The statistical significance filter leads to overoptimistic expectations of replicability

Shravan Vasishth
Department of Linguistics, University of Potsdam, Potsdam, Germany
Daniela Mertzen
Department of Linguistics, University of Potsdam, Potsdam, Germany
Lena A. Jäger
Department of Linguistics, University of Potsdam, Potsdam, Germany
Andrew Gelman
Department of Statistics, Columbia University, New York, USA

January 17, 2018

Abstract

Treating a result as newsworthy, i.e., publishable, because the p-value is less than 0.05 leads to overoptimistic expectations of replicability. The underlying cause of these overoptimistic expectations is Type M(agnitude) error (Gelman & Carlin, 2014): when underpowered studies yield significant results, the effect size estimates are invariably exaggerated and noisy. These effects get published, leading to an illusion that the reported findings are robust and replicable. For the first time in psycholinguistics, we demonstrate the adverse consequences of this *statistical significance filter*. We do this by carrying out direct replication attempts of published results from a recent paper. Six experiments (self-paced reading and eyetracking, 168 participants in total) show that the published (statistically significant) claims are so noisy that even non-significant results are fully compatible with them. We also demonstrate the stark contrast between these small-sample studies and a larger-sample study (100 participants); the latter yields much less noisy estimates but also a much smaller magnitude of the effect of interest. The small magnitude looks less compelling but is more realistic. We suggest that psycholinguistics (i) move its focus away from statistical significance, (ii) attend instead to the precision of their estimates, and (iii) carry out direct replications in order to demonstrate the existence of an effect.

Keywords: Type M error; replicability; surprisal; locality; Bayesian data analysis

Introduction

Imagine that a reading study shows a difference between two means that has an estimate of 77 ms, with standard error 30, that is, with $p = 0.01$. Now suppose instead that the same study had shown an estimate of 40 ms, also with a standard error of 30; this time $p = 0.18$. The usual reporting of these two types of results—as significant and therefore publishable, and not significant and therefore not publishable—is misleading because it implies an inappropriate level of certainty in both cases. Indeed, we believe that this routine attribution of certainty to noisy data is a major contributor to the current replication crisis in science (Open Science Collaboration, 2015; Pashler & Wagenmakers, 2012): the issue is not just the high frequency of failed replications, but also that these failed replications arise in an environment where routine success (defined as $p < 0.05$) is expected.

We will refer to this $p < 0.05$ decision criterion for publication-worthiness as the *statistical significance filter*. One adverse consequence of the statistical significance filter is that it leads to findings that are positively biased (Gelman, 2017). When we use the null hypothesis significance testing (NHST) framework, we tend to focus on Type I error but generally neglect considerations of power—the probability of correctly rejecting the null hypothesis given a particular true value of the effect of interest. Power can be estimated if we know the sample size, the standard deviation of the dependent measure, and have a reasonable guess about the magnitude of the effect, based on, for example, theory or the predictions of a computational model.

NHST can work well when power is relatively high, say 70-80% or higher. But when power is low, published studies that show statistical significance will not only have a positive bias, they will tend to have greatly exaggerated estimates (see Appendix A for a formal argument). This is demonstrated in Figure 1 using simulated data: for a low-power scenario, the estimates from repeated samples tend to fluctuate around the true value, and can also have the wrong sign. When an effect is significant, it is also exaggerated. By contrast, when power is high, the estimates under repeated sampling tend to be close to the true value. Gelman and Carlin (2014) refer to these exaggerated estimates as Type M(agnitude) errors.

Unfortunately, doing a power analysis based on the effect sizes reported in the literature can also be problematic. Whenever an effect in an underpowered study comes out significant, it is necessarily an overestimate. In fields where power tends to be low, these overestimates fill the literature. Now consider what happens when we design a new study. We read the literature and see large effects, which become the basis for our next study. If we ever conduct a formal power analysis based on these exaggerated effects, we are bound to get an exaggerated estimate of power, and can easily convince ourselves that we have an appropriately powered study. Usually, in psycholinguistics, we 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. From this we conclude that the effect must be reliable, and therefore replicable. The reality may be very different.

Although the above observations are well-known in statistics (see the discussion in Wasserstein & Lazar, 2016), they are not widely appreciated in psycholinguistics. Our goal in this paper is to demonstrate—not via simulation but through actual replication attempts

Please send correspondence to vasishth@uni-potsdam.de.

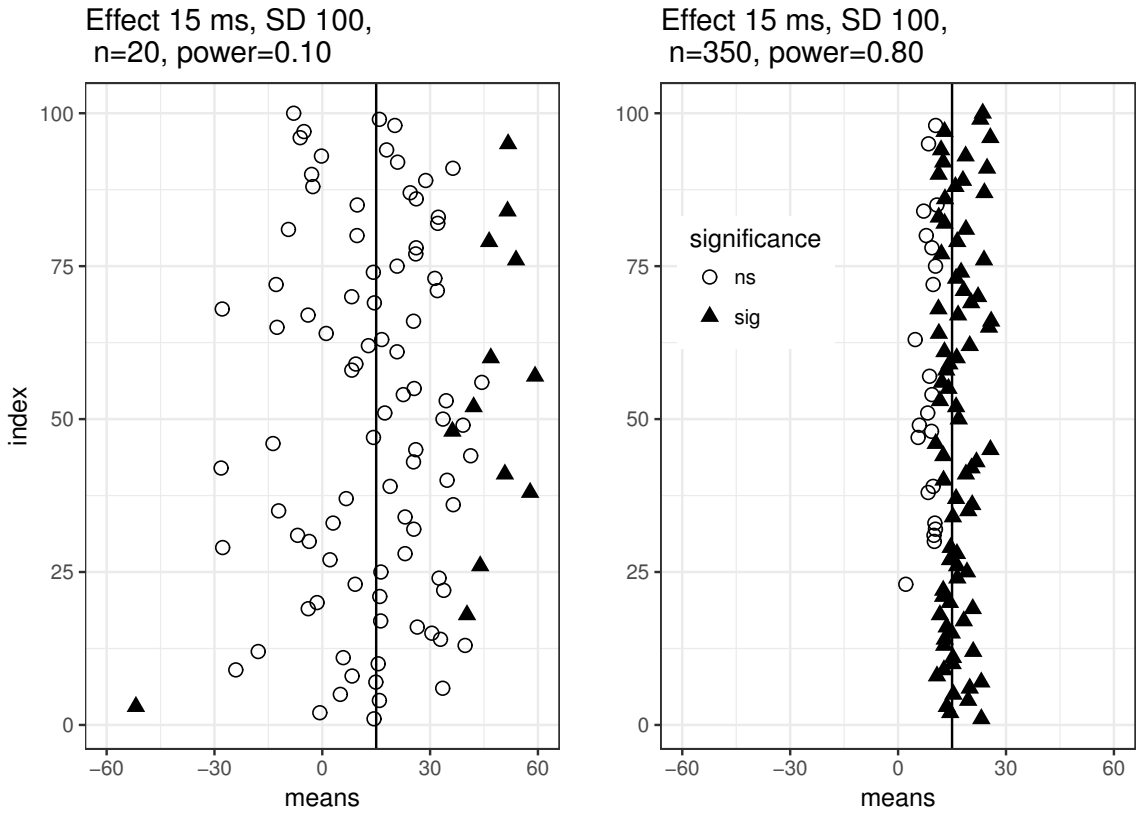

Figure 1. A demonstration of Type M error using simulated data. We assume that the data are generated from a normal distribution with mean 15 ms and standard deviation 100. The true mean is shown in each plot as a vertical line. When power is low, under repeated sampling, whenever the estimates of an effect come out significant, the values are overestimates and can even have the wrong sign. When power is high, significant effects are tightly clustered near the true mean.

of a published empirical claim—that relying exclusively on statistical significance to decide whether a result is newsworthy leads to misleading conclusions.

We show through a case study that small-sample experiments can easily deliver statistically significant results that are exaggerated and non-replicable. For this case study, we chose a paper by Levy and Keller (2013) that investigated expectation and locality effects in sentence comprehension. We selected this particular case study because there are no *a priori* reasons to doubt the results in the paper, as they are theoretically well-founded and have plenty of independent empirical support.

Anticipating our conclusions, we suggest that researchers and journals avoid focusing exclusively on statistical significance to evaluate the validity and reliability of studies. Validity should be established by running as high-precision a study as possible; and reliability should be established through direct replication.

Case study: The effects of expectation vs. memory retrieval in sentence processing

Background

Levy and Keller (2013) published two eyetracking studies in the *Journal of Memory and Language* in which they tested the predictions of two well-established theoretical proposals in sentence processing research: the expectation-based account (Hale, 2001; Levy, 2008) and the memory-based retrieval accounts (Gibson, 1998, 2000; Lewis & Vasishth, 2005).

The expectation-based account, as developed by Levy (2008), predicts that intervening material between, for example, a subject and its verb, facilitates processing at the verb. To illustrate this point, consider the discussion by Levy (2008) of the following sentences from an eyetracking (reading) study conducted by Konieczny and Döring (2003).

(1) a. Die Einsicht, dass [$_{NOM}$ der Freund] [$_{DAT}$ dem Kunden] [$_{ACC}$ das Auto aus
The insight, that the friend the client the car from
Plastik] verkaufte,...
plastic sold,...
'The insight that the friend sold the client the plastic car....'

b. Die Einsicht, dass [$_{NOM}$ [der Freund] [$_{GEN}$ des Kunden]] [$_{ACC}$ das Auto
The insight, that the friend of the client the car from
aus Plastik] verkaufte,...
plastic sold,...
'The insight that the friend of the client sold the plastic car....'

Konieczny and Döring found that regression path durations at the verb *verkaufte* in (1a) were shorter than in (1b) (555 vs. 793 ms). Levy's explanation for this facilitation is that the dative noun phrase (NP) in (1a) sharpens the expectation for the verb to a greater degree than in (1b): in the former, nominative, accusative, and dative NPs narrow the range of possible upcoming verb phrases more than in the latter, where only nominative and accusative NPs have been seen. Levy formalizes this idea in terms of surprisal (Hale, 2001), which essentially states that the conditional probability of the verb phrase appearing given the preceding context determines processing difficulty: the more predictable the verb phrase, the easier it is to process. Using a probabilistic context-free grammar of German, Levy shows that syntactic surprisal is lower in (1a) than (1b) (23.51 vs. 23.91 bits); this suggests that surprisal may be a good explanation for the facilitation effect seen in Konieczny and Döring (2003).

