run an unregistered exploration and then formulate concrete predictions that are tested on a novel data set following preregistered protocol. However, preregistration is more useful for confirmatory research.

4.6 Remaining limitations and observational research

Preregistration is not a panacea for all challenges to empirical sciences. I have already mentioned several limitations of preregistering studies. For example, researchers often face practical limitations as to how they can collect certain data types and how flexible they are with regard to methodological choices in culturally diverse settings or working with different populations. Researchers might also be constrained by limited resources and time, making collecting pilot data not a feasible option. Moreover, preregistration is a work flow designed for confirmatory research, mostly found in experimental fields. Thus, preregistration does not fit all forms of linguistic inquiry. It is clear that a large proportion of linguistic subdisciplines is observational in nature (Grieve this issue), including much of corpus

linguistics, discourse analysis, field linguistics, historical linguistics, and typology. Studies in these fields are usually not testing specific hypotheses. Instead, they are exploratory in nature. For example, if researchers are interested in how a given phonological contrast is phonetically manifested, they do not necessarily test hypotheses, but explore possible relationships between the signal and lexical forms. A corpus linguist who is interested in finding relationships between different semantic fields in the lexicon might also not test specific *a priori* hypotheses. In these cases, preregistration might not be applicable.

Within experimental linguistics, purely exploratory studies, i.e., studies that explicitly "only" generate new hypothesis without testing them, are still difficult to publish and could prohibit or at least substantially slow down publication. One solution would be to explore first and then confirm those exploratory findings on a new data set (Nicenboim et al. 2018b). However, this work flow might not be feasible for certain linguistic projects. The lack of publication value of exploratory analyses in confirmatory fields is deeply rooted in the publication system and must be tackled as a (separate) community-wide effort. A promising way forward relates to changing the incentive structure. Linguistic journals that publish experimental work could explicitly reward exploratory studies by creating respective article types and encourage exploratory analyses (for an example at *Cortex*, see McIntosh 2017). It is important to stress that it is not desirable to make procedures that are used to ensure robustness and generalizability in strictly confirmatory settings obligatory to all types of linguistic studies. This would likely lead to devaluing or marginalizing those studies for which the preregistration procedures do not fit. Instead, linguists should embrace different paths of discovery and value them equally within their journals.

5 Summary

Preregistration, and especially its peer-reviewed version as a Registered Report, is a powerful tool to reduce publication bias (the tendency to predominantly publish confirming and "significant" findings) and it constrains researcher degrees of freedom. With that, preregistration can substantially reduce false discoveries in the publication record, which themselves can have far-reaching consequences, often leading to theoretical claims that may misguide future research (Smaldino and McElreath 2016). Preregistration can increase the robustness of published findings. In turn, a more robust publication record allows experimental linguists to more effectively accumulate knowledge and advance their understanding of human language.

As opposed to commonly articulated concerns, preregistration is possible even if the data is already available (e.g., for corpus analyses). It also must not be considered as restricting researchers, since studies can diverge from a preregistered protocol if they are transparent about the reasons for these changes. Preregistration does not prohibit exploration. It mainly draws a visible line between confirmation and exploration. It is not an additional burden in terms of time or effort, but arguably saves time and resources in the long run. And if the reader is not convinced of preregistration yet, preregistration offers many advantages to you, the individual researcher:

- Preregistration allows others to critically assess your research methods. Shortcomings in your proposed study can be detected beforehand. There might be issues with your design, your choice of sample size, your statistical model specifications, or even its translation into code. All these things can be detected not only before publication, but before data collection. This arguably avoids wasting time and resources on suboptimal empirical endeavors and leads to better attempts to empirically challenge scientific models.
- Preregistration signals confidence. You are not afraid to submit your models to a rigorous test and you are willing to tackle possible sources of bias in a transparent way.
- Registered Reports can protect you from CARKing (Critiquing After the Results are Known, Nosek and Lakens 2014). Reviewers can articulate many reasons for why the results are different from what they expected. However, if reviewers have considered the Registered Report a valid attempt to answer the research question during peer review, criticism of the method after results are known are constrained.
- An "in-principle" accepted Registered Report is academic currency. It signals that your research has already gone through the quality control of peer review and has been found of sufficient quality to be published. Given the discussed biases in the publication system, an accepted Registered Report can be an easier route to publication for early career researchers than the traditional path, especially when the research scrutinizes established views.

