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

July 19, 2018

Author Note

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

Abstract

It is well-known in statistics (e.g., Gelman & Carlin, 2014) that treating a result as publishable just because the p-value is less than 0.05 leads to overoptimistic expectations of replicability. These effects get published, leading to an overconfident belief in replicability. We demonstrate the adverse consequences of this statistical significance filter by conducting seven direct replication attempts (268 participants in total) of a recent paper (Levy & Keller, 2013). 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; the latter generally yields a less noisy estimate but also a smaller effect magnitude, which looks less compelling but is more realistic. We reiterate several suggestions from the methodology literature for improving current practices.

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

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, it has been argued that this routine attribution of certainty to noisy data is a major contributor to the current replication crisis in psychology and other sciences (Open Science Collaboration, 2015; Amrhein, Korner-Nievergelt, & Roth, 2017). For recent examples from psycholinguistics of replication difficulties, see Nieuwland et al. (2018), and Kochari and Flecken (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 one well-known adverse consequence of the statistical significance filter (Gelman, 2018; Lane & Dunlap, 1978), that it leads to findings that are positively biased. We want to stress that none of the statistical points made in this paper are new (for similar arguments, see Hedges, 1984; Button et al., 2013; Dumas-Mallet, Button, Boraud, Gonon, & Munafò, 2017; Goodman, 1992; Ioannidis, 2008; Frank et al., 2017, among others). However, we feel it is necessary to demonstrate through direct replication attempts why significance yields no useful information when statistical power is low. The fact that underpowered studies continue to be treated as informative suggests that such a demonstration is needed.

We assume here that the reader is familiar with the null hypothesis significance testing (NHST) procedure as it is used in psychology today. NHST can work well when power is relatively high. But when power is low, published studies that show statistical significance will have exaggerated estimates (see Appendix A for a formal argument). The effect of low power is demonstrated in Figure 1 using simulated data: for a low-power scenario, the estimates from repeated samples fluctuate wildly 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 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 1 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 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 decrease.

Many researchers, such as Cohen (1962), 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? In medical statistics, prospective power analyses are quite common; not so in psycholinguistics. Suppose that we were to follow this practice from medical statistics and conduct a prospective power analysis based on the effect sizes reported in the literature. Gelman and Carlin (2014), and many others before them, have pointed out

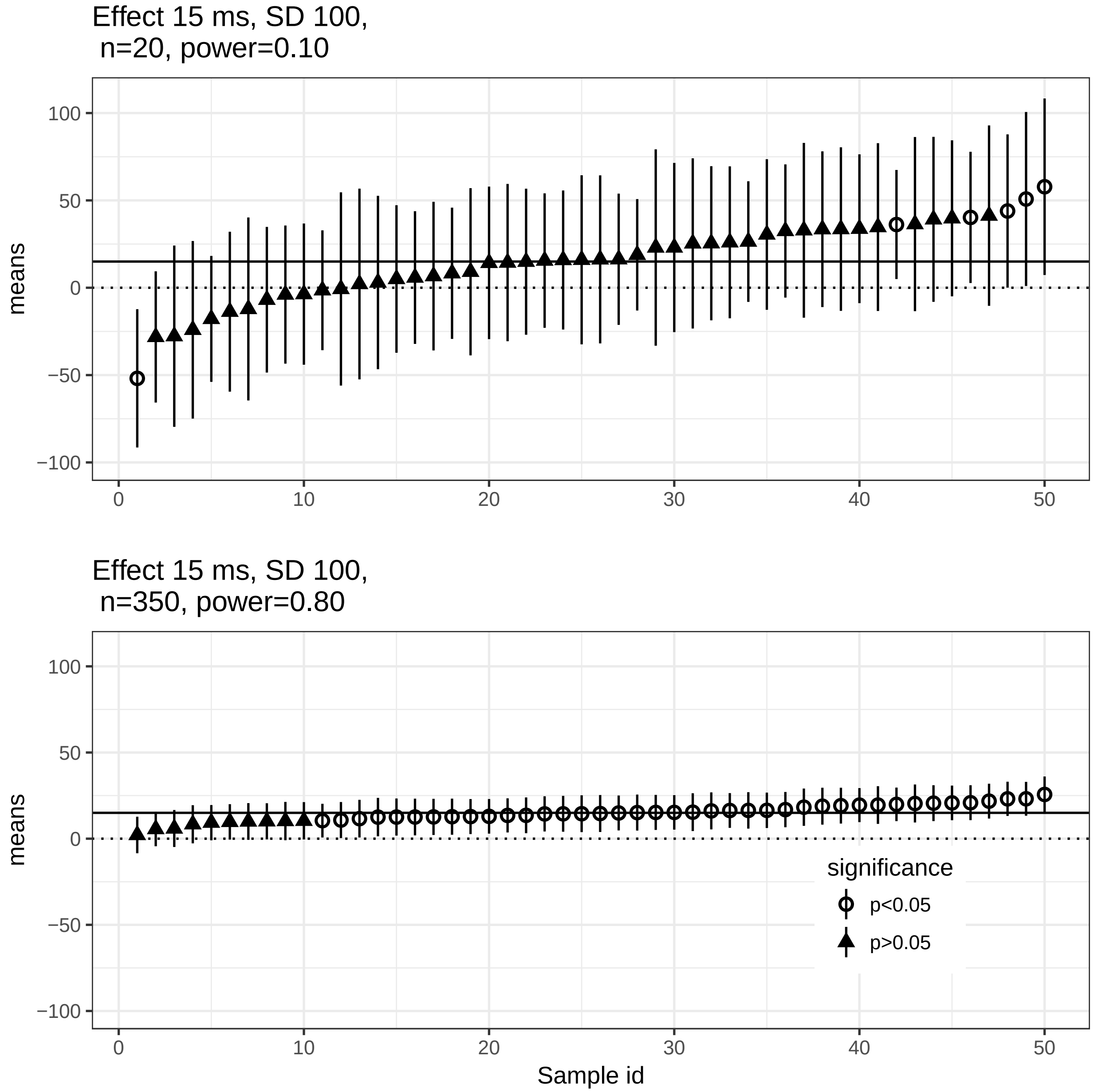

Figure 1. A demonstration of Type M error using simulated data. We assume that the data are generated from a normal distribution with mean 15 ms and standard deviation 100. The true mean is shown in each plot as a solid horizontal line. When power is low, under repeated sampling, whenever the estimates 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.

that this can lead to an interesting problem. 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 will fill the literature. If we base the power analysis on the published literature, we would conclude that the effects are large. A formal power analysis based on such exaggerated estimates is bound to yield an overestimate of power, and we can incorrectly convince ourselves that we have an appropriately powered study.

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.

Although the above observations about power and replications are well-known in statistics and psychology (see the discussion in Wasserstein & Lazar, 2016; Chambers, 2017), 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 overestimate the true effect and are 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 should 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
The insight, that the friend the client the car from
Plastik] verkaufte,. . .
plastic sold,. . .
‘The insight that the friend sold the client the plastic car. . . ’

b. Die Einsicht, dass $[_{NOM}$ [der Freund] $[_{GEN}$ des Kunden]] $[_{ACC}$ das Auto
The insight, that the friend of the client the car
aus Plastik] verkaufte,. . .
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). The Lewis & Vasishth 2005 (LV05) model assumes that completing argument-verb dependencies is affected by similarity-based interference arising from distractor nouns in memory (a related model is by Van Dyke & McElree, 2006). Like the DLT, the LV05 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 treat the DLT and the cue-based retrieval theories as specific instantiations of 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. 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 2, 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]

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

One nice property of the LK design is that the verb position is always constant across conditions being compared: the intervening phrases (dative 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] LK showed in a corpus analysis (summarized in their Table 1 Levy & Keller, 2013, p. 204) that if the dative NP or both the dative NP and the adjunct phrase appeared in the main clause, the main clause verb phrase (the verb phrase that the verb *versteckt*, 'hidden') heads had lower surprisal values. 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 sentences 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 sentences 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.'