A competing class of theories of sentence processing difficulty makes the incorrect prediction for the reading time pattern observed at the verb in the Konieczny and Döring study. For example, the Dependency Locality Theory (Gibson, 2000) assumes that processing difficulty (and therefore reading time) at a verb is a linear function of the distance between the verb and its arguments; distance here is measured in terms of the number of discourse referents intervening between co-dependents. Under such an account, no difference is predicted between the two sentences above, because the same number of new discourse referents intervenes between the subject and verb in (1a) and (1b).

Levy and Keller (hereafter, LK) built on the work of Konieczny and Döring by developing a novel experimental design that cleverly pits the expectation-based and memory-

based accounts against each other. LK's studies are described next, as they form the basis for our replication attempts.

The experiment design by Levy and Keller, 2013

As shown in Table 1, in their items for their Experiment 1, a dative NP and a prepositional adjunct either appeared in a subordinate clause or a main clause. A corpus analysis carried out by LK showed that if either or both of these phrases appeared in the main clause, the main clause verb (*versteckt*, 'hidden') had lower surprisal values. Thus, the critical region in this experiment was the verb *versteckt*; the post-critical region was defined as the two words following the matrix verb (*und somit*, 'and thus', in the example shown in Table 1).

Their Experiment 2 had a similar design, except that syntactic complexity was increased by embedding the main clause of Experiment 1 within a relative clause (see Table 2). Here, the critical region was the head verb of the relative clause and the auxiliary (*versteckt hat*, 'hidden had', in Table 2) and the post-critical region was the noun phrase (here, *die Sache*, 'the affair'). Note that the two experiments take advantage of the head-final property of German: the verb always appears clause-finally in these constructions. Since all the arguments precede the verb, it is easy to investigate the effect of verb predictability conditional on having seen all the arguments.

Predictions for the LK study

As discussed on page 203 of Levy & Keller, 2013, using a corpus analysis, the predictions of the expectation-based account can be derived for both experiments. Because interposing a dative NP or an adjunct sharpens the expectation for a participial verb, reading time at the main verb in condition (b) should be faster than in (a), and condition (d) should be faster than (c). No difference is expected between (b) and (c). It follows that (d) should show the fastest reading time, i.e., the most facilitation (see Figure 2, left panel; our figures here are a reproduction of LK's Figure 1 on their page 2013, which lays out their predictions). Levy (2008) and others refer to such predicted speedups as *expectation effects*.

Memory-based theories make different predictions. Because intervening discourse referents between the subject and the verb should generally lead to greater processing difficulty, placing the dative NP or the adjunct in the main clause should lead to a slowdown at the verb, and placing both the dative NP and the adjunct in the main clause should lead to an even greater slowdown at the verb. This means that reading time at the critical verb in condition (b) should be slower than (a), and condition (d) should be slower than (c); in fact, (d) should show the greatest slowdown in reading time, because it is associated with the highest processing cost (see Figure 2, right panel). Gibson (2000) and others often refer to these slowdowns as *locality effects*.

One nice property of the LK design is that the verb position is always constant across conditions being compared: the intervening phrases (dative and adjunct) always appear in the sentence, either intervening between the subject and verb or at the beginning of the sentence. This resolves a potentially serious confound in such studies; many of the previous studies (Grodner & Gibson, 2005; Konieczny, 2000; Vasishth & Lewis, 2006) had the verb further downstream in the sentence whenever an intervener was present. This positional

confound makes comparisons across conditions difficult to interpret: if a verb appears later in the sentence, this alone may lead to slowdowns or speedups compared to a baseline condition (see Vasishth, 2003 for discussion).

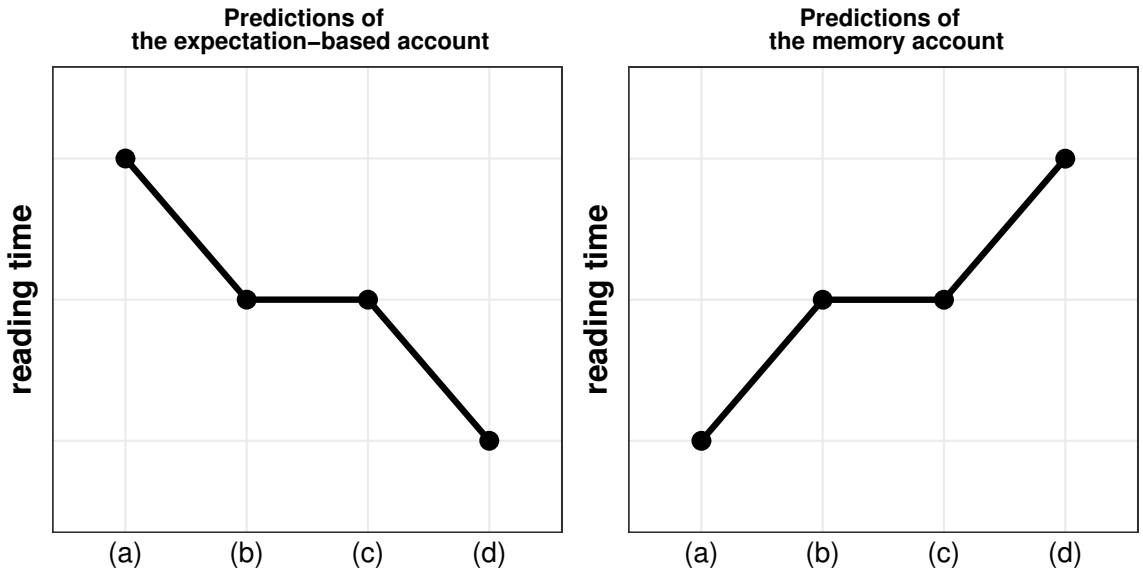

Figure 2. Predictions for the Levy and Keller Experiments 1 and 2: The left panel shows the speedup predicted by the expectation account. The right panel shows the slowdown predicted by memory-based theories. This figure is based on Figure 1 of Levy and Keller (2013).

Table 1

Example items for LK's Experiment 1 (simplified).

a. PP adjunct in subordinate clause, dative NP in subordinate clause

Nachdem	der	Lehrer	[**ADJ** zur Ahndung]	[**DAT** dem Sohn]	...,				
After	*the*	*teacher*	[**ADJ** *as payback*]	[**DAT** *the son*]	...,				
hat	Hans	Gerstner			den	Fußball	**versteckt**,	und	somit...
has	*Hans*	*Gerstner*			*the*	*football*	*hidden*,	*and*	*thus...*

b. PP adjunct in main clause, dative NP in subordinate clause

Nachdem	der	Lehrer		[**DAT** dem Sohn]	...,				
After	*the*	*teacher*		[**DAT** *the son*]	...,				
hat	Hans	Gerstner	[**ADJ** zur Ahndung]		den	Fußball	**versteckt**,	und	somit...
has	*Hans*	*Gerstner*	[**ADJ** *as payback*]		*the*	*football*	*hidden*,	*and*	*thus...*

c. PP adjunct in subordinate clause, dative NP in main clause

Nachdem	der	Lehrer	[**ADJ** zur Ahndung]		...,				
After	*the*	*teacher*	[**ADJ** *as payback*]		...,				
hat	Hans	Gerstner		[**DAT** dem Sohn]	den	Fußball	**versteckt**,	und	somit...
has	*Hans*	*Gerstner*		[**DAT** *the son*]	*the*	*football*	*hidden*,	*and*	*thus...*

d. PP adjunct in main clause, dative NP in main clause

Nachdem	der	Lehrer			...,				
After	*the*	*teacher*			...,				
hat	Hans	Gerstner	[**ADJ** zur Ahndung]	[**DAT** dem Sohn]	den	Fußball	**versteckt**,	und	somit...
has	*Hans*	*Gerstner*	[**ADJ** *as payback*]	[**DAT** *the son*]	*the*	*football*	*hidden*,	*and*	*thus...*

'After the teacher imposed detention classes, Hans Gerstner hid the football from the naughty son of the industrious janitor as additional payback for the multiple wrongdoings, and thus corrected the affair.'

Table 2

Example items for LK's Experiment 2 (simplified)

a. PP adjunct in subordinate clause, dative NP in subordinate clause

Nachdem	der	Lehrer	[**ADJ** zur Ahndung]	[**DAT** dem Sohn]	...,					
After	*the*	*teacher*	[**ADJ** *as payback*]	[**DAT** *the son*]	...,					
hat	der	Mitschüler, der			den	Fußball	**versteckt hat**,	die	Sache...	
has	*the*	*classmate, who*			*the*	*football*	*hidden had*,	*the*	*affair...*	

b. PP adjunct in relative clause, dative NP in subordinate clause

Nachdem	der	Lehrer		[**DAT** dem Sohn]	...,					
After	*the*	*teacher*		[**DAT** *the son*]	...,					
hat	der	Mitschüler, der	[**ADJ** zur Ahndung]		den	Fußball	**versteckt hat**,	die	Sache...	
has	*the*	*classmate, who*	[**ADJ** *as payback*]		*the*	*football*	*hidden had*,	*the*	*affair...*	

c. PP adjunct in subordinate clause, dative NP in relative clause

Nachdem	der	Lehrer	[**ADJ** zur Ahndung]		...,					
After	*the*	*teacher*	[**ADJ** *as payback*]		...,					
hat	der	Mitschüler, der		[**DAT** dem Sohn]	den	Fußball	**versteckt hat**,	die	Sache...	
has	*the*	*classmate, who*		[**DAT** *the son*]	*the*	*football*	*hidden had*,	*the*	*affair...*	

d. PP adjunct in relative clause, dative NP in relative clause

Nachdem	der	Lehrer			...,					
After	*the*	*teacher*			...,					
hat	der	Mitschüler, der	[**ADJ** zur Ahndung]	[**DAT** dem Sohn]	den	Fußball	**versteckt hat**,	die	Sache...	
has	*the*	*classmate, who*	[**ADJ** *as payback*]	[**DAT** *the son*]	*the*	*football*	*hidden had*,	*the*	*affair...*	

'After the teacher imposed detention classes, the classmate who hid the football from the naughty son of the industrious janitor as additional payback for the multiple wrongdoings corrected the affair.'

