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

May 22, 2018

Author Note

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

Abstract

Treating a result as publishable just because the p-value is less than 0.05 leads to overoptimistic expectations of replicability. These overoptimistic expectations arise due to Type M(agnitude) error: when underpowered studies yield significant results, effect size estimates are guaranteed to be exaggerated and noisy. These effects get published, leading to an overconfident belief in replicability. We demonstrate the adverse consequences of this statistical significance filter by conducting six direct replication attempts (168 participants in total) of published results from a recent paper. We show that the published claims are so noisy that even non-significant results are fully compatible with them. We also demonstrate the contrast between such small-sample studies and a larger-sample study (100 participants); the latter generally yields less noisy estimates but also a smaller effect size, which looks less compelling but is more realistic. We make several suggestions for improving best practices in psycholinguistics and related areas.

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

The statistical significance filter leads to overoptimistic expectations of replicability

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—either as significant and therefore "reliable" and publishable, or not significant and therefore either not publishable, or seen as showing that the null hypothesis is true—is misleading because it implies an inappropriate level of certainty in rejecting or accepting the null. Indeed, we believe that this routine attribution of certainty to noisy data is a major contributor to the current replication crisis in science (Pashler & Wagenmakers, 2012; Open Science Collaboration, 2015); for recent examples from psycholinguistics, see Nieuwland et al. (2018), Kochari and Flecken (2018), and Stack, James, and Watson (2018). 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*. We will demonstrate through direct replication attempts that one adverse consequence of the statistical significance filter is that it leads to findings that are positively biased (Gelman, 2018; Lane & Dunlap, 1978).

How null hypothesis significance testing (NHST) works. In order to explain some of the problems with NHST, it is necessary to briefly review the procedure typically adopted. Assume that the data are generated from a random variable Y that has a normal distribution with unknown mean μ and unknown variance σ^2; nothing hinges on this particular distributional assumption. We can write this assumption as $Y \sim Normal(\mu, \sigma^2)$. Hereafter, when we refer to the random variable, we use capital letters (e.g., Y), and when we refer to a vector of observed data-points, we use lower case (e.g., y). Individual data-points are indexed by a subscript i (y_i), with i ranging from 1 to the sample size n.

Suppose that we take an independent and identically distributed sample y of size n, and posit a null hypothesis that μ is some specific point value μ_0, i.e., $\mu = \mu_0$. For example, assume that the null hypothesis is that $\mu = 0$. The typical alternative hypothesis is $\mu \neq 0$; i.e., that μ has any value that is not 0. Notice that, at this stage, the alternative hypothesis is that μ is one of an infinity of positive or negative values that are not 0; we don't have any specific value for μ in mind at this stage. Given the null that $\mu = \mu_0$ and the alternative that $\mu \neq \mu_0$, we then compute the sample mean $\bar{y}$ and the sample variance s^2. These statistics are so-called maximum likelihood estimates (MLEs) of the true mean and variance, respectively. These estimates are computed using so-called estimators, which are functions of the data. The sample mean is computed from the estimator $\bar{Y} = \sum_{i=1}^{n} Y_i/n$, and the sample variance is computed from $S^2 = \sum_{i=1}^{n}(Y_i - \bar{Y})^2/(n-1)$. MLEs have the property of consistency: informally, this means that as the sample size goes to infinity, the estimator will approach the true parameter value with high probability (for details, see Casella & Berger, 2002, pp. 467-468). In general, the estimators that give us the sample mean and variance from a particular data set also have the property that they deliver so-called unbiased estimates of the true mean and true variance. Unbiased here has a technical meaning, namely that the expected value of the estimator is the true value of the corresponding parameter. Unbiasedness does not mean that *every* sample's mean or variance, regardless of sample size, gives us an accurate estimate of the true parameter.

For hypothesis testing using the two-sided one-sample t-test, we proceed as follows. Given that the underlying distribution that generated the data y has a mean μ and variance σ^2, the central limit theorem tells us that the sampling distribution of the sample mean has the distribution $Normal(\mu, \sigma^2/n)$ where n is the sample size. Next, using the sample mean $\bar{y}$, sample standard deviation s, and sample size n, we compute the *observed t-statistic*, $\frac{\bar{y}-\mu_0}{s/\sqrt{n}}$, which comes from a reference t-distribution with degrees of freedom $n-1$ and which we write as $t(n-1)$. If the true mean μ were in fact equal to or near the null hypothesis value μ_0, then we would expect the observed t-statistic to not be far from 0, because $\bar{y}$ (being an

MLE) would be near the hypothesized mean μ_0; i.e., $\bar{y} - \mu_0 \approx 0$. So we compare this observed t-statistic $t_{obs} = \frac{\bar{y}-\mu_0}{s/\sqrt{n}}$ with the distribution $t(n-1)$, which is centered around 0 and thus corresponds to the true mean being identical to the hypothesized null hypothesis mean μ_0; see Figure 1. If t_{obs} lies far enough away from the center of the $t(n-1)$ distribution, i.e., in its left or right tail, we conclude that the observed t-statistic is unlikely to come from the reference t-distribution. More specifically, if the probability of obtaining a value at least large as t_{obs} (the observed statistic) is less than the prespecified probability $\alpha = 0.05$, then we reject the assumption that the reference t-distribution $t(n-1)$ is the true distribution generating the statistic that we observed. Since the reference t-distribution centered around zero expresses the null hypothesis, by extension we reject the null hypothesis as well. This hypothesis testing procedure is identical for likelihood ratio tests in linear mixed models, and F-tests in repeated measures ANOVA. In fact, the t-test, ANOVA, and the likelihood ratio test are all equivalent statistical tests (Casella & Berger, 2002).

It is important to note here that the p-value is the probability of the observed t-statistic or some value more extreme, *assuming that the null is already true*. In other words, when we do the t-test, we have already committed to the assumption that the null is true, and want to know whether the world we are studying is consistent with this assumption. This implies that a low p-value can only furnish evidence against the null hypothesis we posited; it provides no information about any *specific* alternative hypothesis the researcher is investigating. In fact, a positive or negative sample mean that leads to the test-statistic being located in either one of the two tails of the t-distribution would both yield the same absolute value of the t-statistic and the same evidence against the null. A practical implication is that we can obtain a statistically significant result even though the sign of the sample mean is the opposite to that predicted by a given theory. Even if the sign of the sample mean does not match the theoretical prediction, we have still rejected the null hypothesis successfully. In this situation, we have no evidence for or against our given theory from the statistical test itself, we only have evidence against the null. In other words, NHST

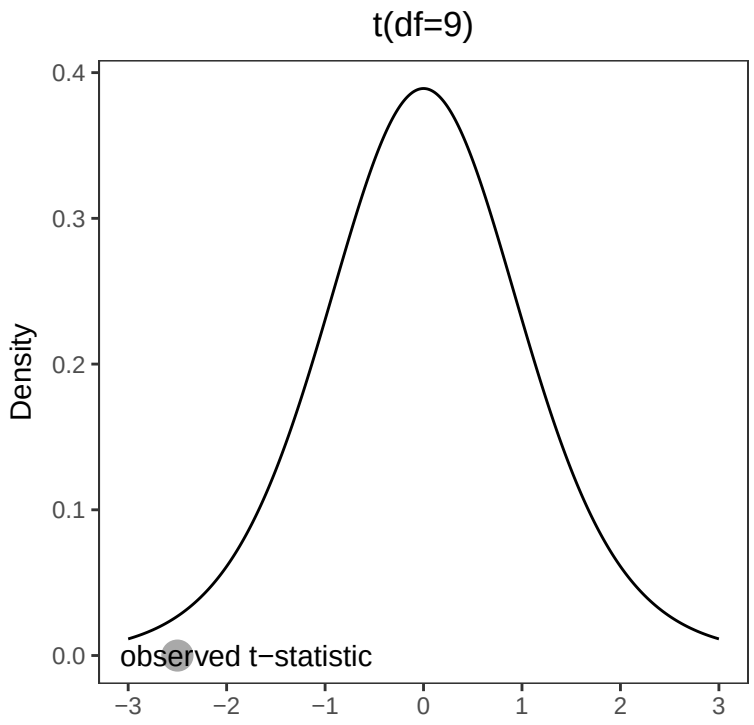

Figure 1. An example showing a reference t(9) distribution against which an observed t-statistic (the dot) is compared. If the observed statistic lies far enough on either side of the center of the reference t-distribution, we reject the null hypothesis. The criterion for rejection is usually that the probability of observing the t-statistic or some value more extreme must be less than 0.05.

does answer *a* question, but it answers a question that we don't really want an answer to.

At this point, the NHST procedure switches to an informal reasoning process: we assume, post-hoc, that the maximum likelihood estimate $\bar{y}$ that we happened to get from our data can now legitimately replace the infinity of possible values that we posited when we stated our alternative hypothesis. It is important to stress that once we reject the null hypothesis based on $p < 0.05$, we not only reject the null but we also argue informally in favor of a *very specific* alternative we adopt *after the fact*, after we have already collected the data. We say that this argument is informal because when we reject the null, what we should be accepting is that the true parameter can be any value except 0. But this is not what we do; we argue that the alternative is that the true parameter is the sample mean that we happened to get. It seems reasonable to take this informal step because the sample

mean, being an MLE, is likely to reflect the true mean. But a great deal depends on whether the particular sample mean that we got is in fact an accurate estimate of the true mean. It would be very convenient indeed if nearly every sample that we take gave us a realistic sample mean which is near the true mean; but it turns out that in certain situations, described below, the sample mean can be a gross overestimate, i.e., highly biased.

Two further important concepts in the NHST framework are Type I error and power. Type I error is the probability of rejecting the null incorrectly, i.e., when the null hypothesis is in fact true. Power is the probability of rejecting the null when true parameter has some specific value different from the null hypothesis parameter value. These two probabilities are intended to be set in advance by the researcher, before they begin collecting data. In psycholinguistics at least, in practice only Type I error is set at 0.05 in advance; power is rarely considered at all.

The NHST procedure can work well when power is relatively high, say 80% or higher. But when power is low (for example, 30% or lower), published studies that show statistical significance are guaranteed to have exaggerated estimates (see Appendix A for a formal argument). This is demonstrated in Figure 2 using simulated data: for a low-power scenario, the estimates from repeated samples fluctuate around the true value, and can also have the wrong sign. Whenever an effect is significant, it is necessarily an overestimate. Gelman and Carlin (2014) refer to these overestimates as Type M(agnitude) errors (when the sign of the effect is incorrect, Gelman and Carlin call this Type S(ign) error). These overestimates occur because the standard error is relatively large in low-power situations; the wider the sampling distribution of the sample mean, the greater the probability of obtaining extreme values. By contrast, when power is high, the estimates under repeated sampling tend to be close to the true value because the standard error is relatively small.

Figure 2 illustrates another important point: when power is high, the estimates have much narrower 95% confidence intervals. We will express this by saying that high-powered studies have higher *precision* than low-powered studies. We borrow the term precision from

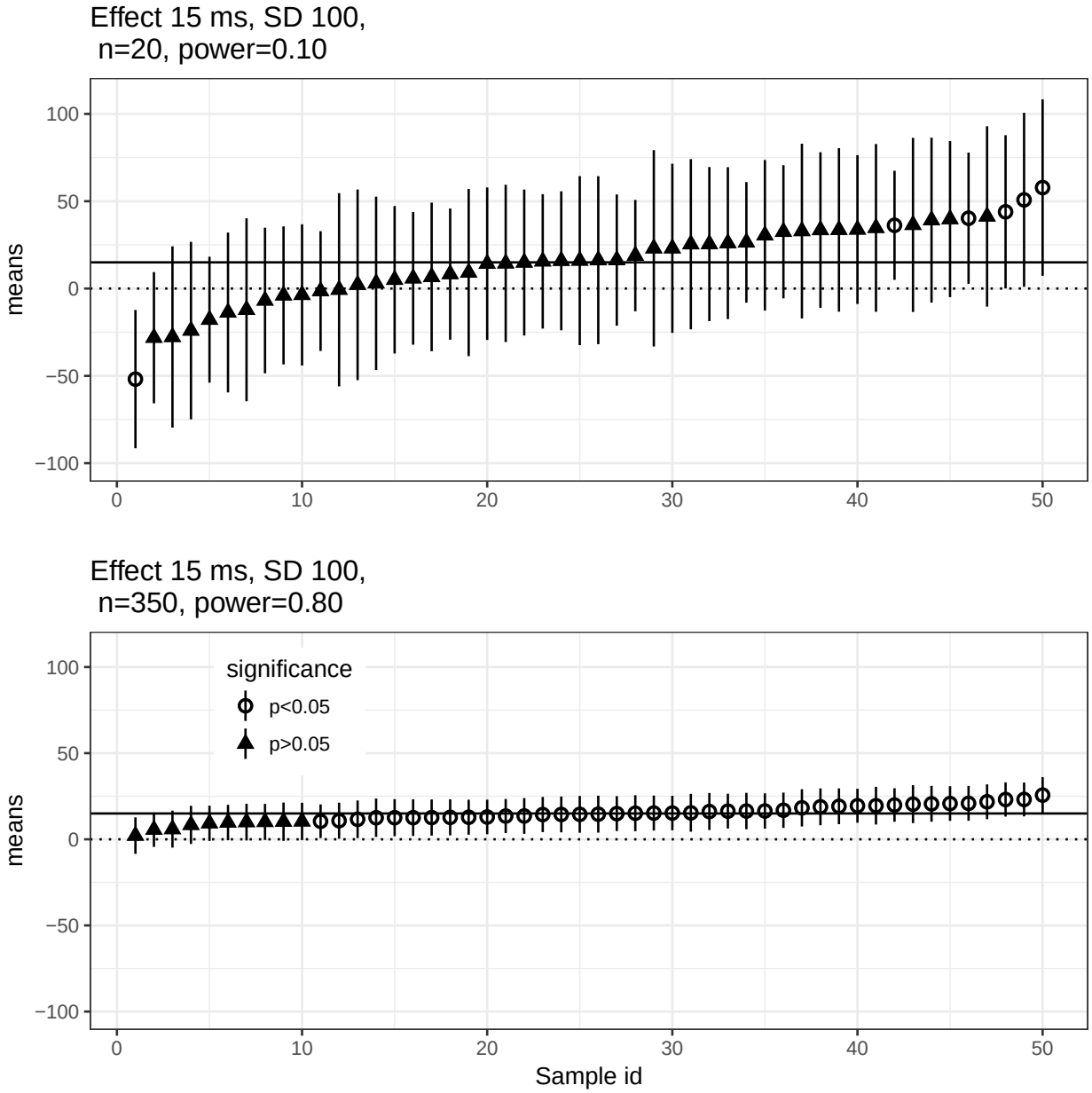

Figure 2. A demonstration of Type M error using simulated data. We assume that the data are generated from a normal distribution with mean 15 ms and standard deviation 100. The true mean is shown in each plot as a solid horizontal line. When power is low, under repeated sampling, whenever the estimates of an effect come out significant, the values are overestimates and can even have the wrong sign. When power is high, significant and non-significant effects will be tightly clustered near the true mean.

