Exploratory and confirmatory analyses in sentence processing: A case study of number interference in German

Bruno Nicenboim
Department of Linguistics, University of Potsdam, Potsdam, Germany
e-mail:bruno.nicenboim@uni-potsdam.de

Shravan Vasishth
Department of Linguistics, University of Potsdam, Potsdam, Germany

Felix Engelmann
School of Psychological Sciences, University of Manchester, Manchester, UK

Katja Suckow
Department of German Studies, University of Göttingen, Göttingen, Germany

Draft of February 13, 2018.

Abstract

Given the replication crisis in cognitive science, it is important to consider what researchers need to do in order to report results that are reliable. We consider three changes in current practice that have the potential to deliver more realistic and robust claims. First, the planned experiment should be divided up into two stages, an exploratory stage and a confirmatory stage. This clear separation allows the researcher to check whether any results found in the exploratory stage are robust. The second change is to carry out adequately powered studies. We show that this is imperative if we want to obtain realistic estimates of effects in psycholinguistics. The third change is to use Bayesian data-analytic methods rather than frequentist ones; the Bayesian framework allows us to focus on the best estimates we can obtain of the effect, rather than rejecting a strawman null. As a case study, we investigate number interference effects in German. Number feature interference is predicted by cue-based retrieval models of sentence processing (Van Dyke & Lewis, 2003; Vasishth & Lewis, 2006), but has shown inconsistent results. We show that by implementing the three changes mentioned, suggestive evidence emerges that is consistent with the predicted number interference effects.

Keywords: exploratory and confirmatory analyses; sentence processing; Bayesian hierarchical modeling; cue-based retrieval; working memory; similarity-based interference; number interference; German

Introduction

From recent work it has become clear that many well-known results from psychology cannot be replicated (e.g., Open Science Collaboration, 2012). Low statistical power and inflated rates of Type I error have been identified as two major causes of non-replicability.

Low statistical power—a low probability of detecting an effect if one exists—has two bad consequences. One is that many null results, i.e., analyses showing a p-value greater than 0.05, will be found. The other consequence, not yet widely appreciated in cognitive science, is that observed effects that turn out to be statistically significant (i.e., the p-value is less than 0.05) tend to be exaggerated and can even have the incorrect sign. Gelman and Carlin (2014) call these Type M(agnitude) and Type S(ign) errors respectively. When low power experiments are run, exaggerated effects that have p-values less than 0.05 will tend to be published; these effects can be much larger than the true effect. Publication bias—publishing only those results that match theoretical claims—prunes away the effects that have the incorrect sign. Once the results with the expected sign are published, an expectation develops that subsequent experiments will also have similarly large effects (Vasishth & Gelman, 2017). As a consequence, smaller (but more realistic) effects, arrived at with higher power, are regarded with suspicion, even though it is the excessively large effects that should be suspect.

The second source of non-replicable effects, Type I error inflation, has many causes (Simmons, Nelson, & Simonsohn, 2011). One is the garden of forking paths (Gelman & Carlin, 2014) and the related concept of researcher degrees of freedom. Given a data-set, many different analytical paths can be taken, at least one of which may well lead to statistical significance. An example from a reading study in sentence processing is presented in Vasishth, Chen, Li, and Guo (2013): a t-test on aggregated data on the raw millisecond scale shows a significant effect, with a t-value of 2.63, but a linear mixed model with a full variance-covariance matrix for participants and items yields a non-significant effect with a t-value of 1.77. Yet more choices for the analysis of these data are discussed in Vasishth, Chen, et al. (2013). Depending on what one wants to claim, one could equally well argue for or against the effect of interest.

In reading studies in psycholinguistics, Type I error inflation occurs because we routinely explore different dependent variables (especially in eyetracking), and explore different regions of interest to check whether an effect appears anywhere at all. These analyses are usually presented as confirmatory, not exploratory. This approach has been criticized by statisticians because it introduces researcher degrees of freedom (Simmons et al., 2011). For a study exploring the consequences of Type I error inflation in eyetracking, see von der Malsburg and Angele (2017). Very similar problems occur in EEG and fMRI research.

A consequence of exercising researcher degrees of freedom is that the p-values reported in published papers in psycholinguistics (and other areas) are usually the result of exploratory and not confirmatory analyses. When the hypothesis tests are carried out after looking at the data, the hypothesis tests cannot be seen as confirmatory tests.

This work was supported by Minerva Foundation, Potsdam Graduate School, the University of Potsdam, and partly by Volkswagen Foundation grant 89 953 to Shravan Vasishth. Bruno Nicenboim was partly funded by the Deutsche Forschungsgemeinschaft grant VA 482/8-1 to Shravan Vasishth and Frank Roesler.

A solution: Separate exploratory and confirmatory analysis

The impulse to explore patterns in the data is completely understandable: flexibly analyzing different dependent measures or different regions can be very informative. However, as discussed above, we must be careful to treat our exploration as exploratory data analysis and not to treat these explorations as prespecified confirmatory tests. The English translation of the article by the psychologist De Groot (1956/2014) reveals that he had noticed this problem long ago, and he had proposed a solution that was never taken up in psychology and related areas:

> If the processing of empirically obtained material has in any way an "exploratory character", i.e. if the attempt to let the material speak leads to ad hoc decisions in terms of processing, ... then this precludes the exact interpretability of possible outcomes of statistical tests. ... One 'is allowed' to apply statistical tests in exploratory research, just as long as one realizes that they do not have evidential impact.

When conducting exploratory analyses, the central problem is that these hypotheses are generated after looking at the data. In eyetracking, if first-pass reading time does not show any expected effects, we look at regression path duration, or re-reading time or total reading time. If the critical region shows no effects, we check if we can find an effect in the critical and post-critical region(s) combined, or in the post-critical region(s) alone; alternatively, we can combine the pre-critical and critical region into one region. Eventually, one can usually find some constellation of dependent measure(s) and/or region of interest(s) which shows an effect. A similar problem holds in self-paced reading studies; what counts as the region to be analyzed is open to interpretation; what data to retain or remove varies from study to study, even when the studies are conducted by the same author.

So how can we freely explore our data to look for effects that we expect based on theory and still carry out valid statistical inference? As De Groot puts it:

> When ... research has such a mixed character, it is still possible to discriminate hypothesis testing parts from exploratory parts; it is also possible, in the text, to separate the discussion of the one type and the other. This is not only possible, this is also highly desirable. Testing and exploration have a different scientific value, they are grounded in different modes of thought, they lead to different certainties, they labor under different uncertainties.

We demonstrate the importance of De Groot's proposal, using as a case study number interference (Suckow & Van Gompel, 2012) in German, described below. In order to avoid the problems mentioned above with underpowered studies, we carry out two relatively high-powered experiments. Experiment 1 (82 participants) is exploratory in that it follows the usual convention of looking at several regions to find out where the effect would be found. Experiment 2 (100 participants) fixes the region to be analyzed a priori based on the exploratory analysis, and follows exactly the same analysis procedure as in Experiment 1. We detail the specific case study in the section entitled: *The research question: Number interference in German.*

Bayesian data analysis for statistical inference

In addition to demonstrating the De Groot approach of using an exploratory analysis to prepare for a confirmatory analysis, we also demonstrate the many advantages of using the Bayesian data analytical approach in psycholinguistics.

Advantages of Bayesian modeling. Bayesian modeling has three major advantages; see Nicenboim and Vasishth (2016) for other benefits of fitting Bayesian models in psycholinguistics.

First, Bayesian methods allow us to report credible intervals rather than confidence intervals. 95% credible intervals demarcate the range within which we can be certain with probability 0.95 that the true value of a parameter lies given the data at hand (Jaynes & Kempthorne, 1976; Hoekstra, Morey, Rouder, & Wagenmakers, 2014; Morey, Hoekstra, Rouder, Lee, & Wagenmakers, 2016). Thus, credible intervals allow us to quantify our uncertainty about the parameter of interest, given the data at hand. Contrast this with the frequentist 95% confidence interval, which depends on the properties of data that we *didn't* collect. The confidence interval has the interpretation that if we were to (counterfactually) carry out the experiment repeatedly across k hypothetical, identical repetitions and computed a 95% confidence interval each time, 95% of the k confidence intervals computed would contain the true parameter value. The one 95% confidence interval that we actually computed from the data we have at hand does not make any statement about the uncertainty of the parameter—indeed it cannot because in the frequentist paradigm the parameter is a point value that has no probability distribution associated with it.

Second, Bayesian procedures allow us to fit virtually any kind of distribution in a straightforward way. In the past, we have fit hierarchical mixture models (Nicenboim & Vasishth, 2017; Vasishth, Nicenboim, Chopin, & Ryder, 2017), and hierarchical measurement error models (Nicenboim, Roettger, & Vasishth, 2017; Vasishth, Beckman, Nicenboim, Li, & Kong, 2017). In this paper, we fit shifted lognormal mixed models, which lie outside the class of generalized linear models. For reading time data, such models can provide more accurate estimates than standard linear mixed models (see: Rouder, 2005; Rouder, Tuerlinckx, Speckman, Lu, & Gomez, 2008; Nicenboim, Logačev, Gattei, & Vasishth, 2016). The justification for assuming a shifted lognormal distribution for the reading times instead of a normal distribution (as in linear mixed models) is that residual reading times in self-paced reading are highly right skewed and have a lower bound greater than zero (i.e., the shift of the distribution). While sometimes reading times are log- or reciprocal transformed to avoid violations of normality assumptions (Box & Cox, 1964), these transformations still assume that the reading times are defined by their location (mean) and scale (standard deviation), and they are unshifted. Unshifted distributions for reading times in self-paced reading are unreasonable, since they do not take into account the fact that there is a minimal amount of time that takes to read a word and press a button on the keyboard, typically around 150–250 ms and that this may vary across participants (see Nicenboim et al., 2016, for self-paced reading; and Logan, 1992; Rouder, 2005; Rouder et al., 2008, for reaction time tasks). If distributions are shifted and analyzed as unshifted, estimates will be affected, and they may influence conclusions (Rouder, 2005), especially when dealing with small effect sizes, as it seems to be the case for the number interference effects we will investigate here (Jäger, Engelmann, & Vasishth, 2017).