A re-analysis of the LK data

The two studies by LK had 28 participants and 24 items each. In their paper, statistical summaries and analyses for the critical and post-critical regions were prepared using the `lme4` package (Bates, Maechler, Bolker, & Walker, 2015) in R. They released their data to us, which allowed us to carry out the same analyses as they did, but within a Bayesian framework (Gelman et al., 2014) using the probabilistic programming language Stan (Carpenter et al., 2016).

Motivation for using Bayesian data analysis. A major advantage of using the Bayesian approach here is that we can compute the posterior distributions of the effects of interest. A posterior distribution furnishes a range for plausible values of an effect given the data and the model. This allows us to focus on a crucial aspect that we wish to discuss in the present paper: the uncertainty of the estimates. A further advantage is that we can also fit so-called "maximal" models with full covariance matrices for by-participant and by-item variance components (Barr, Levy, Scheepers, & Tily, 2013). Such maximal models often fail to converge in `lme4` for small data-sets and yield poor estimates of the variance components (for an example using data from Barr et al., see Bates, Kliegl, Vasishth, & Baayen, 2015).

Throughout this paper, we will summarize the posterior distributions with their mean and the 95% credible interval. This interval demarcates the range over which we are 95% certain (given the data and the model) that the true parameter lies. The credible interval therefore allows us to do something that a frequentist confidence interval cannot: quantify our uncertainty about the estimate of interest. (For differences between the Bayesian credible interval and the frequentist confidence interval, see Hoekstra, Morey, Rouder, & Wagenmakers, 2014).

Statistical methodology. As in the original study, we investigated the main effects of dative position (Dative) and adjunct position (Adjunct) and their interaction, using the same contrast coding that LK employed. The coding is shown in Table 3. A positive coefficient for the main effect of Dative or Adjunct means that a speedup in reading time is seen when the dative NP (respectively, the adjunct) appears within the main clause (Expt 1) or relative clause (Expt 2), i.e., when it is interposed between the grammatical subject and the verb.

Condition			Dat	Adj	DatxAdj
a . . . [Subj . . .		. . . Verb]	0.5	0.5	0.5
b . . . [Subj . . .	ADJ . . .	Verb]	0.5	-0.5	-0.5
c . . . [Subj . . .	DAT	. . . Verb]	-0.5	0.5	-0.5
d . . . [Subj . . .	DAT ADJ . . .	Verb]	-0.5	-0.5	0.5

Table 3
The contrast coding used for main effects of Dat(ive), Adj(unct), and their interaction for the two experiments by Levy and Keller (2013). The structures used in the four conditions are shown schematically; note that the verb was always in the same position because the interveners (Dat and Adj) either intervened between the subject and the verb, or appeared before the subject. In Experiment 1, the subject-verb dependency was in the main clause, and in Experiment 2, it was within a relative clause.

The reading times were log-transformed (Gelman & Hill, 2007; Vasishth & Nicenboim, 2016) and a hierarchical (linear mixed) model was fit with full covariance matrices for participants and for items (the "maximal" model recommended by Barr et al., 2013). All the code and data will be available on publication from http://bit.ly/SSFilter. In all the Stan models, weakly informative priors (Gelman et al., 2014) were used for all parameters and hyperparameters. For all parameters, the prior distribution was defined as the standard normal distribution, $\mathcal{N}(0, 1)$; for variance components these were truncated at 0 (because standard deviations cannot be less than 0). The posteriors are not dependent on these specific priors; other choices (such as a Cauchy prior) lead to similar posterior distributions. For the correlation parameters in the variance-covariance matrix of the random effects, we defined so-called regularizing LKJ priors on the correlation matrix (Stan Development Team, 2016); see Sorensen, Hohenstein, and Vasishth (2016) for a tutorial. For each model, we ran four chains with 2000 iterations, the first half of which were a warm-up and were consequently discarded. Convergence was checked by visually inspecting the chains and via the R-hat convergence diagnostic (Gelman et al., 2014).

The estimates for the main effects and interaction were back-transformed to reading times in milliseconds.

Question-response accuracy in the LK data. Half of the 24 items were followed by comprehension questions that had yes/no responses. Accuracy on the target items was 69% in Experiment 1 and 65% in Experiment 2 (personal communication from Frank Keller).

Reading time results in the LK data. It is standard in eyetracking reading research to argue for an effect if just *any* of several dependent measures examined show an effect. For example, Konieczny and Döring (2003) found their effect only in regression path durations. In the LK studies, which take as a starting point the Konieczny and Döring design, regression path duration showed no effect at all; instead, other measures showed statistically significant effects. We avoid this approach and instead try to reproduce the effect in one dependent measure that LK would consider representative of their claims. The LK paper presents a graphical summary of their effects using total reading times for the two experiments; see LK's Figures 3 and 4 (pages 209 and 214, Levy & Keller, 2013). Because the graphical summary using total reading times was considered by LK to be a representative summary of their overall claims, below we only report the analyses involving total reading times. However, we also analyzed all the dependent measures (critical and post-critical regions) in which they found statistical significance for their main claims. These were first-pass and re-reading times in Experiment 1, and re-reading times, regression probability, and skipping probability in Experiment 2 (in the critical or post-critical region). None of these came out statistically significant.

Inflation of Type I error under multiple analyses is another important reason why one might want to limit the number of dependent measures a priori. Analyzing multiple dependent measures, as is normally done in eyetracking data analyses in psycholinguistics, greatly increases Type I error probability (von der Malsburg & Angele, 2017). For example, LK analyzed eight dependent measures in two regions of interest. Thus, for each experiment, 16 statistical tests were carried out, so for each of the three predictors (the effect of Dative, Adjunct, and their interaction) a total of 32 tests were conducted. Assuming that a p-value less than 0.05 is a statistically significant outcome, Dative showed six significant effects, Adjunct showed one significant effect, and the interaction showed eight significant effects.

As von der Malsburg and Angele (2017) recently showed, in eyetracking, and other types of data where many analyses are carried out, it is vitally important to correct Type I error probability when relying on the frequentist paradigm to carry out statistical inference.

Results of the re-analysis of the LK experiments

Our estimates of total reading times match LK's published results quite closely (see their Tables 6 and 9 on pages 208 and 213). Note that LK's estimates for the interaction term are twice as large as ours; this is only because they multiplied together their main effects, coded ± 0.5, to obtain their interactions, resulting in the interaction in their analyses being coded as ± 0.25.

The results of our re-analysis of the LK Experiments 1 and 2 are summarized in Figure 3. Recall that the critical region is the main clause verb in Experiment 1, and the relative clause verb in Experiment 2. The post-critical region consisted of the two words following the verb. As shown in Figure 3, an analysis of total reading times suggests the following:

1. In Experiment 1, at the critical region, Dative has the estimate 81 ms, with a 95% credible interval [19, 147]. The positive coefficient has the interpretation that interposing the dative NP between the subject and the verb leads to facilitation, as predicted by the expectation-based account. LK explain this result as follows (page 214):

   > "[The main effect of Dative] can be explained by assuming that the presence the [sic] additional preverbal material allows the processor to predict the upcoming verb, which leads to a facilitation effect."

2. In Experiment 2, at the post-critical region, the estimate of the interaction between Dative and Adjunct is 83 ms [19, 151]. LK's interpretation is that having both the Dative and Adjunct interposed between the subject-verb dependency leads to a slowdown. LK explain this outcome in terms of locality effects emerging under high memory load, i.e., when the subject-verb dependency is embedded inside the relative clause (page 214 of Levy & Keller, 2013):

   > "[The interaction] suggests the presence of a locality effect, i.e., the additional material that needs to be integrated at the verb, leading to a distance-based cost. This effect was only present in Experiment 2, which tested relative clauses, rather than main clauses as in Experiment 1. This suggests that locality effects can override expectation effects under conditions of high memory load, as we hypothesized would be most likely to occur in a relative clause."

We were interested in replicating these effects because they are consistent with a large body of evidence for both expectation and memory-based accounts of sentence processing. There is compelling evidence consistent with the expectation-based account proposed by Levy (2008); some examples are the work of Jäger, Chen, Li, Lin, and Vasishth (2015); Linzen and Jaeger (2016); Wu, Kaiser, and Vasishth (2017). Similarly, many studies show

evidence for memory-based effects; see, for example, Bartek, Lewis, Vasishth, and Smith (2011); Grodner and Gibson (2005); Van Dyke and Lewis (2003). Given the literature, it makes sense that we see effects of memory retrieval only under high processing load induced by encountering a relative clause: most of the demonstrations of locality effects in the literature have involved embedded clauses such as those of LK's Experiment 2. Thus, the LK finding that memory load modulates whether expectation effects are observed is quite convincing given theory and existing data.

Although the significant effects are convincing given the prior literature, one striking aspect of the LK estimates is their large uncertainty. The evidence for the first conclusion above comes from an estimate with mean 81 ms, but the 95% credible interval is 19 to 147 ms; and the evidence for the second conclusion comes from an estimate with mean 83 ms, with credible interval 19 to 151 ms. These wide uncertainties imply that the effect could in fact be much smaller.

There is good reason to believe that reading time effects relating to memory-based retrieval may be quite small. Nicenboim, Vasishth, Engelmann, and Suckow (2018) carried out a self-paced reading study investigating number interference in German with 184 participants. They estimated the magnitude of the memory retrieval effect in number interference to be 9 ms with 95% credible interval [0,18]. The meta-analysis by Jäger, Engelmann, and Vasishth (2017) has also shown that studies on memory retrieval effects with 20-40 participants may have power as low as 10-20%. If memory retrieval effects generally have a small magnitude in reading studies, and if a sample size of 28 participants leads to low power, LK's estimates may well be exaggerated. Their estimates have very large standard errors, a characteristic of low-powered studies. For example, if the true effect is in the range 5-20 ms, and if standard deviation is 100 ms (an extremely optimistic estimate for total reading time), using 28 participants, because of Type M error, it would be impossible to obtain statistically significant results *that are also accurate estimates of the effect*. In this scenario, power would lie between 6% and 18%.