Preregistration, however, is not a panacea for all problems. There are other important practices that lead to a more robust and replicable scientific record (e.g., Chambers 2017), including incentivizing openness and transparency of the discovery process (e.g., Munafò et al. 2017) including data sharing (e.g., Berez-Kroeker et al. 2018), incentivizing the publication of null results (e.g., Nosek et al. 2012), direct replications (e.g., Zwaan et al. 2018), and exploratory reports (e.g., McIntosh 2017). All of these developments operate on the community level and their instantiation is arguably slow. Preregistration, however, is a practice that we can integrate into our

work flow right away. There are practical advantages for individual researchers, and the pay-off for the field of linguistics is considerable.

Acknowledgments: I would like to thank Jack Grieve, Melissa Kline, Lukas Sönning, Mathias Stoeber, Valentin Werner, and an anonymous reviewer for their insightful comments on earlier versions of this paper. All remaining errors are my own.

References

Allen, Chris & David M. A. Mehler. 2019. Open science challenges, benefits and tips in early career and beyond. *PLoS Biology* 17(5). e3000246.

Alsius, Agnès, Martin Paré & Kevin G. Munhall. 2018. Forty years after hearing lips and seeing voices: The McGurk effect revisited. *Multisensory Research* 31(1/2). 111–144.

Ameka, Felix. 2006. Real descriptions: Reflections on native speaker and non-native speaker descriptions of a language. In Felix Ameka, Alan Dench & Nicholas Evans (eds.), *Catching language: The standing challenge of grammar writing*, 69–112. Berlin & New York: Mouton de Gruyter.

Anderson, Anne H., Bader Miles, Gurman Bard Ellen, Boyle Elizabeth, Gwyneth Doherty, Simon Garrod, Isard Stephen, Jacquelin Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo & Henry S. Thompson. 1991. The HCRC map task corpus. *Language and Speech* 34(4). 351–366.

Aylett, Matthew & Alice Turk. 2004. The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech* 47(1). 31–56.

Baayen, R. Harald, Vasishth Shravan, Kliegl Reinhold & Bates Douglas. 2017. The cave of shadows: Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language* 94. 206–234.

Barba, Lorena A. 2018. *Terminologies for reproducible research*. arXiv preprint arXiv:1802.03311.

Barr, Dale J., Roger Levy, Christoph Scheepers & Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68(3). 255–278.

Bender, Emily M. 2011. On achieving and evaluating language-independence in NLP. *Linguistic Issues in Language Technology* 6(3). 1–26.

Benjamini, Yoav & Hochberg Yosef. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 57(1). 289–300.

Berez-Kroeker, Andrea L., Lauren Gawne, Smythe Kung Susan, Barbara F. Kelly, Heston Tyler, Gary Holton, Pulsifer Peter, David I. Beaver, Shobhana Chelliah, Dubinsky Stanley, Richard P. Meier, Nick Thieberger, Keren Rice & Anthony C. Woodbury. 2018. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics* 56(1). 1–18.

Box, George E. P. 1976. Science and statistics. *Journal of the American Statistical Association* 71(356). 791–799.

Brugger, Peter. 2001. From haunted brain to haunted science: A cognitive neuroscience view of paranormal and pseudoscientific thought. In Houran James & Rense Lange (eds.), *Hauntings and poltergeists: Multidisciplinary perspectives*, 195–213. NC: McFarland: Jefferson.

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

Chambers, Chris. 2017. *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice*. Princeton, NJ: Princeton University Press.

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

Claerbout, Jon F. & Martin Karrenbach. 1991. Electronic documents give reproducible research a new meaning. In *SEG technical program expanded abstracts*, 601–604. Society of Exploration Geophysicists. https://doi.org/10.1190/1.1822162.

