New directions in statistical analysis for experimental linguistics

Shravan Vasishth

December 11, 2022

Abstract

In recent decades, linguistics has taken an empirical turn–experimental methods have become a standard part of the toolkit for researchers in areas like syntax, semantics, and pragmatics. Because experimental science requires statistical tools, and because experimental data has historically been usually analyzed using frequentist methods, linguists have adopted these standardly used methods. However, by adopting these standard methods, linguistics also imported all the problems that frequentist methods have engendered; the replication crisis is perhaps the most dramatic one of all. Most of these problems arise due to the way the null hypothesis significance testing procedure is set up: a straw-man null hypothesis is rejected that was never of any interest in the first place; binary accept/reject decisions are made based on the p-value, disregarding the uncertainty in the estimates; one is encouraged through conveniently available software to fit canned statistical models with fixed assumptions, even when those assumptions are extremely unreasonable; and there is no way to cumulatively build on previous findings when analyzing data. Meanwhile, there have been important developments recently in statistical computing, particularly in Bayesian methodology. These new statistical tools, which focus on uncertainty quantification using Bayes' rule, are well worth understanding and adopting in linguistics. Bayesian data analysis has several important advantages over the frequentist paradigm: prior knowledge and/or competing prior beliefs can be formally incorporated in a data analysis; the focus shifts to quantifying uncertainty by concentrating on parameter estimation rather than making binary decisions; all sources of variance can be taken into account simultaneously in a model, leading to more conservative inferences; and statistical models can be customized to the research problem at hand. All these advantages lead to more robust inferences. This article explains in detail the problems with standard methods and discusses, with a concrete running example, the advantages of adopting an uncertainty-quantification based approach to statistical inference using modern Bayesian tools. Data and code accompanying this article are available from https://osf.io/kgxpn/.

"Null hypothesis significance testing has been psychology's hammer that made cognition and behaviour look like a nail." Blokpoel and van Rooij (2021).

1 Introduction

In recent decades, linguistics has taken an empirical turn: it is now routine to run experiments to test theoretical questions in areas like syntax (Sprouse et al. 2012), semantics (Hackl et al. 2012), and pragmatics (Chemla 2009). Even those linguists who used to rely on intuition to develop their theories are now quite well-versed in conducting planned experiments using well-executed experiment designs and sophisticated equipment.

As far as data analysis goes, linguistics has historically looked to standard practice in psychology to develop the methodology of statistical inference. Unfortunately, the form of statistical inference that is the norm in psychology, and now in linguistics, has drifted far from its original intent (e.g., Belia et al. 2005). Even psychology textbooks (written by psychologists) propagate a fundamentally incorrect understanding of statistical inference (Cassidy et al. 2019).

As a consequence of such misunderstanding, the statistical inferences that are often reported in published papers end up being at best questionable. We see the misinterpretions of statistical inference playing out in psychology through the replication crisis, whereby an alarmingly high proportion of claimed effects could not be replicated (Open Science Collaboration 2015, Nosek et al. 2022). Similar problems occur in psycholinguistics (e.g., Vasishth et al. 2018).

One major reason for these misinterpretations is that researchers (including the author of this article, when he was a graduate student) often receive only a cursory education in statistical inference, and dive into the nitty-gritty work of data analysis without really knowing what a p-value is or what it can and cannot tell us (Wasserstein and Lazar 2016). Why, despite the focus on empirical methods in linguistics, is statistics education neglected? Through conversations with fellow linguists, it appears that statistics education has received second-class citizen status in linguistics largely because it is considered to be orthogonal to the scientific process itself. However, this is a misunderstanding: As far as experimental linguistics goes, scientific reasoning and statistical inference are tightly connected, and neglecting the latter is likely to lead to invalid scientific conclusions.

In this chapter, I use a practical example from a psycholinguistic data set (Nicenboim et al. 2018) to briefly discuss some of the most serious problems with standard null hypothesis significance testing, and then I present an alternative approach to data analysis that uses Bayesian methods to shift the focus towards uncertainty quantification. One consequence of using Bayesian methods—assuming that they are used as intended—is that statistical inferences will generally be more conservative. For related discussions, see Vasishth and Gelman (2021), Vasishth (2022). One caveat here is that it is of course possible to misuse Bayesian methods as well (Tendeiro et al. 2022); however, one important advantage of Bayesian approaches is that one can directly focus on uncertainty quantification, as I explain below.

This chapter assumes that the reader has some familiarity with experiment design, in particular with repeated measures designs, and have at least a passing acquaintance with standard statistical methods, such as t-tests and the linear mixed model (Winter 2019). If the reader is lacking this background, some material specifically written for (psycho)linguists that will help the reader to acquire the assumed background (Baayen 2008, Baayen et al. 2008, Vasishth and Nicenboim 2016, Vasishth and Gelman 2021, Vasishth 2022, Vasishth, Schad,

Bürki and Kliegl 2022).

Reproducible code and data related to this chapter are available from https://osf.io/kgxpn/.

2 A conventional frequentist data analysis using statistical significance (what could go wrong?)

As a case study, consider the self-paced reading study reported in Nicenboim et al. (2018). (The published paper did not use the frequentist approach for data analysis that I use here; I use the frequentist approach to illustrate the problems that null hypothesis significance testing leads to.)

This experiment investigates a phenomenon called similarity-based interference (Lewis and Vasishth 2005); the essential claim being tested is that when a subject-verb dependency has to be built, the presence of nouns similar to the grammatical subject, but not in subject position, can make it harder to complete the correct dependency. The experiment design in Nicenboim et al. (2018) involves German and consists of two conditions. There is a low-interference condition (1a), where only the subject noun, *Der Wohltäter*, 'the philanthropist', has the same number marking as the auxiliary verb *hatte*, 'had', and a high interference condition (1b), where there are three nouns with the same number marking. In the high-interference condition, it is harder to distinguish the grammatical subject from the other two (distractor) nouns, and as a consequence completing the subject-verb dependency takes more time.

(1) a. Low interference

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

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

b. High interference

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

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

Nicenboim et al. (2018) carried out a self-paced reading experiment with 184 German native-speaker participants, each of whom saw a total of 60 items in a standard Latin square design.

A typical data analysis carried out in such a design would be to isolate the reading time for the critical region (the auxiliary verb *hatte*, 'had'), aggregate the data by subjects, and then carry out a one-sample (equivalently, a paired)

t-test. A repeated measures ANOVA (analysis of variance) would be equivalent to the t-test. One could also do a so-called by-items analysis, but this is omitted here for simplicity.

In this example, I follow standard statistical practice for positive-only dependent measures like reading time by log-transforming reading time; in this particular example, a reciprocal transform would be more approptiate (Kliegl et al. 2010, Box and Cox 1964), but for my current purposes, this is not an important issue, so I will use the log transform.

Such a t-test yields a statistically significant effect of interference: The t-value is 2 (with degrees of freedom 183), the p-value is 0.05. The estimate of the effect (on the log ms scale) is 0.0198 and the 95% confidence interval is [0,0.039]. The conventional conclusion would be that we found evidence for similarity-based interference. This conclusion is actually over-enthusiastic, as explained below.

At this juncture, one might object that the t-test is not appropriate for these data. In fact, in recent years, t-tests and repeated measures ANOVA have increasingly been replaced by the linear mixed model (Pinheiro and Bates 2000, Bates et al. 2015). It is correct that the linear mixed model is the better way to do the statistical analysis in the present case, because all sources of variability can simultaneously be taken into account (cf. Clark 1973). Doing such an analysis using the lme4 package in R yields an estimate of 0.02, with a 95% confidence interval of [-0.0016,0.0409], and a t-value of 1.85.[1]

So, is the result significant or not significant? The t-test suggests the answer is yes, the linear mixed model says no. There are two important observations here that need to be understood in order to interpret these apparently divergent results.

2.1 What could go wrong in frequentist hypothesis tests?

2.1.1 Statistical significance itself is not particularly informative

The first observation is that the difference between significant and not significant may itself not be significant (Gelman and Hill 2007), meaning that the two analyses above are not necessarily showing different "results" (where results refers to significance or non-significance of the effect).

To make this point concrete, we can do a short thought experiment. Suppose that we were to run the above experiment twice, with the same number of subjects each time (n=184) but different subjects in each run. Suppose that in the first run we obtain a t-value of 2.05 with an effect size of 0.02 log ms, with standard deviation 0.1323 (the t-value is computed using the formula: $(0.02-0)/(0.1323/\sqrt{184}) = 2.05$). Then we run the experiment again, and this time the estimate happens to be 0.01 and the standard deviation happens to be 0.14 (this can happen because of random variability). Now, the t-value is 1.02; not significant. Is the difference between these two results significantly different?

[1] When faced with such a null result, a common conclusion one sees in articles is to state that there is evidence *against* the interference effect. This conclusion is also a complete misunderstanding of the hypothesis testing framework: first, absence of evidence is not necessarily evidence of absence, and second, as discussed below, the difference between a significant and non-significant result is itself not necessarily significant (Gelman and Hill 2007).

We see many instances of papers in psycholinguistics and related areas (including one by the author of the present chapter) where researchers conclude that the answer is yes, the two results show meaningful differences (Nieuwenhuis et al. 2011, Levy and Keller 2013, Vasishth and Lewis 2006). But a two-sample t-test can answer that question formally. The difference in effect sizes is 0.02-0.01=0.01, the t-value from the two studies combined is computed using the formula:

$$observed - t = \frac{0.01 - 0}{\sqrt{0.1323^2/184 + 0.14^2/184}} = 0.7 \quad (1)$$

The t-value 0.7 tells us that we don't have a significant difference between the two studies (of course, this does not mean that there is no difference either– we just don't know). And yet, even experienced scientists (e.g., editors-in-chief of major journals in psychology and psycholinguistics) will consider the second study a replication failure, and more generally will interpret the significant vs. non-significant result as pointing to different conclusions.

The main point here is that obtaining a significant or non-significant result by itself is not necessarily going to allow us to make a discovery claim. Doing statistical analyses on experimental data gives the illusion of quantitative rigor; this is why some psycholinguists have started demanding that linguistics always rely on experimental data (Gibson and Fedorenko 2010, 2013). But in fact, the knowledge gleaned from quantitative methods can be very tenuous, and even strong advocates of quantitative methods rarely appreciate this point. One major problem here is statistical power; this is discussed next.

2.1.2 Underpowered studies will be misleading; and studies are often (severely) underpowered

The second observation is that, far more important than a significant or non-significant result is the extent to which the experiment design might overestimate the true effect under repeated sampling (Gelman and Carlin 2014). To understand this point, one must understand the concept of statistical power.

Power is the probability of detecting an effect if it actually exists (has some particular magnitude). Null hypothesis significance testing works as intended when it is used in high power situations; it is likely to lead to misleading results in low power situations. In the present case, even though this experiment was run with 184 subjects (which seems like a lot of subjects), the power of the design is relatively low.

To see this, imagine that we use the estimates from the above design to plan a *new* experiment. How many participants would we need to achieve 80% power (the power level recommended by the American Psychological Association)? Assuming that the effect size on the log ms scale is indeed 0.0198, and that the standard deviation (estimated from the data) is 0.1323 log ms, a power calculation shows that the necessary sample size would be 352 participants.[2]

[2]Incidentally, we are not computing "observed power" here but rather prospective power. Many researchers incorrectly try to determine the power of an already-conducted experiment, referring to this as "observed power"; but as Hoenig and Heisey (2001) show, observed power is just a transform of the p-value, and adds no new information about the current study. The only relevant use of power calculations is to plan a future experiment—prospective power.

What are the consequences of running a study with low power? One important consequence is that a statistically significant effect is likely to be based on an overestimate or may even have the wrong sign; Gelman and Carlin (2014) call this kind of misestimate Type M(agnitude)/Type S(ign) error.

Misestimation of the effect size under low power is one key reason why a statistically significant result such as the one we obtained above with the t-test is not especially informative. It is also unlikely to be replicable if we define replicability as repeatedly finding significant effects when re-running the experiment. In a replication attempt, even if we are lucky enough to detect the effect (through a significant result), the significant effect would again very likely be based on a misestimate.

One can demonstrate this through a simulation. Suppose that the true effect is 0.0198 log ms (so, the null hypothesis of no effect is false), and the standard deviation is 0.1323 log ms. If we repeatedly generate data with a typical sample size of 30 participants (Jäger et al. 2017) from the assumed normal distribution (a normal distribution is what the statistical test assumes), this yields a power of about 14%. We will find that on average, the significant effects will be based on an estimate (with the correct positive sign) that is about three times as large as the assumed true effect of 0.0198 log ms.

It is common in linguistics to run studies with relatively low power (Jäger et al. 2017, Bürki et al. 2020, Vasishth 2022, Vasishth et al. 2018, Jäger et al. 2020). This is not due to any malicious intent, but due to resource and time limitations that researchers are often faced with.

Low power is not a problem that is easy to solve, but what we can change is to move away from the focus on significant/non-significant results (which, as shown above, will be uninformative). But if we don't focus on significance, what *should* we focus on? I discuss this next.

2.2 A proposal from psychology to use frequentist confidence intervals instead of p-values, and the problem with this proposal

One important question we should consider given the above set of analyses (using the t-test and then the linear mixed model) is the following: We know what is different between the two analyses; but what is common to both the analyses? What is common to the two statistical tests is that the 95% confidence intervals are both showing similar values: the paired t-test shows [0,0.039], the linear mixed model [-0.002,0.041]. The linear mixed model estimate is wider because it includes more variance components than the t-test (which artifically reduces sources of variance through aggregation; see Schad, Nicenboim and Vasishth (2022)). Incidentally, the reader might again be tempted to conclude that the confidence intervals are showing different things, but this is the same mistake as treating significant vs. non-significant results as always being meaningful (they can be meaningful, but only when power is high).

Many researchers (e.g., Meehl 1997, Cumming 2014, McShane et al. 2019) have suggested that one should move away from statistical significance and focus instead on estimating and reporting confidence intervals. More generally, these researchers have argued that one should focus on quantifying one's uncertainty of the effect size. Usually, the frequentist confidence interval is used as a way

to quantify this uncertainty.

Under this view, one could just report the confidence interval (here, I use the estimate from the linear mixed effects model): [-0.002,0.041] log ms. Instead of saying that the effect was significant or not significant, we can just say: the observed effect is 0.02 log ms, with 95% CI [-0.0016, 0.0409]; this is consistent with the pattern predicted by the theory being investigated.

There are many advantages to such an approach: for one thing, once enough data accumulates, one can carry out a meta-analysis (Bürki et al. 2020, 2022, Jäger et al. 2017, Nicenboim et al. 2020), which allows us to quantitatively assert (modulo publication bias) what we have learned from existing studies. Another advantage is that other researchers can use the published results to plan a properly powered study.

One technical problem with the confidence interval is that it doesn't quantify uncertainty about the effect size, but has a rather convoluted meaning which is practically useless: if one were (counterfactually) to run the experiment again and again, and compute 95% confidence intervals *each time*, 95% of those repeatedly computed, hypothetical intervals would contain the true mean. This is practically useless because we have only one confidence interval to work with and it either contains the true effect or it doesn't; but we just don't know which of these two possibilities is true!

It is mathematically incorrect to treat the frequentist confidence interval as specifying the range over which we can be 95% certain that the true effect lies; the reason is that the effect is an unknown point value and therefore has no probability density function associated with it. Thus, if the effect is represented as the parameter β (in a linear mixed model, this would be a slope in the fixed effects part of the model), we cannot work out the values "lower" and "upper" such that the probability that β lies within these intervals is 0.95 ($Prob(lower < \beta < upper) = 0.95$). To compute such a probability, we would have to assign a probability density function to β; for example β would need to have a distribution like $Normal(\mu, \sigma)$. As mentioned above, in the frequentist paradigm, the effect is just an unknown point value "out there in nature"; it simply cannot have a probability distribution. It seems that this point has escaped even some psychologists (e.g., Meehl 1997) who argue against p-values as a way to carry out inference (also see Hoekstra et al. 2014).

Figure 1 visualizes the coverage properties of the confidence interval in 100 simulations ; by coverage we mean here the proportion of cases where the true μ is contained in the CI. The data are repeatedly generated from a normal distribution with mean 500 and standard deviation 100. Each confidence interval either contains the true mean 500 or it doesn't; the 95% refers to the probability that the 100 confidence intervals contain the true mean.

So is there some way to quantity uncertainty about the effect? It turns out that this is possible if we switch to a Bayesian way of thinking. I explain this point next.

Until recently, Bayesian methods were very inaccessible to the non-statistician; one reason for this was that sufficiently flexible software did not exist, and complex models were difficult to fit. This situation has changed completely over the last 10 years, and now software like Stan (Carpenter et al. 2017) and JAGS (Plummer 2012) have made it possible to fit relatively complex models quite easily. Moreover, several accessible textbooks, designed for the experimentalist who is not a statistician, have become available (Kruschke 2014, McElreath

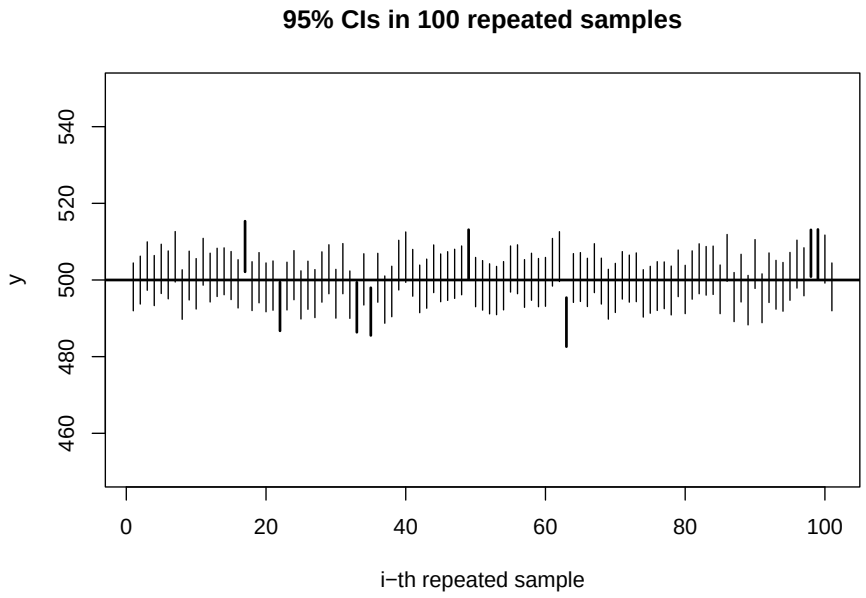

Figure 1: Illustration of the meaning of a 95 percent confidence interval (CI). The thicker bars represent the CIs which do not contain the true mean.

2020, Nicenboim et al. 2022). Because of these developments, it is now relatively easy to switch to a Bayesian methodology.

3 An alternative approach: Uncertainty quantification through Bayesian estimation

In the frequentist approach, the difference in means between two conditions is assumed to be an unknown point value. In our running example, the difference in means between the high and low interference conditions, call it δ, is estimated by computing the difference in sample means between the two conditions; this is the maximum likelihood estimate. After that, the statistical test (the t-test) is carried out by dividing d, the estimate of δ, by the estimated standard error. The 95% confidence interval is then $d \pm t_{crit} \times SE$, where t_{crit} is the critical t-value (which is approximately 2 for sample sizes larger than 20). The standard error only tells us how variable the estimate of the difference in sample means would be under (hypothetical) repeated sampling. The standard error cannot tell us anything about the uncertainty of the effect itself, as the effect δ is a point value by assumption; it has no distribution and therefore no uncertainty associated with it.

By contrast, the Bayesian approach assumes that the true difference in means, δ, has a probability density function associated with it. This is called a prior distribution and represents our prior belief or prior knowledge about this

difference. For example, we could define a prior distribution to δ as follows:

$$\delta \sim Normal(\mu, \sigma) \tag{2}$$

What this means is that is assumed a priori to have a 95% probability of lying between $\mu - 1.96 \times \sigma$ and $\mu + 1.96 \times \sigma$.

Defining such a prior distribution is quite a radical shift from the frequentist approach because, for the first time, we can talk about our prior uncertainty of the effect of interest. A natural question that arises at this point is: how can one come up with a prior distribution on the effect of interest even before running an experiment? Coming up with priors requires a way of thinking that physicists call a Fermi problem (Von Baeyer 1988); it is usually possible to work out reasonable priors for a particular research problem. Formal methods for deriving priors is a well-developed field (O'Hagan et al. 2006, Oakley and O'Hagan 2010, Morris et al. 2014). Actually, linguists are already familiar with prior elicitation: any linguist who has used intuition-based judgements to decide on the grammaticality of a sentence is basically deriving a prior belief about a sentence. For examples from psycholinguistics of how prios can be systematically worked out, see chapter 6 of Nicenboim et al. (2022).

Once we analyze the data in the Bayesian framework, what we obtain is the updated distribution of δ; this is called the posterior distribution of δ. The basic approach is as follows. Suppose that the data are represented by the vector y; then the posterior distribution is the distribution of δ given y: $p(\delta|y)$. The posterior distribution is computed using Bayes' rule, which states that the posterior distribution is proportional to the product of the prior distribution and the likelihood.

To make this concrete, if the prior on δ is $p(\delta) = Normal(\mu, \sigma)$ and the data are assumed to be generated from some likelihood function that takes δ as a parameter (call this likelihood $f(y|\delta)$), then, following Bayes' rule, the posterior distribution is proportional to the product of the likelihood and the prior:

$$p(\delta|y) \propto f(y|\delta)f(\delta) \tag{3}$$

If there is more than one parameter in the model, then a prior is defined for each parameter, and the posterior is then the joint distribution of the parameters given the data. For example, if the likelihood is the normal distribution, the parameters are the mean μ and the standard deviation σ, and the posterior distributions of these parameters are derived by computing:

$$p(\mu, \sigma|y) \propto Normal(y|\mu, \sigma)f(\mu)f(\sigma) \tag{4}$$

Here, $f(\mu)$ and $f(\sigma)$ are prior distributions defined on these parameters.

In our one-parameter example above, the mean of the posterior distribution $p(\delta|y)$ is a compromise between the frequentist maximum likelihood estimate and the mean of the prior distribution. This is a very important difference from the frequentist approach and allows us to build on prior knowledge; for a practical example of an analysis building on prior knowledge, see Vasishth and Engelmann (2022).

Another important consequence of this fact (that the posterior mean is a compromise between the prior mean and the MLE) is that priors serve to regularize the posterior: when the data are sparse and a parameter cannot be

estimated accurately, the posterior mean will be close to the prior mean. This regularization function of priors has the effect that the convergence warnings that one often sees in the lmer function in the lme4 package will not occur (assuming that a regularizing prior is defined). For more discussion, see chapter 5 of Nicenboim et al. (2022).

Usually, in complex linear mixed models, this posterior distribution is computed using Markov Chain Monte Carlo (MCMC) sampling. To carry out this computation, one uses software such as Stan (Carpenter et al. 2017) or its front-end brms (Bürkner 2017), JAGS (Plummer 2012), or the like.

In Bayesian analysis, a radical change from the frequentist approach is that the Bayesian approach allows us to directly talk about the uncertainty of the effect of interest (δ) once we have seen the data: the posterior distribution gives us this information. In other words, we can now say that we are 95% certain (given the statistical model and the data) that the effect lies between a lower and upper bound.

3.1 Bayesian estimation: A concrete example

To make this approach concrete, consider the Bayesian equivalent of the frequentist linear mixed model.

As a baseline, first consider the frequentist linear mixed model. The reader may be familiar with the following lme4 syntax. Here, logrt is a vector containing log-transformed reading times, int is the two-level factor, coded as ± 0.5 (Schad, Vasishth, Hohenstein and Kliegl 2020).

```
m<-lmer(logrt~int + (1+int|subject) +  (1+int|item),dat)
summary(m)

## Linear mixed model fit by REML ['lmerMod']
## Formula: logrt ~ int + (1 + int | subject) + (1 + int | item)
##    Data: dat
##
## REML criterion at convergence: 15324.5
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.6395 -0.5714 -0.1470  0.4158  4.6127
##
## Random effects:
##  Groups   Name        Variance  Std.Dev. Corr
##  subject  (Intercept) 0.1357114 0.36839
##           int         0.0029512 0.05432  -0.04
##  item     (Intercept) 0.0015438 0.03929
##           int         0.0008783 0.02964  -0.65
##  Residual             0.2231627 0.47240
## Number of obs: 10883, groups:  subject, 184; item, 60
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  6.35091    0.02800 226.825
```

```
## int          0.01964    0.01064   1.845
##
## Correlation of Fixed Effects:
##     (Intr)
## int -0.057
```

The model implied here is:

$$y \sim LogNormal(\alpha + u_1 + w_1 + (\beta + u_2 + w_2) \times int, \sigma) \tag{5}$$

where y is the reading time in milliseconds, u_1, u_2 and w_1, w_2 are, respectively, subject-level and item-level adjustments to the fixed effect intercept α and slope β, with both the u and w adjustments coming from bivariate Normal distributions. For example, u_1 is assumed to have a Normal distribution with mean 0 and standard deviation σ_{u1}, u_2 a Normal distribution with mean 0 and standard deviation σ_{u2}, and the correlation between u_1 and u_2 is ρ_u. The bivariate normal distribution is written like this:

$$\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \sim Normal\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{u_1} & \rho_u \sigma_{u_1} \sigma_{u_2} \\ \rho_u \sigma_{u_1} \sigma_{u_2} & \sigma^2_{u_2} \end{pmatrix}\right) \tag{6}$$

Similarly, the by-item adjustments to the intercept and slope, w_1 and w_2, also have a bivariate Normal distribution defined for them:

$$\begin{pmatrix} w_1 \\ w_2 \end{pmatrix} \sim Normal\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{w_1} & \rho_w \sigma_{w_1} \sigma_{w_2} \\ \rho_w \sigma_{w_1} \sigma_{w_2} & \sigma^2_{w_2} \end{pmatrix}\right) \tag{7}$$

This implies that the frequentist model has the following parameters: the fixed effects α, β, and the variance components $\sigma_{u_1}, \sigma_{u_2}, \sigma_{w_1}, \sigma_{w_2}, \sigma, \rho_u$, and ρ_w.

In the Bayesian version of this model, we will need to define prior distributions for each of these parameters. The prior distributions for all parameters except the correlations are on the log scale:

$$\begin{aligned} \alpha &\sim Normal(6, 0.6 \\ \beta &\sim Normal(0, 0.1) \\ \sigma &\sim Normal_+(0, 0.5) \\ \sigma_{u_{1,2}} &\sim Normal_+(0, 0.1) \\ \sigma_{w_{1,2}} &\sim Normal_+(0, 0.1) \\ \rho_u &\sim LKJ(2) \\ \rho_w &\sim LKJ(2) \end{aligned} \tag{8}$$

Why these priors and not some others? The motivation for these priors is discussed in detail in Schad, Betancourt and Vasishth (2020), but for our purposes here, it is enough to state that these can be shown to be reasonable priors for this particular research question.

The priors for the correlation parameters need some discussion. For these correlations, the so-called LKJ prior is available in the Stan programming language. When the LKJ distribution gets the parameter 2, this specifies a prior that is widely spread out between -1 and $+1$ and has mean 0; see Figure 2 for a visualization. A great advantage of this prior on the correlation is that the mean of the posterior distribution of correlation cannot have extreme values like

+1 or −1. Such extreme values are often seen in frequentist linear mixed models, and represent an estimation failure. The LKJ prior prevents such extreme correlations from occurring because of the shape of the LKJ(2) distribution: these extreme values are heavily downweighted. This is what is meant by regularization in Bayesian methods: a priori unlikely values are downweighted by the prior.

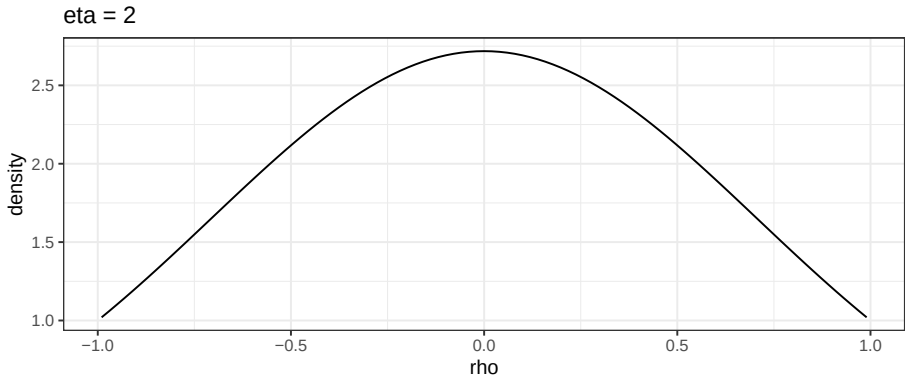

Figure 2: Visualization of the LKJ prior with parameter 2. This is an example of a regularizing prior: extreme values of the correlation value like ±2 are rendered impossible through this prior.

Leaving out the technical details of how the computation is done (see Nicenboim et al. 2022), the estimates of the parameters from the Bayesian linear model are shown in Table 1, with the frequentist estimates shown alongside:

Parameter	Frequentist mean	CI	Bayesian mean	CrI
α	6.35	[6.3, 6.41]	6.35	[6.30, 6.40]
β	0.012	[-0.01, 0.03]	0.02	[-0.00, 0.04]
σ_{u1}	0.37	-	0.36	[0.32, 0.40]
σ_{u2}	0.05	-	0.04	[0.00, 0.08]
ρ_u	-0.04	-	-0.04	[-0.61, 0.54]
σ_{w1}	0.04	-	0.04	[0.03, 0.05]
σ_{w2}	0.03	-	0.03	[0.00, 0.06]
ρ_w	-0.65	-	-0.42	[-0.97, 0.57]
σ	0.47	-	0.47	[0.47, 0.48]

Table 1: Shown are the frequentist and Bayesian estimates from linear mixed models fit in the frequentist and Bayesian linear mixed models. CrI refers to the Bayesian credible interval, and represents the range over which we can be 95% certain that the true value of the parameter lies, given the model and data.

Some important similarities and differences between the frequentist vs. Bayesian estimates:

1. The means of most of the parameters are very similar in both.

2. The mean of correlation parameter for items, ρ_w, is smaller in the Bayesian model; this is an example of the posterior mean being a compromise between the prior mean (0) and the MLE. The posterior mean of ρ_w is being regularized towards 0, because there is not enough data to estimate this parameter accurately. In other words, the frequentist estimate will be most likely an overestimate (Type M error).

3. The Bayesian model provides uncertainty intervals for each parameter; but the lme4 function does not (and cannot even in principle provide such uncertainty intervals, as the parameters are point values and have no distribution). The frequentist model allows us to work out the confidence intervals for the fixed effects, but these intervals are only telling us how variable the sample mean would be under hypothetical repeated sampling; they do not tell us the uncertainty about the true value of the parameters.

On the log scale, the estimate of the effect is 0.019, with 95% credible interval [-0.001, 0.04] log ms. This estimate is not very different from the frequentist one computed using the lme4 package; but the meaning of the credible interval is quite different from that of the confidence interval.

The Bayesian model also allows us to back-transform the posterior distributions of the fixed effects parameters to the millisecond scale (see Nicenboim et al. 2022); this is easier to interpret because a computational model of interference effects makes predictions on the millisecond scale (Vasishth 2020, Vasishth et al. 2019), and because meta-analysis estimates of the interference effect are on the millisecond scale (Jäger et al. 2017). Figure 3 shows this transformed effect estimate from the Bayesian model graphically.

What is interesting about this estimate is that we can now conclude that, given the model and data, the estimate of the interference effect is, with 95% certainty, between -0.6 and 22.85 ms, and the posterior distribution of δ has mean 11.15 ms. The uncertainty interval is called a credible interval. The meta-analysis estimate of this effect (Jäger et al. 2017) is 13 ms, 95% credible interval [2,28]: the observed credible interval in our example data is consistent with the meta-analysis estimate.

The conclusion that the observed posterior distribution of the effect of interest is consistent with predictions is different from claiming that we found *evidence* for the interference effect. Arguing that we have evidence for an effect requires a model comparison using a likelihood ratio test (Royall 1997); I discuss evidence in Bayesian methods below.

A common reaction at this point is to ask: but how can we know that the effect is "reliable" or "real"? As I tried to explain in the first part of this chapter, statistical significance can only answer this question if statistical power is high. In the Bayesian framework, it is in principle possible to carry out a null hypothesis test to attempt to answer this question; this Bayesian test is called the Bayes factor. The Bayes factor is the analog of the frequentist null hypothesis significance test. For authoritative discussions of the Bayes factor, see Lee and Wagenmakers (2014), Wagenmakers, Marsman, Jamil, Ly, Verhagen, Love, Selker, Gronau, Šmíra, Epskamp et al. (2018), Wagenmakers, Love, Marsman, Jamil, Ly, Verhagen, Selker, Gronau, Dropmann, Boutin et al. (2018), Haaf et al. (2019).

The Bayes factor is a ratio that represents the weight of evidence for the effect of interest compared to some null model (such as a model assuming that

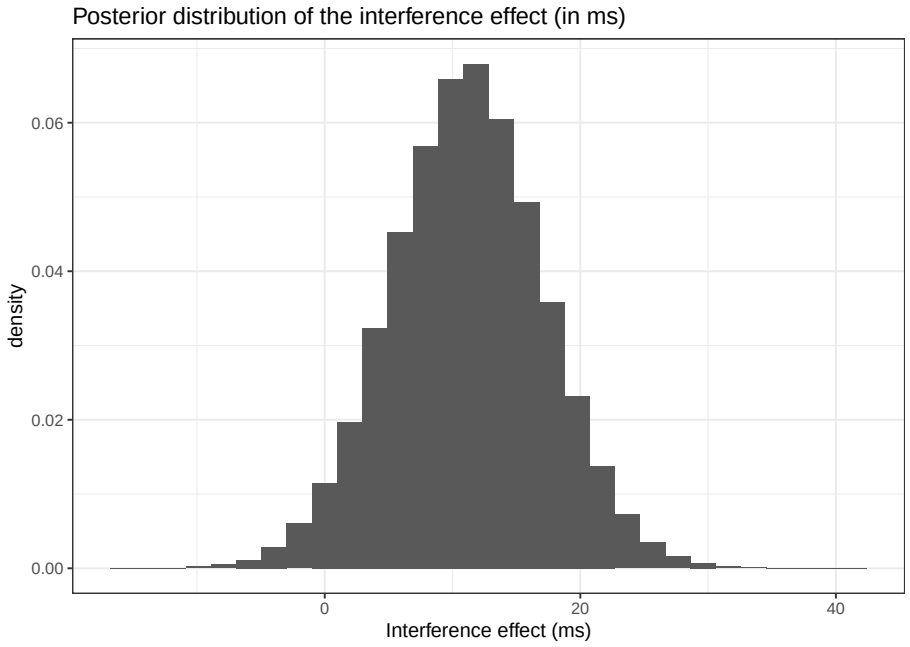

Figure 3: The estimate of the interference effect on the millisecond scale, based on a Bayesian linear mixed model.

the effect is 0). For example, a Bayes factor of 3 means that a model including a parameter representing the effect is three times more likely than a model assuming no effect at all. Some textbooks and articles (e.g., Lee and Wagenmakers 2014) provide a scale for interpreting Bayes factors, but such scales are arbitrary.

The Bayes factor comes at a price (Schad, Nicenboim, Bürkner, Betancourt and Vasishth 2022), the principal one being that it can be very sensitive to the prior specified for the parameter representing the effect (Kass and Raftery 1995). As a consequence, it becomes necessary to report a so-called sensitivity analysis: the Bayes factor is computed under a range of prior specifications for the parameter of interest in the model (in the linear mixed model, this would be the β parameter). Thus, unlike the p-value, a single Bayes factor is almost never informative.

Moreover, the Bayes factor also suffers from the same power problem that we saw with the frequentist p-value; when power is low (e.g., with smaller sample sizes), the Bayes factor can deliver overly strong evidence for an effect (Vasishth, Yadav, Schad and Nicenboim 2022). Further, the Bayes factor can also lead to inconclusive results; for example, a Bayes factor near 1 would be inconclusive. In our running example, the Bayes factor with a relatively constrained prior of Normal(0,0.1) (on the log scale) for the slope parameter β (this represents the interference effect), the Bayes factor is 0.63 in favor of the effect, which is inconclusive. The prior Normal(0,0.1) is relatively constrained because it implies that the effect can range a priori from -115 to $+115$ on the ms scale.

The sensitivity of the Bayes factor to the prior can be illustrated by recomputing the Bayes factor under a range of priors. For example, assume a much wider prior for the β parameter, e.g., Normal(0,1). This prior implies that, a priori, the effect can range from -1345 to $+1345$ ms. Such a prior is sometimes called an uninformative prior. Under such a prior, the Bayes factor is 0.06. This Bayes factor implies that the null hypothesis is 16.3 times more likely than a model assuming that the effect exists! This is an invalid conclusion, and is entirely driven by the a priori assumption that the effect can be in the high hundreds of milliseconds. In general, an uninformative prior will unduly favor the null hypothesis, leading to—as in this case—misleading conclusion.

Despite the limitations of the Bayes factor, when the research question really does boil down to whether the effect is present or absent, the Bayes factor is a good way to evaluate the evidence from the data and is definitely superior to the p-value because it can, in principle, provide evidence for the null (assuming that the study is properly powered). The main issue one must take care of with Bayes factors analyses is to carry out a sensitivity analysis. For more details on this point, see Schad, Nicenboim, Bürkner, Betancourt and Vasishth (2022), Nicenboim et al. (2022). For an example of a sensitivity analysis in psycholinguistics, see Nicenboim et al. (2020).

In summary, a major advantage of adopting the Bayesian approach in experimental linguistics is that uncertainty quantification of the effect of interest–an approach advocated for by prominent psychologists like Meehl, becomes possible. There are of course many other advantages of adopting the Bayesian approach: for example, highly customized and complex models can be fit (Nicenboim et al. 2022).

One final question worth addressing here is: how can one learn enough about Bayesian statistics to be able to use it sensibly in linguistics? Two textbooks accessible to linguists are the following: McElreath (2020),Kruschke (2014). We have also written a textbook, which is available for free online: Nicenboim et al. (2022). I have also created a free online four-week course with video lectures on openhpi.de that covers the first four chapters of Nicenboim et al. (2022).

4 Reproducible code and data

The code and data accompanying this article are available from https://osf.io/kgxpn/.

References

Baayen, R. H. (2008), *Analyzing linguistic data: A practical introduction to statistics using R*, Cambridge University Press.

Baayen, R. H., Davidson, D. J. & Bates, D. M. (2008), 'Mixed-effects modeling with crossed random effects for subjects and items', *Journal of Memory and Language* **59**(4), 390–412.

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

Belia, S., Fidler, F., Williams, J. & Cumming, G. (2005), 'Researchers misunderstand confidence intervals and standard error bars.', *Psychological methods* **10**(4), 389.

Blokpoel, M. & van Rooij, I. (2021), *Theoretical Modeling for cognitive science and psychology*. Retrieved December 4, 2022.
URL: *https://computationalcognitivescience.github.io/lovelace/*

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

Bürki, A., Alario, F.-X. & Vasishth, S. (2022), 'When words collide: Bayesian meta-analyses of distractor and target properties in the picture-word interference paradigm', *Quarterly Journal of Experimental Psychology* . Accepted.

Bürki, A., Elbuy, S., Madec, S. & Vasishth, S. (2020), 'What did we learn from forty years of research on semantic interference? A Bayesian meta-analysis', *Journal of Memory and Language* **114**, 104125.

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

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P. & Riddell, A. (2017), 'Stan: A probabilistic programming language', *Journal of statistical software* **76**(1).

Cassidy, S. A., Dimova, R., Giguère, B., Spence, J. R. & Stanley, D. J. (2019), 'Failing grade: 89% of introduction-to-psychology textbooks that define or explain statistical significance do so incorrectly', *Advances in Methods and Practices in Psychological Science* **2**(3), 233–239.

Chemla, E. (2009), 'Presuppositions of quantified sentences: Experimental data', *Natural language semantics* **17**(4), 299–340.

Clark, H. (1973), 'The Language-as-Fixed-Effect Fallacy: A Critique of Language Statistics in Psychological Research.', *Journal of Verbal Learning and Verbal Behavior* **12**(4), 335–59.

Cumming, G. (2014), 'The new statistics: Why and how', *Psychological Science* **25**(1), 7–29.

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

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

Gibson, E. A. & Fedorenko, E. G. (2010), 'Weak quantitative standards in linguistics research', *Trends in Cognitive Sciences* **14**(6), 233–234.

Gibson, E. & Fedorenko, E. (2013), 'The need for quantitative methods in syntax and semantics research', *Language and Cognitive Processes* **28**(1-2), 88–124.

Haaf, J., Ly, A. & Wagenmakers, E.-J. (2019), 'Retire significance, but still test hypotheses', *Nature* **567**(7749), 461.

Hackl, M., Koster-Hale, J. & Varvoutis, J. (2012), 'Quantification and acd: Evidence from real-time sentence processing', *Journal of Semantics* **29**(2), 145–206.

Hoekstra, R., Morey, R. D., Rouder, J. & Wagenmakers, E.-J. (2014), 'Robust misinterpretations of confidence intervals', *Psychonomic Bulletin and Review* pp. 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**, 19–24.

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. (2020), 'Interference patterns in subject-verb agreement and reflexives revisited: A large-sample study', *Journal of Memory and Language* **111**.

Kass, R. E. & Raftery, A. E. (1995), 'Bayes factors', *Journal of the american statistical association* **90**(430), 773–795.

Kliegl, R., Masson, M. E. & Richter, E. M. (2010), 'A linear mixed model analysis of masked repetition priming', *Visual Cognition* **18**(5), 655–681.

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

Lee, M. D. & Wagenmakers, E.-J. (2014), *Bayesian cognitive modeling: A practical course*, Cambridge University Press.

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

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

McElreath, R. (2020), *Statistical rethinking: A Bayesian course with examples in R and Stan*, CRC Press.

McShane, B. B., Gal, D., Gelman, A., Robert, C. & Tackett, J. L. (2019), 'Abandon statistical significance', *The American Statistician* **73**(sup1), 235–245.

Meehl, P. E. (1997), The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions, *in* L. Harlow, S. Mulaik & J. H. Steiger, eds, 'What if there were no significance tests?', Erlbaum, Mahwah, New Jersey.

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., Schad, D. J. & Vasishth, S. (2022), *Introduction to Bayesian Data Analysis for Cognitive Science*. Under contract with Chapman and Hall/CRC Statistics in the Social and Behavioral Sciences Series.
URL: *https://vasishth.github.io/bayescogsci/*

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**.

Nicenboim, B., Vasishth, S. & Rösler, F. (2020), 'Are words pre-activated probabilistically during sentence comprehension? evidence from new data and a Bayesian random-effects meta-analysis using publicly available data', *Neuropsychologia* **142**.

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.

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Kline Struhl, M., Nuijten, M. B. et al. (2022), 'Replicability, robustness, and reproducibility in psychological science', *Annual review of psychology* **73**, 719–748.

Oakley, J. E. & O'Hagan, A. (2010), *SHELF: The Sheffield Elicitation Framework (version 2.0)*, School of Mathematics and Statistics, University of Sheffield, University of Sheffield, UK.
URL: *http://tonyohagan.co.uk/shelf*

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

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

Pinheiro, J. C. & Bates, D. M. (2000), *Mixed-Effects Models in S and S-PLUS*, Springer-Verlag, New York.

Plummer, M. (2012), 'JAGS version 3.3.0 manual', *International Agency for Research on Cancer. Lyon, France* .

Royall, R. (1997), *Statistical Evidence: A likelihood paradigm*, Chapman and Hall, CRC Press, New York.

Schad, D. J., Betancourt, M. & Vasishth, S. (2020), 'Toward a principled Bayesian workflow in cognitive science', *Psychological Methods* **26**(1), 103–126.

Schad, D. J., Nicenboim, B., Bürkner, P.-C., Betancourt, M. & Vasishth, S. (2022), 'Workflow techniques for the robust use of Bayes factors', *Psychological Methods* .

Schad, D. J., Nicenboim, B. & Vasishth, S. (2022), Data aggregation can lead to biased inferences in Bayesian linear mixed models.

Schad, D. J., Vasishth, S., Hohenstein, S. & Kliegl, R. (2020), 'How to capitalize on a priori contrasts in linear (mixed) models: A tutorial', *Journal of Memory and Language* **110**.

Sprouse, J., Wagers, M. W. & Phillips, C. (2012), 'A test of the relation between working-memory capacity and syntactic island effects', *Language* **88**(1), 82–123.

Tendeiro, J., Kiers, H., Hoekstra, R., Wong, T. K. & Morey, R. D. (2022), 'Diagnosing the use of the Bayes factor in applied research'.

Vasishth, S. (2020), 'Using Approximate Bayesian Computation for estimating parameters in the cue-based retrieval model of sentence processing', *MethodsX* .
URL: *10.1016/j.mex.2020.100850*

Vasishth, S. (2022), Some right ways to analyze (psycho)linguistic data. Submitted.

Vasishth, S. & Engelmann, F. (2022), *Sentence Comprehension as a Cognitive Process: A Computational Approach*, Cambridge University Press, Cambridge, UK.
URL: *https://vasishth.github.io/RetrievalModels*

Vasishth, S. & Gelman, A. (2021), 'How to embrace variation and accept uncertainty in linguistic and psycholinguistic data analysis', *Linguistics* **59**, 1311–1342.

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., Mertzen, D., Jäger, L. A. & Gelman, A. (2018), 'The statistical significance filter leads to overoptimistic expectations of replicability', *Journal of Memory and Language* **103**, 151–175.
URL: *https://osf.io/eyphj/*

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., Engelmann, F. & Burchert, F. (2019), 'Computational models of retrieval processes in sentence processing', *Trends in Cognitive Sciences* **23**, 968–982.

Vasishth, S., Schad, D. J., Bürki, A. & Kliegl, R. (2022), *Linear Mixed Models for Linguistics and Psychology: A Comprehensive Introduction*. Under contract with Chapman and Hall/CRC Statistics in the Social and Behavioral Sciences Series.
URL: *https://vasishth.github.io/Freq_CogSci/*

Vasishth, S., Yadav, H., Schad, D. & Nicenboim, B. (2022), 'Sample size determination for Bayesian hierarchical models commonly used in psycholinguistics', *Computational Brain and Behavior* .

Von Baeyer, H. C. (1988), 'How Fermi would have fixed it', *The Sciences* **28**(5), 2–4.

Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Selker, R., Gronau, Q. F., Dropmann, D., Boutin, B. et al. (2018), 'Bayesian inference for psychology. Part II: Example applications with JASP', *Psychonomic bulletin & review* **25**(1), 58–76.

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., