The illusion of power:
How the statistical significance filter leads to overconfident expectations of replicability

Shravan Vasishth (vasishth@uni-potsdam.de)
Department of Linguistics, University of Potsdam, Potsdam 14476, Germany, and
Centre de Recherche en Mathématiques de la Décision, CNRS, UMR 7534, Université Paris-Dauphine,
PSL Research University, Paris, France.

Andrew Gelman (gelman@stat.columbia.edu)
Department of Statistics, Columbia University, New York, NY 10027, USA.

Abstract

We show that publishing results using the statistical significance filter—publishing only when the p-value is less than 0.05—leads to a vicious cycle of overoptimistic expectation of the replicability of results. First, we show through a simple derivation that when true statistical power is relatively low, computing power based on statistically significant results will lead to overestimates of power. Then, we present a case study using 10 experimental comparisons drawn from a recently published meta-analysis in psycholinguistics (Jäger et al., 2017). We show that the statistically significant results yield an illusion of replicability, i.e., an illusion that power is high. This illusion holds even if the researcher doesn't conduct any formal power analysis but just uses statistical significance to informally assess robustness of results.

Keywords: Statistical significance; p-values; power; replicability

"...variation in the strength of initial evidence (such as original P value) was ...predictive of replication success..." (Open Science Collaboration, 2015, p. 943)

"'...in [an]...academic environment that only publishes positive findings and rewards publication, an efficient way to succeed is to conduct low power studies. Why? Such studies are cheap and can be farmed for significant results, especially when hypotheses only predict differences from the null, rather than precise quantitative differences and trends." (Smaldino & McElreath, 2016, p. 5)

Introduction

The statistical significance filter tells us that significant results—those findings in which the p-value is less than 0.05—are positively biased. The statistically significant estimate is, by definition, more than t standard errors away from zero, where t is some critical value determined by a statistical test (such as the t-test) and the pre-specified Type I error (the probability of incorrectly rejecting the null hypothesis).

Statistical power is the probability of correctly rejecting the null hypothesis assuming some true effect μ.[1] It is well-known that when statistical power is low, the effect (the sample mean) will tend to be exaggerated; these are referred to as

Type M errors by Gelman and Carlin (2014) (also see Gelman & Tuerlinckx, 2000). This exaggeration of effects has been noticed before (Hedges, 1984; Lane & Dunlap, 1978), most recently in neuroscience and epidemiology: Button et al. (2013) refer to the exaggeration of effects in neuroscience as the "winner's curse" and "the vibration of effects", and Ioannidis (2008) discusses this exaggeration of effects in epidemiological studies in terms of the vibration ratio: the ratio of largest to smallest observed effects.

These overestimates get published and fill the literature. Now consider what happens when researchers design a new study. They read the literature and see all these big effects, then plan their next study. They do a power calculation based on these big effects and get an exaggerated estimate of power, and can easily convince themselves that they have a high powered study. Alternatively—and this is probably the more common route in many fields, such as psychology—they don't do a formal power analysis, but just rely on the informal observation that most of the previously published results had a significant effect and so the effect must be present.

As the first quote at the beginning of this paper shows, the replication studies carried out in the Open Science Collaboration (2015) also found that the magnitude of the published p-values were predictive of replication success. From this, it seems reasonable to conclude that lower p-values are generally more predictive of replication success. In other words, a lower p-value suggests a higher probability that the effect can be detected in future repeated studies. Reasoning in this way, we could fix our sample size based on the sizes previous researchers have used. But in doing so, we fall prey to the illusion of power. We elaborate on this point by first showing that if statistical significance is used as a filter for publishing a result and then use the observed effect to compute power, this will lead to an overestimate. Then, we present a case study involving 10 reading studies in psycholinguistics that illustrates this illusion of power.

The relationship between p-values and power

Assume for simplicity the case that we carry out a one-sided statistical test where the null hypothesis is that the null hypothesis mean is $\mu_0 = 0$ and the alternative is that $\mu > 0$.[2]

---

[1] In order to compute power, we need to have an estimate of the true effect, the sample size, and an estimate of the standard deviation.

[2] The presentation below generalizes to the two-sided test.

Given some continuous data $x_1, \ldots, x_n$, we can compute the t-statistic and derive the p-value from it. For large sample size n, a normal approximation allows us to use the z-statistic, $Z = \frac{\bar{X} - \mu_0}{\sigma_X / \sqrt{n}}$, to compute the p-value. Here, $\bar{X}$ is the mean, σ_X the standard deviation, and n the sample size.

The p-value is the probability of observing the z-statistic or a value more extreme assuming that the null hypothesis is true. The p-value is a random variable P with the probability density function (Hung, O'Neill, Bauer, & Kohne, 1997):

$$g_\delta(p) = \frac{\phi(Z_p - \delta)}{\phi(Z_p)}, \quad 0 < p < 1 \tag{1}$$

where

- $\phi(\cdot)$ is the pdf of the standard normal distribution (Normal(0,1)).

- Z_p, a random variable, is the (1-p)th percentile of the standard normal distribution.

- $\delta = \frac{\mu - \mu_0}{\sigma_X / \sqrt{n}}$ is the true point value expressed as a z-score. Here, μ is the true (unknown) point value of the parameter of interest.

Hung et al. (1997) further show that the cumulative distribution function (cdf) of P is:

$$G_\delta(p) = \int_0^p g_\delta(x)\,dx = 1 - \Phi(Z_p - \delta), \quad 0 < p < 1 \tag{2}$$

where $\Phi(\cdot)$ is the cdf of the standard normal.

Once we have observed a particular z-statistic z_p, the cdf $G_\delta(p)$ allows us to compute not only the p-value, but also the power based on the z-statistic (Hoenig & Heisey, 2001). To compute the p-value given that the null hypothesis is true, let the true value be $\mu = 0$. It follows that $\delta = 0$. Then:

$$p = G_0(p) = 1 - \Phi(z_p) \tag{3}$$

To compute power from the observed z_p, set δ to be the observed statistic z_p, and let the critical z-score be z_α, where α is the Type I error (typically 0.05). The power is therefore:

$$G_{z_p}(\alpha) = 1 - \Phi(z_\alpha - z_p) \tag{4}$$

In other words, power computed from the observed statistic is a monotonically increasing function of the observed z-statistic: the larger the statistic, the higher the power estimate based on this statistic (Figure 1). Together with the common practice that only statistically significant results get published, and especially results with a large z-statistic, this leads to overestimates of power. As mentioned above, one doesn't need to actually compute power in order to fall prey to the illusion; merely scanning the statistically significant z-scores gives an impression of consistency and invites the inference that the effect is replicable and robust. The word "reliable" is frequently used in psychology, presumably with the meaning that the result is replicable.
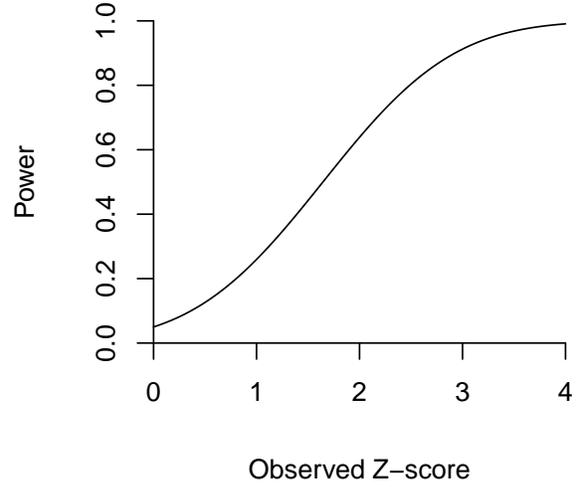

Figure 1: The relationship between power and the observed z-score. The larger z-scores are easier to publish due to the statistical significance filter, and these published studies therefore give a mistaken impression of higher power.

A direct consequence of Equation 4 is that overestimates of the z-statistic will lead to overestimates of power. For example, if we have 36 data points and the true effect is 0.1 on some scale and standard deviation is 1, then statistical power is 15%. This is because the observed z-score is:

$$z_p = \frac{0.1 - 0}{1/\sqrt{36}} = 0.6 \tag{5}$$

Noting that $z_\alpha = \Phi(1 - \alpha) = \Phi(1 - 0.05) = 1.64$, it follows that power is:[3]

$$G_{z_p}(\alpha) = 1 - \Phi(z_\alpha - z_p) = 1 - \Phi(1.64 - 0.6) = 0.15 \tag{6}$$

If we now re-run the same study, collecting 36 data points each time, and impose the condition that only statistically significant results with Type I error $\alpha = 0.05$ are published, then only observed z-scores larger than 1.64 (for a one-sided test) would be published and the power estimate based on these z-scores must have a lower bound of

$$G_{Z_\alpha}(\alpha) = 1 - \Phi(1.64 - 1.64) = 0.5 \tag{7}$$

Thus, in a scenario where the real power is 15%, and only z-scores greater than or equal to z_α are published, power based

---

[3]This result can also be confirmed by running the following command using R (R Core Team, 2014): `power.t.test(delta=0.1,sd=1,n=36,alternative = "one.sided",type="one.sample")`.

on the z-score will be overestimated by at least a factor of 0.5/0.15=3.33. Call this ratio the Power Inflation Index. Because studies with lower p-values are more likely to be published because they are regarded as more "reliable", the PII will tend to be even higher than this lower bound.

We turn next to a case study involving psycholinguistic data that illustrates the illusion of power.

Case study: Interference effects in reading studies

To illustrate the illusion of power and of replicability, we consider the 10 experiments that were reviewed in the literature review and meta-analysis presented in Jäger, Engelmann, and Vasishth (2017). These were psycholinguistic studies in which the dependent measure was reading time of words in milliseconds, and the experimental manipulation involved pairs of sentence types where one type was easier to read than the other. Thus, an appropriate statistical test is the paired t-test (alternatively, one could do a one-sided t-test, although this is less common in psycholinguistics).

We had the raw data from these 10 studies and so were able to carry out the pairwise comparison. As discussed in detail in Jäger et al. (2017), theory predicts an effect with a negative sign. The original results as published were analyzed on the raw milliseconds scale, but here we analyze the data on the log milliseconds scale because reading time data are log-normally distributed.

A summary of the pairwise t-test is shown in Table 1. From the table, it is clear that the studies consistently found negative values for the coefficient; this consistent result raises our confidence in the reproducibility of the result. A formal power analysis based on these studies, also shown in the last column of the table, leads to estimates of power ranging from 17 to 60%.

	t	d	n	se	s	pval	power
1	-1.88	-0.06	40	0.03	0.21	0.07	0.26
2	-3.13	-0.07	32	0.02	0.13	0.00	0.59
3	-1.45	-0.05	32	0.03	0.19	0.16	0.17
4	-2.08	-0.04	32	0.02	0.10	0.05	0.31
5	-1.71	-0.04	32	0.02	0.12	0.10	0.22
6	-2.58	-0.08	28	0.03	0.16	0.02	0.43
7	-1.56	-0.05	60	0.03	0.23	0.12	0.19
8	-3.18	-0.08	44	0.03	0.17	0.00	0.60
9	-1.87	-0.06	60	0.03	0.24	0.07	0.26
10	-2.64	-0.04	114	0.02	0.18	0.01	0.46

Table 1: Results from the paired t-tests for the 10 experimental comparisons. Shown are the t-score, the effect d in log ms, the sample size n, the standard error se, the standard deviation s, and the p-value. Also shown is the power estimated from each study.

Using a Bayesian random-effects meta-analysis to estimate the power function

In Table 1, we calculated power based on the individual studies. As discussed above, these will tend to be overestimates because there is a preference to publish effects with low p-values. How can we check this for the 10 studies? True power is unknown so we have no basis for comparing the power estimates from individual studies with a true value for power.

One way to arrive at a conservative estimate of the true power given these 10 studies is to carry out a Bayesian random-effects meta-analysis (Gelman et al., 2014). This hierarchical modelling approach allows us to determine the posterior distribution of the effect, which can then be used for computing an estimate of power. As discussed in Button et al. (2013), using estimates from a meta-analysis yields a more conservative estimate of power; in the random-effects meta-analysis, this conservativity arises due to the shrinkage property of hierarchical models: Larger sample studies receive a greater weighting in determining the posterior than smaller sample studies. Note, however, that even here the power may be an overestimate due to the fact that the studies that go into the meta-analysis are likely to have publication bias. But as we show below, the estimates of power from individual studies tend to be ever larger.

The random-effects meta-analysis model was set up as follows. Let y_i be the effect size in log milliseconds in the i-th study, where i ranges from 1 to n. Let μ be the true (unknown) effect in log ms, to be estimated by the model, and μ_i the true (unknown) effect in each study. Let σ_i log ms be the true standard deviation of the sampling distribution; each σ_i is estimated from the sample standard error from study i. The standard deviation parameter τ represents between-study variability.

Then, our model for n studies is as follows. The model assumes the i-th data point (the effect observed on the log ms scale) y_i is generated from a normal distribution with mean μ_i and some standard error σ, estimated from the sample's standard error. Each of the true underlying means μ_i are assumed to be generated from a normal distribution with true mean μ and between-study standard deviation τ. We assign Cauchy(0,2.5) priors to the parameters μ and μ_i, and a truncated Cauchy(0,2.5) prior for the between-study standard deviation τ, truncated so that τ is greater than 0. The model can be stated mathematically as follows:

$$
\begin{aligned}
\text{Likelihoods:} \\
y_i \mid \mu_i, \sigma_i^2 &\sim Normal(\mu_i, \sigma_i^2) \quad i = 1, \ldots, n \\
\mu_i \mid \theta, \tau^2 &\sim Normal(\mu, \tau^2), \\
\text{Priors:} \\
\mu &\sim Cauchy(0, 2.5), \\
\mu_i &\sim Cauchy(0, 2.5), \\
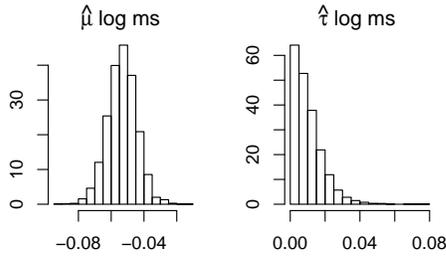\tau &\sim Cauchy(0, 2.5), \tau > 0
\end{aligned}
\tag{8}
$$

Figure 2: Posterior distributions of the estimated effect ($\hat{\mu}$), and the standard deviation of estimate of the between-study variability ($\hat{\tau}$) in the random-effects meta-analysis.

We fit the model using Stan 2.14.1 (Stan Development Team, 2016), running four chains with 2000 iterations (half of which were warm-ups). Convergence was successful, as diagnosed using the $\hat{R}$ diagnostic (Gelman et al., 2014). The posterior distributions of $\hat{\mu}$ and of the between-study standard deviation $\hat{\tau}$ are shown in Figure 2. The posterior mean of the effect is -0.05 log ms, with 95% credible interval [-0.07,-0.04]. Next, we use this estimate of the posterior distribution to compute a power distribution.

Computing the power distribution using the posterior distribution of the effect   An analysis of reading studies, including the ones considered here, showed that the precisions (the inverse of the variance) in reading time studies have mean values 16.3 and standard deviation 7.07 (the unit for precision is 1/log ms^2). Since precision can be modelled as a Gamma distribution, we assumed that precisions are distributed as $Gamma(\alpha = 5.3, \beta = 0.3)$. These parameters of the Gamma distribution were computed by taking the mean $\bar{x}$ and standard deviation s of the precisions, and then deriving the parameters of the Gamma distribution by solving for α and β. We use the fact that for a random variable generated from a Gamma distribution with parameters α and β, the expectation μ and variance σ^2 are:

$$E(X) = \frac{\alpha}{\beta} = \mu \quad \text{and} \; Var(X) = \frac{\alpha}{\beta^2} = \sigma^2 \qquad (9)$$

Having obtained the estimate of the effect (through the meta-analysis) and the distribution of the precisions, we used these estimates to carry out 10,000 Monte Carlo simulations to derive a power distribution for different sample sizes ($n = 20, \ldots, 50$) in the following manner. For each sample size, we repeatedly computed power after obtaining:

- one sample for the effect by sampling from the distribution $Normal(-0.05, 0.009)$; this is the posterior distribution of the effect derived from the random-effects meta-analysis;

- one sample for the precision by sampling from the $Gamma(5.3, 0.3)$, and then converting this to a standard deviation.

Such a Monte Carlo sampling procedure gives a probability distribution of power values and allows us to quantify our uncertainty about the estimated power by taking all sources of uncertainty into account—the uncertainty regarding the effect, and the uncertainty regarding the standard deviation.

Figure 3 shows the resulting power distributions with 95% credible intervals for power given different sample sizes. These power distributions are of course only estimates, not the true power; and as Button et al. (2013) point out, are probably slight overestimates if the studies themselves have publication bias.

The power distributions illustrate two important points. First, the range of plausible power values is remarkably low for typical sample sizes used in psycholinguistic reading experiments relating to interference effects (see Table 1). As an aside, we note that our estimates are similar to those from a recent review of 44 meta-analyses of research in social and behavioural sciences published between 1960-2011; they report a mean power of 0.24 with most studies suggesting power to be below 0.4 (Smaldino & McElreath, 2016, p. 6, Fig. 1). The second observation is that the power values computed from individual studies (the red dots) tend to be overestimates. The power distribution represents a conservative estimate of plausible values of power, and the power computed from individual studies tends to appear on or beyond the far right end of the power distribution. Of course, if statistical power were very high (approximately 80% or higher), then the overestimation problem would disappear or at least be negligible.

We can quantify the overestimation of power by computing the Power Inflation Index: the ratio of the power computed from individual studies to the power distribution computed using Monte Carlo simulations. If power is overestimated, then the distribution of the PII will be such that the mean ratio will be greater than 1. These distributions of PIIs are computed for a typical sample size used in psycholinguistic studies (n=20, 30, 40) in Table 2. Here, we can see that the PII can be a value as high as 12.

Discussion

We have shown that if statistical significance is used to decide whether to publish a result, overestimates of the effect will be published, leading to an overenthusiastic belief in the replicability of the effect.

Recently, the replication project reported by Open Science Collaboration (2015) showed that only 47% of the studies they investigated could be replicated. One factor causing these failures to replicate could have been low power in the original studies. Even before the replication project, Cohen (1962, 1988) and others have repeatedly warned against running low-powered studies. Despite these injunctions, many researchers do not believe that there is a problem of low
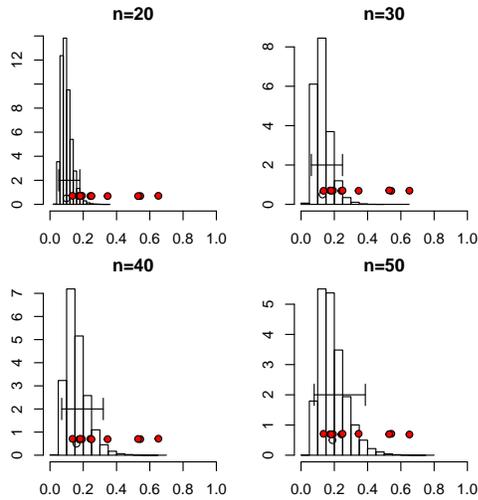
Figure 3: Power distributions for different sample sizes (log reading times). The histogram shows the power distribution (generated through Monte Carlo sampling; see text for details) along with 95% credible intervals. The red dots show power estimates from the 10 individual experimental comparisons considered in this case study.

	n=20		n=30		n=40	
Study	2.5%	97.5%	2.5%	97.5%	2.5%	97.5%
1	1.37	4.64	0.98	3.95	0.76	3.47
2	3.67	12.45	2.63	10.61	2.04	9.30
3	1.08	3.67	0.78	3.13	0.60	2.74
4	1.95	6.61	1.40	5.64	1.08	4.94
5	1.41	4.76	1.01	4.06	0.78	3.56
6	3.06	10.37	2.19	8.83	1.70	7.75
7	0.76	2.59	0.55	2.20	0.42	1.93
8	2.99	10.12	2.14	8.62	1.65	7.56
9	0.98	3.34	0.71	2.84	0.55	2.49
10	1.02	3.47	0.73	2.96	0.57	2.59

Table 2: The 95% credible intervals of the Power Inflation Index for each of the 10 experimental comparisons, for different sample sizes. The Power Inflation Index can be as large as 12.

power. For example, Gilbert, King, Pettigrew, and Wilson (2016) contested the 47% replication rate and argued that the replication rate may be much higher, perhaps even "statistically indistinguishable from 100%." The objections of Gilbert et al. (2016) were largely based on arguments about the lack of fidelity to the original design, but it is possible that, in addition to concerns about fidelity, Gilbert et al. are, like many researchers, generally overconfident about the replicability and robustness of their results. This overconfidence is also evident in reading research in psycholinguistics, where it is routine to run experiments with sample sizes ranging from 20 to 40 participants. Recent work has argued that sample sizes such as these may be too low for reading studies on interference (Jäger et al., 2017). We are hopeful that future work will take this finding into account when planning studies.

Currently, the replication problems in psycholinguistics are serious. In experimental work carried out in the first author's lab, although successful replications have been carried out (Bartek, Lewis, Vasishth, & Smith, 2011; Safavi, Husain, & Vasishth, 2016; Frank, Trompenaars, & Vasishth, 2015), the number of replication failures is remarkable. For example, in recent work (Mertzen, Jäger, & Vasishth, 2017) we carried out six replication attempts of two eyetracking experiments published in the *Journal of Memory and Language* (Levy & Keller, 2013). The two original studies had sample size 28 each, and each of our replication attempts had the same number and used the same items that were in the original studies. We were unable to replicate any of the claims in the paper. We have also failed to replicate published results (Paape & Vasishth, 2016; Husain, Vasishth, & Srinivasan, 2014) from our own lab. There is an urgent need to attempt to replicate published results, and not just in psycholinguistics. For example, Makel, Plucker, and Hegarty (2012) present a quantitative analysis of the low rate of successful replications in psychology (1%).

This low replication rate is suggestive that the concerns about low power, at least in reading studies, are not misplaced. Other fields are also affected. For example, Button et al. (2013) have shown that in neuroscience studies, power may also be quite low, ranging from 8 to 31%. Smaldino and McElreath (2016) have shown through a 50-year meta-analysis in behavioural science that power has not improved (mean power: 24%). In biomedical sciences, approximately 50% of studies have power in the 0-10% or 11-20% range (Dumas-Mallet, Button, Boraud, Gonon, & Munafò, 2017).

Despite these indications, many researchers remain overconfident about the robustness of their results. This overconfidence may be in part due to the statistical significance filter leading to illusion of power.

Concluding remarks

We have shown that the statistical significance filter directly leads to over-optimistic expectations of replicability of published research: disproportionately more statistically signifi-

cant results tend to get published, and power computed from these studies will be overestimates. Even if the researcher doesn't conduct any formal power analyses, they can fall prey to this illusion because of the informal assessment of replicability afforded by the statistical significance filter. We illustrate the illusion of power through a case-study involving 10 published experimental comparisons.

Acknowledgements

References

Bartek, B., Lewis, R. L., Vasishth, S., & Smith, M. (2011). In search of on-line locality effects in sentence comprehension. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *37*(5), 1178–1198.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376.

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

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

Dumas-Mallet, E., Button, K. S., Boraud, T., Gonon, F., & Munafò, M. R. (2017). Low statistical power in biomedical science: A review of three human research domains. *Royal Society Open Science*, 160254. doi: 10.1098/rsos.160254

Frank, S. L., Trompenaars, T., & Vasishth, S. (2015). Cross-linguistic differences in processing double-embedded relative clauses: Working-memory constraints or language statistics? *Cognitive Science*, n/a.

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

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

Gelman, A., & Tuerlinckx, F. (2000). Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics*, *15*(3), 373–390.

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on "estimating the reproducibility of psychological science". *Science*, *351*(6277), 1037–1037. doi: 10.1126/science.aad7243

Hedges, L. V. (1984). Estimation of effect size under non-random sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational Statistics*, *9*(1), 61–85.

Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician*, *55(1)*, 19–24.

Hung, H. J., O'Neill, R. T., Bauer, P., & Kohne, K. (1997). The behavior of the p-value when the alternative hypothesis is true. *Biometrics*, 11–22.

Husain, S., Vasishth, S., & Srinivasan, N. (2014). Strong expectations cancel locality effects: Evidence from Hindi. *PLoS ONE*, *9*(7), 1–14.

Ioannidis, J. P. (2008). Why most discovered true associations are inflated. *Epidemiology*, *19*(5), 640–648.

Jäger, L. A., Engelmann, F., & Vasishth, S. (2017). Similarity-based interference in sentence comprehension: Literature review and Bayesian meta-analysis. *Journal of Memory and Language*. (In Press) doi: 10.1016/j.jml.2017.01.004

Lane, D. M., & Dunlap, W. P. (1978). Estimating effect size: Bias resulting from the significance criterion in editorial decisions. *British Journal of Mathematical and Statistical Psychology*, *31*(2), 107–112.

Levy, R. P., & Keller, F. (2013). Expectation and locality effects in German verb-final structures. *Journal of Memory and Language*, *68*(2), 199–222.

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science*, *7*(6), 537–542.

Mertzen, D., Jäger, L. A., & Vasishth, S. (2017). The importance of replication in psycholinguistics. In *Proceedings of the 30th Annual CUNY Conference on Sentence Processing.* Boston, USA.

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

Paape, D., & Vasishth, S. (2016). Local coherence and preemptive digging-in effects in German. *Language and Speech*, *59*, 387-403.

R Core Team. (2014). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org

Safavi, M. S., Husain, S., & Vasishth, S. (2016). Dependency resolution difficulty increases with distance in Persian separable complex predicates: Implications for expectation and memory-based accounts. *Frontiers in Psychology*, *7*. doi: 10.3389/fpsyg.2016.00403

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

Stan Development Team. (2016). Stan modeling language users guide and reference manual, version 2.12 [Computer software manual]. Retrieved from http://mc-stan.org/