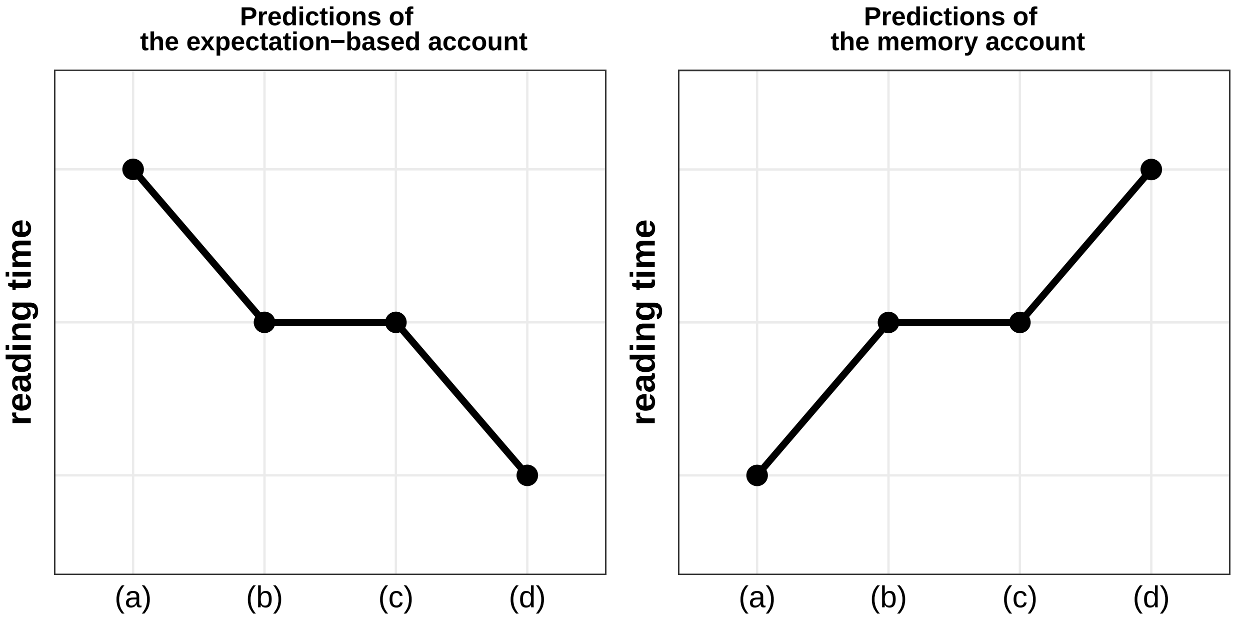

Figure 2. Predictions for the Levy and Keller Experiments 1 and 2: The left panel shows the speedup predicted by the expectation account. The right panel shows the slowdown predicted by memory-based accounts. This figure is based on Figure 1 of Levy and Keller (2013).

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 (R Core Team, 2018). 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. Briefly, our main interest here is in quantifying uncertainty about the parameter estimates of interest. We elaborate on this point next.

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 likelihood function and the priors). Technically, this cannot be done with 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., in physics, the speed of light in a vacuum), but is a fictional construct in areas like psychology and psycholinguistics. For example, there exists no single number representing the increase in reading time in object vs. subject relatives in English. The Bayesian approach allows us to focus on the uncertainty of the estimates of interest. A further, although more peripheral, advantage is that we can always fit so-called "maximal" models with full covariance matrices for by-participant and by-item variance components (Schielzeth & Forstmeier, 2009; 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 priors next.

Prior specification in Bayesian models. In the Bayesian approach, it is common to use so-called mildly and weakly informative priors that have a regularizing effect on the posteriors.[3] A weakly informative prior allows a wide range of plausible values; regularizing means that we downweight extreme parameter 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 3. When the nu (ν) parameter in the built-in Stan function for an LKJ prior is less than 1, intermediate values are downweighted; such a situation is the opposite of what we mean here by regularizing priors. When ν is higher than 1, extreme values are downweighted. Regularizing priors are also 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 (2015) 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.[4] This equal-tailed interval demarcates the range over which

[3] Ideally, one should use a "community of priors" to conduct an analysis, so that all opinions on a topic are taken into account. This approach is used sometimes in areas like medicine (Spiegelhalter, Abrams, & Myles, 2004).

[4] Kruschke (2015) uses highest posterior density intervals. As Kruschke (2015, p. 87) 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." 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.

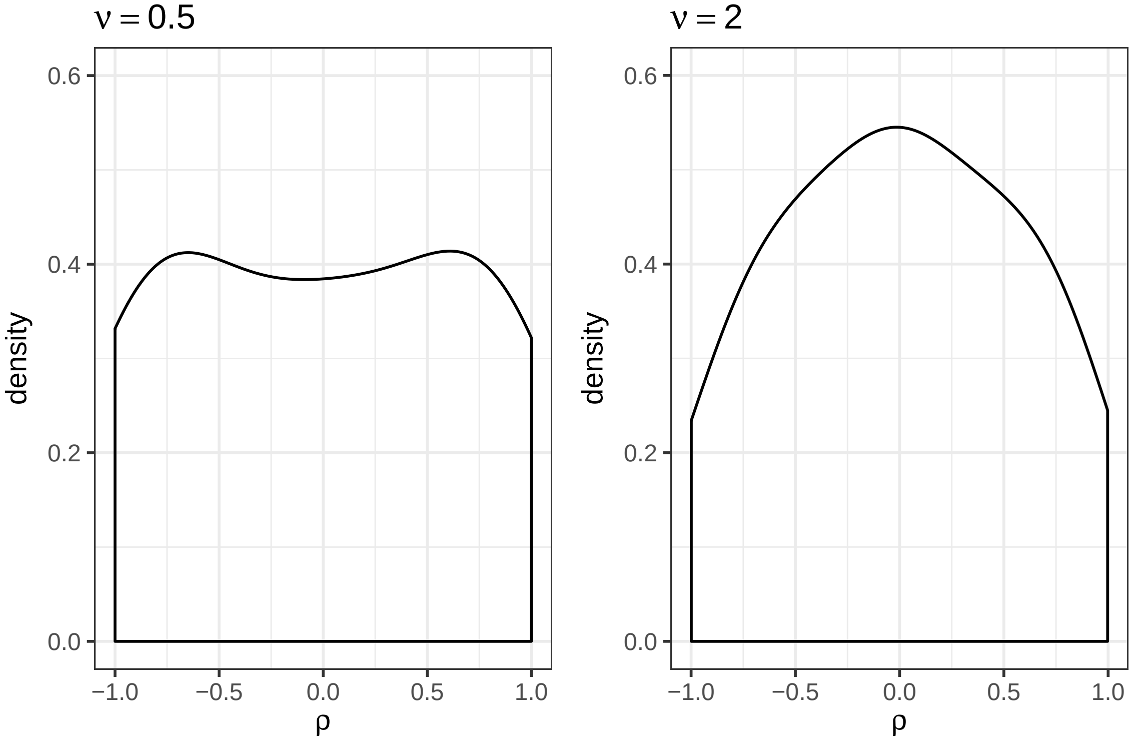

Figure 3. Example showing two different prior distributions using LKJ priors on a correlation parameter ρ. When the ν (nu) parameter in the LKJ function is 2, this downweights extreme values such as ±1. The LKJ(2) prior can be used to define priors for arbitrarily large correlation matrices, not just for a single correlation parameter.

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, in the NHST way of thinking, a parameter in the frequentist paradigm is an unknown point value. Once one assumes that a parameter can only have a fixed but unknown point value, the parameter cannot have a probability distribution associated with it. All estimation in the frequentist paradigm is done with reference to the sampling distribution of the estimator, $\hat{\mu}$, which is a function that gives us the sample mean for a particular data set. The estimate of the standard error of the sample mean (SE) is a function of the estimated standard deviation $\hat{\sigma}$ and the sample size n: $SE = \hat{\sigma}/\sqrt{n}$, is quantifying the uncertainty of the distribution of the sample mean returned by the estimator $\hat{\mu}$ under hypothetical repeated sampling. Given a data set consisting of a vector of data points x, the sample mean $\bar{x}$ serves as an estimate of the unknown point value μ. The confidence interval is then computed as $\bar{x} \pm 2 \times SE$. Thus, a *particular* 95% confidence interval from a single data set either contains the true, unknown μ or it doesn't. Second, the meaning of the confidence interval is so convoluted that it is difficult to understand or communicate: 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 anyway. For further discussion of confidence intervals, see Hoekstra, Morey, Rouder, and Wagenmakers (2014), Kruschke (2015). In typical psycholinguistic experiments, the confidence interval and the Bayesian credible interval will look very similar (for examples, see Bates, Kliegl, Vasishth, & Baayen, 2015); in some cases, the Bayesian interval may be slightly wider than the confidence interval. As a consequence of this rough equivalence, the confidence interval is often interpreted as if it were a Bayesian credible interval, even though this is technically incorrect. In the past, due to limitations of software for carrying out Bayesian analyses, the confidence interval was much easier to compute than the Bayesian credible interval, so equating Bayesian and frequentist intervals was arguably a reasonable approximation. However, today, packages like `brms` (Bürkner, 2017, in press) make Bayesian linear mixed models relatively easy to fit, and so it is now quite straightforward to compute Bayesian credible intervals.

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 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 https://osf.io/eyphj/. 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 (except the parameters of the correlation matrix for the random effects), the prior distribution was defined as the standard normal distribution, $Normal(0,1)$; 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 warm-up samples and were discarded. Convergence was checked by visually inspecting the chains and via the R-hat convergence diagnostic (Gelman et al., 2014).

The posterior distributions for the main effects and interaction on the log scale were back-transformed to posterior distributions in milliseconds. This was done as follows. Suppose that the fixed effects part of the model is defined as:

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

with the effects coded as ±0.5. Then, we can obtain the posterior distribution of the difference in means between the two levels of Dative by computing: $exp(\beta_0 + \beta_1 \times 0.5) - exp(\beta_0 - \beta_1 \times 0.5)$. This computation is done within Stan, taking as input each of the posterior samples of $\beta_0, \dots \beta_3$, and returning as output the posterior distribution of the difference in means on the raw ms scale. This transformation of the posterior distributions from the log scale to the ms scale allows us to compute credible intervals on the raw scale as well. Analogous calculations were done for the other factors.

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.[5]

Limiting the dependent measure to total reading times (in both our reanalyses of LK's original studies, and in the analyses of our replication attempts) 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 Dat(ive), Adj(unct), and their interaction, DatxAdj) 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).

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 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 4. 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 4, an analysis of total reading times suggests the following:

1. In Experiment 1, at the critical region, the mean of the posterior for the effect of

[5] 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.

Dative is 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 claim that memory load modulates whether expectation effects are observed is compelling given theory and existing data.

Although the claimed effects are compelling 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 ranges from 16 to 153 ms; and the evidence for the second conclusion comes from an estimate with mean 82 ms, with a credible interval ranging from 19 to 146 ms. These wide uncertainties imply that values as small as, for example, 20 ms are also plausible.

There is good reason to believe that reading time effects relating to memory-based retrieval may be closer to 20 ms than 80 ms. 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 posterior mean 13 ms. Similarly, recently published estimates of facilitation in reading time (total reading time) due to memory misretrieval are approximately -20 ms, with credible intervals ranging approximately from -1 to -40 ms (Cunnings & Sturt, 2018). 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. In this scenario, power for 28 participants and 24 items would be about 13 to 41% (see Appendix B for full details). Because of Type M error, with 28 participants it would be essentially 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 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 interval of the original estimate. We will consider both possible ways to interpret a replication attempt.

Experiments 1-4

We conducted two self-paced reading and two eyetracking studies; the correspondence to the original LK experiments is as shown in Table 4.

Our experiment	Original experiment	participants	items
Expt 1 (SPR)	LK Expt 1	28	24
Expt 2 (ET)	LK Expt 1	28	24
Expt 3 (SPR)	LK Expt 2	28	24
Expt 4 (ET)	LK Expt 2	28	24

Table 4
The correspondence between our experiments and those of Levy & Keller (2013).

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 https://osf.io/eyphj/.

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 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 https://osf.io/eyphj/.

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

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

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

sampling rate of 500 Hz using Eyetrack software[8] for the experimental presentation.

In LK's Experiment 1, the materials were presented in a non-monospaced font (Times New Roman, font size 20), whereas in their Experiment 2 the materials were presented in a monospaced font (Lucida Console, font size 14). The position on the screen of the critical verb differed in their two experiments: In LK's Experiment 1, the critical verb appeared in the middle of either the third or fourth line of the presented text, whereas in their Experiment 2 the critical verb was always the fourth word of the fourth line.

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, 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 Experiment 1-4

Question-response accuracies. The question-response accuracies for Experiments 1 and 2 were 66 and 64, respectively, and for Experiments 3 and 4 they were 61% and 60%, respectively. These are comparable to LK's 69% and 65% in their

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

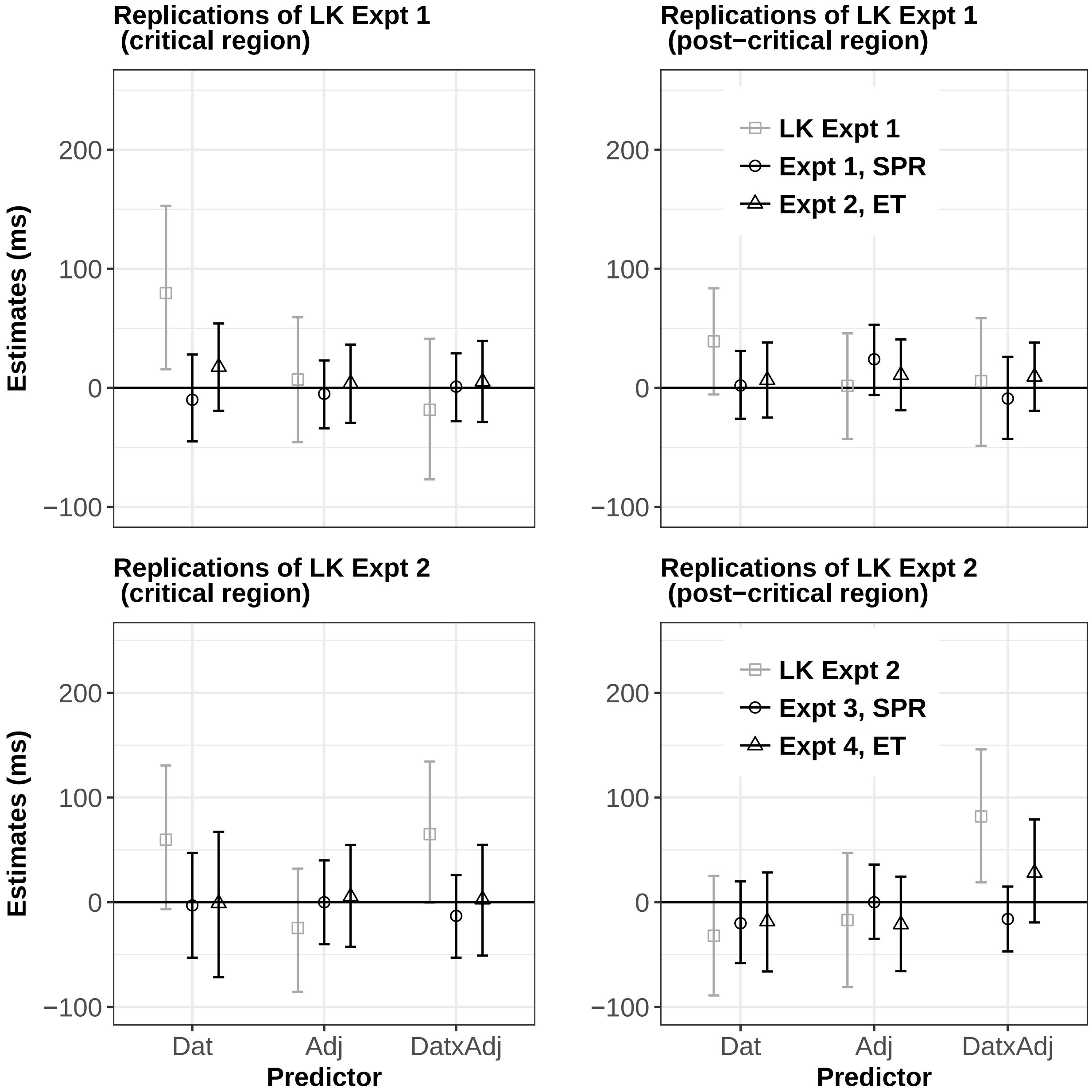

Figure 4. The effects of Dat(ive) and Adj(unct) interposition (and their interaction, DatxAdj) at the critical and post-critical regions. Shown are the mean and 95% credible intervals from the original LK Experiments 1 and 2, and the two replication attempts. SPR stands for self-paced reading, and ET stands for eyetracking.