Cooper, William E., Stephen J. Eady & Pamela R. Mueller. 1985. Acoustical aspects of contrastive stress in question–answer contexts. *The Journal of the Acoustical Society of America* 77(6). 2142–2156.

DeAngelis, Catherine D., Jeffrey M. Drazen, Frank A. Frizelle, Charlotte Haug, John Hoey, Richard Horton, Sheldon Kotzin, Christine Laine, Ana Marusic, A. John P. M. Overbeke, Torben V. Schroeder, Hal C. Sox & Martin B. Van der Weyden. 2005. Clinical trial registration: A statement from the International Committee of Medical Journal Editors. *Archives of Dermatology* 141(1). 76–77.

de Groot, Adrianus Dingeman. 2014 [1956]. The meaning of "significance" for different types of research [Trans. and annotated by Eric-Jan Wagenmakers, Eric-Jan Borsboom, Josine Verhagen, Rogier Kievit, Marjan Bakker, Angelique Cramer, Dora Matzke, Don Mellenbergh & Han L. J. van der Maas]. *Acta Psychologica* 148. 188–194.

Fafchamps, Marcel & Julien Labonne. 2017. Using split samples to improve inference on causal effects. *Political Analysis* 25(4). 465–482.

Fanelli, Daniele. 2012. Negative results are disappearing from most disciplines and countries. *Scientometrics* 90(3). 891–904.

Fischhoff, Baruch. 1975. Hindsight is not equal to foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance* 1(3). 288.

Franco, Annie, Neil Malhotra & Gabor Simonovits. 2014. Publication bias in the social sciences: Unlocking the file drawer. *Science* 345(6203). 1502–1505.

Gelman, Andrew & Eric Loken. 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*. Department of Statistics, Columbia University Unpublished paper, rt.

Gigerenzer, Gerd, Stefan Krauss & Vitouch Oliver. 2004. The null ritual. In David Kaplan (ed.), *The Sage handbook of quantitative methodology for the social sciences*, 391–408. Thousand Oaks, CA: Sage.

Gil, David. 2001. Escaping eurocentrism. In Paul Newman & Martha Ratcliff (eds.), *Linguistic fieldwork*, 102–132. Cambridge: Cambridge University Press.

Goddard, Cliff & Anna Wierzbicka. 2014. Semantic fieldwork and lexical universals. *Studies in Language* 38(1). 80–127.

Goldin-Meadow, Susan. 2016. Presidential column: Why preregistration makes me nervous. *APS Observer* 29(5/6). https://www.psychologicalscience.org/observer/why-preregistration-makes-me-nervous (accessed 1 September 2020).

Gordon, Matthew & Timo Roettger. 2017. Acoustic correlates of word stress: A cross-linguistic survey. *Linguistics Vanguard* 3(1). https://doi.org/10.1515/lingvan-2017-0007.

Havron, Naomi, Christina Bergmann & Sho Tsuji. 2020. Preregistration in infant research: A Primer. *Infancy* 25(5). 734–754.

Henrich, Joseph, Steven J Heine & Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and Brain Sciences* 33(2/3). 61–83.

John, Leslie K., George Loewenstein & Drazen Prelec. 2012. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science* 23(5). 524–532.

Kerr, Norbert L. 1998. HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review* 2(3). 196–217.

Kirby, James & Morgan Sonderegger. 2018. Mixed-effects design analysis for experimental phonetics. *Journal of Phonetics* 70. 70–85.

Lehmann, Erich L. 1990. Model specification. *Statistical Science* 5. 160–168.

Levisen, Carsten. 2018. Biases we live by: Anglocentrism in linguistics and cognitive sciences. *Language Sciences* 76. 101173.

Lindquist, Everet Franklin. 1940. *Statistical analysis in educational research*. Oxford: Houghton Mifflin.

Majid, Asifa & Stephen C. Levinson. 2010. WEIRD languages have misled us, too. *Behavioral and Brain Sciences* 33(2/3). 103.