The shifted lognormal distribution may help us to avoid anti-conservative conclusions, but mainly should produce more accurate estimates by fitting our data with a model that resembles the process that generates the data. The main disadvantages of this type of model are that it is not available out of the box (but we provide code to implement it[1]) and that it takes much longer to converge than a (generalized) linear mixed model (between one and two hours for our datasets).

Third, Bayesian methods allow us to fit fully hierarchical models without convergence issues with the so-called "maximal random effect structure" justified by the design (Schielzeth & Forstmeier, 2009; Barr, Levy, Scheepers, & Tily, 2013).[2] Fitting models with full variance covariance matrices gives us estimates of uncertainty about the parameters that take all potential sources of variance into account.

Interpreting the estimates from a Bayesian model. The most important information one obtains from a Bayesian model is the posterior distribution of the parameter of interest. We summarize the posterior distribution by presenting the posterior probability of the parameter being positive given the data, i.e., $P(\beta > 0)$, and its 95% credible interval (CrI). We also show graphically the 95% and 80% credible intervals, and posterior distributions of the estimates on the raw scale (milliseconds for the reading times and proportion for the comprehension accuracy). One way to interpret the results is as follows: If zero lies outside the 95% credible interval, we assume there is evidence for an effect; if zero is included within the interval but the probability of the parameter being greater than zero, $P(\beta > 0)$, (or less than zero depending on the expected sign) is relatively high, we assume that there is only weak evidence for an effect; and if the probability for a parameter being greater or less than zero is near 50%, we conclude that we found no evidence for an effect (see also Nicenboim & Vasishth, 2016). These sharp distinctions between evidence, weak evidence and no evidence are over-simplifications. In reality, the posterior distribution serves to quantify our uncertainty about an estimate of the effect; the estimate, along with its uncertainty, should be interpreted with reference to existing knowledge. We demonstrate this by comparing our posterior distributions with the best estimates of the number interference effect available at this time (Jäger et al., 2017).

The rest of the paper is structured as follows. We first describe the motivation for investigating our research question (number interference). Then, we carry out an exploratory analysis (Experiment 1) with a relatively large sample size (82 participants), looking at the critical and post-critical regions for an effect. Next, we re-run the same study with a larger sample size (100 participants), and examine only the regions where we found an effect in Experiment 1. Then, we obtain the most precise estimates possible given the data from the two experiments and discuss the implications of these estimates for the statistical power in previously published experiments. We conclude with a discussion of our findings.

The research question: Number interference in German

Many theories have been developed to explain dependency completion processes in parsing. One of these is broadly referred to as the cue-based retrieval account. This is actually a class of closely related theories (McElree, 2000; Van Dyke & Lewis, 2003; Lewis

[1] https://osf.io/mmr7s

[2] For a discussion on the advisability of fitting maximal models in a frequentist setting, see Matuschek, Kliegl, Vasishth, Baayen, and Bates (2017) and Bates, Kliegl, Vasishth, and Baayen (2015).

& Vasishth, 2005; Lewis, Vasishth, & Van Dyke, 2006; Van Dyke & McElree, 2006). One assumption common to these theories is that grammatical relations are created via the retrieval of necessary constituents from memory. Grammatical heads, such as verbs, provide retrieval cues (e.g., grammatical case, thematic role, semantic properties) that are used to identify the required constituent (i.e., the TARGET of the retrieval process) and to differentiate it from other constituents stored in memory (i.e., the COMPETITORS or DISTRACTORS). In sentence (1), for example, feature information such as NP, ANIMATE, and SUBJECT could be the retrieval cues that are used to find the target (the dependent, *the resident*).

(1) The worker was surprised that the resident who was living near the dangerous *warehouse/neighbor* was complaining about the investigation.

Van Dyke (2007) investigated the difficulty that readers have in identifying the target of the retrieval triggered at *was complaining* when multiple competitors are present. This work showed, inter alia, that in (1), question-response accuracy was lower and reading times at the verb *was complaining* were longer, when the noun *warehouse* within the preposition phrase *near the dangerous ...* was replaced by an animate noun such as *neighbor*.

Van Dyke explained these differences in reading times and accuracy in terms of SIMILARITY-BASED INTERFERENCE. The number of competitor nouns sharing the animacy feature with the target is higher in (1) when *the resident* and *the neighbor* precede the verb *was complaining*. This makes the target harder to retrieve compared to the case where only one animate noun (*the resident*) is present, leading to longer reading times at the verb, and lower question-response accuracy. This increase in processing difficulty in examples such as (1) has been referred to as INHIBITORY INTERFERENCE (Dillon, Mishler, Sloggett, & Phillips, 2013; Engelmann, 2016). This is in contrast to FACILITATORY INTERFERENCE (sometimes called intrusion; Drenhaus, Saddy, & Frisch, 2005), which refers to a facilitation that is observed in certain ungrammatical sentences (for details, see Jäger et al., 2017).

Much of the evidence that is consistent with inhibitory interference comes from configurations where similar syntactic and semantic features on multiple nouns make the target noun difficult to identify (Van Dyke & Lewis, 2003; Van Dyke, 2007; Van Dyke & McElree, 2006; Van Dyke & McElree, 2011; Van Dyke, Johns, & Kukona, 2014). Furthermore, the literature review and meta-analysis of Jäger et al. (2017) shows overall evidence for inhibitory interference for studies which examined interference of semantic and syntactic cues in grammatical sentences ($\hat{\beta} = 13$ ms, 95% CrI$= [2, 28]$). In contrast, research that included configurations with number feature interference in grammatical sentences, such as (2), has in general shown non-significant results, where cue-based retrieval accounts would also predict inhibitory effect (e.g., Wagers, Lau, & Phillips, 2009; Dillon et al., 2013; Lago, Shalom, Sigman, Lau, & Phillips, 2015). In *ungrammatical* sentences with a similar configuration, such as (3), however, a facilitatory effect is in general found (e.g., Wagers et al., 2009; Dillon et al., 2013; Lago et al., 2015). This facilitation is predicted by cue-based retrieval (for details, see Jäger et al., 2017; Engelmann, 2016; but see Wagers et al., 2009, for the discussion of other explanations for the facilitation). In addition, Jäger and colleagues' show that the meta-analytic estimate from studies investigating the number feature in ungrammatical sentences supports cue-based retrieval ($\hat{\beta} = -22$ ms, 95% CrI $= [-36, -9]$), while in grammatical sentences is inconclusive ($\hat{\beta} = -7$ ms, 95% CrI

$= [-16, 4]$). Wagers et al. (2009) hypothesized that it is possible that cue-based retrieval, at least using the number feature, is triggered only when a reanalysis is needed. Under Wagers and colleagues' account, cue-based retrieval would be a last resort strategy, which is used in the ungrammatical sentences, due to the lack of agreement between the subject and the verb.

To sum up, there is evidence for similarity-based interference consistent with cue-based retrieval for animacy or syntactic features (e.g., subjecthood) in grammatical sentences and for the number feature in ungrammatical sentences, but not for the number feature in grammatical sentences.

(2) a. The key$_{+sing}$ to the cabinets$_{+plur}$ is in the box.
b. The key$_{+sing}$ to the cabinet$_{+sing}$ is in the box.

(3) a. * The key$_{+sing}$ to the cabinet$_{+sing}$ are in the box.
b. * The key$_{+sing}$ to the cabinets$_{+plur}$ are in the box.

However, it may be the case that the inhibitory interference effect is also present for the number feature in grammatical sentences, but it is harder to detect. This could be because either (i) the weight of the number feature as a cue for retrieval might be smaller than other semantic or syntactic cues (for details about cue weights in cue-based retrieval, see Nicenboim et al., 2016), (ii) facilitatory interference has a different cause than inhibitory interference (for details, see Jäger et al., 2017; Engelmann, 2016), or (iii) there are other mechanisms at play that may counteract the effect of interference (e.g., feature percolation/movement: Bock & Miller, 1991; Pearlmutter, Garnsey, & Bock, 1999; and see Patson & Husband, 2016, for evidence of the faulty number representation of nouns in number interference configurations).

Regardless of the cause for the asymmetry between grammatical and ungrammatical sentences in number interference, and between the different features that cause interference, if number interference exists and has a small magnitude, a relatively high-power experiment must be run to detect such an effect; as Jäger et al. (2017) show in their Appendix B, experiments on interference usually have remarkably low power. Finding number interference effects in grammatical sentences would strengthen the evidence for cue-based retrieval as a general mechanism to build dependencies (as opposed to Wagers and colleagues' proposal that cue-based retrieval is a last resort strategy in number interference configurations) and was the motivation for the experiments presented below.

We turn to a description of the self-paced reading experiments next.