Experiments 1 and 2, respectively.

Reading time results. Figure 4 summarizes the results of our four experiments. As mentioned earlier, we only report the analyses of total reading time data.[9]

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 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:

- Expt 1 (SPR replication of LK Expt 1): Effect of Dative in critical region -10 ms [-45,28].
- Expt 2 (Eyetracking replication of LK Expt 1): Effect of Dative in critical region 18 ms [-19,54].
- Expt 3 (SPR replication of LK Expt 2): Interaction of Dative and Adjunct in post-critical region -16 ms [-47,15].
- Expt 4 (Eyetracking replication of LK 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

[9] In our data, we also analyzed all the dependent measures (critical and post-critical regions) in which LK found statistical significance in their data. 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 dependent measures came out statistically significant in our data.

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 their Experiment 2, but the claim to be investigated involves the patterns seen across their Experiments 1 and 2, and this was not checked. We evaluate this claimed interaction next.

Table 5

Example sentences (simplified) for investigating the Load-Distance interaction by combining the 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 5 for the design. This analysis tested for the main effects of Load, Distance, and their interaction. As shown in Table 6, 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 6
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 5, 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.

Experiments 5, 6: Replication attempts of the Load-Distance interaction

We carried out two attempts to reproduce the Load-Distance interaction. As discussed above, we designed the experiment to pit Load and Distance against each other by taking conditions (c) and (d) of the original LK Experiment 1 (which we will refer to as the low memory load conditions) and conditions (c) and (d) of Experiment 2 (high memory load conditions). We conducted a self-paced reading study and an eyetracking study, each with the same sample size as the original experiments (28 participants, 24 items). The procedure was as described for the preceding studies.

As shown in Figure 5, both replication attempts showed that the estimate for Load in the critical region had a positive sign:

- Expt 5 (SPR): 76 ms [42,111]
- Expt 6 (ET, total reading times) 152 ms [104,200].

These effects suggest that increasing load (the relative clause conditions (c) and (d) in Table 5) 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 under the LK account.

The estimate for Distance is close to 0 ms: 3 ms [-30,39] in Expt 5; and 1 ms [-38,39] in Expt 6. Finally, the interaction between Load and Distance is not far from 0 ms; 5 ms [-30,43] in Expt 5, and -14 ms [-48,20] in Expt 6.

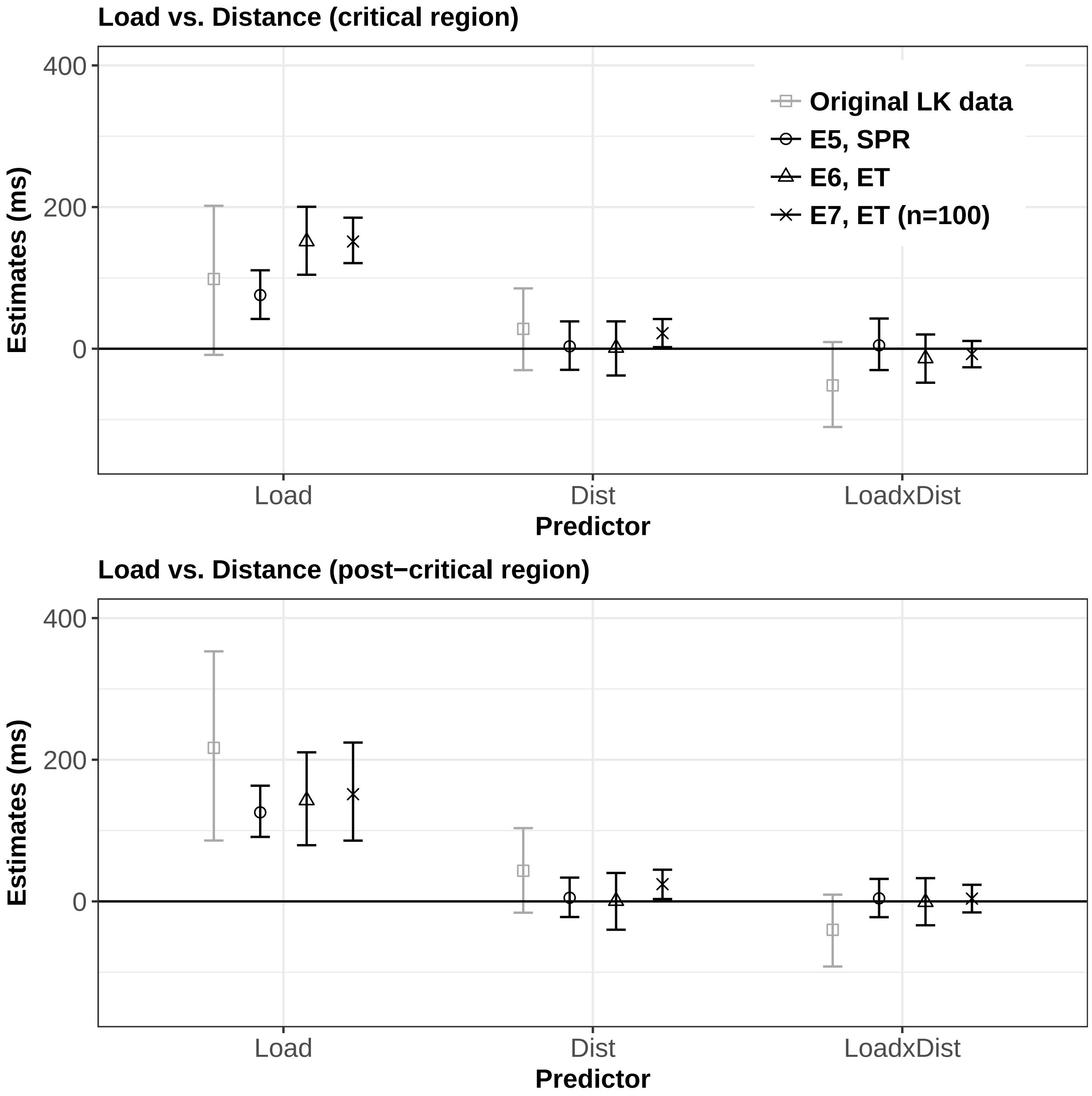

Figure 5. Load and Distance effects at the critical and post-critical regions. Shown are the mean and 95% credible intervals from conditions (c) and (d) of the two original LK Experiments 1 and 2; and from our three replication attempts (Expts 5-7). SPR stands for self-paced reading, and ET stands for eyetracking.

An interesting question arises here. If we were to run the experiment with a larger sample size, would we perhaps detect the Load-Distance interaction? After all, the interaction claimed by LK is very well-motivated both theoretically and empirically. We turn to this larger-sample study next.

Expt 7 (Eyetracking): A larger-sample replication attempt of the Load-Distance interaction

Before we discuss the results of Experiment 7, we first explain how we decided on sample size. We used an approach that Kruschke (2015) refers to as the region of practical equivalence (ROPE). Below, 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 incrementally 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" as a region rather than a point value. 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 (Cornfield, 1966; Armitage, 1989; Spiegelhalter et al., 2004; Berry, Carlin, Lee, & Muller, 2010; Freedman, Lowe, & Macaskill, 1984; Spiegelhalter, Freedman, & Parmar, 1994). Kruschke (2015) re-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. We will use Kruschke's terminology here.