How can we determine whether the effects in the LK studies are the result of Type M error? One obvious way is to re-run the original experiments repeatedly *with the same sample size as the original studies*. If power is high, we should be able to reproduce the effect consistently. On the other hand, if power is low, under repeated sampling, we would fail to detect the effect in the majority of cases. Thus, it will be very informative to actually conduct direct replication attempts of the LK experiments using the same sample size that they had.

We began by trying to replicate the two significant effects found by LK: the main effect of Dative in Experiment 1 (critical region), and the interaction between Dative and Adjunct in Experiment 2 (post-critical region). We did this by conducting four experiments: two self-paced reading studies of the two LK studies, and two eyetracking studies. We chose these two methods because they are the two standard behavioral approaches for studying cognitive processing costs in reading, and most of the previous research on expectation-based effects and memory effects has relied on either self-paced reading or eyetracking.

Four replication attempts (two self-paced reading and two eyetracking studies)

Participants. For each of the two self-paced reading experiments and the two eyetracking studies, we used the same participant sample size as LK (28). Thus, the total number of participants in these four studies was 112. Participants were native German undergraduate students from the University of Potsdam who were permitted to take part in only one of the replication studies. All had normal or corrected-to-normal vision, and received 7 Euros or course credit for their participation.

Experimental design and materials. We followed the 2×2 fully-crossed factorial design of the original study. We used the same 24 experimental items, 48 filler items, and the same comprehension questions that were employed in LK's Experiments 1 and 2.

In the eyetracking experiments, the critical and post-critical regions were as in the LK studies; in the self-paced reading studies, due to an oversight, the post-critical region consisted of only one word (in the LK studies, the post-critical region consisted of two words).

In two experimental items, a non-critical part of the sentence was changed; one due to a plausibility issue and another due to a repetition of an NP within one sentence. One comprehension question following one of the experimental items was replaced due to ambiguity of the question. For further details on these changes, please see the supplementary materials, available from http://bit.ly/SSFilter.

Procedure: Self-paced reading studies. Experimental items were presented word-by-word in a centered self-paced reading experiment using Linger.[1] As in the original studies, half the items were followed by yes/no questions. Due to the length of the sentences, non-critical regions were presented phrase-by-phrase. The experiment began after four practice trials. Participants were required to press the space bar on a keyboard to move on to each subsequent word or phrase; in trials with comprehension questions, they recorded a response via a button press. The experimental procedure lasted approximately 35 minutes. For the purposes of future direct replication, all materials and relevant software settings can be obtained from http://bit.ly/SSFilter.

Procedure: Eyetracking studies. The experimental procedure was identical in all of our eyetracking experiments. Participants' eye movements (right eye monocular tracking) were recorded with an EyeLink 1000 eye-tracker (SR Research[2]) with a desktop-mounted camera system at a sampling rate of 1000 Hz. The participant's head was stabilized using a chin/forehead rest. Stimuli were presented on a 22-inch monitor with a 1680×1050 screen resolution. The eye-to-screen distance measured approximately 66 cm. For the experimental presentation, SR Research Experiment Builder software was used. Stimuli were presented in a monospaced font (Courier new) with font size 24 and were arranged on the presentation screen such that the critical region always appeared in the same position (fourth word on the fourth and final line). Each session began with the calibration of the eyetracker and four practice trials preceding the experimental materials. Re-calibrations were carried out when necessary. In 50% of the trials, a comprehension question had to be answered by pressing a button on a gamepad. The entire procedure lasted approximately 40 minutes.

Our procedure differed from the one used by LK in the following aspects. The original

[1] See http://tedlab.mit.edu/~dr/Linger/.

[2] http://www.sr-research.com/eyelink1000.html

LK experiments were run with an SR Research Eyelink II eyetracker with a head-mounted camera system at a sampling rate of 500 Hz using Eyetrack software[3] for the experimental presentation. In LK's Experiment 1, the materials were presented in a non-monospaced font (Times New Roman, font size 20), whereas in their Experiment 2 the materials were presented in a monospaced font (Lucida Console, font size 14). The position of the critical verb differed in their two experiments: In LK's Experiment 1, the critical verb appeared in the middle of either the third or fourth line of the presented text, whereas in their Experiment 2 the critical verb was always the fourth word of the fourth line.

Results of the four replication attempts

The question-response accuracies for the SPR replications of LK's Experiments 1 and 2 were 66 and 61%; and for the eye-tracking replications, they were 64 and 60%.

Figure 3 summarizes the results of our four experiments. Before we discuss these, it is necessary to define what counts as a successful replication. A successful replication can mean that a statistically significant result in the original study is also found to be significant in the replication attempt. Alternatively, a successful replication could mean that the estimated means from a replication attempt fall within the 95% credible intervals of the original estimates.

If statistical significance is taken as a criterion for successful replication, we failed to replicate the two key effects in the LK studies: the main effect of Dative in Experiment 1 (critical region), and the interaction of Dative and Adjunct in Experiment 2 (post-critical region). If a frequentist p-value were to be computed for these effects, none would come out even close to significant in any of the four attempts. The means and 95% credible intervals for the critical comparisons in each experiment are as follows:

- SPR replication of Expt 1: Effect of Dative in critical region -10 ms [-44,25].
- Eyetracking replication of Expt 1: Effect of Dative in critical region 18 ms [-20,55].
- SPR replication of Expt 2: Interaction of Dative and Adjunct in post-critical region -16 ms [-50,15].
- Eyetracking replication of Expt 2: Interaction of Dative and Adjunct in post-critical region 29 ms [-16,80].

However, the replication attempts can also be seen as a near-complete success: *all* the total reading times estimates from the eyetracking studies (and most of the estimates from self-paced reading) fall within the 95% credible intervals of the original studies.

The crucial point here is that the original estimates are so noisy that, despite the fact that some of the effects in the original paper were statistically significant, the wide credible intervals are consistent with the true value being very small, and with the effect being near 0 ms.

When the estimates are noisy, the p-value furnishes no information about reliability (i.e., that the effect is true) or replicability (i.e., that the finding can be reproduced if the study is repeated).

[3] https://blogs.umass.edu/eyelab/software/

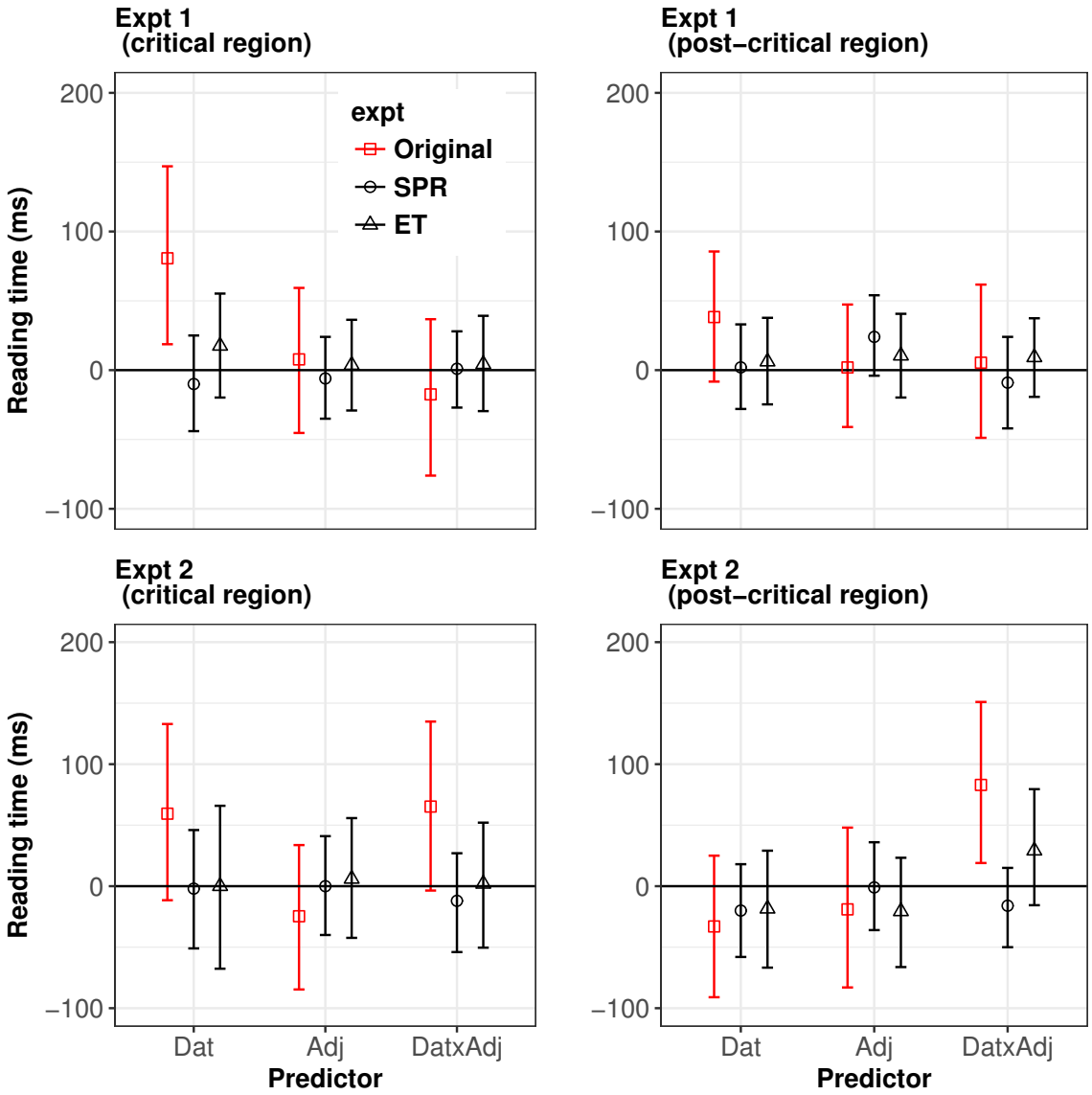

Figure 3. The effects of Dative and Adjunct interposition (and their interaction) at the critical and post-critical regions. Shown are the mean and 95% credible intervals from the original LK Experiments 1 and 2, and the two replication attempts.

In these first four small-sample replication attempts above, we aimed to show that the original estimates are noisy and therefore uninformative, despite being statistically significant.

Next, we demonstrate how a larger-sample study can yield a much more informative conclusion. Our starting point was one of the conclusions that LK drew from their study (page 214 of Levy & Keller, 2013). The emphasis is ours.

> "[The interaction] suggests the presence of a locality effect, i.e., the additional material that needs to be integrated at the verb, leading to a distance-based cost. *This effect was only present in Experiment 2, which tested relative clauses, rather than main clauses as in Experiment 1.*"

Here, LK are pointing to the fact that the interaction between Dative and Adjunct was found in Experiment 2 but not in Experiment 1. We will refer to this difference between the two experiments as the **load-distance interaction**. Our goal here is to show how the estimates of the effect change under a larger-sample replication attempt.

Investigating the load-distance interaction

LK describe the load-distance interaction in their General Discussion in the following manner:

> "[Experiment 1 showed] that the presence of a dative noun phrase led to decreased reading time at the corresponding verb, compared to a condition in which there is no preceding dative noun phrase.
>
> "Experiment 2 showed an interaction of adjunct position and dative position, with the verb more difficult to process when both the adjunct and the dative phrase were present than when only one was present.
>
> "[O]urs is the first demonstration to our knowledge that both expectation and locality effects can occur in the same structure in the same language, and that the two effects interact with each other."

This claimed interaction between expectation and locality can be investigated in several different ways. One way to interpret the interaction is in terms of the contrast in reading time patterns in their Experiment 1 vs 2. LK's Figures 3 and 4 (pages 209 and 214 in Levy & Keller, 2013), which summarize total reading times at the critical region, clearly show that Experiment 1 exhibits a speedup in (d) vs (c), whereas Experiment 2 exhibits a slowdown in these conditions (see Tables 1 and 2 for the items). Although visual inspection of the figures does suggest a cross-over interaction between load and distance, as Nieuwenhuis, Forstmann, and Wagenmakers (2011) have pointed out, the interaction must be formally tested. Such an interaction would allow us to conclude, as LK did, that "*. . . both expectation and locality effects can occur in the same structure in the same language, and that the two effects interact with each other*".

Table 4

Example items (simplified) for investigating the load-distance interaction by combining the conditions (c) and (d) of LK's Experiment 1 and of Experiment 2.

a [E1 c]. PP adjunct in subordinate clause, dative NP in main clause									
Nachdem	der	Lehrer	[**ADJ** zur Ahndung]						…,
After	*the*	*teacher*	[**ADJ** *as payback*]						…,
hat	Hans	Gerstner		[**DAT** dem Sohn]	den	Fußball	**versteckt**,	und	somit…
has	*Hans*	*Gerstner*		[**DAT** *the son*]	*the*	*football*	*hidden*,	*and*	*thus…*
b [E1 d]. PP adjunct in main clause, dative NP in main clause									
Nachdem	der	Lehrer							…,
After	*the*	*teacher*							…,
hat	Hans	Gerstner	[**ADJ** zur Ahndung]	[**DAT** dem Sohn]	den	Fußball	**versteckt**,	und	somit…
has	*Hans*	*Gerstner*	[**ADJ** *as payback*]	[**DAT** *the son*]	*the*	*football*	*hidden*,	*and*	*thus…*
c [E2 c]. PP adjunct in subordinate clause, dative NP in relative clause									
Nachdem	der	Lehrer	[**ADJ** zur Ahndung]						…,
After	*the*	*teacher*	[**ADJ** *as payback*]						…,
hat	der	Mitschüler, der		[**DAT** dem Sohn]	den	Fußball	**versteckt**	**hat**, die	Sache…
has	*the*	*classmate, who*		[**DAT** *the son*]	*the*	*football*	*hidden*	*had*, *the*	*affair…*
d [E2 d]. PP adjunct in relative clause, dative NP in relative clause									
Nachdem	der	Lehrer							…,
After	*the*	*teacher*							…,
hat	der	Mitschüler, der	[**ADJ** zur Ahndung]	[**DAT** dem Sohn]	den	Fußball	**versteckt**	**hat**, die	Sache…
has	*the*	*classmate, who*	[**ADJ** *as payback*]	[**DAT** *the son*]	*the*	*football*	*hidden*	*had*, *the*	*affair…*

'After the teacher imposed detention classes, Hans Gerstner/the classmate (who) hid the football from the naughty son of the industrious janitor as additional payback for the multiple wrongdoings corrected the affair.'

Re-analysis of conditions (c) and (d) of LK's Experiments 1 and 2. We therefore investigated the interaction statistically by combining the original LK data from conditions (c) and (d) of each experiment; see Table 4 for the design. This analysis tested for the main effects of Load, Distance, and their interaction. As shown in Table 5, a positive coefficient for Load would imply that processing a verb within a relative clause is more difficult than in a main clause; note that this effect is not interesting because the verb phrase (*versteckt hat*) in conditions c and d of Experiment 2 is longer than the verb phrase (*versteckt*) in conditions c and d of Experiment 1. More interesting is the effect of Distance. A positive coefficient for Distance would imply that increasing subject-verb distance by interposing an adjunct (which contains a new discourse referent) in addition to a dative NP will lead to longer reading times at the verb; this is as predicted by memory-based accounts such as the Dependency Locality Theory (Gibson, 2000). A negative sign would support the expectation-based account of Levy (2008), as discussed earlier. Finally, a negative coefficient for the load-distance interaction would confirm the cross-over interaction seen visually in Figures 3 and 4 of LK's paper (their paper's pages 209 and 214): interposing a dative NP and an adjunct vs a dative NP alone should lead to a slowdown only in the relative clause conditions.

Condition	Load	Dist	Load×Dist
E1 c ... [$_{MC}$ Subj ... DAT ... Verb]	-0.5	-0.5	-0.5
E1 d ... [$_{MC}$ Subj ... DAT ADJ ... Verb]	-0.5	0.5	0.5
E2 c ... [$_{RC}$ Subj ... DAT ... Verb]	0.5	-0.5	0.5
E2 d ... [$_{RC}$ Subj ... DAT ADJ ... Verb]	0.5	0.5	-0.5

Table 5
The contrast coding used for main effects of Load, Dist(ance), and their interaction in the two experiments by Levy and Keller (2013). The first two conditions here are conditions c and d of Experiment 1, and the last two conditions are conditions c and d of Experiment 2.

Results: The load-distance interaction in the LK data. As shown in Figure 4, in the LK data the estimates for the interaction in the critical region are -51 ms [-111,6]; and in the post-critical region, -41 ms [-92,13]. Here again, even though the interaction has the predicted sign, we have very noisy estimates; the credible intervals have a width of about 100 ms. If a significance test were to be conducted here, the interaction would not come out significant. However, significance is not interesting for us here. We wanted to know whether we can obtain estimates for the load-distance interaction in our replication attempts that have the same sign as the original LK experiments, and whether our estimates are plausible given the wide credible intervals in the LK data. As before, we had 28 participants for two experiments, a self-paced reading study and an eyetracking study. We also carried out a third, larger-sample study (100 participants) to illustrate what changes when we have higher precision estimates.

Two replication attempts of the load-distance interaction

We first carried out two attempts to reproduce the load-distance interaction. As discussed above, we designed the experiment to pit load and distance against each other by

taking conditions (c) and (d) of the original LK Experiment 1 (which we will refer to as the low memory load conditions) and conditions (c) and (d) of Experiment 2 (high memory load conditions). We conducted a self-paced reading study and an eyetracking study, each with the same sample size as the original experiments (28 participants, 24 items). The procedure was as described for the preceding studies.

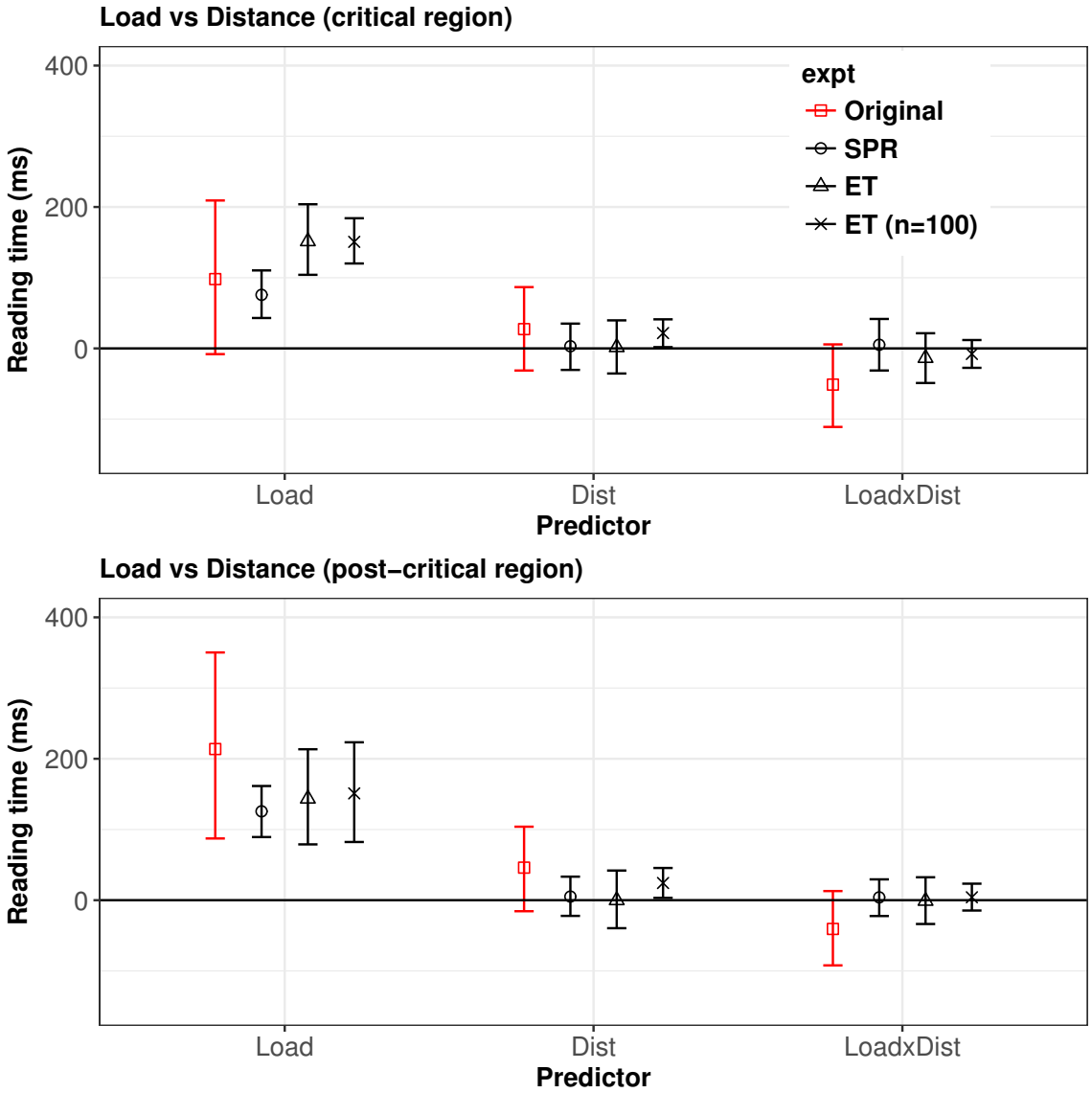

Figure 4. Load and distance effects at the critical and post-critical regions. Shown are the mean and 95% credible intervals from conditions c and d of the two original LK Experiments 1 and 2; and from our three replication attempts.

As shown in Figure 4, both replication attempts showed that the estimate for Load in the critical region had a positive sign; SPR 76 ms [43,110]; and eyetracking (total reading times) 151 ms [104,204]. These effects suggest that increasing load (the relative clause conditions (c) and (d) in Table 4) leads to increased processing difficulty. However, recall

that the effect of load is not interesting because the verb length differs in the two sets of conditions. Therefore, we disregard this load effect, even though theoretically the sign of the effect makes sense.

The estimate for Distance is close to 0 ms; SPR: 3 ms [-30,35]; and eyetracking (total reading times) 2 ms [-35,40]. Finally, the interaction between Load and Distance is not far from 0 ms; SPR: 5 ms [-31,42]; and eyetracking (total reading times) -14 ms [-49,21].

An interesting question arises here: if we were to run the experiment with a larger sample size, would we perhaps detect the load-distance interaction? After all, the interaction claimed by LK is very well-motivated both theoretically and empirically. A larger sample size here effectively means a larger number of participants; between-item variability in planned experiments is generally relatively low compared to between-participant variability, and this was true in the present experiments as well. Thus, the standard error estimate of the effect of interest is largely affected by the number of participants. In order to increase precision, we therefore increased the number of participants to 100 (from 28). The sample size was based on a precision analysis recorded in a pre-registration (https://osf.io/dgewb/): we ran the study until the estimate, especially for the load-distance interaction, had a 95% credible interval of no larger than approximately 40 ms.

A larger-sample replication attempt of the load-distance interaction

The results of this 100-participant study are summarized in Figure 4. This time, the estimate of Load at the critical region is 151 ms [120,184]; the effect of Distance is 22 ms [2,41]; and the Load-Distance interaction is -8 ms [-27,12].

The positive coefficient for Distance suggests that increasing subject-verb distance by interposing an adverb in addition to a dative NP led to slower reading times at the verb. A follow-up analysis using nested contrast coding shows that in the critical region, the distance effect in the low-load conditions is 14 [-13,42]; and in the high-load conditions, it is 29 [1,57]. The expectation account incorrectly predicts a negative coefficient in the low-load conditions; but the locality account correctly predicts the positive coefficient in the high-load conditions.

It is worth considering how our estimates from this 100-participant study would differ from a study that has only 28 participants. This can be demonstrated by repeatedly sampling 28 participants pseudo-randomly from this larger-sample data-set, and then fitting a maximal linear mixed model using Stan. We carried out this repeated sampling 100 times. The mean and 95% credible intervals for the effect of distance are shown in Figure 5, along with the mean and credible interval from the 100-participant study. The wide credible intervals and the fluctuation around the larger sample's estimated mean illustrates the problem that arises with low-precision studies: wide uncertainty of the estimate and fluctuation of means under repeated sampling. Because of this fluctuation, those estimates that happen to come out significant in a frequentist test will, due to Type M error, necessarily be wild overestimates relative to the reference point of the mean and credible intervals estimated from the full data. For a similar demonstration investigating similarity-based interference using a larger data-set, see Nicenboim et al. (2018).

In conclusion, in this 100-participant study we don't see any grounds for claiming an interaction between Load and Distance. The most that we can conclude is that the data are consistent with memory-based accounts such as the Dependency Locality Theory (Gibson,

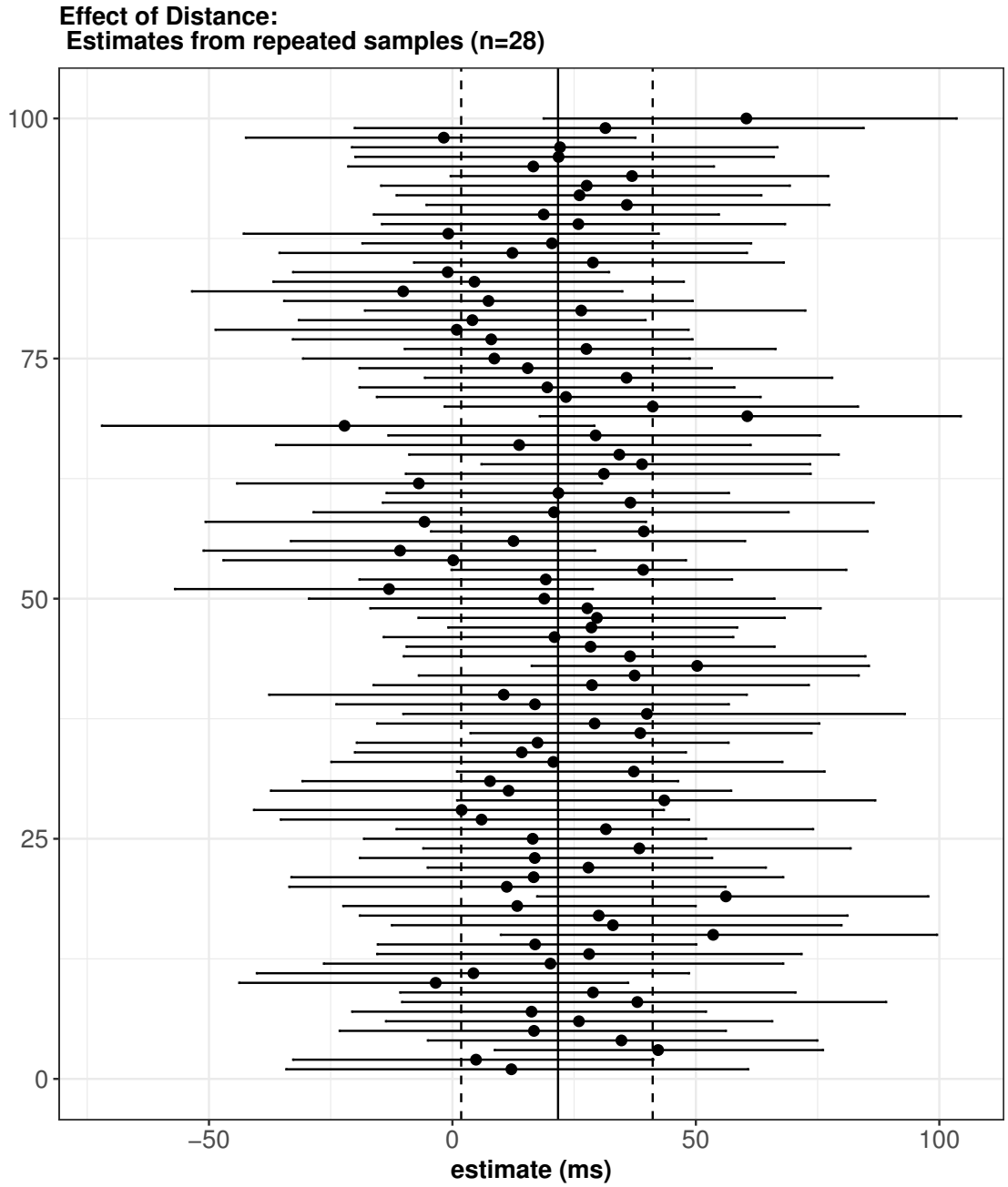

Figure 5. A demonstration of the fluctuation in the estimates for the effect of distance when we choose 28 participants randomly from the 100-participant experiment. The solid vertical lines are the estimated means from the 100-participant data, and the vertical broken lines show the corresponding 95% credible intervals. The points show the means and 95% credible intervals when randomly sampling from the 100-participant data-set.

2000), which predict increased processing difficulty when subject-verb distance is increased.

General Discussion

Our first six replication attempts showed that the statistically significant effects found in Levy and Keller (2013) are noisy enough that a broad range of possible outcomes can be seen as consistent with the original studies' estimates. The noisiness of the estimates in the original LK study, expressed in the wide credible intervals, implies low power, which can—and in this case did—lead to exaggerated effects through Type M error. Had we carried out statistical significance tests on these replication attempts, we would have found that the original results would not be replicable, if by replicable we mean that significance should be found consistently. Our point here is not that the locality and expectation effects found by LK are not true; on the contrary, we believe that ample empirical evidence exists in the literature to support the claims. Rather, our aim is to draw attention to the point that we cannot learn much from a low-precision experiment, regardless of whether statistically significant effects are found.

Our seventh experiment showed that a larger-sample study delivers, as one would expect, much narrower credible intervals. It also delivers smaller estimates, which are probably more realistic. This study also shows that the key claim of a load-distance interaction in LK's original experiments has no support. One interesting conclusion from this 100-participant study is that the locality effect that is predicted by the Dependency Locality Theory (Gibson, 2000) is validated. Since this is, to our knowledge, the first time that locality effects have been seen in German, clearly further investigation is needed. Locality effects have been reported for other head-final languages such as Hindi, Husain, Vasishth, & Srinivasan, 2015, and Persian, Safavi, Husain, & Vasishth, 2016; but it remains to be seen if head-final languages in general also show locality effects.

Noisiness is not an isolated property of the LK study considered here. Reading studies on other well-established effects also have similar issues to those discussed here. One example is the difference in reading times at the head noun of subject vs. object relative clauses in Chinese. A meta-analysis of 12 studies (Vasishth, Chen, Li, & Guo, 2013) showed that the estimates of the effect across different studies fluctuate widely, from -123 to 100 ms, with confidence intervals ranging in width from 80 to 320 ms. A more recent example is so-called number agreement attraction. Here, ungrammatical sentences like the following are investigated: *The key to the cabinet/cabinets are on the table.* For theoretical reasons that don't concern us here, faster reading times are expected at the auxiliary when the preceding noun agrees in number with the auxiliary's number marking (i.e., *cabinets are* is read faster than *cabinet are*). Several studies have been published showing a significant effect in the direction predicted by the theory. Hence, this effect is considered very reliable in psycholinguistics. We re-analyzed the data from 10 published experiments, most of which showed a significant effect. We found that the uncertainty of the estimates in the data is remarkably high (see Appendix B). In the agreement attraction case as well, higher precision replication attempts need to be carried out to determine accurate estimates of the effect.

The central problem is that underpowered studies can yield a statistically significant result due to Type M error, and these significant results will be overestimates. Given that significant results are favored by journals and reviewers, when power is low, effects reported

in the literature will be overestimates. They will also be seen as very convincing because of their large magnitude. A large effect like 200 ms with a large standard error of 80 ms, leading to a t-value of 2.5, seems more convincing than a small effect of 9 ms with a small standard error of 4.5 ms and a t-value of 2. But in fact, the former is consistent with a wide range of values, including values near 0; the smaller estimate with narrow credible intervals may reflect reality better. Thus, when power is low, using significance to decide whether to publish a result leads to a proliferation of exaggerated estimates in the literature.

What is a reasonable alternative? We offer two suggestions. First, we can carry out a precision analysis (see chapter 13, Kruschke, 2014) before running an experiment to decide how much uncertainty of the estimate is acceptable. For example, a 95% credible interval of 40 ms is one option we chose in our final experiment, but this was only for illustration purposes; depending on the resources available, one could aim for even higher precision. For example, 184 participants in the Nicenboim et al. (2018) study had a 95% credible interval of 20 ms. Note that the goal here should not be to find an interval that does not include an effect of 0 ms; that would be identical to applying the statistical significance filter and is exactly the practice that we criticize in this paper. Rather, the goal is to achieve a particular precision level of the estimate.

Once we have fixed the precision that is acceptable to us, we can run the experiment until we reach this desired level. This has at least two advantages over a conventional power analysis. First, there is no need to define a stopping criterion in advance of running our experiment. In psycholinguistics, running more participants until a desired outcome (statistical significance with a particular sign of the effect) is reached is a fairly common practice. But this stopping criterion is known to inflate Type I error (Pocock, 2013). In the Bayesian framework, there is no concept of hypothetical replications; the data are not interpreted in the light of imagined repeated sampling Gelman et al. (2014). The data are what they are. We can therefore check the precision of our estimates while running the experiment, and stop the experiment when the desired precision (as opposed to the desired or expected sign of the effect becoming significant) is reached.

A second advantage of using precision as a guide to data collection is that we can shift the focus to what really matters: quantifying our uncertainty about the estimate of interest. A conventional power analysis assumes a good guess about the magnitude of the true effect; such a guess is often difficult to arrive at. In a precision-based analysis, the focus is on the amount of uncertainty in the estimate that we are willing to tolerate. The estimate itself is much more important theoretically than just counting the number of significant vs. not significant results, as is commonly done to decide whether an effect is "present" vs. "absent" (for an example of a voting-based approach to deciding whether an effect is present or absent, see Phillips, Wagers, & Lau, 2011). Having higher-precision estimates allows for formal model comparison of competing quantitative models. For example, the first comprehensive quantitative evaluation involving 77 published results (Engelmann, Jäger, & Vasishth, 2018) of the computational memory-retrieval model of Lewis and Vasishth (2005) was only possible because the estimates (and their uncertainty) were available from a preceding meta-analysis (Jäger et al., 2017). The results of this quantitative evaluation might look quite different if the estimates from the published data had higher precision.

Our second suggestion is that we should attempt to directly replicate our experiments. Every major claim should be either accompanied by a direct replication, or even better,

other researchers from competing labs should be encouraged to replicate the original result. Leading journals could trigger this positive change by introducing a special article type (e.g., a Replication Report) for direct replication attempts. Currently, direct replications are not considered to be novel enough to be worth publishing, and novelty of results is given disproportionate weight. However, replication is an important tool for establishing reliability. This is something that a p-value, especially a p-value computed from an underpowered study, cannot ever deliver. Increasing precision and conducting direct replications are vital for any empirically rigorous science.

A potential response to our proposals is: why should it matter that overestimates are published, as long as the claims turn out to be true? First, ignoring the magnitude of the estimated effect will have a consequence when evaluating the predictions of computational models. For example, the activation-based model (Engelmann et al., 2018) delivers reading time predictions, even for different eye-tracking measures (Engelmann, 2016; Engelmann, Vasishth, Engbert, & Kliegl, 2013). If the empirical data consist of overestimates, model evaluation based on these overestimates will be misleading. A second problem with ignoring the magnitude of published effects is that, due to publication bias arising from applying the statistical significance filter, the effect may in fact be close to 0 ms. Thus, for theory development, it is vital that our published estimates are as accurate as possible. The two-pronged approach—high precision studies and direct replications—will go a long way towards achieving this goal.

In sum, a novel contribution of the present paper is to demonstrate through a case study that published results—even results published in top journals—may not be all that newsworthy. Too often, published empirical results are treated as a novel contribution simply because of the application of the statistical significance filter. How many strong published claims are actually newsworthy remains to be seen. For example, the recent failure to find significant effects in anticipatory processing by Nieuwland et al. (2017) suggests that replicability problems arising from the statistical significance filter may run deep in psycholinguistics. Our suggestions, to aim at higher precision and to conduct direct replications, will help in improving the reliability of published results.

Acknowledgements

We are grateful to Roger Levy and Frank Keller for their openness in sharing their data and code with us; without their assistance and cooperation, this paper would not have been possible. This project began in December 2015 as part of a demonstration of a replication for a course that the first author taught at the University of Tokyo, Japan; Doug Roland and Yuki Hirose provided many useful comments at this initial stage. Our grateful thanks to Johanna Thieke, lab manager of Vasishth Lab at the University of Potsdam, Germany, for carrying out the experiments over a space of two years. We also thank Reinhold Kliegl, Christian Robert, Titus von der Malsburg, and Bruno Nicenboim for helpful discussions. For partial support of this research, we thank the Volkswagen Foundation through grant 89 953, the Deutsche Forschungsgemeinschaft, Collaborative Research Center (SFB) 1287 (*Limits of Variability in Language*), project Q (PIs Shravan Vasishth and Ralf Engbert), which partly funded Lena Jäger, and B03 (PIs Ralf Engbert and Shravan Vasishth), which partly funded Daniela Mertzen; and the U.S. Office of Naval Research through grant N00014-15-1-2541.

Appendix A

How the statistical significance filter leads to inflated estimates of power

Assume for simplicity the case that we carry out a one-sided statistical test where the null hypothesis is that the true mean is $\mu_0 = 0$ and the alternative is that $\mu > 0$.[4] Given some continuous data $x_1, \ldots, x_n$ (such as reading times), we can compute the t-statistic and derive the p-value from it. For a 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 estimated from the data, 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 itself 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 observe 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 estimate power based on the z-statistic (Hoenig & Heisey, 2001). To estimate the p-value in the case where the null hypothesis is in fact true, let the true value be $\mu = 0$. It follows that $\delta = 0$. Then:

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

To estimate 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 estimated 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 A1). Together with the common practice that only statistically

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

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 estimate 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 and reflects the reality.

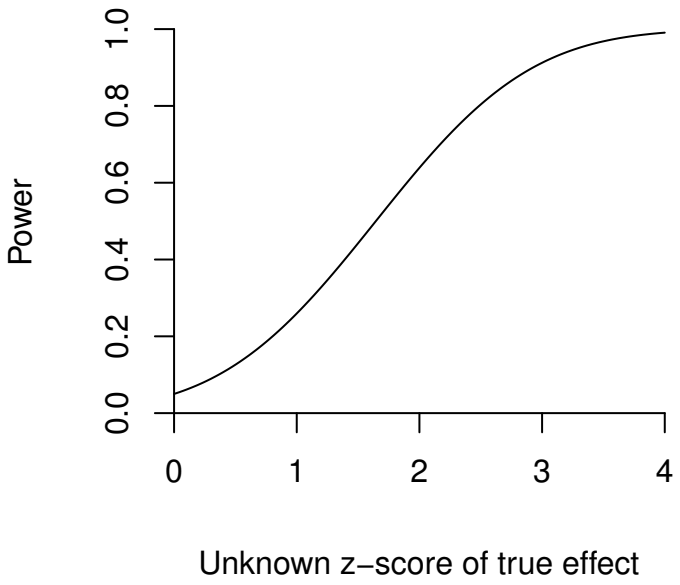

Figure A1. The relationship between power and the unknown z-score of the true effect. Larger z-scores are easier to publish due to the statistical significance filter, and these studies therefore give a mistaken impression of higher power.