The items of the experiments included high and low interference conditions followed by one multiple choice question targeting one of the dependencies of the sentence. High and low interference conditions were created by manipulating the number feature of two intervening competitor NPs (*the assistants/s of the director/s*); while the target subject (for both verbs) remains singular (*The philanthropist*). In the high interference condition, the two competitors share the feature singular (*sg*) with the target (*The philanthropist*), while in the low interference condition the competitor NPs have, in contrast, the feature plural (*pl*).

(4) a. High Interference

Der Wohltäter, der den Assistenten
The.sg.nom philanthropist, who.sg.nom the.sg.acc assistant (of)
des Direktors **begrüßt hatte**, saß später im
the.sg.gen director **greeted had.sg**, sat.sg later in the
Spendenausschuss.
donations committee.

'The philanthropist, who had greeted the assistant of the director, sat later in the donations committee.'

b. Low Interference

Der Wohltäter, der die Assistenten
The.sg.nom philanthropist, who.sg.nom the.pl.acc assistant(s) (of)
der Direktoren **begrüßt hatte**, saß später im
the.pl.gen director(s) **greeted had.sg**, sat.sg later in the
Spendenausschuss.
donations committee.

'The philanthropist, who had greeted the assistants of the directors, sat later in the donations committee.'

This design should increase the magnitude of the interference effect by using two competitor nouns rather than one (a design adapted from the one in English presented in Suckow & Van Gompel, 2013). In addition, our design avoids the possible spillover confound in studies where the critical region immediately follows the last distractor (Wagers et al., 2009).[3]

To summarize, our case study aims to detect an interference effect predicted by cue-based retrieval accounts by carrying out a relatively high-powered experiment. Furthermore, we aim to avoid the problems of researcher degrees of freedom by running the experiment once and doing an exploratory analysis, and then doing the experiment again with new participants. In the second experiment, we only carry out a confirmatory analysis that attempts to reproduce the results of the exploratory analysis. In the next two sections, we present the experiments.

Experiment 1: Exploratory analysis

Method

Stimuli. The stimuli for the experiments (both exploratory and confirmatory phases) consisted of 60 items with two conditions in Latin square design as depicted in (4). After reading each experimental sentence, one question from the three possibilities in (5) was displayed. This means that six lists were generated (two conditions times three questions).

[3] In the Pearlmutter (2000) study, although two competitor nouns were used, one of them was placed immediately before the critical region confounding attraction with differential spillover effects. In addition, there at least two papers that had two competitors in language production experiments: Franck, Vigliocco, and Nicol (2002) and Gillespie and Pearlmutter (2011). It is unclear, however, what the extent of the overlap is between the processes that give rise to interference in production and comprehension; see also Tanner, Nicol, and Brehm (2014).

Given that we were interested in the number feature as a retrieval cue, we wanted to prevent the target of the dependency from being identified already at the participle verb (e.g., *begrüßt*, "greeted") before the inflected auxiliary verb. We designed our experimental items so that the three noun phrases in each sentence could in principle be the subject of both verbs (or the object of the first verb). Since we did not norm the sentences, there is the risk that in some of the experimental items one of the nouns would be a more likely subject (or object) than the others, rendering the number cue less (or not) important. This is especially important because such semantic cues always appear before the number cue in head-final constructions in German (at the participle verb, e.g., "greeted"). Semantic cues could have an overwhelmingly larger weight in comparison with number cues. This is a concern because semantic cues have been shown to modulate agreement attraction in production and comprehension (e.g., Thornton & MacDonald, 2003). However, since our statistical models are fully hierarchical, the degree in which semantic cues can help to identify the target of retrieval and attenuate number interference effects should be accounted by the by-items slope of the random effects: Our statistical models take into account that the plausibility of the agent (to which degree being a plausible "greeter" identifies the target of the retrieval in comparison to the competitors in the previous example) may attenuate the effect of interference in some of the items. See the Supplementary Material B for the complete stimuli and see Fig. A1 in the Appendix which shows the variability in the by-items slope.

Regarding the question types in our stimuli, they were built such that one targeted the subject of the matrix verb (*sat* in ex. 4) (MV), one targeted the subject of the embedded verb (*had greeted* in ex. 4) (EVS), and one the object of the embedded verb (EVO). The answers were provided in multiple-choice format as in (5), such that *(i) the philanthropist(s)* was the correct answer for the questions targeting both the subject of the matrix and the embedded verb (MV and EVS); and *(ii) the assistant(s)* was the correct answer for the question targeting the object of the embedded verb (EVO). Even though option *(iii) the director(s)* was never the answer for the experimental sentences, it was for some fillers with similar genitive constructions (*den Assistenten des Direktors*). These fillers had questions equivalent to *Wozu/Zu wem gehört der Assistent?* "Whose was the Assistant?". For all the questions participants had the option to answer *(iv) I don't know*, when they did not remember or could not answer. Only one question type appeared after each experimental sentence, and its answers appeared in random order.

We avoided Yes/No questions to encourage readers to pay attention to the targets of the dependencies and avoid good-enough processing (Ferreira, Bailey, & Ferraro, 2002). Given the evidence that readers may not try to achieve a fully specified representation of the sentences especially when they expect superficial comprehension questions (Swets, Desmet, Clifton, & Ferreira, 2008; Nicenboim, Vasishth, Gattei, Sigman, & Kliegl, 2015), questions that target all the dependencies ensure that participants would have to pay attention to the whole representation of the sentence and avoid shallow parsing or heuristics where they focus on only one dependency.

(5) QUESTION

a. (MV) Wer saß später im Spendenausschuss?
   Who sat in the donations committee?

b. (EVS) Wer hatte jemanden begrüßt?

Who had greeted someone?

c. (EVO) Wen hatte jemand begrüßt?

Whom had someone greeted?

(6) MULTIPLE-CHOICE OPTIONS

a. (i) der/die Wohltäter (MV-EVS); (ii) der/die Assistent/en (EVO); (iii) der/die Direktor/en; (iv) Ich weiß es nicht

(i) the philanthropist(s) (MV-EVS); (ii) the assistant(s) (EVO); (iii) the director(s); (iv) I don't know

The 60 experimental items were presented together with 110 fillers designed specifically for the experiments. The filler sentences presented included 30 items with subject relative clauses and a similar structure to the experimental items but with more variety of nouns (in gender and number); 30 items similar to the aforementioned fillers but with object relative clauses; and 50 items with two verbs but without relative clauses. The fillers included more variety of questions that also targeted other aspects of the sentences. We removed one item of the low interference condition because of a typo in the precritical region.

Participants. All participants reported to be native speakers of German and were naïve to the purpose of the study. We ran this first experiment with 84 subjects aged between 19–39 years old (mean 25.2 years) who were recruited using ORSEE (Greiner, 2004) at the University of Potsdam, Germany. The data from two participants were removed because of technical issues.

Procedure. Subjects were tested individually using a PC. Participants completed a moving window self-paced reading task (Just, Carpenter, & Woolley, 1982). In addition, participants performed tests to assess the individual differences in reading fluency, cognitive control, and working memory capacity. Given that the individual differences in interference were inconclusive, for ease of exposition, we report the description of the procedures and results in the Supplementary Material A.

For the self-paced reading task (Just et al., 1982) all sentences were displayed word by word on a single line using Linger software (http://tedlab.mit.edu/~dr/Linger/) in a PC. Masked words were presented as a series of underscores and the current word was presented in 18 pt Arial font; participants used the space bar of a PS/2 keyboard to unmask the next word.

We want to emphasize that there are clear advantages in the use of self-paced reading in comparison with the less artificial eye-tracking-while-reading method. Self-paced reading is a very well-tested method used in a large body of foundational research and with relatively clear, replicable results (e.g., garden path effects, good enough processing, etc.). It is admittedly a simplification of the reading process, but using eyetracking introduces new complexity, such as highly correlated multiple measures (first fixation duration, single fixation duration, gaze time, etc.) which are analyzed as if they were separate sources of information (von der Malsburg & Angele, 2017), and individual strategies (von der Malsburg & Vasishth, 2013; Nicenboim et al., 2015). This is, however, not to argue against the merit of eyetracking-while-reading, especially when there are clear hypotheses related to

the oculomotor control (e.g., Engelmann, Vasishth, Engbert, & Kliegl, 2013). See Vasishth, von der Malsburg, and Engelmann (2013) for the role of eye-movements for investigating human sentence processing and its limitations.

Data Analysis

The data analysis was conducted in the R programming environment (R Core Team, 2016), using Bayesian hierarchical models in Stan (Stan Development Team, 2016b) with the R package *RStan* (Stan Development Team, 2016a). For details on fitting Stan models, see Nicenboim and Vasishth (2016) and Sorensen, Hohenstein, and Vasishth (2016). In all the models, the interference condition was sum coded (-1 for low interference and 1 for high interference condition) and covariates presented in the Supplementary Material A were scaled and centered. Data and code are available in https://osf.io/mmr7s. We fit the models with four chains and 3000 iterations, half of which were the burn-in or warm-up phase. In order to assess convergence, we verified that there were no divergent transitions, that the $\hat{R}$s (the between- to within-chain variances) were close to one, and we also visually inspected the chains (Gelman, Carlin, Stern, & Rubin, 2014). Thus, the convergence diagnostics indicated that all the models had converged.