Marsden, Emma, Kara Morgan-Short, Sophie Thompson & David Abugaber. 2018a. Replication in second language research: Narrative and systematic reviews and recommendations for the field. *Language Learning* 68(2). 321–391.

Marsden, Emma, Kara Morgan-Short, Trofimovich Pavel & Nick C. Ellis. 2018b. Introducing registered reports at language learning: Promoting transparency, replication, and a synthetic ethic in the language sciences. *Language Learning* 68(2). 309–320.

Massaro, Dominic W. & David G. Stork. 1998. Speech recognition and sensory integration. *American Scientist* 86(3). 236–244.

Matuschek, Hannes, Reinhold Kliegl, Shravan Vasishth, Harald Baayen & Bates Douglas. 2017. Balancing Type I error and power in linear mixed models. *Journal of Memory and Language* 94. 305–315.

McGurk, Harry & John MacDonald. 1976. Hearing lips and seeing voices. *Nature* 264(5588). 746.

McIntosh, Robert D. 2017. Exploratory reports: A new article type for Cortex. *Cortex* 96. A1–A4.

Morgan-Short, Kara, Emma Marsden, Jeanne Heil, Bernard I. Issa Ii, Ronald P. Leow, Anna Mikhaylova, Sylwia Mikołajczak, Nina Moreno, Roumyana Slabakova & Paweł Szudarski. 2018. Multisite replication in second language acquisition research: Attention to form during listening and reading comprehension. *Language Learning* 68(2). 392–437.

Munafò, Marcus R., Brian A. Nosek, Dorothy V. M. Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie Du Sert, Uri Simonsohn, Eric-Jan Wagenmakers,

Jennifer J. Ware & John P. A. Ioannidis. 2017. A manifesto for reproducible science. *Nature Human Behaviour* 1(1). 0021.

Nicenboim, Bruno, Timo B. Roettger & Shravan Vasishth. 2018a. Using meta-analysis for evidence synthesis: The case of incomplete neutralization in German. *Journal of Phonetics* 70. 39–55.

Nicenboim, Bruno, Shravan Vasishth, Felix Engelmann & Katja Suckow. 2018b. Exploratory and confirmatory analyses in sentence processing: A case study of number interference in German. *Cognitive Science* 42. 1075–1100.

Nickerson, Raymond S. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology* 2(2). 175–220.

Nieuwland, Mante S., Stephen Politzer-Ahles, Evelien Heyselaar, Katrien Segaert, Emily Darley, Nina Kazanina, Sarah von Grebmer zu Wolfsthurn, Federica Bartolozzi, Vita Kogan, Aine Ito, Diane Mézière, J Dale, Guillaume A. Rousselet Barr, Heather J. Ferguson, Simon Busch-Moreno, Xiao Fu, Jyrki Tuomainen, Eugenia Kulakova, E. Matthew Husband, David I. Donaldson, Zdenko Kohút, Shirley-Ann Rueschemeyer & Huettig Falk. 2018. Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *eLife* 7. e33468.

Nosek, Brian A., Emorie D. Beck, Lorne Campbell, Jessica K. Flake, Tom E. Hardwicke, David T. Mellor, Anna E. van't Veer & Simine Vazire. 2019. Preregistration is hard, and worthwhile. *Trends in Cognitive Sciences* 23(10). 815–818.

Nosek, Brian A., Charles R. Ebersole, Alexander C. DeHaven & David T. Mellor. 2018. The preregistration revolution. *Proceedings of the National Academy of Sciences* 115(11). 2600–2606.

Nosek, Brian A. & Daniël Lakens. 2014. Registered reports: A method to increase the credibility of published results. *Social Psychology* 45. 137–141.

Nosek, Brian A. & D. Stephen Lindsay. 2018. Preregistration becoming the norm in psychological science. *APS Observer* 31. https://www.psychologicalscience.org/observer/preregistration-becoming-the-norm-in-psychological-science (accessed 1 September 2020).

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

Nuzzo, Regina. 2015. Fooling ourselves. *Nature* 526(7572). 182.

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

Popper, Karl R. 1963. Science as falsification. *Conjectures and refutations* 1. 33–39.

Roettger, Timo B. 2019. Researcher degrees of freedom in phonetic sciences. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 10(1). 1.

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

Roettger, Timo B., Bodo Winter & Harald Baayen. 2019. Emergent data analysis in phonetic sciences: Towards pluralism and reproducibility. *Journal of Phonetics* 73. 1–7.

Rosenthal, Robert. 1966. *Experimenter effects in behavioral research*. New York, NY: Appleton-Century-Crofts.

Scheel, Anne M., Mitchell Schijen & Daniël Lakens. 2021. An excess of positive results: Comparing the standard Psychology literature with Registered Reports. *Advances in Methods and Practices in Psychological Science*, https://doi.org/10.1177/25152459211007467.

Simmons, Joseph P., Leif D. Nelson & Uri Simonsohn. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science* 22(11). 1359–1366.

Smaldino, Paul E. & Richard McElreath. 2016. The natural selection of bad science. *Royal Society Open Science* 3(9). 160384.

Stack, Caoimhe M. Harrington, Ariel N. James & Duane G. Watson. 2018. A failure to replicate rapid syntactic adaptation in comprehension. *Memory & Cognition* 46(6). 864–877.

Sterling, Theodore D. 1959. Publication decisions and their possible effects on inferences drawn from tests of significance – or vice versa. *Journal of the American Statistical Association* 54(285). 30–34.

Stone, Mervyn. 1974. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)* 36(2). 111–133.

Tomaschek, Fabian, Peter Hendrix & R. Harald Baayen. 2018. Strategies for addressing collinearity in multivariate linguistic data. *Journal of Phonetics* 71. 249–267.

Tukey, John Wilder. 1977. *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Tversky, Amos & Kahneman Daniel. 1974. Judgment under uncertainty: Heuristics and biases. *Science* 185(4157). 1124–1131.

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

von der Malsburg, Titus & Bernhard Angele. 2017. False positives and other statistical errors in standard analyses of eye movements in reading. *Journal of Memory and Language* 94. 119–133.

Wagenmakers, Eric-Jan, Ruud Wetzels, Denny Borsboom, Han L. J. van der Maas & Rogier A. Kievit. 2012. An agenda for purely confirmatory research. *Perspectives on Psychological Science* 7(6). 632–638.

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.

Weston, Sarah J., David Mellor, Marjan Bakker, Olmo van den Akker, Lorne Campbell, Stuart J. Ritchie, William J. Chopik, Rodica I. Damian, Jessica Kosie, Courtney K. Soderberg, Charles R. Ebersole, Brian Brown, Pamela Davis-Kean, Andrew Hall, Elliott Kruse, Jerome Olsen, K. D. Valentine, Thuy-vy Nguyen. 2019. Secondary data preregistration. https://www.osf.io/x4gzt (accessed 13 September 2019).

Wicherts, Jelte M., Coosje L. S. Veldkamp, Hilde E. M. Augusteijn, Marjan Bakker, Robbie van Aert & Marcel A. L. M. van Assen. 2016. Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology* 7. 1832.

Wieling, Martijn, Josine Rawee & Gertjan van Noord. 2018. Reproducibility in computational linguistics: Are we willing to share? *Computational Linguistics* 44(4). 641–649.

Wierzbicka, Anna. 2009. Overcoming Anglocentrism in emotion research. *Emotion Review* 1(1). 21–23.

Winter, Bodo. 2011. Pseudoreplication in phonetic research. In Sum Lee Wai & Eric Zee (eds.), *Proceedings of the 17th international congress of phonetic science*, 2137–2140. Hong Kong: City University of Hong Kong.

Zwaan, Rolf A., Alexander Etz, Richard E. Lucas & M. Brent Donnellan. 2018. Making replication mainstream. *Behavioral and Brain Sciences* 41. E120.