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, the true effect is 0.1 on some scale, and standard deviation is 1, then statistical power is 15%.[5]

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 probability (α) 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

$$G_{Z_\alpha}(\alpha) = 1 - \Phi(1.64 - 1.64) = 0.5 \tag{5}$$

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

[5]This can 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",strict=TRUE)`.

to z_α are published, the power estimate based on the z-score will be inflated by at least a factor of 0.5/0.15=3.33.

Now, lower p-values are widely regarded as more "reliable" than p-values near the Type I error probability of 0.05.[6] This incorrect belief among researchers has the effect that studies with lower p-values are more likely to be reported and published, with the consequence that the inflation in power will tend to be even higher than the lower bound discussed here.

Appendix B

Estimates from 10 reading studies on agreement attraction

Figure B1 shows the relatively large uncertainty in the 95% credible intervals for ten agreement attraction studies that were part of the meta-analysis in Jäger et al. (2017). All these studies (one was an eyetracking study and the rest were self-paced reading) investigated ungrammatical sentences such as *The key to the cabinet/cabinets are on the table.* The reading time was either recorded at the critical (the auxiliary *are*) or post-critical region. Theory predicts a facilitation effect at the auxiliary or the following region when the noun preceding the auxiliary is *cabinets* vs *cabinet*. Study 1 is the ungrammatical agreement data from Experiment 1 of Dillon, Mishler, Sloggett, and Phillips (2013); studies 2-5 are the experiments reported in Lago, Shalom, Sigman, Lau, and Phillips (2015), and 6-10 are from Wagers, Lau, and Phillips (2009). The estimates shown in the figure were computed by fitting a linear mixed model (with full covariance matrices for random effects) in Stan using log-transformed reading times, and then by back-transforming the estimate of the number agreement effect to milliseconds. Our estimates are different from the original published estimates because we did not remove any data. The ten studies' mean estimates range from -40 to -4, with credible intervals ranging in width from 30 to 91 ms. Again, our interest here is not in whether effects were significant or not significant; rather, what's remarkable here is the wide variation in the estimates of the mean effect, and the large uncertainty in many of the estimates expressed by the 95% credible intervals.

[6] Treating lower p-values as furnishing more evidence against the null hypothesis reflects a misunderstanding about the meaning of the p-value; given a continuous dependent measure, when the null hypothesis that $\mu = 0$ is true, under repeated sampling the p-value has a uniform distribution. This has the consequence that, when the null is true, a p-value near 0 is no more surprising than a p-value near 0.05.

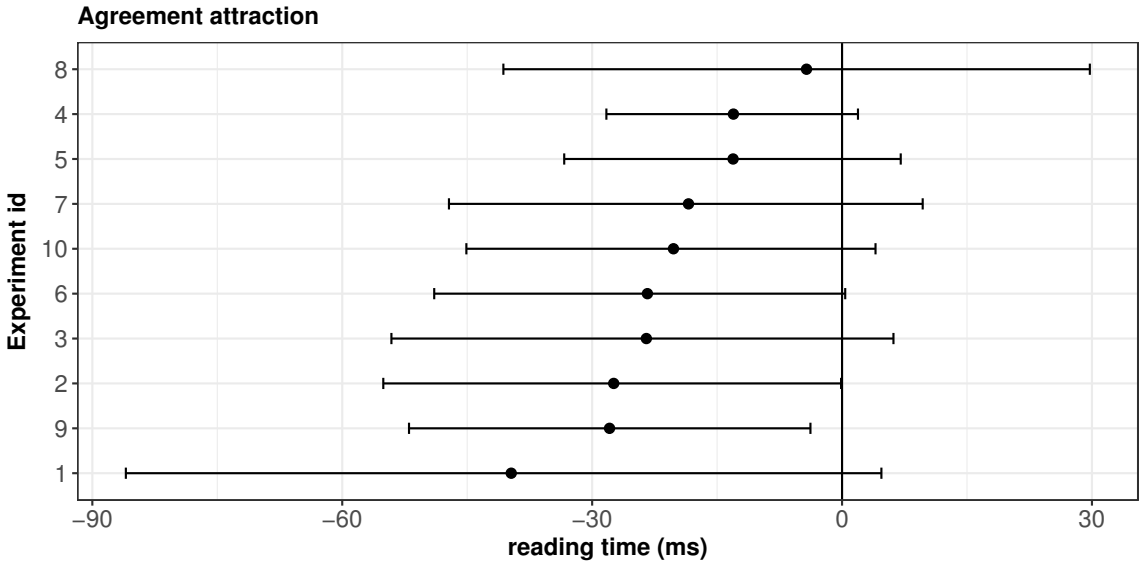

Figure B1. The means and 95% credible intervals of the predicted facilitation effect from 10 published studies on number agreement.

References

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

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.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). *Parsimonious mixed models.* (Unpublished manuscript)

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2016). Stan: A probabilistic programming language. *Journal of Statistical Software*, *20*, 1–37.

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

Engelmann, F. (2016). *Toward an integrated model of sentence processing in reading* (Doctoral thesis). Universität Potsdam, Germany.

Engelmann, F., Jäger, L. A., & Vasishth, S. (2018). *The effect of prominence and cue association in retrieval processes: A computational account.* (Manuscript submitted to Cognitive Science)

Engelmann, F., Vasishth, S., Engbert, R., & Kliegl, R. (2013). A framework for modeling the interaction of syntactic processing and eye movement control. *Topics in Cognitive Science*, *5*(3), 452-474.

Gelman, A. (2017). The failure of null hypothesis significance testing when studying incremental changes, and what to do about it. *Personality and Social Psychology Bulletin*.

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.). Boca Raton, FL: Chapman and Hall/CRC.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models.* Cambridge, UK: Cambridge University Press.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, *68*, 1–76.

Gibson, E. (2000). Dependency locality theory: A distance-based theory of linguistic complexity. In A. Marantz, Y. Miyashita, & W. O'Neil (Eds.), *Image, Language, Brain: Papers from the First Mind Articulation Project Symposium.* Cambridge, MA: MIT Press.

Grodner, D., & Gibson, E. (2005). Consequences of the serial nature of linguistic input. *Cognitive Science*, *29*, 261–290.

Hale, J. T. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics.* Pittsburgh, PA.

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, 1–8.

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. (2015). Integration and prediction difficulty in Hindi sentence comprehension: Evidence from an eye-tracking corpus. *Journal of Eye Movement Research*, *8(2)*, 1–12.

Jäger, L. A., Chen, Z., Li, Q., Lin, C.-J. C., & Vasishth, S. (2015). The subject-relative advantage in Chinese: Evidence for expectation-based processing. *Journal of Memory and Language*, *79–80*, 97–120.

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*, *94*, 316-339.

Konieczny, L. (2000). Locality and parsing complexity. *Journal of Psycholinguistic Research*, *29(6)*, 627–645.

Konieczny, L., & Döring, P. (2003). Anticipation of clause-final heads: Evidence from eye-tracking and SRNs. In *Proceedings of Joint International Conference on Cognitive Science (ICCS/ASCS)* (pp. 13–17).

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

Lago, S., Shalom, D. E., Sigman, M., Lau, E. F., & Phillips, C. (2015). Agreement attraction in Spanish comprehension. *Journal of Memory and Language*, *82*, 133–149.

Levy, R. P. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*, 1126–1177.

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

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, *29*, 1–45.

Linzen, T., & Jaeger, T. F. (2016). Uncertainty and expectation in sentence processing: Evidence from subcategorization distributions. *Cognitive Science*, *40*, 1382-1411.

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

Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E.-J. (2011). Erroneous analyses of interactions in neuroscience: A problem of significance. *Nature Neuroscience*, *14*(9), 1105–1107.

Nieuwland, M., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., ... others (2017). Limits on prediction in language comprehension: A multi-lab failure to replicate evidence for probabilistic pre-activation of phonology. *BioRxiv*, 111807.

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

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*(6), 528–530.

Phillips, C., Wagers, M. W., & Lau, E. F. (2011). Grammatical illusions and selective fallibility in real-time language comprehension. *Experiments at the Interfaces*, *37*, 147–180.

Pocock, S. J. (2013). *Clinical trials: A practical approach*. John Wiley & Sons.

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*.

Sorensen, T., Hohenstein, S., & Vasishth, S. (2016). Bayesian linear mixed models using Stan: A tutorial for psychologists, linguists, and cognitive scientists. *Quantitative Methods for Psychology*, *12*(3), 175-200. Retrieved from http://www.ling.uni-potsdam.de/~vasishth/statistics/BayesLMMs.html

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

Van Dyke, J., & Lewis, R. L. (2003). Distinguishing effects of structure and decay on attachment and repair: A cue-based parsing account of recovery from misanalyzed ambiguities. *Journal of Memory and Language*, *49*, 285–316.

Vasishth, S. (2003). *Working memory in sentence comprehension: Processing Hindi center embeddings.* New York: Garland Press. (Published in the Garland series Outstanding Dissertations in Linguistics, edited by Laurence Horn)

Vasishth, S., Chen, Z., Li, Q., & Guo, G. (2013, 10). Processing Chinese relative clauses: Evidence for the subject-relative advantage. *PLoS ONE*, *8*(10), 1–14.

Vasishth, S., & Lewis, R. L. (2006). Argument-head distance and processing complexity: Explaining both locality and antilocality effects. *Language*, *82*(4), 767-794.

Vasishth, S., & Nicenboim, B. (2016). Statistical methods for linguistic research: Foundational ideas – Part I. *Language and Linguistics Compass*, *10*(8), 349–369.

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

Wagers, M., Lau, E., & Phillips, C. (2009). Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language*, *61*(2), 206–237.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's Statement on p-Values: Context, Process, and Purpose. *The American Statistician*, *70*(2), 129-133.

Wu, F., Kaiser, E., & Vasishth, S. (2017). Effects of early cues on the processing of Chinese relative clauses: Evidence for experience-based theories. *Cognitive Science*.