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, what we count as "no effect" is a range of possible values: -20 to 20 ms. This range can be seen as representing a 95% credible interval over a distribution (say a normal distribution with mean 0) of plausible values. 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.[10] 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 1) showed 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 Experiment 1 (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) found 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 likely to be 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%

[10] 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. The essential point here is to use the information available to arrive at reasonable estimates; this is not very different from the elicitation of expert opinion in Bayesian data analysis of clinical data (O'Hagan et al., 2006; Morris, Oakley, & Crowe, 2014).

credible intervals $[2, 28]$ ms. Since these estimates were based on SPR data and first-pass reading times (FPRT), it is reasonable to assume that in total reading times (TRT) the effect of interference would be larger; from experience with eyetracking data, we can say that the effect is approximately twice as large, i.e., 30 ms (TRTs are a sum of FPRT and re-reading times, so they are bound to be larger than FPRT). Given these assumptions, for TRT we fixed ±20 ms around 0 ms as counting as effectively a null effect for the LK studies.

Our estimates of the region of practical equivalence (ROPE) are based on an 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 phenomena for which no data exist, some initial experiments could be used to establish a ROPE, and/or quantitative predictions from a computational process model could be used as a guide (an example is discussed in Engelmann et al., 2018).

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/eyphj/). 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.

For interpreting the results, the ROPE method can be used as follows. As shown in Figure 6 (adapted from Spiegelhalter et al., 2004, p. 184), once the data with the pre-determined precision have been collected, there are five possible scenarios. For illustration purposes, we assume that a positive sign on a parameter validates some theory X, and a negative sign validates a competing theory Y. A concrete example of such opposing predictions is the expectation vs. locality question discussed by LK in their paper.

The five outcome scenarios are as follows:

- A, B: data's credible interval falls clearly outside the null region. Decision: reject the null region, and conclude that theory X or Y is validated (depending on the sign).
- C,D: data's credible interval overlaps with the null region. Decision: if the sign is positive, reject theory Y; if the sign is negative, reject theory X.
- E: data's credible interval falls within the null region. Decision: conclude that the data are consistent with "no effect." Note that we do not say here that the decision is that we have "proved" that the null is true, but merely that the data are consistent with the posited ROPE. Only direct replications can establish whether an estimate and its 95% credible interval consistently falls within the ROPE.

Incidentally, the ROPE method can also be used for affirming a theory's predictions, if the theory makes quantitative predictions. A stringent test of a theory's predictions would be that the posterior's credible interval falls within the range predicted by 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 of equivalence. But this is no worse than the way NHST is used; subjective decisions are routinely taken in NHST and there are no fixed standards for these (Chambers, 2017). Another obvious objection is that the ROPE approach can be misused. For example, one could first run the study and compute the standard error and then retroactively define the null region as four times the estimated 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 the same in NHST, where a prospective power calculation must be done before conducting the study.

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.

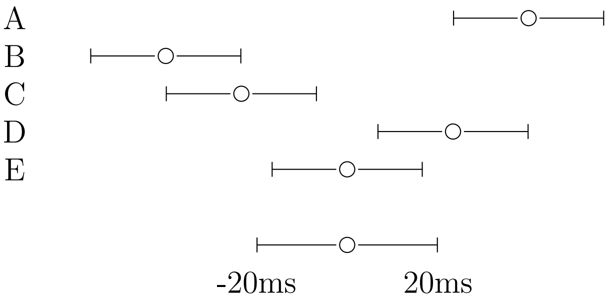

Figure 6. The five possible outcomes when using the null region or "region of practical equivalence" method for decision-making (Kruschke, 2015). Outcome A supports a theory X that predicts a positive sign on the parameter; B supports a theory Y that predicts a negative sign. Outcome C rejects theory X, and D rejects Y; and the estimate E is consistent with the null region. The width of the null region here is 40 ms, but would depend on the dependent measure, and the measurement precision achievable by the measurement instrument.

We now turn to Experiment 7, in which we investigated the Load-Distance interaction with a larger sample.

Results of Experiment 7. The estimates are summarized in Figure 5. 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 Experiment 7, 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 high-load conditions is compatible with the LK argument in their paper that locality effects outweigh

[11] At first glance, it may be surprising that in the post-critical region, the 95% credible interval for the effect of Load in Experiment 7 is as wide as that of Experiment 6, which had 28 participants. One might expect that a larger-sample study always yields a narrower credible interval. But this need not necessarily be true in a particular sample; the credible interval is dependent on the estimates of the variance components, which will vary from study to study.

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.

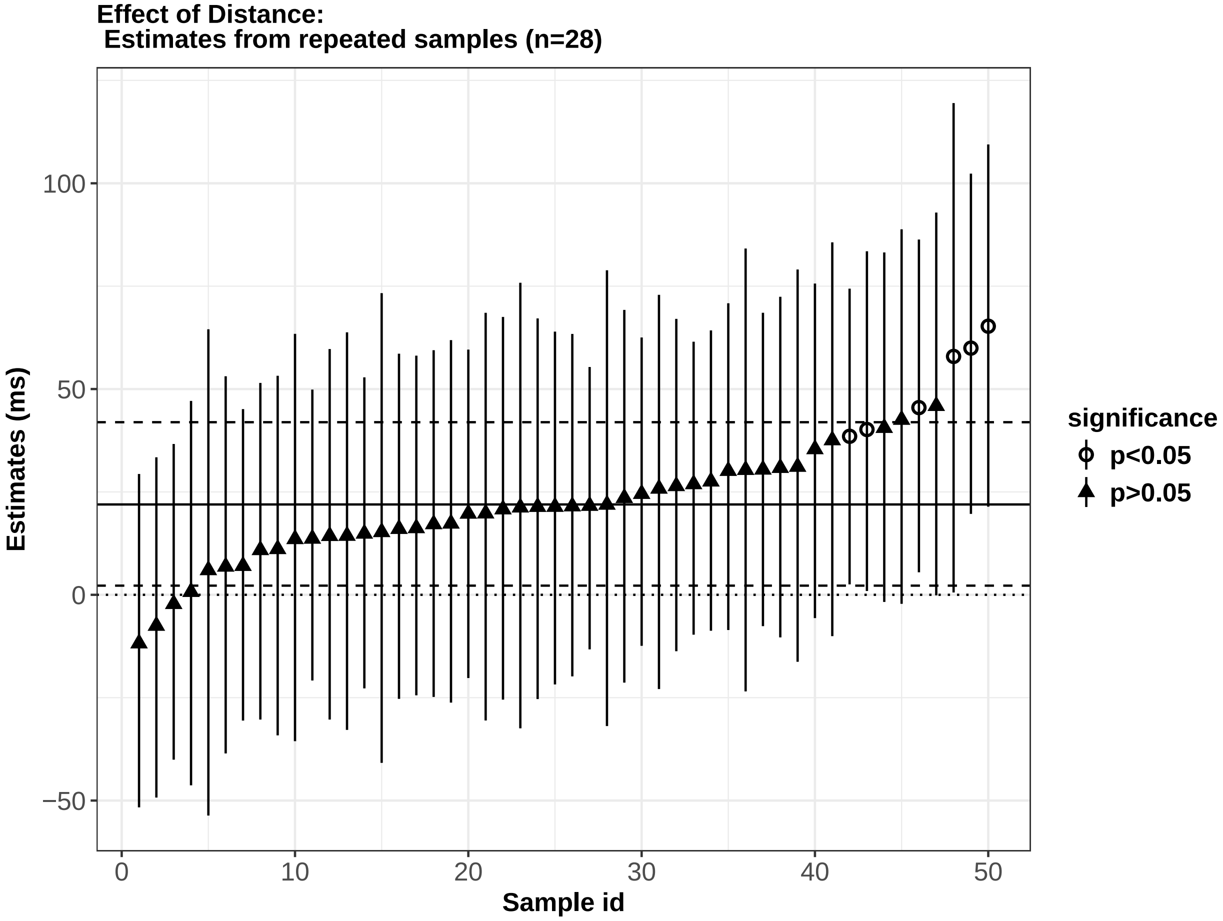

Figure 7. A demonstration of the fluctuation in the estimates for the effect of Distance when we choose 28 participants pseudo-randomly from the 100-participant experiment. The solid horizontal line is the estimated mean from the 100-participant data set, and the broken lines show the corresponding 95% credible intervals. The points show the means and 95% credible intervals when randomly sampling from the 100-participant data set.

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

General Discussion

Experiment 1-6 showed that the statistically significant (or nearly-significant) effects found in Levy and Keller (2013) are noisy enough that a broad range of possible outcomes—including no effect—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. 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. One cannot definitively conclude much from the original studies and our replication attempts. 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.

Experiment 7 showed that a larger-sample study generally delivers narrower credible intervals than the 28-participant studies. It also delivers a smaller estimate of the posterior mean for the Load-Distance interaction compared to the original study; the larger-sample estimate is probably more realistic. Experiment 7 suggests that the key claim of a Load-Distance interaction in LK's original experiments may be consistent with no effect. One interesting suggestion from this 100-participant study is that the locality effect that is predicted by account such as the Dependency Locality Theory (Gibson, 2000) may have some weak support. Since this is, to our knowledge, the first time that any evidence for locality has 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 these and other head-final languages consistently show locality. An important line of research would be to attempt to replicate the published results for head-final languages like German, Hindi, and Persian, and to investigate other head-final languages like Japanese and Korean. If the LK experiment design is followed up on in future work, it would be advisable to choose simpler sentences than the ones LK used; comprehension accuracy needs to be better than in the studies discussed in the present paper.

A legitimate concern at this point is that most of the effects investigated in our seven experiments showed results consistent with no effect. Could it be that there is something fundamentally wrong with our experimental methodology? In order to address this worry, we checked whether we could recover well-known word length and frequency effects from the filler items in the four eyetracking experiments. The details are discussed in Appendix D, but briefly, all four experiments show the expected effects. Thus, the methodology does not seem to have any fundamental problems. Of course, with null results one cannot be certain that no effect is present, especially when power is low. In future work, other labs should attempt to replicate LK's and our reported estimates of effects.

We return now to the point that the significant effects in LK's experiments are too noisy to interpret using statistical significance. Noisiness is not a property that is unique to the LK study considered here. Reading studies on other well-established effects also have issues similar 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 is around −26 ms for most parameter configurations; if the model parameters are varied over a narrow range, the predicted facilitation varies from approximately −10 to −57 ms.[12] Several studies have been published showing statistically

[12] Engelmann et al. (2018) provide a detailed investigation of the range of predictions that the model makes for the facilitatory interference discussed here. They varied the latency factor F, the noise parameter ANS, the maximum associative strength MAS, the mismatch penalty MP, and the retrieval threshold θ, and computed model predictions for the range of parameter combinations given by $F \in \{0.01, 0.02, \ldots, 0.6\}, ANS \in \{0.1, 0.2, 0.3\}, MAS \in \{1, 2, 3, 4\}, MP \in \{0, 1, 2\}, \theta \in \{-2, -1.5, \ldots, 0\}$. In order to derive approximate upper and lower bounds of model predictions, we take the median together with the first and third quartiles of Engelmann et al.'s simulated facilitatory interference effects (6000 iterations for each parameter combination), with the difference that we only used parameter configurations with a latency factor of 0.05 to 0.6. The

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; in one study, total reading time from 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 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 -57 ms), in the sense that the credible intervals from these 10 studies overlap with the theoretically predicted range. But these data are not strongly consistent with the Lewis and Vasishth model predictions. If these estimates had been more precise (i.e., had much narrower credible intervals) and their 95% credible intervals had fallen within the predicted range, this would have been a stronger validation of the model's predictions. With such wide credible intervals in the data, a broad range of outcomes is compatible with the data, including effectively no facilitatory effect at all. Thus, even in the relatively clear agreement attraction case, in future work 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, effects reported in the literature are *guaranteed* to be overestimates when power is low. 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

calculations can be reproduced using the Shiny app: https://engelmann.shinyapps.io/inter-act/.

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. There is in principle no harm in publishing low-powered studies in top journals, as long as strong claims are avoided. This is what statisticians mean when they suggest that researchers “accept uncertainty and embrace variation” (McShane, Gal, Gelman, Robert, & Tackett, 2017). Currently, in psycholinguistics and other areas, we are taught to have the expectation that every experiment be a “win.” Under this prior belief in routine success, even null results from low-powered studies start to look informative.

It is of course possible to publish more informative studies by simply running higher-power experiments. But 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, ±20 ms around 0 ms) and to run the experiment until the desired precision was 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, Vasishth, et al. (2018) self-paced reading study had a 95% credible interval of 20 ms. Note that the goal here should not be to find a credible 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.[13] 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.

[13] Doing hypothesis testing with Bayes factors would lead to similar problems unless one can specify a fully generative model (Gelman et al., 2014).

Once we have fixed the region of practical equivalence, we effectively also fix the precision (the 95% credible interval) that is theoretically meaningful to us. Now we can run the experiment until we reach this desired level of precision. 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 the properties of data from imagined repeated sampling. The Bayesian framework rather obeys the likelihood principle, which states that all the information from the data is contained in the likelihood function (Gelman et al., 2014; Lee, 2012). We can therefore check the precision of our estimates while running the experiment, and stop the experiment when the desired precision is reached (as opposed to the desired effect becoming significant).

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 summarize the literature in narrative reviews, and 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 that 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-powered studies should be treated as inconclusive rather than proving that the null hypothesis is true; unfortunately, this detail is not widely appreciated. In sum, simple vote-counting would be highly misleading when power is low and publication bias exists.

Many researchers have pointed out that we should aim for higher-precision estimates and focus on estimation rather than only focusing on statistical significance (e.g., Claridge-Chang & Assam, 2016; Greenland et al., 2016). Focusing on estimation will allow for better-quality meta-analyses and better quantitative model comparisons of competing computational models. The first comprehensive quantitative evaluation of the computational memory-retrieval model of Lewis and Vasishth (2005) involved comparing model predictions to estimates from 77 published results on retrieval processes (Engelmann et al., 2018). This evaluation 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. However, the results of the quantitative evaluation by Engelmann et al. (2018) would have been more informative if the estimates from the published individual studies had had higher precision.

In addition to fixing precision in advance, a second suggestion (Chambers, 2017) 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 Chambers (2017) explains, pre-registration involves defining in advance the analysis that is planned and depositing this in an embargoed repository like OSF (osf.io) or aspredicted.org. OSF time-stamps the pre-registration, which serves as a transparent way to demonstrate that the analysis plan was defined before the data were collected. Pre-registering will also minimize problems like p-hacking, HARKing (hypothesizing after the results are known), and the garden-of-forking paths problem (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, but pre-registration is a valuable tool that clearly separates the prior analysis plan from the exploratory part (De Groot, 1956/2014). 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, so 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 (Vasishth et al., 2013). 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, driven exclusively by the search for significance (an example is discussed in Vasishth & Nicenboim, 2016). 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. These decisions are often not reported in the published paper. Pre-registration would remove these degrees of freedom and thereby ensure a clear separation between confirmatory and exploratory analyses (De Groot, 1956/2014).[14]

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 other cases the crucial information—the data itself—are not available. For example, Nieuwland et al. (2018) tried but failed to obtain the data for the published result (DeLong, Urbach, & Kutas, 2005) that they attempted to replicate. Many researchers have generously released their data to us in connection with the present and other replication attempts. But attempts to obtain data from published studies are often unsuccessful. Wicherts, Borsboom, Kats, and Molenaar (2006) report an attempt to obtain data from 141 articles from major psychology journals, which had a total of 249 experiments. Of these, 73% of the data were not released. Wicherts and colleagues report that this is approximately the same non-response rate as in 1962. The continued absence of reproducible code and data seriously harms cumulative progress in science; evidence synthesis (meta-analysis) needs accurate estimates from published papers. Our experience with meta-analysis (Nicenboim, Roettger, & Vasishth, 2018; Jäger et al., 2017) shows that published summaries are usually far from adequate because they often don't contain the minimal information (the estimated mean of the parameter of interest and the standard error) needed to conduct the meta-analysis. Sometimes the published analysis is incorrect and as a result the published statistics are unusable (some examples are discussed in Nicenboim, Roettger, & Vasishth, 2018). Obtaining and reanalyzing the original data (using the originally used code) is the most reliable way to obtain accurate estimates for evidence synthesis.

[14] 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. However, this would just be scientific fraud; pre-registration is not designed to solve that problem.

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 on 26 November 2015, and ended on 29 September 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. Clearly, expectations regarding publication rate need to change. In closing, a contribution of the present paper is to demonstrate through direct replication attempts the fact (well-known in some scientific communities) 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 in psycholinguistics 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) and Kochari and Flecken (2018) suggests that replicability problems arising from the statistical significance filter could run deep in psycholinguistics. 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. The reliability of published results may improve if we finally start to follow the best practices that have been advocated again and again by statisticians but which have been largely ignored by psychology and other areas: aim at higher precision, conduct direct, pre-registered replications, and release data and code. These changes will 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.

History has shown that any suggestions to improve power and to replicate results have generally not been adopted in psychology (see discussion in Lane & Dunlap, 1978). We nevertheless reiterate some proposals that many others have made in the past (e.g., Bakan, 1966; Amrhein, Trafimow, & Greenland, 2018; McShane et al., 2017; Chambers, 2017). Researchers should (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 aspredicted.org; 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. This needs to become the default.

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 in scope. Roger Levy provided detailed comments on the paper; we are very grateful for his advice. 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 period of nearly two years. We also thank Reinhold Kliegl, Sander Greenland, Steven Goodman, Christian Robert, Titus von der Malsburg, Daniel Schad, Sandra Hanne, Sol Lago, Jan Vanhove, Tatjana Scheffler, Michael C. Frank, João Veríssimo, Dario Paape, Bruno Nicenboim, Steven Goodman, Keith O'Rourke, Daniel Lakeland, and Martha Smith 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

Amrhein, V., Korner-Nievergelt, F., & Roth, T. (2017). The earth is flat (p> 0.05): Significance thresholds and the crisis of unreplicable research. *PeerJ*, *5*, e3544.

Amrhein, V., Trafimow, D., & Greenland, S. (2018). Abandon statistical inference. *PeerJ PrePrints*.

Armitage, P. (1989). Inference and decision in clinical trials. *Journal of Clinical Epidemiology*, *42*(4), 293–299.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, *66*(6), 423–437.

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

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

Berry, S. M., Carlin, B. P., Lee, J. J., & Muller, P. (2010). *Bayesian adaptive methods for clinical trials.* Boca Raton, FL: Chapman and Hall/CRC Press.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28.

Bürkner, P.-C. (in press). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal.*

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

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.

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

Claridge-Chang, A. & Assam, P. N. (2016). Estimation statistics should replace significance testing. *Nature Methods*, *13*(2), 108.

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

Cornfield, J. (1966). Sequential trials, sequential analysis and the likelihood principle. *The American Statistician*, *20*(2), 18–23.

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.

Dumas-Mallet, E., Button, K. S., Boraud, T., Gonon, F., & Munafò, M. R. (2017). Low statistical power in biomedical science: A review of three human research domains. *Royal Society Open Science*, *4*(2), 160254.

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.

Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., ... Levelt, C., et al. (2017). A collaborative approach to infant research: promoting reproducibility, best practices, and theory-building. *Infancy*, *22*(4), 421–435.

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., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd Ed.). Boca Raton, FL: Chapman and Hall/CRC Press.

Gelman, A. & Loken, E. (2016). The statistical crisis in science. In M. Pitici (Ed.), *The best writing on mathematics 2015* (pp. 305–318). Princeton, NJ: Princeton University Press.

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

Gibson, E. (2000). Dependency locality theory: A distance-based theory of linguistic complexity. In A. Marantz, Y. Miyashita, & W. O'Neil (Eds.), *Image, Language, Brain: Papers from the First Mind Articulation Project Symposium* (pp. 95–126). Cambridge, MA: MIT Press.

Goodman, S. N. (1992). A comment on replication, p-values and evidence. *Statistics in Medicine*, *11*(7), 875–879.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, p values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology*, *31*(4), 337–350.

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.

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

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.

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

Jäger, L. A., Engelmann, F., & Vasishth, S. (2017). Similarity-based interference in sentence comprehension: Literature review and Bayesian meta-analysis. *Journal of Memory and Language*, *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. (2015). *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.

Lee, P. M. (2012). *Bayesian statistics: An introduction*. Chichester, UK: John Wiley & Sons.

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: Chapman and Hall/CRC Press.

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2017). Abandon statistical significance. *arXiv preprint arXiv:1709.07588.*

Morris, D. E., Oakley, J. E., & Crowe, J. A. (2014). A web-based tool for eliciting probability distributions from experts. *Environmental Modelling & Software*, *52*, 1–4.

Nicenboim, B., Roettger, T. B., & Vasishth, S. (2018). Using meta-analysis for evidence synthesis: The case of incomplete neutralization in German. *Journal of Phonetics*. Accepted.

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*, *42*.

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.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., ... Rakow, T. (2006). *Uncertain judgements: Eliciting experts' probabilities*. Chichester, UK: John Wiley & Sons.

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*, *13*(6), e0198620.

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, UK: John Wiley & Sons.

R Core Team. (2018). *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*, 403.

Schielzeth, H. & Forstmeier, W. (2009). Conclusions beyond support: Overconfident estimates in mixed models. *Behavioral Ecology*, *20*(2), 416–420.

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.

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., Abrams, K. R., & Myles, J. P. (2004). *Bayesian approaches to clinical trials and health-care evaluation.* Chichester, UK: John Wiley & Sons.

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)*, *157*(3), 357–416.

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

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.

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist*, *61*(7), 726.

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 $\mu > 0$.[15] 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.

One informal definition of the p-value is the following: "A p-value is the probability under a specified statistical model that a statistical summary of the data (e.g., the sample mean difference between two compared groups) would be equal to or more extreme than its observed value." (Wasserstein & Lazar, 2016). 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:

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

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

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

Once we have observed a particular z-statistic z_p, the cdf $G_\delta(p)$ allows us to estimate power based on the z-statistic (Hoenig & Heisey, 2001). To estimate the p-value in the case where the null hypothesis is in fact true, let the true value be $\mu = 0$. It follows that $\delta = 0$. Then:

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

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%.[16]

If we now re-run the same study, collecting 36 data points each time, and impose the

[16] This can be confirmed by running the following command using R (R Core Team, 2018): `power.t.test(delta=0.1,sd=1,n=36,alternative = "one.sided",type="one.sample",strict=TRUE)`.

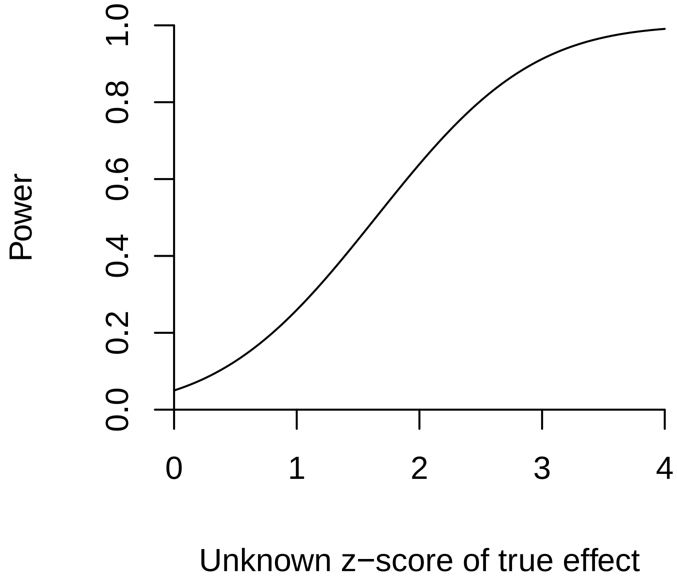

Figure A1. The relationship between power and the unknown z-score of the true effect. Larger z-scores are easier to publish due to the statistical significance filter, and these studies therefore give a mistaken impression of higher power.

condition that only statistically significant results with Type I error probability (α) 0.05 are published, then only observed z-scores larger than 1.64 (for a one-sided test) would be published and the power estimate based on these z-scores must have a lower bound of

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

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.[17] This incorrect belief among researchers has the effect that

[17] 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

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.

that, when the null is true, a p-value near 0 is no more surprising than a p-value near 0.05.

Appendix B

Prospective power analysis for repeated measures designs

We show here how power can be computed for data that are analyzed using linear mixed models, with crossed random effects for participants and items. 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 prospective power *for a future study* using a range of plausible effects. In other words, this is not intended to be a post-hoc power analysis; that would provide no new information beyond the p-value (Hoenig & Heisey, 2001).

When we do such a prospective 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 13 to 41%; see Table B1. If the true effect were as large as 80 ms (this is the estimate we obtained in the LK Experiment 1 for the effect of Dative), a sample size of 28 participants and 24 items would lead to approximately 75% power. If the true effect is even smaller than 30 ms, obtaining power greater than 80% would require hundreds of participants and more items.

Effect (ms)	Power (percentage)
30	13
50	41
80	75

Table B1
Estimates of prospective 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.

How exactly did we compute these power estimates? For a repeated measures design, one convenient way is to use fake-data simulation. As an illustration, we consider how we would compute prospective power for a future replication of 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 (hereafter, the parameter 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 the parameter estimates from the step above, 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 design, we first define a function, `gen_fake_lnorm2x2`; see Listing 1. This function generates log-normally distributed data because the dependent variable is reading time data, and this is often assumed to be generated from a log-normal distribution. We start by setting the number of participants and items to those used in the LK experiments (28 participants, 24 items).

Then, we fit a linear mixed model to the Levy and Keller Experiment 1 data (log-transformed) to obtain estimates of all the variance components and fixed effects. These estimates will then be used for the power analysis. See Listing 2 for the parameter estimates from the LK Experiment 1 data.

Next, we set the parameters for the fake-data simulation using the above model's results. See Listing 3.

```
library(MASS)
gen_fake_lnorm2x2 <- function(nitem=24,nsubj=28,
                         beta=NULL,
                         Sigma_u=NULL,Sigma_w=NULL,sigma_e=NULL){
  ## prepare data frame for four condition in a latin square design:
  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$item)),
            mu=c(0,0,0,0),Sigma=Sigma_w)
  ## generate data row by row:
  N<-dim(fakedat)[1]
  rt<-rep(NA,N)
  for(i in 1:N){
    rt[i] <- rlnorm(1,beta[1] +
                      u[fakedat[i,]$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}
```

Listing 1: Function for generating log-normally distributed fake data.

```
Linear mixed model fit by REML ['lmerMod']
Formula:
log(region7) ~ dat + adj + int + (dat + adj + int | subj) + (dat +
    adj + int | item)
   Data: reading_time_nozeros

REML criterion at convergence: 1079.8

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.70628 -0.66295  0.02944  0.62573  3.10073

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 subj     (Intercept) 0.147897 0.38457
          dat         0.016867 0.12987   0.28
          adj         0.007536 0.08681   0.15  0.88
          int         0.014489 0.12037  -0.47  0.69  0.77
 item     (Intercept) 0.025317 0.15911
          dat         0.033824 0.18391   0.11
          adj         0.013480 0.11610   0.33 -0.89
          int         0.019574 0.13991  -0.34  0.13 -0.16
 Residual             0.233152 0.48286
Number of obs: 660, groups:  subj, 28; item, 24

Fixed effects:
            Estimate Std. Error t value
(Intercept)  6.27331    0.08180  76.690
dat          0.14966    0.05855   2.556
adj          0.01296    0.04740   0.273
int         -0.03357    0.05245  -0.640

Correlation of Fixed Effects:
    (Intr) dat    adj
dat  0.132
adj  0.112 -0.159
int -0.257  0.170  0.069
```

Listing 2: Output of the linear mixed model fit to Levy and Keller's experiment 1 data.

```
# Extract parameter estimates:
beta<-round(summary(m)$coefficients[,1])
sigma_e<-round(attr(VarCorr(m),"sc"))

## assemble variance covariance matrix for subjects:
subj_ranefsd<-round(attr(VarCorr(m)$subj,"stddev"))
subj_ranefcorr<-round(attr(VarCorr(m)$subj,"corr"),1)
## choose some intermediate values for correlations:
corr_matrix<-(diag(4) + matrix(rep(1,16),ncol=4))/2
Sigma_u<-SIN::sdcor2cov(stddev=subj_ranefsd,corr=corr_matrix)

## assemble variance covariance matrix for items:
item_ranefsd<-round(attr(VarCorr(m)$item,"stddev"))
Sigma_w<-SIN::sdcor2cov(stddev=item_ranefsd,corr=corr_matrix)
```

Listing 3: Fix parameters for fake-data simulation based on the linear mixed model fit of LK Experiment 1.

Finally, we simulate data 100 times, for a range of effect magnitudes (30, 50, and 80 ms), 28 participants and 24 items, and estimate power for each effect magnitude. See Listing 4.

```
set.seed(4321)
nsim<-100
## effect size ranging from 30 to 80 ms on log scale:
## e.g.: exp(6.3+.056/2)-exp(6.3-.056/2)=30 ms
(beta2<-c(0.056,0.095,0.155))
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_lnorm2x2(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(log(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)
## estimate power:
pow<-rep(NA,length(beta2))
for(k in 1:length(beta2)){
  pow[k]<-mean(abs(tvalsc1[k,])>2,na.rm=TRUE)
}
```

Listing 4: Simulate data and compute power for different effect sizes.

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.

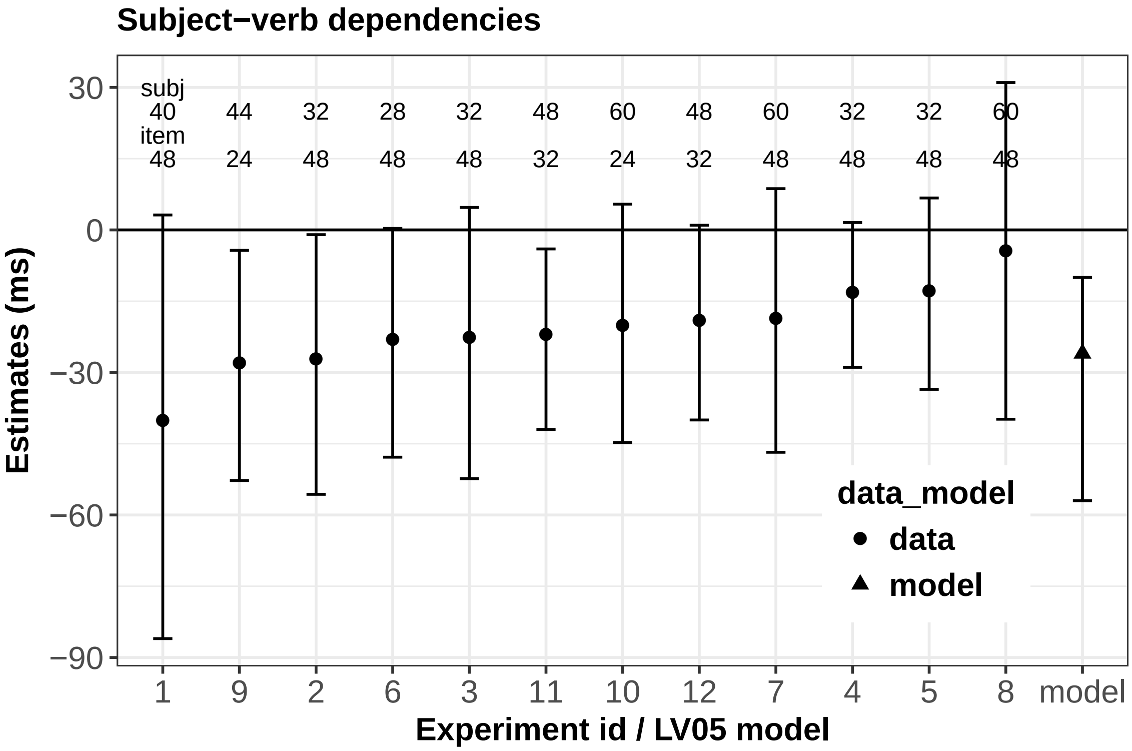

Figure C1. The means and 95% credible intervals of the predicted facilitation effect from 10 published studies on subject-verb dependencies with number agreement (Dillon, Mishler, Sloggett, & Phillips, 2013; Lago, Shalom, Sigman, Lau, & Phillips, 2015; Wagers, Lau, & Phillips, 2009), and two studies 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.