Bayesian statistics, where it has a specific meaning: the inverse of the variance. Here, we are using the term precision to stand for the uncertainty about our estimate of interest (the sample mean, or a difference in sample means). This uncertainty is expressed in frequentist statistics in terms of the standard error of the sample mean. The standard error decreases as a function of the square root of the sample size; hence, if power is increased by increasing sample size, standard error will go down.

Many researchers, such as Cohen (1988) and Gelman and Carlin (2014), have pointed out that a prospective power analysis should be conducted before we run a study; after all, why would one want to spend money and time running an experiment where the probability of detecting an effect is 30% or less? Although there is no tradition of prospective power analysis in psycholinguistics, suppose that we were to follow this advice and conduct a power analysis based on the effect sizes reported in the literature. This can lead to an interesting problem. As discussed above, 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 if we design a new study based on such a power analysis. 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, we are bound to get an overestimate of power, and can easily convince ourselves that we have an appropriately powered study.

Of course, in psycholinguistics, usually we do no power analyses at all. We 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. However, we rarely check the replicability of a result through direct replications (Simons, 2014). In psycholinguistics, conceptual replication attempts are more common. The term conceptual replication refers to an experiment that is similar to the original study, but has some novel manipulation as well. By contrast, a direct replication involves re-doing the original experiment, usually using the same method but with new participants. Conceptual

replications are more common because of the demand from journals for novelty; direct replications are generally seen as not adding much value. In fact, direct replications are actively discouraged by journals. A recent example is the report on Retraction Watch (http://bit.ly/NatureNeuro) of a rejection by Nature Neuroscience of a multi-lab replication failure of DeLong, Urbach, and Kutas (2005) by Nieuwland et al. (2018).

Although the above observations about power and replications 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 of a published empirical result—that relying exclusively on statistical significance to decide whether or not 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 overestimated 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 paper 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 (we explain this later in the paper); and reliability should be established through direct replication using pre-registration.

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 Plastik] verkaufte,...
The insight, that the friend the client the car from 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 aus Plastik] verkaufte,...
The insight, that the friend of the client the car from 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).[1]

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 or DLT (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 new 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). A closely related account is a computational model of cue-based retrieval (Lewis & Vasishth, 2005; Engelmann, Jäger, & Vasishth, 2018; Nicenboim & Vasishth, 2018). This model assumes that completing argument-verb dependencies is affected by similarity-based interference arising from distractor nouns in memory (a related model is Van Dyke & McElree, 2006). Like the DLT, this model predicts that interposing nouns between the argument(s) and verb in grammatical sentences will increase processing difficulty at the verb. For the Konieczny and Döring data, this model also predicts no difference in processing difficulty between the two conditions (1a) and (1b). We will refer to both the DLT and the cue-based retrieval theories as the memory-based account.

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

[1] A reviewer, Roger Levy, points out that these values are almost certainly overestimates of "true" comprehender surprisal for these cases, because the probabilistic context free grammar used for the calculations encodes much less information than human comprehenders would deploy.

the basis for our replication attempts.

The experiment design by Levy and Keller (2013)

As shown in Table 1, in their sentences 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 the dative NP or both the dative NP and the adjunct phrase appeared in the main clause, the main clause verb phrase (in our example, the verb phrase that heads the 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 design similar to Experiment 1, with one difference: 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

LK lay out the predictions of the expectation-based account as follows (Levy & Keller, 2013):

> . . . [condition (a)] (neither dative nor adjunct in the main clause) should be hardest to process, while [condition (d)] should be easiest (both dative and adjunct in the main clause). [Conditions (b) and (c)] should be in between (one phrase in the main clause). (p. 202)

The reasoning behind these predictions is that interposing material sharpens the expectation for a participial verb. For a graphical summary of the predictions, see Figure 3, left panel; this figure is a reproduction of LK's Figure 1. As mentioned above, Levy (2008) and others refer to such predicted speedups as expectation effects.[2]

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 3, 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 NP 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 (Konieczny, 2000; Grodner & Gibson, 2005; Vasishth & Lewis, 2006) had the verb further downstream in the sentence whenever an additional 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 (for discussion, see Ferreira & Henderson, 1993).

[2]However, note that in their corpus analysis, summarized in their Table 1 (Levy & Keller, 2013, p. 204), the participial verb had higher predictability in only condition (b), where the dative NP intervenes, and condition (d), where both the dative NP and adjunct intervene. Thus, according to the corpus analysis, conditions (a) and (c) would be predicted to be read slower than conditions (b) and (d). In the present paper, we follow the predictions laid out in LK's Figure 1.

Table 1

Example items for LK's Experiment 1 (simplified). The abbreviations mean the following: ADJ: adjunct; DAT: dative; PP: prepositional phrase; NP: noun phrase.

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). The abbreviations mean the following: ADJ: adjunct; DAT: dative; PP: prepositional phrase; NP: noun phrase.

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.'

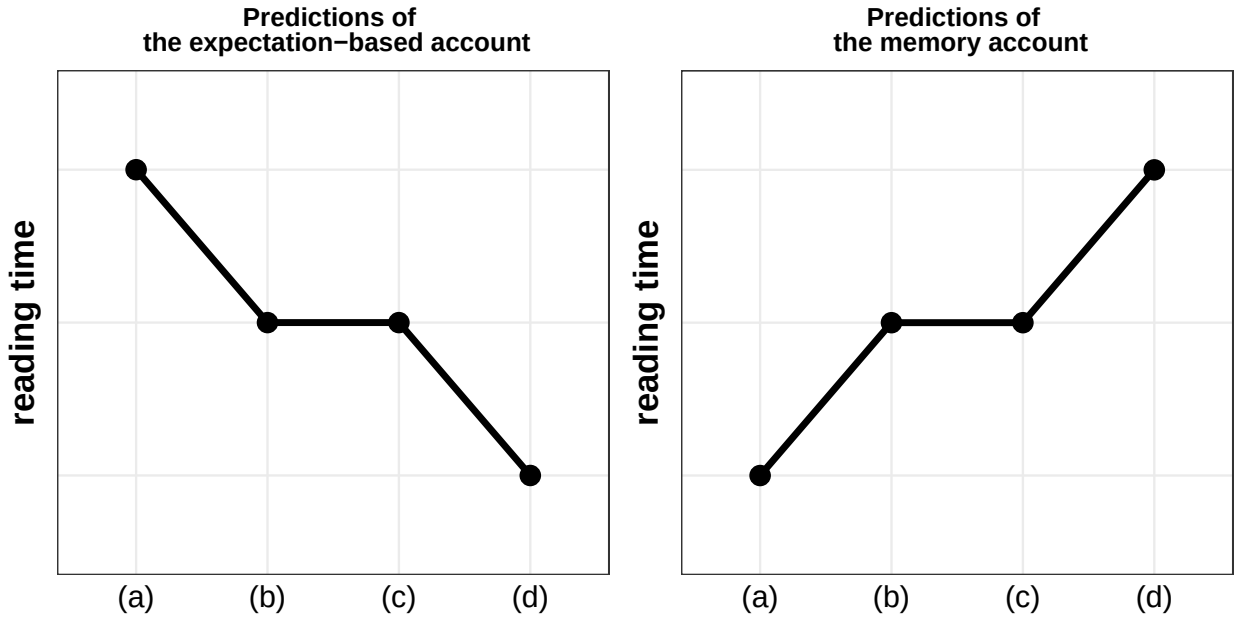

Figure 3. Predictions for the Levy and Keller Experiments 1 and 2: The left panel shows the speedup predicted by the expectation account. The right panel shows the slowdown predicted by memory-based accounts. This figure is based on Figure 1 of Levy and Keller (2013).

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). Below, we explain our reasons for using the Bayesian data-analytic approach.

Motivation for using Bayesian data analysis. In the Bayesian framework, all the parameters in the model, which can be represented as a vector θ, are assumed to have some prior distribution of plausible values, $p(\theta)$. Given the prior, and a likelihood function for the data $p(data \mid \theta)$, Bayes' rule is used to compute the posterior distribution of the parameters: $p(\theta \mid data)$. Bayes' rule states that the posterior is proportional to the prior multiplied by the likelihood: $p(\theta \mid data) \propto p(\theta)p(data \mid \theta)$. Thus, the application of Bayes' rule furnishes a posterior distribution representing plausible values of a parameter given the data and model (the model subsumes the priors). This is very different from the frequentist approach, where each parameter is assumed to be an unknown point value. Such a point value may represent an invariant number in some fields (e.g., the speed of light in physics), but is a fictional construct in areas like psychology and psycholinguistics. With the Bayesian approach, we can focus on a crucial aspect that we wish to discuss in the present paper: the uncertainty of the estimates of interest. A further advantage is that we can always 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 unrealistic estimates of the variance components (see Vasishth, Nicenboim, Beckman, Li, & Kong, 2018, for an example). Fitting a maximal model has the advantage that we can make the most conservative possible claim about the parameters given the data and model. The reason that Bayesian methods allow us

to fit essentially arbitrarily complex random effects variance components is the involvement of prior information in the model. We discuss this next.

Prior specification in Bayesian models. In the Bayesian approach, it is common to use so-called mildly/weakly informative priors that have a regularizing effect on the posteriors. A weakly informative prior allows a wide range of plausible values; regularizing means that we downweight extreme values that are a priori unlikely to occur. A simple example is a prior on correlations or correlation matrices; Stan allows us to define a so-called LKJ prior (Lewandowski, Kurowicka, & Joe, 2009) on even large correlation matrices such that the prior downweights -1 and $+1$ as possible values. This is illustrated in Figure 4. When the nu (ν) parameter in the built-in Stan function for an LKJ prior is less than or equal to 1, extreme values are weighted higher than intermediate values; these cases are the opposite of what we mean here by regularizing priors. When $\nu = 1$ or higher, extreme values are downweighted. Such regularizing priors are defined for all other parameters in the model. For detailed tutorials specifically intended for psycholinguistics, see Vasishth et al. (2018), Nicenboim and Vasishth (2016), Sorensen, Hohenstein, and Vasishth (2016). More general introductory book-length treatments suitable for psychologists and psycholinguists are Kruschke (2014) and McElreath (2016). An advanced treatment is in Gelman et al. (2014).

Throughout this paper, we will summarize the posterior distributions with their mean and the 95% credible interval.[3] This equal-tailed 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 parameter of interest. The frequentist confidence interval cannot quantify uncertainty about the estimate of interest for two reasons. First, a

[3]Kruschke (2014) uses highest posterior density intervals. As Kruschke puts it: "the HDI summarizes the distribution by specifying an interval that spans most of the distribution, say 95% of it, such that every point inside the interval has higher credibility than any point outside the interval." Kruschke, 2014, p. 87. This interval is identical to the credible interval when the posterior distribution is symmetric about its mean. When the posterior is asymmetric, the HPDI and the credible interval will have a large overlap, but the lower and upper end-points will differ. In our data, the posteriors of interest are symmetric.

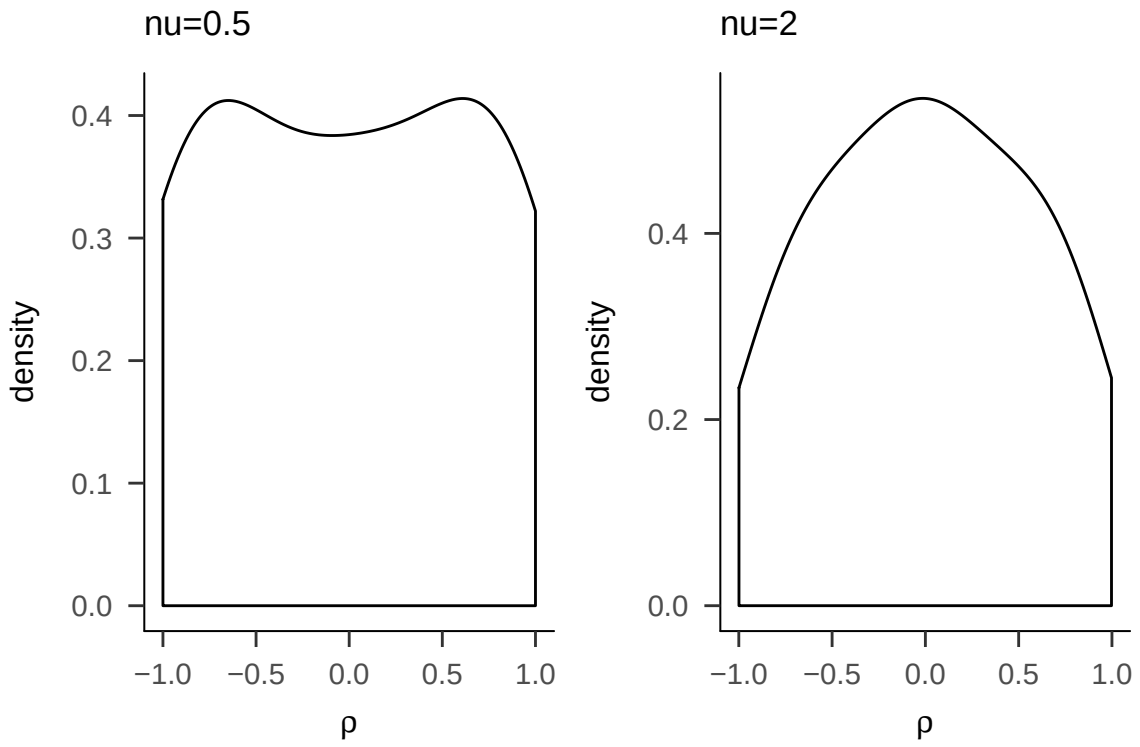

Figure 4. Example showing two different prior distributions using LKJ priors on a correlation parameter ρ. When the nu parameter in the LKJ function is 2, this downweights extreme values such as ± 1. The LKJ(2) prior can be used to define priors for arbitrarily large correlation matrices, not just for a single correlation parameter.

parameter in the frequentist paradigm is an unknown point value. A point value cannot have any uncertainty associated with it. All estimation in the frequentist paradigm is done with reference to the sampling distribution of the estimator, $\hat{\mu}$. Given a particular data set consisting of a vector of data points x, the sample mean $\bar{x}$ serves as an estimate of the point value μ. Thus, a particular confidence interval either contains the true, unknown μ or it doesn't. Second, the meaning of confidence intervals is so convoluted that statistics textbooks intended for psychology and linguistics routinely misrepresent it. The meaning of a confidence interval is that if one were to—counterfactually— repeat the same experiment multiple times and compute a 95% confidence interval each time, then 95% of those hypothetical confidence intervals would contain the true parameter μ. No probability statement can be made from any single confidence interval. For further discussion of

confidence intervals, see Hoekstra, Morey, Rouder, and Wagenmakers (2014), Kruschke (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. Their contrast 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 by Levy and Keller (2013) for main effects of Dat(ive), Adj(unct), and their interaction for the two experiments. 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) 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 are available from http://bit.ly/SSFilter. Because a reviewer requested it, all models were also refit using raw reading times with `lme4`. The results do not change depending on whether one log-transforms or not. In all the Stan models, regularizing, 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 regularizing LKJ priors on the correlation matrix (Stan Development Team, 2016). For each model, we ran four chains with 2000 iterations each. The first half of these were a warm-up and were 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. This was done as follows. Suppose that the fixed effects part of the model is defined as:

$$\log(rt) = \beta_0 + \beta_1 Dative + \beta_2 Adjunct + \beta_3 Dative \times Adjunct \quad (1)$$

with the effects coded as ± 0.5. Then, we can obtain the difference in means between the two levels of Dative by computing: $exp(\beta_0 + \beta_1 \times 0.5) - exp(\beta_0 - \beta_1 \times 0.5)$. Analogous calculations were done for the other factors. Posterior distributions on the raw ms scale were generated within the Stan model.

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 (Levy & Keller, 2013, pp. 209, 214). 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.[4] 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, the proportion of first-pass regressions, and skipping proportions in Experiment 2 (in the critical or post-critical region). None of these came out statistically significant.

Limiting the dependent measure to total reading times had a second motivation: Analyzing multiple dependent measures 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 models were fit, so for each of the three predictors (the effect of Dative, Adjunct, and their interaction) a total of 32 statistical tests were conducted for both experiments combined. 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. Because of the inflated probability of incorrectly rejecting the null when multiple dependent measures are analyzed, it is vitally important to correct Type I error probability, e.g., via the Bonferroni correction, to compensate for the inflated false positive rate (von der Malsburg & Angele, 2017). We avoid having to do this correction by only investigating total reading times.

Our estimates of total reading times match LK's published results quite closely (see their Tables 6 and 9, pp. 208, 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. Some estimates (e.g., the effect of Dative in Experiment 1) differ slightly between

[4] We attempted to obtain the Konieczny and Döring estimates for total reading time in order to compare them with the LK estimates, but were unsuccessful.

LK's analysis and ours, because we analyze on log-transformed data and back-transform to raw reading times, whereas LK analyzed raw reading times.

Our re-analysis of the LK Experiments 1 and 2 is summarized in Figure 5. 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 5, an analysis of total reading times suggests the following:

1. In Experiment 1, at the critical region, Dative has the estimate 80 ms, with a 95% credible interval [16, 153]. 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:

   > "[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." (p. 214)

2. In Experiment 2, at the post-critical region, the estimate of the interaction between Dative and Adjunct is 82 ms [19, 146]. LK's interpretation is that having both the dative NP 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 (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." (p. 214)

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 Linzen & Jaeger, 2016; Kwon, Lee, Gordon, Kluender, & Polinsky, 2010; Demberg & Keller, 2008). Similarly, many studies show evidence for memory-based effects; see, for example, Grodner and Gibson (2005), Van Dyke and Lewis (2003), Van Dyke and McElree (2006), Van Dyke and McElree (2011). Given the literature, it makes sense that we see effects of memory retrieval only under high processing load induced by encountering a relative clause: all demonstrations of locality effects in the literature (e.g., Hsiao & Gibson, 2003; Grodner & Gibson, 2005; Bartek, Lewis, Vasishth, & Smith, 2011) 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 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 80 ms, but the 95% credible interval is 16 to 153 ms; and the evidence for the second conclusion comes from an estimate with mean 82 ms, with credible interval 19 to 146 ms. These wide uncertainties imply that values as small as 20 ms are also plausible.

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 $[0, 18]$. A meta-analysis by Jäger, Engelmann, and Vasishth (2017) has also shown that similarity-based interference effects as reported by Van Dyke and colleagues have a 95% probability of lying between 2 to 28 ms, with mean 13 ms. If memory retrieval effects generally have a small magnitude in reading

studies, and if a sample size of 28 participants and 24 items 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, assume that the true effect in the LK studies is 30 and 50 ms, and that the standard deviation is 200 ms (this is a reasonable estimate for total reading time). In this scenario, power for 28 participants would be about 12 to 25%.[5] Because of Type M error, with 28 participants it would be impossible to obtain statistically significant results *that are also accurate estimates of the effect*.

But how can we determine whether the effects in the LK studies are the result of Type M error? If the LK results were not due to a Type M error and LK's effect sizes were in fact as large as LK's estimates, conducting a replication with 28 participants should have sufficient power to detect them reliably and we should be able to reproduce the effect consistently. However, if the LK results were due to Type M error leading to an overestimate of the true effect, we should 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 was used in the original study.

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 (SPR) studies of the two LK studies, and two eyetracking (ET) studies. We chose these two methods because they are the two standard behavioral approaches for studying cognitive processing costs in reading, and the previous research on expectation-based effects and memory effects has largely relied on either self-paced reading or eyetracking.

Two definitions of replication success. Before we discuss the replication attempts, 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

[5]The same calculation for power can be done in a more sophisticated way, using linear mixed models, with very similar results. See Appendix B for full details.

to be significant in the replication attempt. Alternatively, a successful replication could have the interpretation that the estimated mean from a replication attempt falls within the 95% credible intervals of the original estimates. We will consider both possible ways to interpret a replication attempt.

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 numbers of participants and items as LK (28 and 24, respectively). 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 within-participants factorial design of the original study. The factors were Dative (in main or subordinate clause) and Adjunct (in main or subordinate clause). We used the same 24 experimental items as LK from their Experiment 1 and 2, and 48 filler items. The yes/no comprehension questions that followed the items targeted various dependencies; these were also identical to the questions employed in the LK experiments. For the example in Table 1, the question for condition (a) was 'Did the teacher impose something on the naughty son?' (*'Hat der Lehrer dem ungezogenen Sohn etwas verhängt?'*) and the question for condition (b) was 'Did the teacher impose detention classes?' (*'Hat der Lehrer den Strafunterricht verhängt?'*). For a list of all experimental and filler items with their respective comprehension question, see 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.[6] 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

[6]See http://tedlab.mit.edu/∼dr/Linger/.

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[7]) 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.

Differences between the LK studies and ours. Our procedure and participants differed from the one used by LK in the following way. The original 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[8] for the experimental presentation. We used Eyelink 1000 and the Experiment Builder software.

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 on the screen of the critical

[7] http://www.sr-research.com/eyelink1000.html.
[8] https://blogs.umass.edu/eyelab/software/.

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.

In the eyetracking experiments, the critical and post-critical regions were the same 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). Finally, 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 an ambiguity in the question. For details on these changes, please see the supplementary materials.

LK had 44 filler items in each of their Experiments 1 and 2, but not all were identical across the experiments. We combined their fillers from their two experiments to assemble 48 filler items, which were then held constant across all the experiments we conducted.

Finally, the population of participants differed significantly between the original LK studies and ours. Our participants were native speakers of German who were undergraduates at the University of Potsdam, whereas LK's participants were native speakers of German living in Edinburgh (Levy & Keller, 2013, p. 204).

Results of the four replication attempts

Question-response accuracies. The question-response accuracies for the SPR replications of LK's Experiments 1 and 2 were 66 and 61%, respectively; and for the eye-tracking replications, they were 64 and 60%. These are comparable to LK's 69% and 65% in their Experiments 1 and 2, respectively.

Reading time results. Figure 5 summarizes the results of our four experiments. Recall that a successful replication can either mean that a significant effect found in an original study is found to be significant in a replication attempt; or it can mean that the

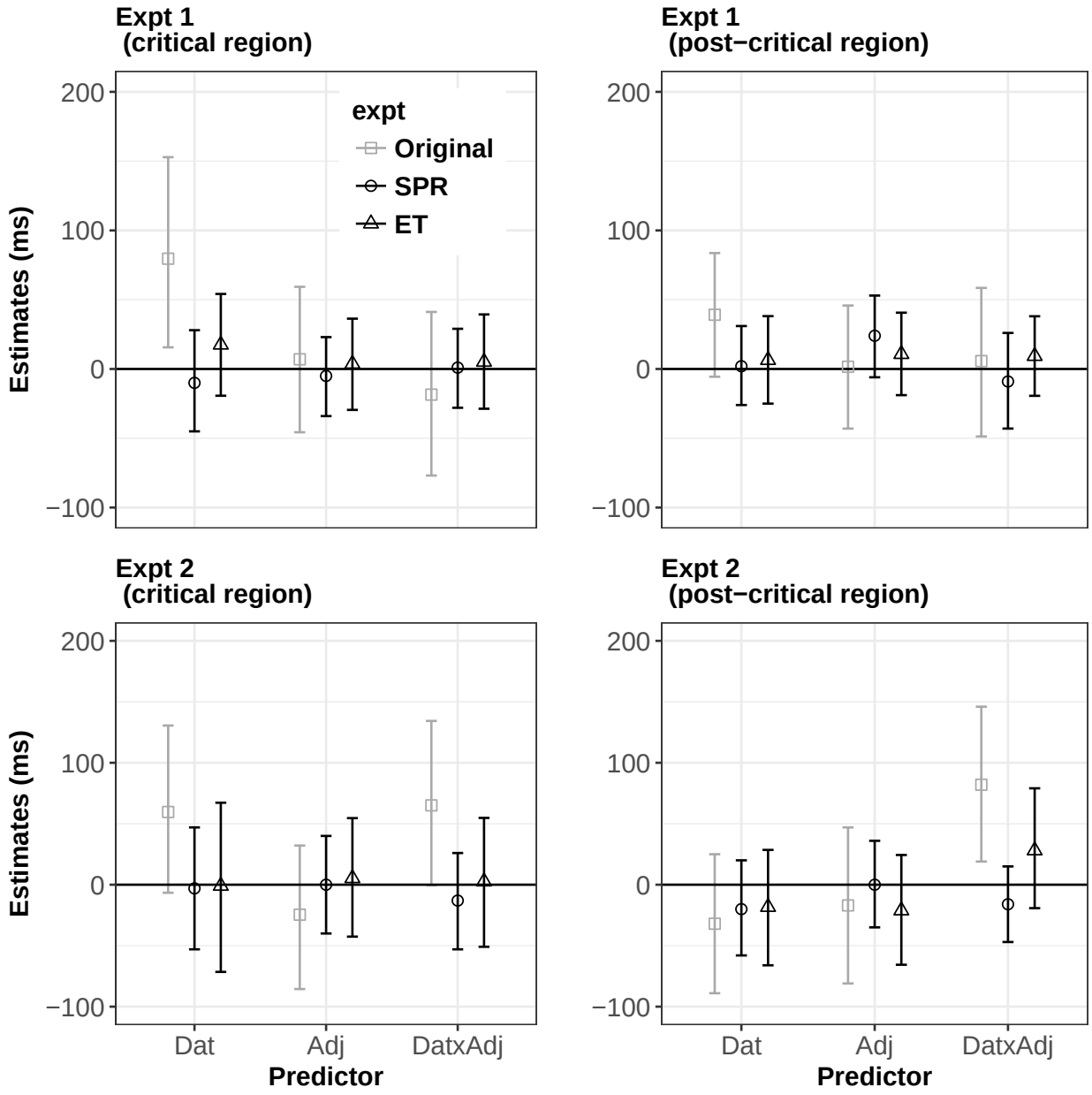

Figure 5. The effects of Dat(ive) and Adj(unct) interposition (and their interaction) at the critical and post-critical regions. Shown are the mean and 95% credible intervals from the original LK Experiments 1 and 2, and the two replication attempts. SPR stands for self-paced reading, and ET stands for eyetracking.

estimated means from the replication attempt fall within the 95% credible interval 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 [-45,28].
- Eyetracking replication of Expt 1: Effect of Dative in critical region 18 ms [-19,54].
- SPR replication of Expt 2: Interaction of Dative and Adjunct in post-critical region -16 ms [-47,15].
- Eyetracking replication of Expt 2: Interaction of Dative and Adjunct in post-critical region 28 ms [-19,79].

However, the replication attempts can also be seen as a near-complete success: *all* the total reading times estimates from the eyetracking studies (and 9 of the 12 of the estimates computed in the self-paced reading experiments) 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 effect being near 0 ms. When the estimates are noisy, the p-value furnishes little information about reliability (i.e., that the effect is true) or replicability (i.e., that the significant effect can be reproduced if the study is repeated). Of course, even when estimates are not noisy, the only way to establish replicability is to actually replicate the effect.

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 turned our attention to one of the conclusions that LK drew from their study (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.*" (p. 214)

The emphasis is ours. 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 across the two experiments 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 (Levy & Keller, 2013, pp. 209, 214), 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 our 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*". LK did investigate the expectation-locality interaction in Experiment 2, but the claim to be investigated involves the patterns seen across Experiments 1 and 2, and this was not checked.

Table 4

Example items (simplified) for investigating the Load-Distance interaction by combining conditions (c) and (d) of LK's Experiment 1 and of Experiment 2. The abbreviations mean the following: ADJ: adjunct; DAT: dative; PP: prepositional phrase; NP: noun phrase.

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 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: 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 6, in the LK data the estimates for the interaction in the critical region are -52 ms [-110,9]; and in the post-critical region, -40 ms [-92,10]. 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, the interaction would not come out significant. However, significance is not interesting for us. 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.

Two replication attempts of the Load-Distance interaction

We first carried out two attempts to reproduce the Load-Distance interaction, using the original sample size of 28 participants. 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.

As shown in Figure 6, both replication attempts showed that the estimate for Load in the critical region had a positive sign; 76 ms [42,111] for SPR; and 152 ms [104,200] for eyetracking (total reading times). 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. Differently put, the Load effect could at least partly be due to the word length effect. Therefore, we disregard the Load effect, even though theoretically the sign of the effect makes sense.

The estimate for Distance is close to 0 ms: 3 ms [-30,39] for SPR; and 1 ms [-38,39] for eyetracking (total reading times). Finally, the interaction between Load and Distance is not far from 0 ms; 5 ms [-30,43] for SPR and -14 ms [-48,20] for eyetracking (total reading times).

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

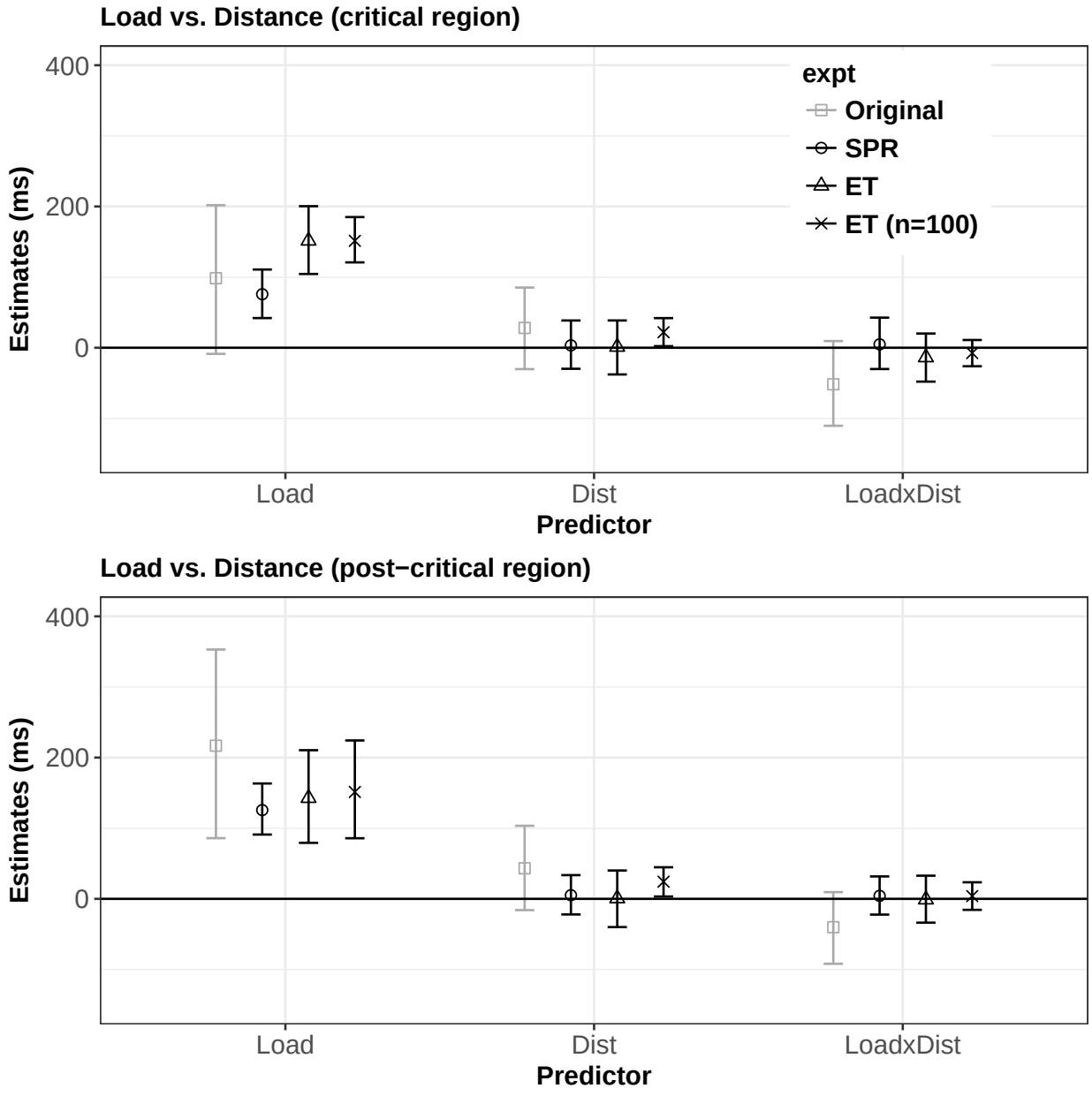

Figure 6. Load and Distance effects at the critical and post-critical regions. Shown are the mean and 95% credible intervals from conditions (c) and (d) of the two original LK Experiments 1 and 2; and from our three replication attempts. SPR stands for self-paced reading, and ET stands for eyetracking.

interaction claimed by LK is very well-motivated both theoretically and empirically. We turn to this larger-sample study next.

A larger-sample replication attempt of the Load-Distance interaction

Before we discuss the results of the larger-sample study, we first explain how we determined the target precision for our study in order to decide on sample size. This uses an approach that Kruschke (2014) refers to as the region of practical equivalence (ROPE). We also discuss how the ROPE approach can be used to make decisions about the research question.

Determining sample size using a Bayesian approach. The Bayesian framework allows us to determine how many participants we should run in order to make a decision about our research question. One way to do this is to define what constitutes "no effect." This approach was developed in the context of clinical trials, where it is essential to stop the trial if the treatment is turning out to harm the patients, or when it is immediately clear that the treatment is superior to the control (Freedman, Lowe, & Macaskill, 1984; Spiegelhalter, Freedman, & Parmar, 1994). Kruschke (2014) introduced this idea into psychology, but it has not yet been widely adopted. This approach serves both as a stopping rule, and for deciding whether one has evidence for one's theory.

As mentioned above, the starting point is to define what counts as "no effect." Instead of the frequentist approach of asserting a point null value, we can define a *region of practical equivalence* that counts as a null region. For example, in LK Experiment 1, we start by asserting that in total reading times, an effect magnitude that has mean 0 and lies between -20 and 20 ms with probability 95% constitutes "no effect." Note that if we were investigating first-pass reading times, the range would be much smaller, because effects in first-pass reading time will be smaller in magnitude.

How did we decide on the width of 40 ms for the region of practical equivalence? This decision is subjective but not arbitrary. It is based on estimates derived from what is already

known and well-established empirically.[9] For clear grammaticality violations that the reader is immediately consciously aware of, total reading time effects (at the word where the ungrammaticality is detected) can show effect magnitudes of approximately 100 to 150 ms. For example, the data in Dillon, Mishler, Sloggett, and Phillips (2013) (their Experiment 2) show a 41 ms $[23, 58]$ effect of ungrammaticality (n=40) in first-pass reading time (FPRT), and a 100 ms $[69, 134]$ effect in total reading time (TRT). In a large-sample (n=181) replication attempt of Dillon et al.'s work (Jäger, Mertzen, Van Dyke, & Vasishth, 2018), we found an effect of 55 ms $[45, 65]$ in FPRT, and an effect of 121 ms $[100, 141]$ in TRT. We consistently find this magnitude of effect or smaller effects when the sentence is ungrammatical; for example, Wagers, Lau, and Phillips (2009) and Lago, Shalom, Sigman, Lau, and Phillips (2015) also showed the effect of (un)grammaticality in SPR with estimates similar to those found by Dillon and colleagues. Sometimes we see even larger effects for ungrammaticality; for example, an eyetracking study by Paape, Hemforth, and Vasishth (2018) shows that total reading times at the moment that an ungrammaticality was registered in French was 176 ms, with 95% credible intervals 84 and 264 ms. Now, if we consider more subtle experimental manipulations in sentence processing, the effects in total reading time are in a lower range than effects of grammaticality. As an example, we mentioned earlier that a meta-analysis showed that the similarity-based interference effects found by Van Dyke and colleagues have a posterior mean of about 13 ms, with 95% credible intervals $[2, 28]$ ms. Since these estimates were based on SPR data and first-pass reading times, it is reasonable to assume that in total reading time the effect of interference would be about twice as large; say 30 ms. This is because effects are larger in total reading time than first-pass reading time (TRTs are a sum of first-pass reading time and re-reading time). Given these assumptions, for total reading times we can fix ± 20 ms around 0 ms as counting as effectively a null effect for the LK studies.

[9] This estimation approach is sometimes called Fermi-zation (Tetlock & Gardner, 2016). The name comes from Fermi's skill in obtaining rough but accurate estimates for physical phenomena; an example is the 1945 nuclear detonation conducted as part of the Manhattan project (the Trinity test). Fermi obtained remarkably accurate estimates of the blast's force before the data were available.

Our estimates of the region of practical equivalence (ROPE) are based on a empirical argument, but are of course open to challenge. We cannot provide a one-size-fits-all recommendation for deciding on a null region for specific phenomena, but we believe that for the present question, our estimates are reasonable. For subtle subtle phenomena for which no data exist, some initial experiments would be necessary to establish the ROPE.

Once we have decided on a null region, the goal should be to collect data until the 95% credible interval of the parameter of interest is at most as wide as the null region; in the above example, it should be at most 40 ms wide. This is how we established our stopping rule in our pre-registration of the larger-sample study (which is available from: https://osf.io/dgewb/). Note that, unlike the frequentist power analysis, we do not fix a sample size in advance, but rather run the experiment until a certain precision is reached: until the 95% credible interval of the posterior distribution has width 40 ms or less.

Using the ROPE method to evaluate the research hypothesis. The ROPE method can also be used for making decisions about the research question. As shown in the figure below, once the data with the appropriate precision have been collected, there are five possible scenarios.

- A, B: data's credible interval falls clearly outside the null region. Decision: reject the null region.
- C,D: data's credible interval overlaps with null region. Decision: not conclusive.
- E: data's credible interval falls within the null region. Decision: conclude that the data are consistent with "no effect".

Kruschke (2014, p. 337) points out that, apart from these five scenarios, it is possible that the parameter's credible interval falls entirely within the null region, but excludes the zero value. Kruschke points out that this situation is rare, and that it indicates that the null region was defined to be too wide: the data can be collected with much higher precision than the ROPE expressed by the null region. For further discussion of the ROPE approach, see

B. P. Carlin and Louis (2008) and Kruschke (2014). Incidentally, the ROPE method can also be used for affirming a theory's predictions. A stringent test of a theory's predictions would be that the posterior's credible interval falls within the predicted range of predictions from theory; weaker evidence for a theory would involve overlap with the predicted range of values; and a rejection of a theory would involve a credible interval from data that falls completely outside a predicted range of values. In the General Discussion, we give an example of how ROPE can be used for model evaluation.

An obvious objection to the ROPE approach is its subjectivity. One can empirically justify a region of practical equivalence, but different researchers could define different regions. By contrast, the NHST approach is seen as being objective, with a fixed decision criterion of $p < 0.05$. But this objectivity of NHST is just an illusion. When a desired significant result is not obtained, the strict significance criterion is routinely circumvented by relaxing the requirement that $p < 0.05$. It is common to use phrases like 'marginally significant', 'a trend towards significance', etc.[10] Other subjective decisions are routinely taken in NHST and there are no fixed standards for these. Example are deleting (or deciding not to delete) values that are 2.5 or 3.5 standard deviations from the mean; and deciding to remove subjects whose accuracies fall below a certain level or who answered the questions incorrectly, or deciding not to remove any subjects at all. These and other subjectivities are never questioned when using the NHST procedure. The only reason that the ROPE proposal sounds subjective is that it involves a subjective decision we are not used to taking.

Another objection could be that the ROPE approach can be misused. For example, one could first run the study and compute the standard error and then state that the null region corresponded to four times the standard error. However, we are assuming here that the definition of the null region will be decided on before the experiment is conducted. This is no different than computing prospective power before conducting an experiment.

[10]For a list of over 300 phrases commonly used to circumvent the supposedly strict criterion of $p < 0.05$, see https://mchankins.wordpress.com/2013/04/21/still-not-significant-2.

Finally, this null region approach does not solve the problem of demonstrating replicability; whatever the outcome of an experiment, one would still need to replicate the effect. The only way to establish replicability is to actually conduct pre-registered direct replications. We discuss pre-registration and replication in the general discussion.

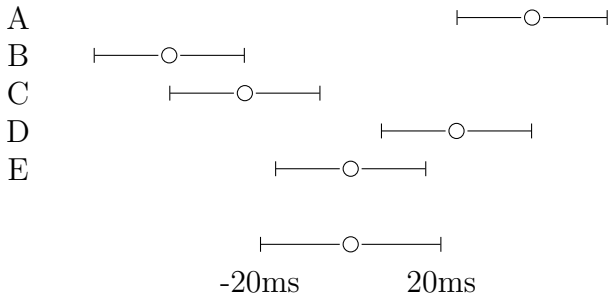

Figure 7. The five possible outcomes when using the null region or "region of practical equivalence" method for decision-making (Kruschke, 2014). The estimates A and B would lead to a decision to assuming that the effect of interest is not null; estimates C and D are inconclusive; and the estimate E is consistent with the null region. The width of the region here is 40 ms, but would depend on the dependent measure, and the measurement precision achievable by the instrument.

We now turn to the results of our larger-sample study, in which we investigated the Load-Distance interaction.

Results of the larger-sample study. The results of this 100-participant study are summarized in Figure 6. This time, the estimate of Load at the critical region is 151 ms [121,185]; the effect of Distance is 22 ms [2,42]; and the Load-Distance interaction is -8 ms [-26,11].[11]

Discussion. In this larger-sample study, the positive coefficient for Distance suggests that increasing subject-verb distance by interposing an adjunct 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 ms [-14,43]; and in the high-load conditions, it is 29 ms [2,55]. The larger distance effect in the

[11] At first glance, it may be surprising that in the post-critical region, the 95% credible interval for the effect of Load in the 100-participant study is as wide as that of the 28-participant eyetracking study. One might expect that a larger-sample study always yields a narrower credible interval. But this need not necessarily be true; the credible interval is dependent on the estimates of the variance components, which will vary from study to study.

high-load conditions is compatible with the LK argument in their paper that locality effects outweigh expectation effects when memory load is high. However, the expectation account incorrectly predicts a negative coefficient in the low-load conditions. One possible explanation for the smaller distance effect in low-load conditions could be that expectation and locality act in opposite directions. Such an explanation is compatible with the LK proposal, and is consistent with the data. However, note that when we use the region of practical equivalence approach, both the two nested contrasts and the main effect of Distance are not conclusive because the 95% credible interval of the respective estimates overlap with the ROPE of ±20 ms centered around 0 ms.

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 8, 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 overestimates relative to the reference point of the mean and credible intervals estimated from the full data set. 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, 2000), which predict increased processing difficulty when subject-verb distance is increased. However, this Distance effect yields estimates that are also consistent with our posited null

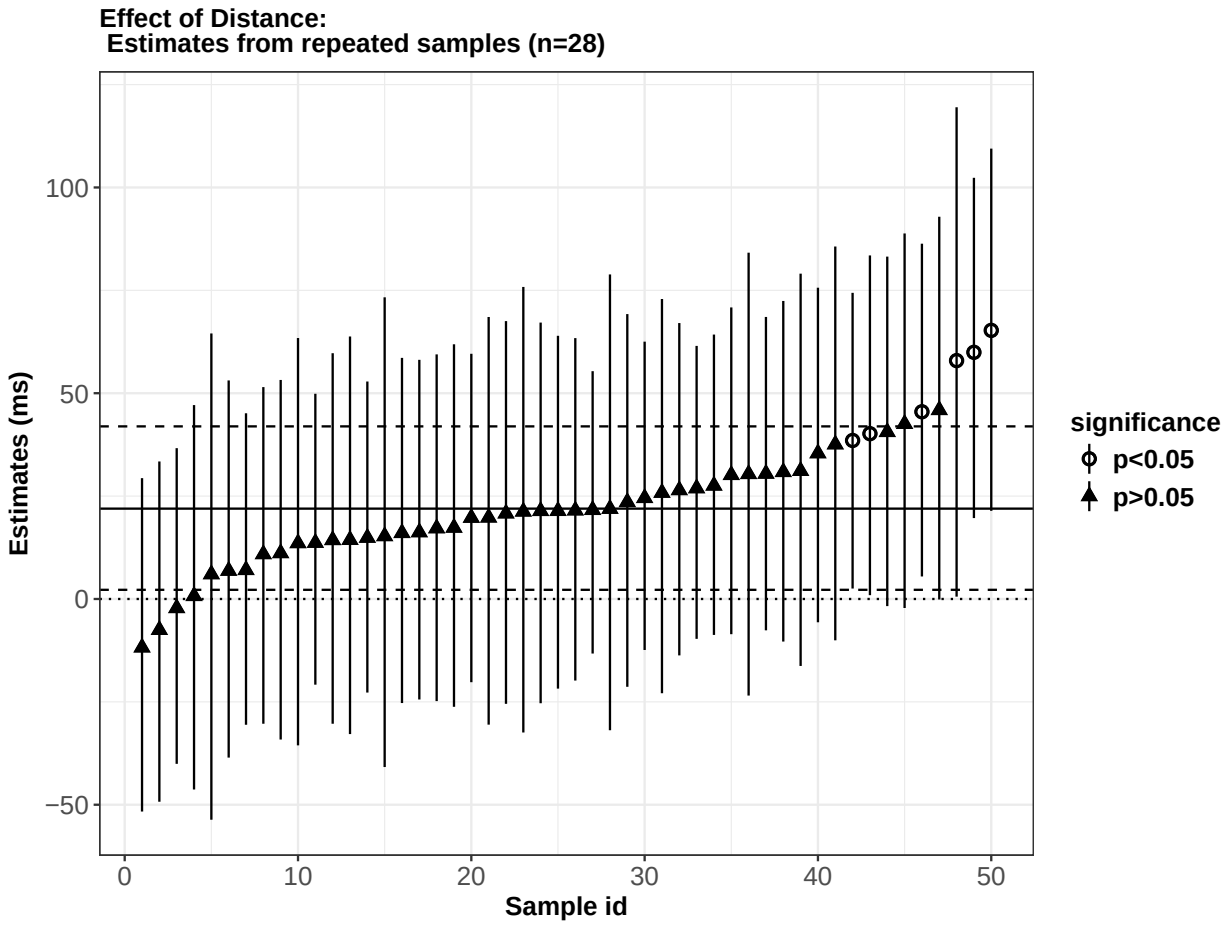

Figure 8. A demonstration of the fluctuation in the estimates for the effect of Distance when we choose 28 participants pseudo-randomly from the 100-participant experiment. The solid horizontal line is the estimated mean from the 100-participant data set, and the broken lines show the corresponding 95% credible intervals. The points show the means and 95% credible intervals when randomly sampling from the 100-participant data set.

region; so the evidence for the Distance effect cannot be considered convincing.

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 in the original studies 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.

Regarding the absence of locality and expectation effects in our experiments, our point here is not that the 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 or not statistically significant effects are found.

Our seventh experiment showed that a larger-sample study generally delivers narrower credible intervals. It also delivers smaller estimates of the mean, 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) has some weak support. 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 whether head-final languages also consistently show these effects. An important line of research would be to attempt to replicate the published results for head-final languages like German, Hindi, and Persian.

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 (from self-paced reading and eyetracking) across different studies fluctuate quite a lot, from -123 to 100 ms, with confidence intervals ranging in width from 80 to 320 ms (also see Vasishth, 2015). 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 (see Engelmann et al., 2018), faster reading times are expected at the auxiliary when the preceding noun agrees in number with the auxiliary's number marking (i.e., the auxiliary verb in *cabinets are* is read faster than in *cabinet are*). One theory, the Lewis and Vasishth (2005) cue-based retrieval model, predicts that the mean expected facilitation of −25 ms with the original parameter settings used in the original Lewis and Vasishth model. If the model parameters are varied over a narrow range, the mean expected facilitation varies from −10 to −50 ms.[12] Several studies have been published showing statistically significant facilitation effects, as predicted by theory. Because of the repeated significant effects found, this facilitation effect is considered very reliable in psycholinguistics. We re-analyzed the data (self-paced reading or total reading time in eyetracking) from 10 published experiments, 8 out of 10 reported a significant effect. We fit Bayesian linear mixed models with full variance-covariances matrices for all random effects, and the same regularizing, weakly informative priors that we used in the LK data. Unlike the original studies, we did not delete extreme values; rather, we modeled the reading-time data as being generated from a log-normal distribution and back-transformed the estimates to the milliseconds (see Appendix C for details). We find that the uncertainty of the estimates in the data is quite high: The ten studies' mean estimates range from -40 to -4 ms, with credible intervals ranging in width from 30 to 89 ms. These empirical estimates (along with their 95% credible intervals) are all consistent with the model predictions (−10 to −50 ms), in the sense that the credible intervals from these 10 studies overlap with the theoretically predicted range. But these data do not strongly validate the Lewis and Vasishth model predictions. If these estimates had been more precise, i.e., had much narrower credible intervals, they might have fallen within the predicted range; this would have been a stronger validation of model predictions. With such wide credible intervals, a wide range of outcomes is compatible with

[12] An online Shiny app provides the quantitative predictions: https://engelmann.shinyapps.io/inter-act/ (Engelmann et al., 2018).

the data, including no facilitatory effect at all. Thus, even in the relatively clear agreement attraction case, higher precision replication attempts need to be carried out to determine better estimates of the facilitation effect.

A 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 are *guaranteed* to 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. In fact, with a null region defined under the region of practical equivalence approach, both results could be consistent with there being "no effect." However, the smaller estimate with narrower 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. We are not suggesting that low-powered studies that show significant results should now no longer be published in journals. We are only suggesting that such results should be accompanied with an appropriate level of awareness of what they do and do not tell us. Strong claims should be avoided in such cases.

It is of course possible to run higher-power studies. How can we decide what constitutes a higher-powered study? Frequentist statistics has several proposals for sequential testing (e.g., Frick, 1998), which avoid running unnecessarily large numbers of participants. A Bayesian approach that we used in this paper is to define a region of practical equivalence for total reading time (specifically, $\pm$20 ms around 0 ms) and to run the experiment until the desired precision is reached. Our choice of a 95% credible interval width of 40 ms 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 for the estimate, and to use the region of practical equivalence for interpreting the results, possibly alongside the p-value if that makes the researcher feel more comfortable .

Once we have fixed the precision that is theoretically meaningful 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, in the Bayesian framework, 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 within the frequentist paradigm, this stopping criterion will inflate Type I error (e.g., Pocock, 2013). In the Bayesian framework, there is no concept of hypothetical replications; the data at hand are not interpreted in the light of imagined repeated sampling (Gelman et al., 2014). 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, and this guess is often based on previously published data. As we have shown here, when the sample sizes are small and there is a bias to only publish statistically significant effects, effect magnitudes will be overestimated by a large amount. Using these estimates leads to a large underestimation of the sample size needed for high-powered replications. In a precision-based analysis, the focus is on the amount of uncertainty in the estimate that we are willing to tolerate. The magnitude of the estimate, together with its uncertainty, are much more important theoretically than just counting the number of significant vs. not

significant results in the literature. Such a vote-counting approach is commonly adopted to decide whether an effect is "present" vs. "absent." The voting-based approach would be fine if there were no publication bias at all and if power were sufficiently high in published studies. For an example of a voting-based approach to deciding whether an effect is present or absent, see Phillips, Wagers, and Lau (2011). There, when discussing whether reflexives show similarity-based interference effects, the authors conclude: "Thus, most evidence suggests that the processing of simple argument reflexives in English is insensitive to structurally inappropriate antecedents, indicating that the parser engages a retrieval process that selectively targets the subject of the current clause." If power in the studies Phillips and colleagues base their conclusions on is low, then many null results are to be expected. It is well-known in statistical theory that null results from low-power studies should be treated as inconclusive rather than proving that the null hypothesis is true; unfortunately, this detail has been lost in translation when statistical methods were adopted in the psychological sciences. In sum, simple vote-counting would be highly misleading when power is low and publication bias exists.

Having higher-precision estimates will allow for better-quality formal model comparison of competing quantitative models. The first comprehensive quantitative evaluation of the computational memory-retrieval model of Lewis and Vasishth (2005) involved comparing model predictions to 77 published results on retrieval processes (Engelmann et al., 2018). This was only possible because the estimates (and their uncertainty) were available from a meta-analysis (Jäger et al., 2017). The meta-analysis provided estimates based on all relevant reading-time studies which were then compared with the model predictions. Although the meta-analytic estimates are likely to be biased (due to publication bias and Type M error in individual studies), they are more precise than the estimates from individual studies because the meta-analysis aggregates data from multiple studies after weighting them by their precision—the meta-analysis allows us to take into account accumulated knowledge in a quantitative manner. The results of the

quantitative evaluation by Engelmann et al. (2018) would have looked very different if the estimates from the published individual studies had had higher precision.

In addition to fixing precision in advance, our second suggestion is that we should attempt to conduct direct, pre-registered replications of experiments, because there is no guarantee that a result reflects reality just because it is statistically significant. Every major claim should be either accompanied by a pre-registered direct replication, or even better, other researchers from competing labs should be encouraged to replicate the original result. Direct replications are necessary even for higher-precision studies, because population differences, lab practices, etc. can easily bias an individual result. As discussed in Chambers (2017), pre-registration involves defining in advance the analysis that is planned and depositing this in an embargoed repository like OSF (osf.io), which time-stamps the pre-registration. This step avoids or at least minimizes problems like p-hacking and the garden-of-forking paths (Gelman & Loken, 2016; Forstmeier, Wagenmakers, & Parker, 2017; Simmons, Nelson, & Simonsohn, 2011) that have plagued psychology and other areas. With pre-registration, the researcher is still free to explore their data after the fact, but pre-registration clearly separates the prior analysis plan from the exploratory part (De Groot, 1956/2014; Nicenboim et al., 2018). Currently, due to the unreasonable pressure to publish fast and to report novel results in top journals, crucial data-analysis decisions are often made after examining the data. For example, the same researcher will often include or exclude data on different criteria, such that it eventually passes the statistical significance filter. Sometimes, excluding or including a few data points can make the difference between significance and non-significance. Another example is region-of-interest selection in reading studies: researchers often change the region of interest from study to study or even within a study, in order to "tell the best story." Another common approach is to run the study, check for significance, then either run more participants if significance is desired but not reached, or stop collecting data if a null result is desired. Pre-registration would remove these degrees of freedom and thereby ensure a clear separation between confirmatory and exploratory

analyses (De Groot, 1956/2014).[13]

Our third suggestion is that data and code be released mandatorily along with the published paper. Some authors are happy to share their data and code, but in many cases the crucial information—the data itself—are not available. For example, Nieuwland et al. (2018) tried but failed to obtain the original data that they attempted to replicate. Many researchers, such as Levy and Keller and Colin Phillips' lab in the present case, generously released their data in connection with the LK and other replication attempts; without the raw data, we would not have been able to conduct a complete analysis. But the first author's experience has been that 25-30% of the attempts to obtain raw data are unsuccessful. This seriously hampers scientific progress.

Leading journals could trigger a positive change by requiring data and code release for all articles, and introducing a special article type (e.g., a pre-registered 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.

There is clearly a downside to focusing on higher precision and direct replications. Perhaps the biggest one is that carrying out experiments towards the aim of increasing precision would take much longer. For example, the experiments in the present paper were started in December 2015, and ended around July 2017, a period of nearly two years. This means that at least in smaller universities, where recruiting participants is not easy, internet experiments may serve as a partial solution (but this comes with other disadvantages). Another obvious side-effect is that the speed with which we can publish papers will go down.

[13] A common objection we hear is that anyone could defeat the purpose of pre-registration by first collecting the data and then depositing a fake pre-registration. But this would just be scientific fraud; pre-registration is not designed to solve that problem.

Clearly, expectations regarding publication rate need to change.

In closing, a 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 because they may be consistent with effectively no effect and may not be replicable in the sense that significant effects may not be found to be significant under replication. Too often, published empirical results are treated as a novel contribution simply because of the application of the statistical significance filter. How many published claims actually reflect reality remains to be seen. Big effects involving, e.g., grammaticality violations or strong garden paths, are likely to be replicable, but more subtle effects may not be. For example, the recent failure to find significant effects in anticipatory processing by Nieuwland et al. (2018), Kochari and Flecken (2018) suggests that replicability problems arising from the statistical significance filter could run deep in psycholinguistics. Another recent example is the claimed failure to find statistically significant adaptation effects (Stack et al., 2018). Of course, the issues are not limited to psycholinguistics and extend to all other scientific disciplines that use this decision criterion to decide whether or not to publish results. Our suggestions, to aim at higher precision, to conduct direct, pre-registered replications, and to release data and code, may contribute towards improving the reliability of published results.

Conclusion

In sentence processing, many results, such as the classical garden-path findings (Frazier & Rayner, 1982), have large and robust effects. These are very likely to be easily replicable. But the low-hanging fruit has long been picked. Subtle manipulations require designs and sample sizes that deliver accurate estimates.

We suggest that researchers (i) move their focus away from statistical significance and attend instead to increasing the precision of their estimates (e.g., by increasing sample size, or improving the quality of measurements, or designing stronger manipulations); (ii) carry

out direct (not just conceptual) replications in order to demonstrate the existence of an effect; (iii) pre-register their designs and planned analyses and deposit them in venues like osf.io; and (iv) release their data and code upon publication. Journals can encourage these practices by favoring pre-registered analyses, introducing a short-article type featuring direct replications, and mandating open data and code release upon publication. Some of the leading journals already require data and code release upon publication, and in some cases during the review process.

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. We also owe a debt of gratitude to Colin Phillips, Matt Wagers, Brian Dillon, and Sol Lago for generously sharing their data; without their data, our project would have been considerably reduced on scope. 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, Daniel Schad, Sandra Hanne, Sol Lago, Jan Vanhove, Tatjana Scheffler, João Veríssimo, Dario Paape, 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.

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

Carlin, B. P. & Louis, T. A. (2008). *Bayesian methods for data analysis*. Boca Raton, FL: CRC Press.

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.

Casella, G. & Berger, R. L. (2002). *Statistical inference*. Duxbury Pacific Grove, CA.

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

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

Cunnings, I. & Sturt, P. (2018). Retrieval interference and sentence interpretation. *Journal of Memory and Language*, *102*, 16–27.

De Groot, A. (1956/2014). The meaning of "significance" for different types of research [translated and annotated by Eric-Jan Wagenmakers, Denny Borsboom, Josine Verhagen, Rogier Kievit, Mar1 Bakker, Angelique Cramer, Dora Matzke, Don Mellenbergh, and Han LJ van der Maas]. *Acta Psychologica*, *148*, 188–194.

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, *8*(8), 1117.

Demberg, V. & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*(2), 193–210.

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., Jäger, L. A., & Vasishth, S. (2018). *The effect of prominence and cue association in retrieval processes: A computational account*. Unpublished Manuscript.

Ferreira, F. & Henderson, J. M. (1993). Reading processes during syntactic analysis and reanalysis. *Canadian Journal of Experimental Psychology*, *47*, 247–275.

Forstmeier, W., Wagenmakers, E.-J., & Parker, T. H. (2017). Detecting and avoiding likely false-positive findings–a practical guide. *Biological Reviews*, *92*(4), 1941–1968.

Frazier, L. & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, *14*, 178–210.

Freedman, L., Lowe, D., & Macaskill, P. (1984). Stopping rules for clinical trials incorporating clinical opinion. *Biometrics*, *40*, 575–586.

Frick, R. W. (1998). A better stopping rule for conventional statistical tests. *Behavior Research Methods, Instruments, & Computers*, *30*(4), 690–697.

Gelman, A. (2018). The failure of null hypothesis significance testing when studying incremental changes, and what to do about it. *Personality and Social Psychology Bulletin*, *44*(1), 16–23.

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., Aki, & Rubin, D. B. (2014). *Bayesian data analysis* (3rd 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.

Gelman, A. & Loken, E. (2016). The statistical crisis in science. *The Best Writing on Mathematics 2015*, 305.

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* (pp. 95–126). 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 K. Knight (Ed.), *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics*. Stroudsburg, PA: The Association for Computational Linguistics.

Heister, J., Würzner, K.-M., Bubenzer, J., Pohl, E., Hanneforth, T., Geyken, A., & Kliegl, R. (2011). dlexDB – Eine lexikalische Datenbank für die psychologische und linguistische Forschung. *Psychologische Rundschau*, *62*(1), 10–20.

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, *21*, 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.

Hsiao, F. P.-F. & Gibson, E. (2003). Processing relative clauses in Chinese. *Cognition*, *90*, 3–27.

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*, *53*, 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., 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.

Jäger, L. A., Mertzen, D., Van Dyke, J. A., & Vasishth, S. (2018). *Contrasting facilitation profiles for agreement and reflexives revisited: A large-scale empirical evaluation of the cue-based retrieval model*. MS in preparation.

Klein, W. & Geyken, A. (Eds.). (2016). *Das digitale Wörterbuch der deutschen Sprache (DWDS)*. Available from http://www.dwds.de. Berlin-Brandenburg Academy of Science.

Kliegl, R., Nuthmann, A., & Engbert, R. (2006). Tracking the mind during reading: The influence of past, present, and future words on fixation durations. *Journal of Experimental Psychology: General*, *135*(1), 12–35.

Kochari, A. & Flecken, M. (2018). *Lexical prediction in language comprehension: a replication study of grammatical gender effects in Dutch*. Available from PsyArXiv: https://osf.io/k6b9u/.

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). Sydney, Australia: University of New South Wales.

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

Kwon, N., Lee, Y., Gordon, P., Kluender, R., & Polinsky, M. (2010). Cognitive and linguistic factors affecting subject/object asymmetry: An eye-tracking study of pre-nominal relative clauses in Korean. *Language*, *86*(3), 546–582.

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.

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. (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.

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, *100*(9), 1989–2001.

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.

McElreath, R. (2016). *Statistical rethinking: A Bayesian course with examples in R and Stan*. Boca Raton, FL: CRC Press.

Nicenboim, B. & Vasishth, S. (2016). Statistical methods for linguistic research: Foundational ideas – Part II. *Language and Linguistics Compass*, *10*, 591–613.

Nicenboim, B. & Vasishth, S. (2018). Models of retrieval in sentence comprehension: A computational evaluation using Bayesian hierarchical modeling. *Journal of Memory and Language*, *99*, 1–34.

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*.

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. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., ... Huettig, F. (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *eLife*, *7*, e33468.

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

Paape, D., Hemforth, B., & Vasishth, S. (2018). Processing of ellipsis with garden-path antecedents in French and German: Evidence from eye tracking. *PLoS ONE*. Accepted.

Pashler, H. & Wagenmakers, E. (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*. Chichester, West Sussex: John Wiley & Sons.

R Core Team. (2014). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria.

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*.

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

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science*, *9*(1), 76–80.

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.

Spiegelhalter, D. J., Freedman, L. S., & Parmar, M. K. (1994). Bayesian approaches to randomized trials. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 357–416.

Stack, C. M. H., James, A. N., & Watson, D. G. (2018). A failure to replicate rapid syntactic adaptation in comprehension. *Memory & Cognition*, 1–14.

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

Tetlock, P. E. & Gardner, D. (2016). *Superforecasting: The art and science of prediction*. New York, NY: Random House.

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.

Van Dyke, J. & McElree, B. (2006). Retrieval interference in sentence comprehension. *Journal of Memory and Language*, *55*, 157–166.

Van Dyke, J. & McElree, B. (2011). Cue-dependent interference in comprehension. *Journal of Memory and Language*, *65*(3), 247–263.

Vasishth, S. (2015). *A meta-analysis of relative clause processing in Mandarin Chinese using bias modelling*. MSc dissertation, School of Mathematics and Statistics, Sheffield University, UK. Sheffield, UK.

Vasishth, S., Chen, Z., Li, Q., & Guo, G. (2013). 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., Beckman, M. E., Li, F., & Kong, E. J. (2018). Bayesian data analysis in the phonetic sciences: A tutorial introduction. *Journal of Phonetics*. Accepted.

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.

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 $\mu > 0$.[14] 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, 1):

$$g_\delta(p) = \frac{\phi(Z_p - \delta)}{\phi(Z_p)}, \quad 0 < p < 1 \tag{2}$$

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, 1) further observe that the cumulative distribution function (cdf) of P is:

$$G_\delta(p) = \int_0^p g_\delta(x)\, dx = 1 - \Phi(Z_p - \delta), \quad 0 < p < 1 \tag{3}$$

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

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

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:

$$p = 1 - \Phi(z_p) \tag{4}$$

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:

$$G_{z_p}(\alpha) = 1 - \Phi(z_\alpha - z_p) \tag{5}$$

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 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 reality.

A direct consequence of Equation 5 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%.[15]

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

[15]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)`.

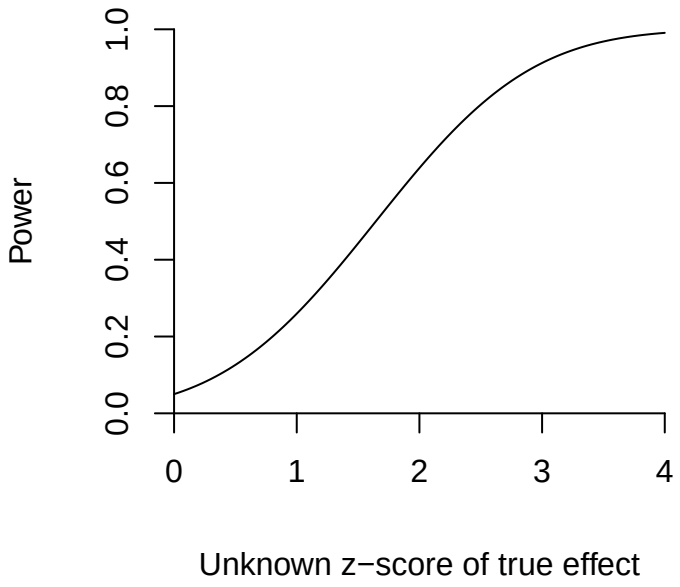

Figure A1. The relationship between power and the unknown z-score of the true effect. Larger z-scores are easier to publish due to the statistical significance filter, and these studies therefore give a mistaken impression of higher power.

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

$$G_{Z_\alpha}(\alpha) = 1 - \Phi(1.64 - 1.64) = 0.5 \tag{6}$$

Thus, in a scenario where the real power is 15%, and only z-scores greater than or equal 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.[16] This incorrect belief among researchers has the effect that

[16]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.

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

Approaches for conducting prospective power analysis for repeated measures designs. Regarding our power analysis of LK's Experiment 1 in the introduction, the reader may object that the simplified example of power analysis is artificial, because we usually fit linear mixed models, where power could be much higher due to the many variance components partitioning sources of variance in a more nuanced manner. However, this is not true. Consider the LK Experiment 1 data; we can estimate all effects and variance components from this 2×2 design by fitting a "maximal" linear mixed model and then estimating power for a range of plausible effects. When we do such a power analysis, for an effect of 30 to 50 ms, which is close to the estimates from our meta-analysis of memory retrieval effects (Jäger et al., 2017), power is around 12 to 29%; see Table B1. If the true effect were as large as 100 ms (this is the estimate reported by LK Experiment 1 for the effect of Dative), a sample size of 40 participants (and 24 items) would lead to approximately 82% power; but if the true effect is 80 ms (our estimate from the LK Experiment 1 data), 40 participants and 24 items would give approximately only 51% power. If the true effect is even smaller, obtaining power greater than 80% would require hundreds of participants and many more items. Details of these calculations, along with repoducible code, are shown in Appendix B. Thus, conducting power calculations using estimates from a linear mixed model, instead of our simpler calculations in the introduction, doesn't yield very different estimates of power.

Effect (ms)	Power (percentage)
30	12
50	29
80	51

Table B1
Estimates of power for different effect magnitudes for Levy and Keller's Experiment 1. These estimates of power use estimates of variance components computed from the Levy and Keller data; see Appendix B for details.

How exactly did we compute these power estimates? For a repeated measures design, one convenient way to compute sample size via prospective power analysis is by using

fake-data simulation. As an illustration, we consider how we would compute prospective power for the LK13 Experiment 1.

1. Fit a "maximal" linear mixed model to existing data. As an example, we fit the model to the LK Experiment 1 data below.

2. Extract all variance component estimates and fixed effects estimates from the fitted model. For the fixed effect of interest, choose a range of effect magnitudes that are considered realistic (this is discussed below in detail).

3. Using these estimated means, and the assumed effect magnitude, repeatedly generate 100 fake data sets with a particular number of participants and items, and compute the proportion of times that the relevant predictor is "significant" at the specified α value (here, 0.05). This is the estimated prospective power for a future study.

4. For sample size calculations with the goal of achieving 80% power, given a range of effect magnitudes, increase the number of participants and/or items until you have 80% power.

We illustrate this procedure next. In order to generate fake data from a 2×2 repeated measures design with a Latin square, we first define a function, `gen_fake_norm2x2`.

```
library(MASS)
## assumes that no. of subjects and no. of items is divisible by 4.
gen_fake_norm2x2<-function(nitem=24,
                           nsubj=40,
                           beta=c(660,102,16,-48),
                           Sigma_u=Sigma_u, # subject vcov matrix
                           Sigma_w=Sigma_w, # item vcov matrix
                           sigma_e=359){
```

```
## prepare data frame for four condition latin square:
g1<-data.frame(item=1:nitem,
               cond=rep(letters[1:4],nitem/4))
g2<-data.frame(item=1:nitem,
               cond=rep(letters[c(2,3,4,1)],nitem/4))
g3<-data.frame(item=1:nitem,
               cond=rep(letters[c(3,4,1,2)],nitem/4))
g4<-data.frame(item=1:nitem,
               cond=rep(letters[c(4,1,2,3)],nitem/4))



## assemble data frame:
gp1<-g1[rep(seq_len(nrow(g1)),
            nsubj/4),]
gp2<-g2[rep(seq_len(nrow(g2)),
            nsubj/4),]
gp3<-g3[rep(seq_len(nrow(g3)),
            nsubj/4),]
gp4<-g4[rep(seq_len(nrow(g4)),
            nsubj/4),]
fakedat<-rbind(gp1,gp2,gp3,gp4)


## add subjects:
fakedat$subj<-rep(1:nsubj,each=nitem)


## add contrast coding:
## main effect 1:
```

```
fakedat$c1<-ifelse(fakedat$cond%in%c("a","b"),-1/2,1/2)
## main effect 2:
fakedat$c2<-ifelse(fakedat$cond%in%c("a","c"),-1/2,1/2)
## interaction:
fakedat$c3<-ifelse(fakedat$cond%in%c("a","d"),-1/2,1/2)


## subject random effects:
u<-mvrnorm(n=length(unique(fakedat$subj)),
           mu=c(0,0,0,0),Sigma=Sigma_u)


## item random effects
w<-mvrnorm(n=length(unique(fakedat$item)),
           mu=c(0,0,0,0),Sigma=Sigma_w)


## generate data row by row:
N<-dim(fakedat)[1]
rt<-rep(NA,N)
for(i in 1:N){
  rt[i] <- rnorm(1,beta[1] +
                   u[fakedat[i,]$subj,1] +
                   w[fakedat[i,]$item,1] +
                   (beta[2]+u[fakedat[i,]$subj,2]+
                      w[fakedat[i,]$item,2])*fakedat$c1[i]+
                   (beta[3]+u[fakedat[i,]$subj,3]+
                      w[fakedat[i,]$item,3])*fakedat$c2[i]+
                   (beta[4]+u[fakedat[i,]$subj,4]+
                      w[fakedat[i,]$item,4])*fakedat$c3[i],
```

```
                  sigma_e)
  }
  fakedat$rt<-rt
  fakedat$subj<-factor(fakedat$subj)
  fakedat$item<-factor(fakedat$item)
  fakedat}
```

Then, we fit a linear mixed model to the Levy and Keller Expt 1 data to obtain estimates of all the variance components and fixed effects. These will then be used for the power analysis.

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: region7 ~ dat + adj + int + (dat + adj + int | subj) + (dat +
##     adj + int | item)
##    Data: reading_time_nozeros
##
## REML criterion at convergence: 9731.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.9898 -0.5546 -0.1336  0.3812  5.7260
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  subj     (Intercept)  77493   278.38
##           dat          10042   100.21    0.49
##           adj           1216    34.88   -0.37 -0.03
##           int           7450    86.31   -0.89 -0.28  0.74
```

```
##  item     (Intercept)  11119   105.45
##           dat           9707    98.52    0.13
##           adj           4554    67.48    0.33 -0.82
##           int           4782    69.15   -0.84  0.14 -0.30
##  Residual             129189   359.43
## Number of obs: 660, groups:  subj, 28; item, 24
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   660.57      58.54  11.284
## dat           102.32      39.34   2.601
## adj            16.35      31.90   0.512
## int           -48.39      35.36  -1.368
##
## Correlation of Fixed Effects:
##     (Intr) dat    adj
## dat  0.233
## adj -0.017 -0.184
## int -0.494 -0.033  0.013
```

Next, we set the parameters for the fake-data simulation using the above model's results.

```
## set true parameter values:
(beta<-round(summary(m)$coefficients[,1]))

## (Intercept)         dat         adj         int
##         661         102          16         -48
```

```
(sigma_e<-round(attr(VarCorr(m),"sc")))

## [1] 359

(subj_ranefsd<-round(attr(VarCorr(m)$subj,"stddev")))

## (Intercept)         dat         adj         int
##         278         100          35          86

(subj_ranefcorr<-round(attr(VarCorr(m)$subj,"corr"),1))

##             (Intercept)  dat  adj  int
## (Intercept)         1.0  0.5 -0.4 -0.9
## dat                 0.5  1.0  0.0 -0.3
## adj                -0.4  0.0  1.0  0.7
## int                -0.9 -0.3  0.7  1.0

## choose some intermediate values for correlations:
(corr_matrix<-(diag(4) + matrix(rep(1,16),ncol=4))/2)

##      [,1] [,2] [,3] [,4]
## [1,]  1.0  0.5  0.5  0.5
## [2,]  0.5  1.0  0.5  0.5
## [3,]  0.5  0.5  1.0  0.5
## [4,]  0.5  0.5  0.5  1.0

## assemble variance matrix for subjects:
Sigma_u<-SIN::sdcor2cov(stddev=subj_ranefsd,corr=corr_matrix)

(item_ranefsd<-round(attr(VarCorr(m)$item,"stddev")))
```

```
## (Intercept)         dat         adj         int
##         105          99          67          69

## assemble variance matrix for items:
Sigma_w<-SIN::sdcor2cov(stddev=item_ranefsd,corr=corr_matrix)
```

Finally, we simulate data 100 times to compute power, for a range of effect magnitudes (30, 50, and 80 ms), 28 participants and 24 items.

```
set.seed(4321)
nsim<-100
## effect size ranging from 30 to 80 ms:
(beta2<-c(30,50,80))

## [1] 30 50 80

tvalsc1<-tvalsc2<-tvalsc3<-matrix(rep(NA,nsim*length(beta2)),ncol=nsim)
failed<-matrix(rep(0,nsim*length(beta2)),ncol=nsim)
for(j in 1:length(beta2)){
for(i in 1:nsim){
  beta[2]<-beta2[j]
  dat<-gen_fake_norm2x2(nitem=24,
                        nsubj=28,
                       beta=beta,
                       Sigma_u=Sigma_u,
                       Sigma_w=Sigma_w,
                      sigma_e=sigma_e)

## no correlations estimated to avoid convergence problems:
```

```
m<-lmer(rt ~ c1+c2+c3 + (c1+c2+c3||subj) +
          (c1+c2+c3||item), data=dat)


## ignore failed trials
if(any( grepl("failed to converge", m@optinfo$conv$lme4$messages) )){
  failed[j,i]<-1
} else{
tvalsc1[j,i]<-summary(m)$coefficients[2,3]
tvalsc2[j,i]<-summary(m)$coefficients[3,3]
tvalsc3[j,i]<-summary(m)$coefficients[4,3]
}
}
}
## proportion of convergence failures:
rowMeans(failed)

## [1] 0.00 0.00 0.02
```

Then, we estimate power for each effect magnitude:

```
pow<-rep(NA,length(beta2))
for(k in 1:length(beta2)){
  pow[k]<-mean(abs(tvalsc1[k,])>2,na.rm=TRUE)
}
```

The power estimates for different effect sizes are shown in Table B2.

We can also use this method to show how much power one gains if one log-transforms the data:

Sample size estimation for increasing power. We can also compute

	effect	power
1	30	11
2	50	28
3	80	51

Table B2
Power estimates for different effect sizes, using estimates from the linear mixed model fit to the LK13 Experiment 1 data.

approximately how much power we would have with a sample size of 40 participants and 24 items, and an effect of 50 or 80 ms for the main effect of Dative (other effects can be investigated similarly).

```
nsim<-100
## effect size
beta2<-c(50,80)
failed<-tvalsc1<-tvalsc2<-
  tvalsc3<-
  matrix(rep(NA,nsim*length(beta2)),ncol=nsim)
failed<-matrix(rep(0,nsim*length(beta2)),ncol=nsim)
for(j in 1:length(beta2)){
for(i in 1:nsim){
  beta[2]<-beta2[j]
  dat<-gen_fake_norm2x2(nitem=24,
                        nsubj=40,
                       beta=beta,
                      Sigma_u=Sigma_u,
                      Sigma_w=Sigma_w,
                      sigma_e=sigma_e)

## no correlations estimated to avoid convergence problems:
```

```
m<-lmer(rt ~ c1+c2+c3 + (c1+c2+c3||subj) +
          (c1+c2+c3||item), data=dat)


## ignore failed trials
if(any( grepl("failed to converge", m@optinfo$conv$lme4$messages) )){
  failed[j,i]<-1
} else{
tvalsc1[j,i]<-summary(m)$coefficients[2,3]
tvalsc2[j,i]<-summary(m)$coefficients[3,3]
tvalsc3[j,i]<-summary(m)$coefficients[4,3]
}
}
}
## proportion of convergence failures:
rowMeans(failed)

## [1] 0 0
```

Power estimates for the above cases:

```
pow50<-mean(abs(tvalsc1[1,])>2,na.rm=TRUE)
round(100*pow50,digits=0)

## [1] 20

pow80<-mean(abs(tvalsc1[2,])>2,na.rm=TRUE)
round(100*pow80,digits=0)

## [1] 66
```

Thus, with 40 participants and 24 items, for an effect of 50 ms, we would only have 20% power, whereas for an effect of 80 ms, power would be 66%. The above approach can be adapted for different designs.

Increasing number of items in a power analysis. In some cases, doubling the number of items without increasing the number of participants can also increase power substantially. For the present example, the power calculation for an increased number of items can be done quite easily. Here, we increase the number of items to 48, holding the number of participants unchanged at 40. We again consider two effect sizes: 50 and 80 ms:

```
nsim<-100
## effect size
beta2<-c(50,80)
failed<-tvalsc1<-tvalsc2<-
  tvalsc3<-
  matrix(rep(NA,nsim*length(beta2)),ncol=nsim)
failed<-matrix(rep(0,nsim*length(beta2)),ncol=nsim)
for(j in 1:length(beta2)){
for(i in 1:nsim){
  beta[2]<-beta2[j]
  dat<-gen_fake_norm2x2(nitem=48,
                        nsubj=40,
                       beta=beta,
                      Sigma_u=Sigma_u,
                      Sigma_w=Sigma_w,
                      sigma_e=sigma_e)

## no correlations estimated to avoid convergence problems:
m<-lmer(rt ~ c1+c2+c3 + (c1+c2+c3||subj) +
```

```
          (c1+c2+c3||item), data=dat)


## ignore failed trials
if(any( grepl("failed to converge", m@optinfo$conv$lme4$messages) )){
  failed[j,i]<-1
} else{
tvalsc1[j,i]<-summary(m)$coefficients[2,3]
tvalsc2[j,i]<-summary(m)$coefficients[3,3]
tvalsc3[j,i]<-summary(m)$coefficients[4,3]
}
}
}
## proportion of convergence failures:
rowMeans(failed)

## [1] 0 0
```

Power estimates for simulations with 48 items:

```
pow50<-mean(abs(tvalsc1[1,])>2,na.rm=TRUE)
round(100*pow50,digits=0)

## [1] 41

pow80<-mean(abs(tvalsc1[2,])>2,na.rm=TRUE)
round(100*pow80,digits=0)

## [1] 81
```

Two simpler approximations of power. When the items designed for an experiment are very uniform and do not induce much variability, it can happen that

item-level variance is much smaller than subject-level variance. In those situations, quick approximations of power can be done using the `power.t.test` function. By-subject standard deviation can be estimated from the data by aggregating over items, by condition. We do this below for LK Experiment 1:

```
dat<-reading_time_nozeros[,c(1,11,14)]
means<-with(dat,aggregate(region7,by=list(subj,condition),mean,na.rm=TRUE))
colnames(means)<-c("subj","cond","trt")
## by-condition standard deviations:
round(with(means,tapply(trt,cond,sd)))

##   a   b   c   d
## 299 386 338 258

## optimistically take the smallest sd:
power.t.test(d=80,n=40,sd=250,
             alternative="two.sided",
             type="one.sample",
             strict=TRUE)

##
##      One-sample t test power calculation
##
##               n = 40
##           delta = 80
##              sd = 250
##       sig.level = 0.05
##           power = 0.5056057
##     alternative = two.sided
```

This quick approximation of course ignores the different sources of variance that we explicitly took into account in the simulation-based example above, but it can be a reasonable way to get ballpark estimates.

A third approach is to use the `retrodesign` function provided by Gelman and Carlin (2014). The original function takes as input the effect estimate (e.g., from a previous study) and the standard error of the estimate, and returns the power, Type M and S error. Here, we use a simplified version, which we call `retrodesign_power`, that only returns the power estimate. This function is enough for computing power because the standard error of a fixed effect in a linear mixed model contains all the information needed to compute power.

```
## Adapted from Gelman and Carlin 2014:
retrodesign_power <- function(A, s, alpha=.05, df=Inf, n.sims=10000){
z <- qt(1-alpha/2, df)
p.hi <- 1 - pt(z-A/s, df)
p.lo <- pt(-z-A/s, df)
power <- p.hi + p.lo
round(power*100)
}
```

For example, in LK's Experiment 1, assuming that the effect of Dative is 80 ms with SE 39 (taken from the model fit). This gives us 54% power:

```
retrodesign_power(A=80,s=39)

## [1] 54
```

In other words, the linear mixed model-based power analysis with 40 participants and 24 items gives a power estimate of 66%; power.t.test gives an estimate of 51%; and the retropower function gives an estimate of 54%. These are all quite similar estimates.

Appendix C

Estimates from 12 reading studies on facilitation effects, and model predictions

Figure C1 shows the 95% credible intervals for ten agreement attraction studies that were part of the meta-analysis in Jäger et al. (2017); and two recently published studies on semantic plausibility effects (Cunnings & Sturt, 2018). The number agreement 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 (Engelmann et al., 2018) predicts a facilitation effect at the auxiliary or the following region when the noun preceding the auxiliary is *cabinets* vs. *cabinet.* The two semantic plausibility studies investigated by Cunnings and Sturt (2018) involved implausible sentences like *Sue remembered the letter that the butler with the cup/tie accidently shattered today in the dining room.* These are implausible because letters can't shatter. Here, theory predicts a facilitation effect at *shattered* due to misretrieval of the non-subject *cup* (vs. *tie*) (details are discussed in Engelmann et al., 2018). In the number agreement experiments, study 1 is the ungrammatical agreement data from Experiment 1 of Dillon et al. (2013); studies 2-5 are the experiments reported in Lago et al. (2015), and 6-10 are from Wagers et al. (2009). Studies 11 and 12 are from Cunnings and Sturt (2018). 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 facilitation effect to milliseconds. Our estimates may be slightly different from the original published estimates in some cases 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 89 ms. Again, our interest here is not in whether effects were significant or not significant—only one of these 10 studies would show a significant effect if a p-value were to be computed. 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.

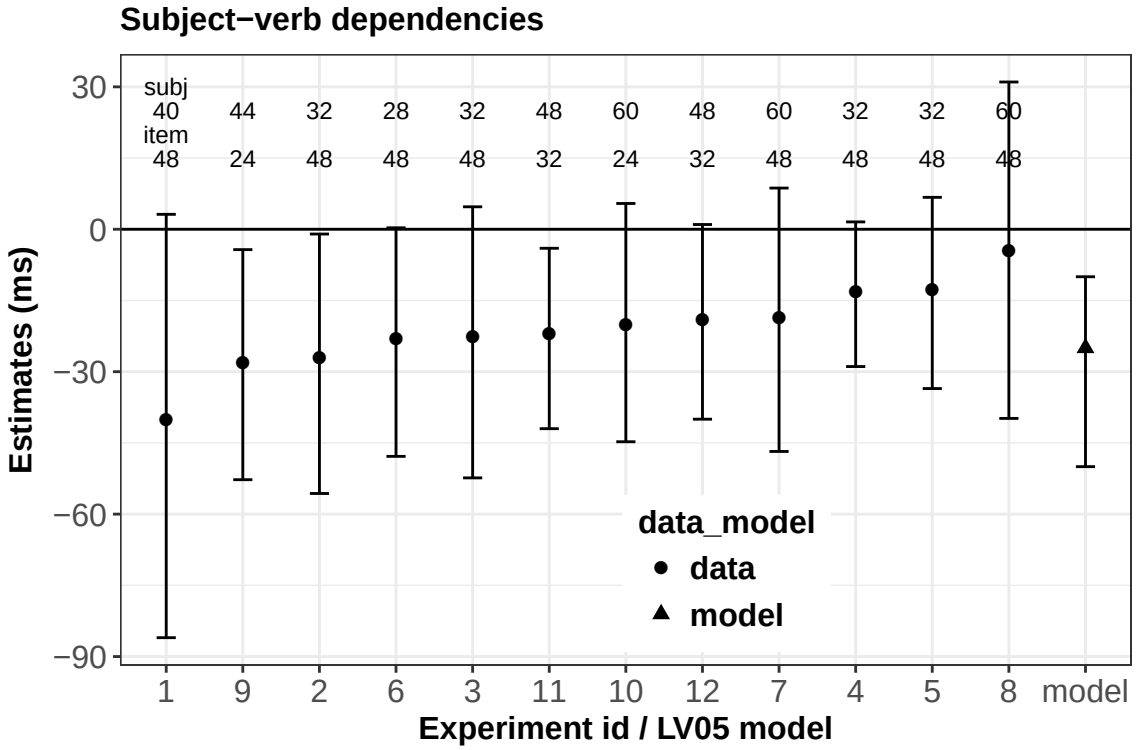

Figure C1. The means and 95% credible intervals of the predicted facilitation effect from 10 published studies on subject-verb dependencies with number agreement (Dillon, Mishler, Sloggett, & Phillips, 2013; Lago, Shalom, Sigman, Lau, & Phillips, 2015; Wagers, Lau, & Phillips, 2009), and two studies on subject-verb dependencies with a semantic plausibility manipulation (Cunnings & Sturt, 2018), and model predictions from the Lewis & Vasishth (2005) model for these configurations. Also shown are the number of participants (subj) and the number of items (item) in each study.

Appendix D

Word length and frequency effects in the eyetracking data

Because we found almost no effects in the eyetracking studies, a legitimate concern is that there may have been a systemic problem in the data-collection. We therefore checked whether the well-known word length and word frequency effects on reading time (Kliegl, Nuthmann, & Engbert, 2006) can be seen in all the four eyetracking data sets. If word length and frequency effects cannot be found, then there would be something fundamentally wrong with the data. We extracted type-frequencies (occurrences of a type per million tokens) of all words occurring in a filler item from the dlexDB database (Heister et al., 2011), which is based on the reference corpus underlying the Digital Dictionary of the German Language (DWDS) (Klein & Geyken, 2016). We only investigated first-pass reading time. Linear mixed models were fit using `lme4` with centered log frequency and centered word length as predictors, with all variance components but without intercept-slope correlations for random effects. The results are shown in Table D1; there are clear effects of word length and frequency, in the expected directions. Thus, our data do have the basic characteristics of eyetracking data. Obviously, we cannot entirely rule out that there may be important systematic differences between the original studies and ours that could explain why only effects in the original work passed the statistical significance filter. But this is a limitation of any replication attempt.

ET Experiment	Predictor	Estimate	Std.Error	t-value
Expt 1 (LK13 Expt 1)	Freq	-3	1	-3
	Len	22	2	15
Expt 2 (LK13 Expt 2)	Freq	-2	1	-3
	Len	22	1	16
Expt 3 (LoadxDist n=28)	Freq	-4	1	-3
	Len	21	1	16
Expt 4 (LoadxDist n=100)	Freq	-2	1	-2
	Len	24	1	17

Table D1

The effect of centered word frequency and centered word length on first-pass reading times in the four eyetracking studies. Experiment 1 is the replication attempt of LK's Experiment 1; Experiment 2 is the replication attempt of LK's Experiment 2; and Experiments 3 and 4 are small and larger-sample experiments investigating the Load-Distance interaction.