Results

Comprehension accuracy. Comprehension accuracy was relatively high (notice that chance level is 33% since there are three valid options): Participants answered correctly on average 80% (SE = 4) comprehension probes including fillers. Generalized linear mixed models (where the correct answer was coded as 1 and any type of incorrect answer was coded as 0) showed no evidence for lower accuracy due to interference ($\hat{\beta} = 0$, 95% CrI = $[-0.07, 0.08]$, $P(\beta > 0) = 0.52$). See Fig. 3 for the summary of comprehension accuracy (on the raw scale) for Experiments 1 and 2, and for the pooled data of both experiments.

Reading Times. In order to identify the critical region(s) where the effect may appear, we first fit a single model for several regions (see the top part of Fig. 1 for the log-transformed mean reading times of every region in Experiment 1 and see Fig. 2 for the results of the statistical model). We coded the different regions with Helmert contrasts, since this contrast allows us to compare each region with the average of the previous ones, such that it is possible to discover a change in the pattern of the effect and distinguish whether the effect originated in a given region or in the previous ones. In other words, this analysis is meant to identify whether there is evidence for a difference in reading times between high and low interference conditions in a given region (the cross in Fig. 2) in comparison with the average difference between conditions across the previous regions (the horizontal black bar in Fig. 2). This is meant to prevent results that are due to spurious differences in reading times between conditions before the region of interest. In order to account for the correlations between the regions in a single sentence, we included random effects by sentences as well as the usual random effects by participants and items (the random effects structure included intercept and slopes for the experimental condition as well as their correlation).

The model shows evidence for a change of pattern caused by the experimental condition at four different points in comparison with the average of the previous regions (notice that *prev* is different for every region and it is represented in Fig. 2 with a thick black line and that a positive difference refers to reading times at the high interference condition being larger than at the low interference one): (i) a slowdown at the accusative noun (e.g., *Assistenten*, *condition:prev-N.gen*) ($\hat{\beta} = 0.003$, 95% CrI $= [-0.001, 0.007]$, $P(\beta > 0) = 0.91$), (ii) a speedup at the participle verb belonging to the relative clause (e.g. *begrüßt*, *condition:prev-V.part.RC*) ($\hat{\beta} = -0.003$, 95% CrI $= [-0.007, 0.001]$, $P(\beta > 0) = 0.05$), (iii) a slowdown at the embedded auxiliary verb (*hatte*, *condition:prev-V.aux.RC*) ($\hat{\beta} = 0.003$, 95% CrI $= [-0.001, 0.006]$, $P(\beta > 0) = 0.94$), and finally (iv) a slowdown at the second spillover of the matrix verb (*condition:prev-spill2*) ($\hat{\beta} = 0.002$, 95% CrI $= [0.001, 0.004]$, $P(\beta > 0) \approx 1$). See Fig. 2 for the summary of the main results.

We further inspected the last two results since we hypothesized that these regions may be reflecting inhibitory interference at the retrieval triggered by the embedded auxiliary verb, and by the matrix verb but detected at the second spillover region (henceforth, matrix verb + 2). Notice that the first two results cannot be attributed to relevant regions since they appear before the first number cue provided by the verbs. Looking ahead, the confirmatory analysis provided by Experiment 2 will show that we have evidence only for the effect at the embedded auxiliary verb.

A further inspection of the last two results of Experiment 1 consisted in, first, taking the embedded auxiliary verb and the matrix verb + 2 as potential critical regions and then fitting separate models to these two regions to evaluate whether the slowdowns were not only due to a speedup in the average of the previous region. In these models, we included scores for individual differences as covariates, which we report in the Supplementary Material A. The exploratory models are consistent with interference effects at the auxiliary verb: $\hat{\beta} = 0.02$, 95% CrI $= [-0.01, 0.05]$, $P(\beta > 0) = 0.94$, and at the matrix verb + 2 region: $\hat{\beta} = 0.02$, 95% CrI $= [0.01, 0.04]$, $P(\beta > 0) \approx 1$. See Fig. 4 for the summary of the results on the raw scale (milliseconds) of Experiments 1 and 2, and for the pooled data of both experiments. We delay the discussion of the results until after the presentation of the results of Experiment 2 and of the pooled data.

Experiment 2: Confirmatory analysis

Experiment 2 served to confirm or falsify the findings of Experiment 1. It was an exact replication of the first experiment, but we only fit models at the regions where we found some evidence for interference in Experiment 1.

Method

Participants. The confirmatory phase was run with 100 subjects aged between 18–40 years old (mean 24.2 years) recruited from the same pool of participants as Experiment 1 at the University of Potsdam, Germany. None of these participants took part in Experiment 1.

Procedure. Subjects of the confirmatory experiment were tested at the same lab with the same settings and following the same procedure as in Experiment 1.

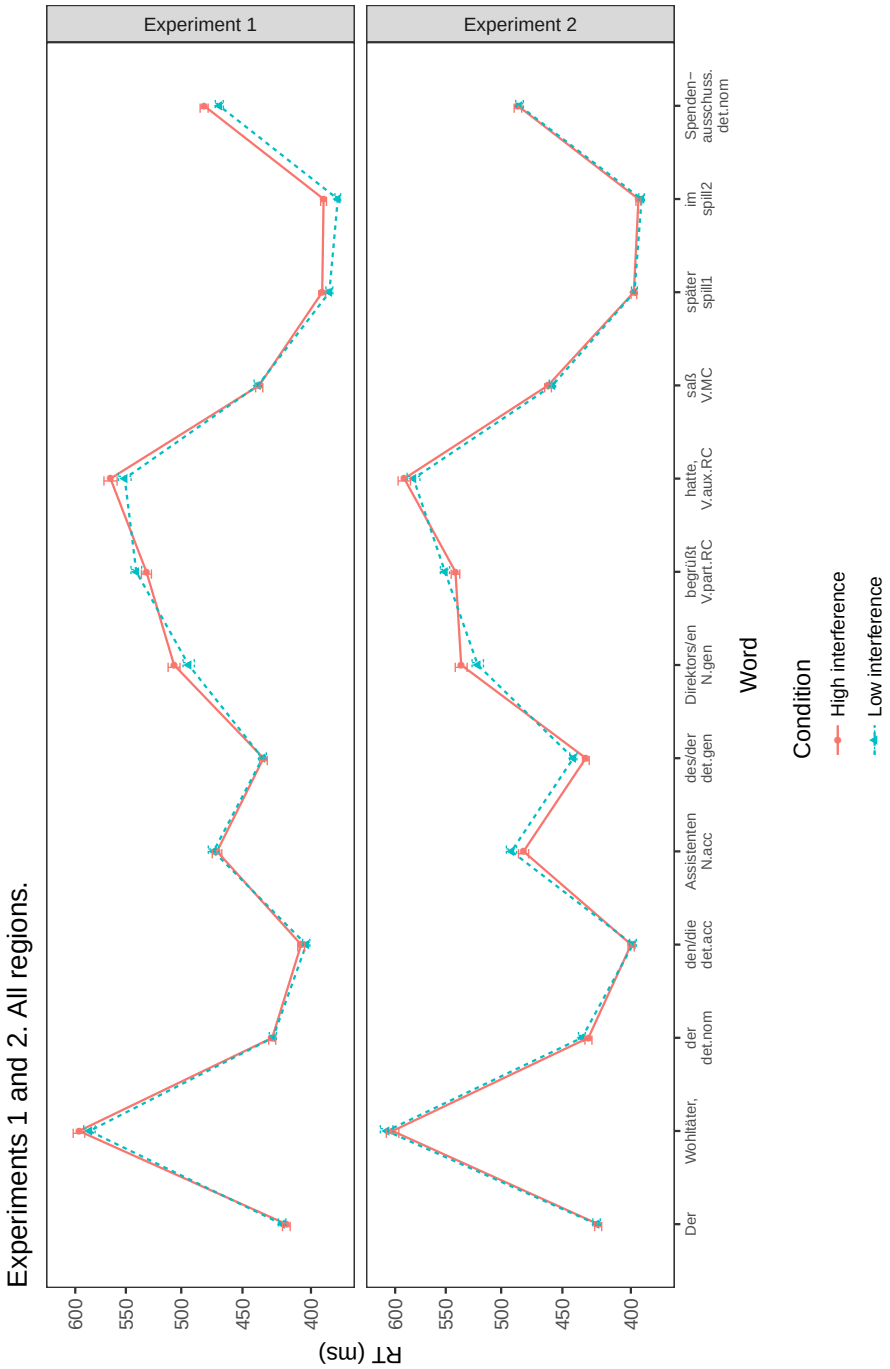

Figure 1. Summary of log(RTs) in all regions averaged by subjects, error bars are two standard errors from the log-transformed means.

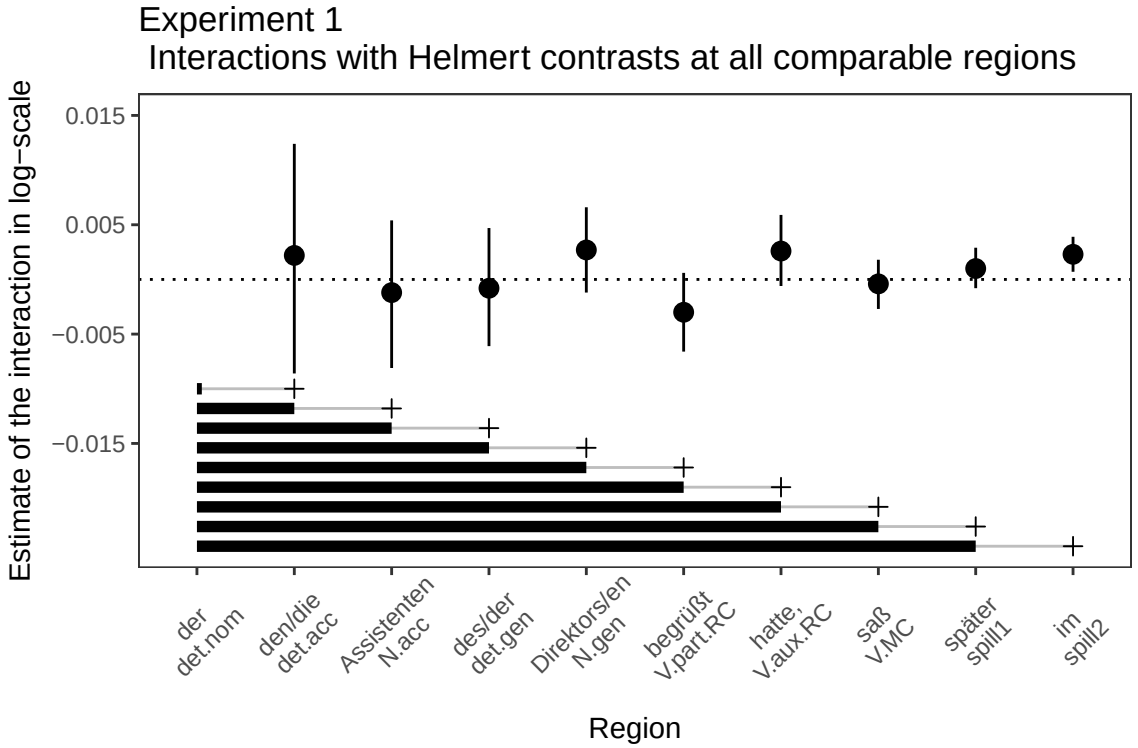

Figure 2. Means and 95% credible intervals for the posterior distributions of the estimates of the interactions between condition and the Helmert coded regions for the exploratory data (each region vs. the average of the previous regions is depicted with a thick black line).

Results

Comprehension accuracy. As for Experiment 1, accuracy was relatively high: participants answered correctly on average 82% (SE = 4) comprehension probes of the whole experiment including fillers. As before, we fit a generalized linear mixed model (where the correct answer was coded as 1 and any type of incorrect answer was coded as 0). In contrast to Experiment 1, this model showed some evidence for an effect of interference ($\hat{\beta} = -0.07$, 95% CrI $= [-0.14, 0]$, $P(\beta > 0) = 0.03$); we will return to this discrepancy with Experiment 1 in the Discussion section. See Fig. 3 for the summary of comprehension accuracy (on the raw scale) for Experiments 1 and 2, and for the pooled data of both experiments.

Reading Times. For this experiment, we repeated the analysis of the embedded auxiliary verb and the matrix verb + 2 regions. Both the model fit at the embedded auxiliary verb ($\hat{\beta} = 0.01$, 95% CrI $= [-0.01, 0.04]$, $P(\beta > 0) = 0.8$), and to the model fit at the matrix verb + 2 region ($\hat{\beta} = 0.004$, 95% CrI $= [-0.01, 0.019]$, $P(\beta > 0) = 0.72$) furnished only weak support for interference. We evaluate the evidence for inhibitory interference of these two regions in the Discussion section.

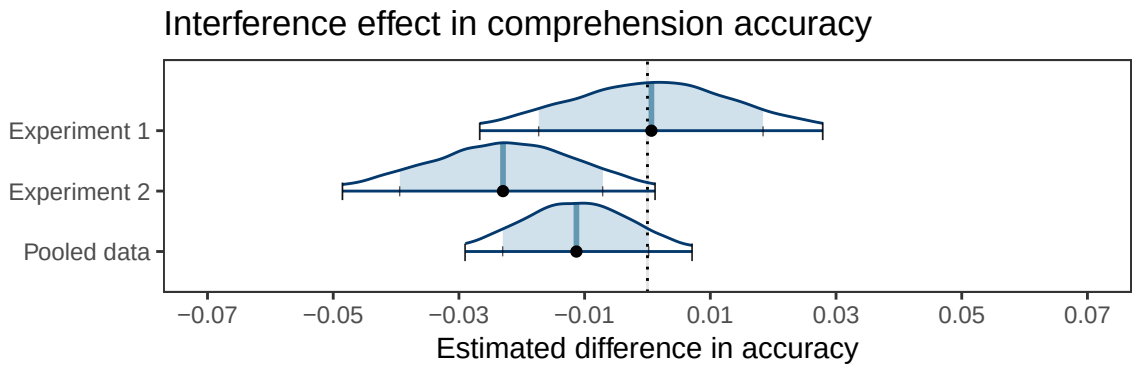

Figure 3. Posterior distributions of the estimated difference in accuracy between conditions (on the raw scale) for the generalized linear mixed models for Experiments 1 and 2, and for the pooled data. The full vertical lines indicate the mean of the posteriors, the outer error bars demarcate the 95% credible intervals, and the inner error bars and filled section of the distributions the 80% credible intervals. The broken vertical line indicates a difference of zero.

Analysis of the pooled data

The analysis of pooled data is meant to answer the following question: what is the best estimate we can obtain of the interference effect given *all* the data? This question is posed from the perspective of the Bayesian framework, in which there is no concept of Type I or II error. The data at hand are what they are, and in order to answer a research question, we should attempt to obtain the most precise estimates we can. This is only possible by considering all the data available. Of course, since we first identified the strongest evidence we can find by exploring the data from Experiment 1, and then pooled that strongest evidence with the data in Experiment 2, our estimates from the pooled data are likely to be overestimates. Only a future replication of the experiment, with sample size 180 or 200, can tell us how much of an overestimate we have here.

In the pooled data, the estimate for the effect in comprehension accuracy was the following: $\hat{\beta} = -0.03$, 95% CrI $= [-0.08, 0.02]$, $P(\beta > 0) = 0.1$. The estimates of the effect in reading times were the following: at the auxiliary verb, $\hat{\beta} = 0.02$, 95% CrI $= [0, 0.03]$, $P(\beta > 0) = 0.97$, and at the matrix verb + 2 region, $\hat{\beta} = 0.01$, 95% CrI $= [0, 0.02]$, $P(\beta > 0) = 0.98$.

In Fig. 3 and 4, we show graphically the differences between conditions, back-transformed to percentages (accuracy) or milliseconds (reading times). The objective of the meta-analytic estimates is to provide a reference for future work, and to be able to compare the results of the current study with previous results from the literature.

Discussion

In comprehension accuracy, Experiment 1 showed no evidence for the interference effect ($\hat{\beta} = 0.002$, 95% CrI $= [-0.072, 0.075]$, $P(\beta > 0) = 0.52$), but Experiment 2 showed some evidence consistent with the predictions of interference theories ($\hat{\beta} = -0.07$, 95% CrI $= [-0.14, 0]$, $P(\beta > 0) = 0.03$). In addition, the difference in accuracy between conditions

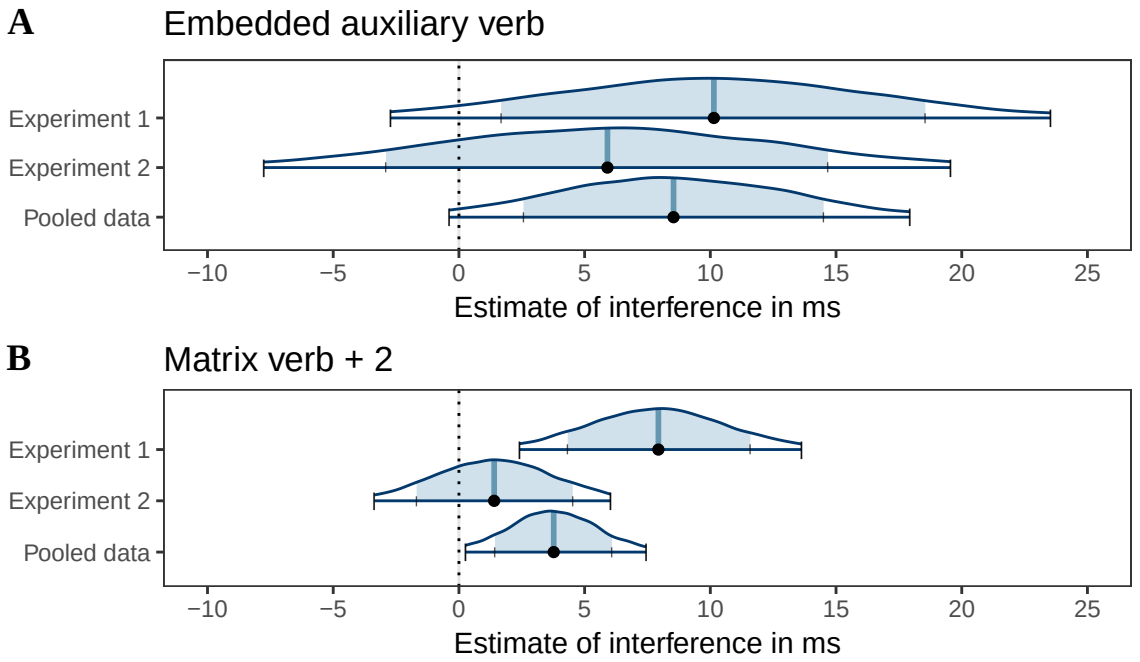

Figure 4. Posterior distributions of the estimated difference in reading times between conditions (in milliseconds) for the models fit for Experiments 1 and 2, and for the pooled data. The vertical lines indicate the mean of the posteriors, the outer error bars demarcate the 95% credible intervals, and the inner error bars and filled section of the distributions the 80% credible intervals.

on the raw scale for the pooled data, that is, the influence of interference in the proportion of correct responses, reveals a reduction of just -1.1% (95% CrI = $[-2.9, 0.7]$) between conditions; see Fig. 3. This casts doubt on whether the difference in accuracy is meaningful.

Regarding the effects in reading times, a visual inspection of Fig. 4 reveals that the posteriors of the effect at the auxiliary verb in Experiment 1 and Experiment 2 are quite similar. In addition, the pooled raw estimate of the effect at the auxiliary verb (9 ms 95% CrI = $[0, 18]$) is within the range of plausible values derived in the meta-analysis reported by Jäger et al. (2017) across non-agreement interference studies (13 ms 95% CrI=$[2, 28]$).

For the effect in reading times at the matrix verb + 2 region, however, Fig. 4 shows very little overlap between posteriors. In addition, our estimate on the raw scale is smaller than the effect at the auxiliary verb and than the average effect reported in the literature. Our model yields a difference between conditions of 4 ms (95% CrI = $[2, 14]$) (in comparison with the meta-analytic estimate of 13 ms; 95% CrI = $[2, 28]$ of Jäger et al., 2017).

Whereas the use of the first experiment for hypothesis generation eliminates much of the multiple comparisons problem, the second experiment still analyzes two separate regions in the confirmatory analysis. This inflates the chances of finding a spurious inhibitory effect in one of these regions. A decisive way to establish the inhibitory interference effect would be to carry out a new high-power replication of the experiment, and to analyze the effect only at the auxiliary verb. In general, replication is the only way to establish robustness of an effect.

In our particular experiment, it is possible that the multiple comparisons problem is not too severe. This is because the two regions, the embedded verb and the matrix verb, trigger two different retrieval events, and therefore represent two partially independent inhibitory effects. We acknowledge that ideally the relationship between these two events should be modeled, but it is still possible that only one retrieval (at the embedded verb) shows an interference effect. The lack of clear evidence for interference at the matrix verb could be because, in our stimuli, the target of the retrieval at the matrix verb is a noun (the subject), which does not have competitor nouns belonging to the same matrix clause, and thus it may have an extra feature, namely MATRIX-CLAUSE; this could differentiate the target from the competitors. Another possibility is that if the subject noun has already been retrieved by the embedded verb, it may be salient in memory (due to reactivation: Vasishth & Lewis, 2006; or being in the focus of attention: McElree, Foraker, & Dyer, 2003) so that the retrieval at the matrix verb might be less costly regardless of interference.

In sum, taken together, Experiment 1 and 2 furnish some weak evidence for an interference effect, and only at the embedded auxiliary verb.

Interestingly, Wagers et al. (2009) argued that they found no evidence for number interference in constructions similar to those in our experiments. Based on their null results, they drew the strong conclusion that number interference does not occur in grammatical sentences. However, the absence of evidence may well be a consequence of low statistical power. We can demonstrate this point by considering what would have happened if, instead of running a large sample study, we had repeatedly run experiments on number interference with small sample sizes, and carried out the conventional analyses.

Repeated sampling with small sample sizes. In order to investigate how likely it would be to find significant effects in datasets with typical sizes, we repeatedly sampled from our entire dataset, assuming either 30 or 60 participants. We sampled 10,000 times the relevant number of participants and 24 items from our complete data, and we reanalyzed the subsets of the data at the auxiliary verb (where we had found evidence for interference). Notice that we reduced both the number of participants and items to emulate typical datasets. However, in planned experiments, between-participant variability is usually the largest source of variance. This means that in planned experiments, the number of participants will typically affect the statistical power the most.

For each of the sampled datasets, we performed the same trimming procedure as Wagers et al. (2009) did, and we followed their procedure of using untransformed raw reading times as the dependent variable. Our models had condition as predictor (high interference coded as .5 vs. low interference coded as $-.5$; in this way, for non-transformed data, β represents the interference effect in milliseconds), and the random effect structure included by-participants and by-items intercepts and slopes without their correlation to ensure convergence. In order to assess significance ($p < .05$), we ran frequentist linear mixed models (with the package *lme4*; Bates, Mächler, Bolker, & Walker, 2015) and we used likelihood ratio tests to compare each model with a baseline model that lacked the predictor condition. In case of convergence failure, we removed the model from the simulation (0.2% of the models for each run of 10,000 iterations).

The repeated sampling reveals the following: With 30 subjects, we would find significant effects only 11.4% of the times. Moreover, 11% of those significant results yielded a *speedup* in high interference conditions, which is the opposite direction of our results. The

results that were consistent with inhibitory interference (a slowdown in the high interference condition) showed differences between conditions that ranged from 40 to 284 ms (vs. our estimate of 9 ms).

With 60 subjects, the situation did not improve much, we found significant results only 16% of the times with 4.9% of those having the opposite sign. The range of differences between conditions for inhibitory interference was from 36 to 148 ms.

Fig. 5 shows 700 randomly chosen iterations of the 10,000 runs. The figure shows that if we run a low-power study on number interference, we are very likely to obtain uninformative null results. Thus, the small sample sizes that Wagers et al. (2009) had in their experiments don't justify their making strong claims in favor of null results.

Fig. 5 also reveals another interesting and disturbing aspect of running low-power experiments: low-power studies can yield impressively large significant effects, which are statistically significant but are not realistic (Gelman & Carlin, 2014). The figure shows that whenever significant effects are found under repeated sampling, the estimate tends to be much larger than the 9 ms value in the full dataset. Such overestimates have the consequence that a low precision result with a mean of 180 ms with standard error 50 ms can be "more" significant (t-value=3.6) than a high precision result like 9 ms with standard error 4.5 (t-value=2). The former admits a wide range of plausible values and the estimated mean could well be the result of Type M error, whereas the latter is much more precise and could be more realistic. And yet, the low-precision estimate with a larger t-value would be regarded as better reflecting reality.

In sum, low-power experiments have two harmful consequences: they can lead to invalid conclusions about the null being true, and they can lead to overestimates of the true effect. In either of these two scenarios, one does not learn much about the effect of interest.

General Discussion

In this paper, we make three methodological contributions that can help to mitigate the replication crisis in cognitive science. First, we show how separating an experiment into an exploratory phase and a confirmatory phase yields a more realistic estimate of the effect of interest. Second, we demonstrate the importance of conducting appropriately powered studies in order to obtain accurate estimates of effects. Third, adopting a Bayesian data-analytical approach allows us (i) to fit statistical models that better reflect the underlying generative process of the data, and (ii) to focus on quantifying our uncertainty about the magnitude of the effect of interest, rather than making binary decisions about an effect being present or absent.

In our case study, we found weak evidence for an inhibitory interference effect during the retrieval of the subject when it shares the number feature with other competitor nouns, suggesting that retrieval in high interference conditions took longer than in low interference conditions. The evidence for this effect is stronger at the embedded auxiliary verb than at the matrix verb.

The inhibitory effect of interference is consistent with the predictions of models (Lewis & Vasishth, 2005; Vasishth & Lewis, 2006; McElree, 2000; Van Dyke & McElree, 2006) that assume a cue-based retrieval mechanism to create non-local dependencies using number as a cue. Our findings are difficult to reconcile with Wagers et al.'s strong claim (2009) that dependency completion via cue-based retrieval only occurs in ungrammatical sentences.

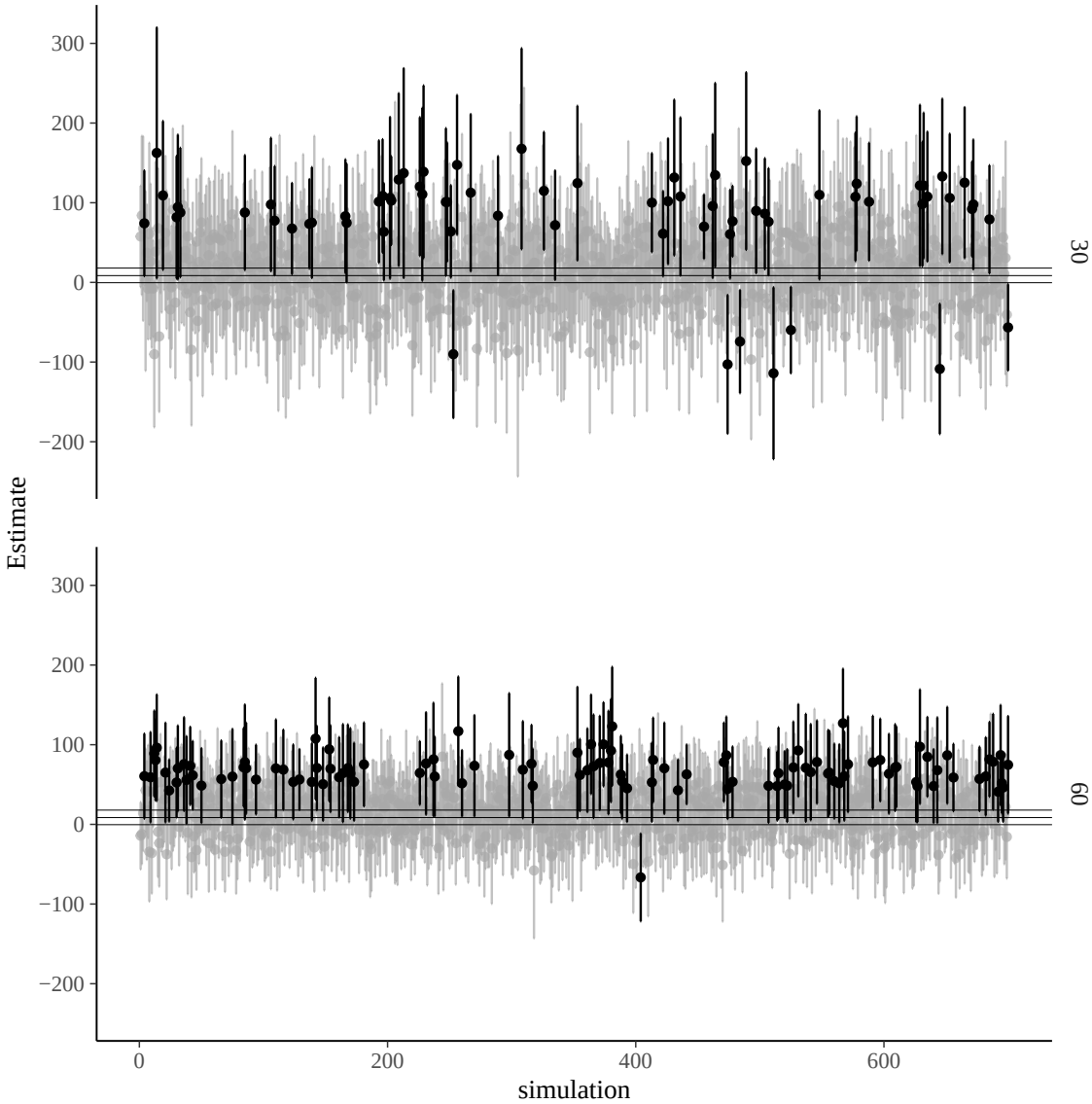

Figure 5. The figure shows in black the means and 95% confidence intervals of statistically significant effects under 700 repeated samples (with replacement) from all the data in the present paper, assuming that the number of participants is 30 or 60. Null results are shown in gray. The horizontal lines show our estimate of the effect from all the data, 9 ms with credible interval 0-18 ms.

Our results provide some evidence in favor of dependency creation via retrieval even in grammatical sentences, consistent with the assumption that parsing depends on cue-based retrieval mechanisms that are available in the general memory system.

Conclusion

From the methodological perspective, our work makes three novel contributions.

First, we demonstrate one way that the replication crisis in the psychological sciences can be addressed: by actually attempting to replicate an effect found in one's study. We demonstrate how we can implement De Groot's recommendation to first carry out an exploratory analysis to flexibly look for an effect. Because such an analysis, which relies on researcher degrees of freedom, has no evidentiary force, the researcher then follows up on this exploratory step by carrying out a confirmatory analysis that allows for no researcher degrees of freedom.

Second, we show that running as high-powered a study as logistically feasible is vital for obtaining realistic estimates of effects. One cause for the replication crisis in the psychological sciences is that many studies have low power; this leads either uninterpretable null results, or exaggerated significant effects due to Type M error (Gelman & Carlin, 2014) coupled with publication bias.

Third, we demonstrate the considerable advantages that a Bayesian analysis provides for interpreting data. As a case study, we investigated number interference, which has in previous work been argued to show null results. We show, through a Bayesian hierarchical data analysis, that there is weak evidence consistent with the predictions of cue-based retrieval theories. These theories predict an inhibitory interference effect during the retrieval of the subject when it shares the number feature with other competitor nouns, suggesting that retrieval in high interference conditions took longer than in low interference conditions. It is clear from our results, however, that number as a cue does not play a major role for the parser: our pooled analysis shows a very small magnitude for the estimate of interference effect at the embedded auxiliary verb, (9 ms 95% CrI = $[0, 18]$, between conditions).

References

Barr, D. J., Levy, R. P., 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. doi:10.1016/j.jml.2012.11.001

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). *Parsimonious mixed models*. ArXiv e-print.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. version 1.1-12. doi:10.18637/jss.v067.i01

Bock, K. & Miller, C. A. (1991). Broken agreement. *Cognitive Psychology*, *23*(1), 45–93. doi:10.1016/0010-0285(91)90003-7

Box, G. E. & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 211–252.

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. doi:10.1016/j.actpsy.2014.02.001

Dillon, B., Mishler, A., Sloggett, S., & Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: Experimental and modelling evidence. *Journal of Memory and Language*, *69*, 85–103. doi:10.1016/j.jml.2013.04.003

Drenhaus, H., Saddy, D., & Frisch, S. (2005). Processing negative polarity items: When negation comes through the backdoor. In S. Kepser & M. Reis (Eds.), *Linguistic evidence – Empirical, theoretical, and computational perspectives* (pp. 145–165). Studies in Generative Grammar 85. Mouton de Gruyter Berlin, Germany.

Engelmann, F. (2016). *Toward an integrated model of sentence processing in reading* (Doctoral dissertation, University of Potsdam).

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

Ferreira, F., Bailey, K. G. D., & Ferraro, V. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science*, *11*(1), 11–15. doi:10.1111/1467-8721.00158

Franck, J., Vigliocco, G., & Nicol, J. L. (2002). Subject-verb agreement errors in French and English: The role of syntactic hierarchy. *Language and Cognitive Processes*, *17*(4), 371–404.

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., & Rubin, D. B. (2014). Bayesian Data Analysis. (Third, Chap. 11). Boca Raton, FL: Taylor & Francis.

Gillespie, M. & Pearlmutter, N. J. (2011). Hierarchy and scope of planning in subject–verb agreement production. *Cognition*, *118*(3), 377–397. doi:10.1016/j.cognition.2010.10.008

Greiner, B. (2004). An online recruitment system for economic experiments. *Forschung und wissenschaftliches Rechnen*, *63*, 79–93.

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, *21*(5), 1157–1164. doi:10.3758/s13423-013-0572-3

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.

Jaynes, E. T. & Kempthorne, O. (1976). Confidence intervals vs. Bayesian intervals. In W. L. Harper & C. A. Hooker (Eds.), *Foundations of probability theory, statistical inference, and statistical theories of science* (Vol. 6b, pp. 175–257). The University of Western Ontario Series in Philosophy of Science. Dordrecht: Springer Netherlands. doi:10.1007/978-94-010-1436-6_6

Just, M. A., Carpenter, P. A., & Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, *111*(2), 228–238. doi:10.1037/0096-3445.111.2.228

Lago, S., Shalom, D., Sigman, M., Lau, E., & Phillips, C. (2015). Agreement processes in Spanish comprehension. *Journal of Memory and Language*, *82*, 133–149. doi:10.1016/j.jml.2015.02.002

Lewis, R. L. & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, *29*(3), 375–419. doi:10.1207/s15516709cog0000_25

Lewis, R. L., Vasishth, S., & Van Dyke, J. A. (2006). Computational principles of working memory in sentence comprehension. *Trends in Cognitive Sciences*, *10*(10), 447–454. doi:10.1016/j.tics.2006.08.007

Logan, G. D. (1992). Shapes of reaction-time distributions and shapes of learning curves: A test of the instance theory of automaticity. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *18*(5), 883–914.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, R. H., & Bates, D. (2017). Balancing Type I Error and Power in Linear Mixed Models. *Journal of Memory and Language*, *94*, 305–315. doi:10.1016/j.jml.2017.01.001

McElree, B. (2000). Sentence comprehension is mediated by content-addressable memory structures. *Journal of Psycholinguistic Research*, *29*(2), 111–123. doi:10.1023/A:1005184709695

McElree, B., Foraker, S., & Dyer, L. (2003). Memory structures that subserve sentence comprehension. *Journal of Memory and Language*, *48*(1), 67–91. doi:10.1016/s0749-596x(02)00515-6

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *23*(1), 103–123. doi:10.3758/s13423-015-0947-8

Nicenboim, B., Logačev, P., Gattei, C., & Vasishth, S. (2016). When high-capacity readers slow down and low-capacity readers speed up: Working memory and locality effects. *Frontiers in Psychology*, *7*(280). doi:10.3389/fpsyg.2016.00280

Nicenboim, B., Roettger, T. B., & Vasishth, S. (2017). *Using meta-analysis for evidence synthesis: The case of incomplete neutralization in German*. Submitted to Journal of Phonetics.

Nicenboim, B. & Vasishth, S. (2016). Statistical methods for linguistic research: Foundational Ideas - Part II. *Language and Linguistics Compass*, *10*(11), 591–613. doi:10.1111/lnc3.12207. eprint: https://arxiv.org/abs/1602.00245

Nicenboim, B. & Vasishth, S. (2017). Models of retrieval in sentence comprehension: A computational evaluation using Bayesian hierarchical modeling. *Journal of Memory and Language*. Accepted.

Nicenboim, B., Vasishth, S., Gattei, C., Sigman, M., & Kliegl, R. (2015). Working memory differences in long-distance dependency resolution. *Frontiers in Psychology*, *1*. doi:10.3389/fpsyg.2015.00312

Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science*, *7*(6), 657–660. PMID: 26168127. doi:10.1177/1745691612462588

Patson, N. D. & Husband, E. M. (2016). Misinterpretations in agreement and agreement attraction. *The Quarterly Journal of Experimental Psychology*, *69*(5), 950–971. doi:10.1080/17470218.2014.992445

Pearlmutter, N. J. (2000). Linear versus hierarchical agreement feature processing in comprehension. *Journal of Psycholinguistic Research*, *29*(1), 89–98. doi:10.1023/A:1005128624716

Pearlmutter, N. J., Garnsey, S. M., & Bock, K. (1999). Agreement processes in sentence comprehension. *Journal of Memory and language*, *41*(3), 427–456. doi:10.1006/jmla.1999.2653

R Core Team. (2016). *R: A language and environment for statistical computing*. R version 3.3.2. R Foundation for Statistical Computing. Vienna, Austria.

Rouder, J. N. (2005). Are unshifted distributional models appropriate for response time? *Psychometrika*, *70*(2), 377–381. doi:10.1007/s11336-005-1297-7

Rouder, J. N., Tuerlinckx, F., Speckman, P. L., Lu, 6., & Gomez, P. (2008). A hierarchical approach for fitting curves to response time measurements. *Psychonomic Bulletin & Review*, *15*(6), 1201–1208. doi:10.3758/pbr.15.6.1201

Schielzeth, H. & Forstmeier, W. (2009). Conclusions beyond support: Overconfident estimates in mixed models. *Behavioral Ecology*, *20*(2), 416–420. doi:10.1093/beheco/arn145

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. doi:10.1177/0956797611417632

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.

Stan Development Team. (2016a). RStan: the R interface to Stan. R package version 2.14.1.

Stan Development Team. (2016b). Stan: A C++ library for probability and sampling, version 2.14.0.

Suckow, K. & Van Gompel, R. P. G. (2012). Does number interference occur during sentence processing? In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th annual conference of the cognitive science society* (pp. 2357–2362). Cognitive Science Society. Austin, TX.

Suckow, K. & Van Gompel, R. P. G. (2013). Distinguishing effects of number interference and number attraction in sentence processing. In *AMLaP*. Marseille, France. doi:10.14293/P2199-8442.1.SOP-LING.PFPWIG.v1

Swets, B., Desmet, T., Clifton, C., & Ferreira, F. (2008). Underspecification of syntactic ambiguities: Evidence from self-paced reading. *Memory & Cognition*, *36*(1), 201–216. doi:10.3758/MC.36.1.201

Tanner, D., Nicol, J. L., & Brehm, L. (2014). The time course of feature interference in agreement comprehension: Multiple mechanisms and asymmetrical attraction. *Journal of Memory and Language*, *76*, 195–215.

Thornton, R. & MacDonald, M. C. (2003). Plausibility and grammatical agreement. *Journal of Memory and Language*, *48*(4), 740–759. doi:10.1016/S0749-596X(03)00003-2

Van Dyke, J. A. (2007). Interference effects from grammatically unavailable constituents during sentence processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(2), 407. doi:10.1037/0278-7393.33.2.407

Van Dyke, J. A., Johns, C. L., & Kukona, A. (2014). Low working memory capacity is only spuriously related to poor reading comprehension. *Cognition*, *131*(3), 373–403. doi:10.1016/j.cognition.2014.01.007

Van Dyke, J. A. & 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*(3), 285–316. doi:10.1016/S0749-596X(03)00081-0

Van Dyke, J. A. & McElree, B. (2006). Retrieval interference in sentence comprehension. *Journal of Memory and Language*, *55*(2), 157–166. doi:10.1016/j.jml.2006.03.007

Van Dyke, J. A. & McElree, B. (2011). Cue-dependent interference in comprehension. *Journal of Memory and Language*, *65*(3), 247–263. doi:10.1016/j.jml.2011.05.002

Vasishth, S., Beckman, M. E., Nicenboim, B., Li, F., & Kong, E. J. (2017). *Bayesian data analysis in the phonetic sciences: A tutorial introduction*. Submitted to Journal of Phonetics.

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. & Gelman, A. (2017). The statistical significance filter leads to overconfident expectations of replicability. In *Proceedings of cognitive science conference*. London, UK.

Vasishth, S. & Lewis, R. L. (2006). Argument-head distance and processing complexity: Explaining both locality and antilocality effects. *Language*, *82*(4), 767–794. doi:10.1353/lan.2006.0236

Vasishth, S., Nicenboim, B., Chopin, N., & Ryder, R. (2017). *Bayesian hierarchical finite mixture models of reading times: a case study*. submitted to Psychological Review. doi:10.17605/OSF.IO/FWX3S

Vasishth, S., von der Malsburg, T., & Engelmann, F. (2013). What eye movements can tell us about sentence comprehension. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*(2), 125–134.

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. doi:10.1016/j.jml.2016.10.003

von der Malsburg, T. & Vasishth, S. (2013). Scanpaths reveal syntactic underspecification and reanalysis strategies. *Language and Cognitive Processes*, *28*(10), 1545–1578. doi:10.1080/01690965.2012.728232. eprint: http://dx.doi.org/10.1080/01690965.2012.728232

Wagers, M. W., Lau, E. F., & Phillips, C. (2009). Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language*, *61*(2), 206–237. doi:10.1016/j.jml.2009.04.002

Appendix

Stimuli

Fig. A1 shows the variability in the by-items slope of the model with the pooled data, that is, the unaccounted contribution to the interference effect of each experimental item.

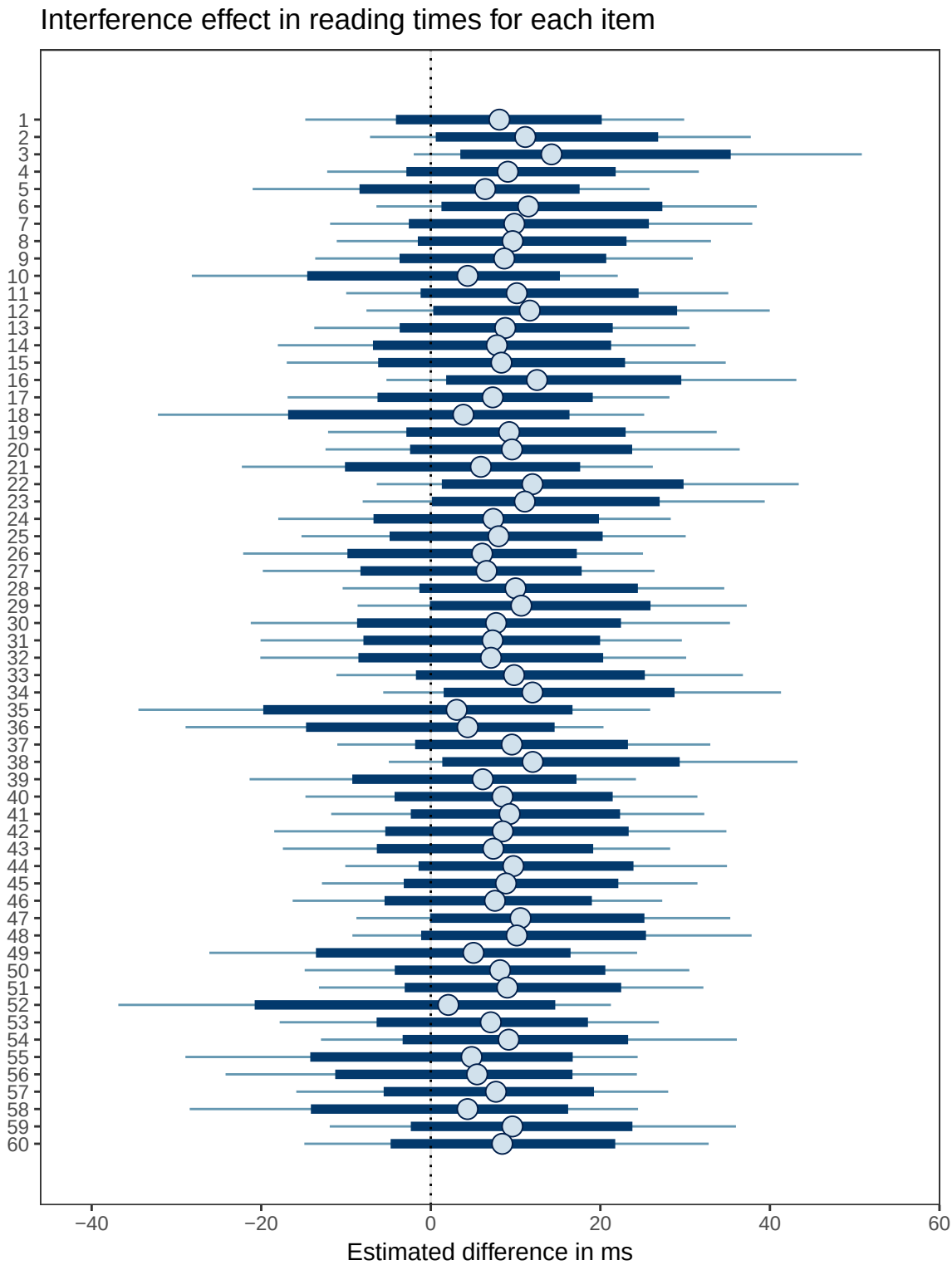

Figure A1. Means and 95% credible intervals for the posterior distributions of the unaccounted variation in milliseconds at the embedded auxiliary verb (*hatte*, *had.sg*) contributed by each item.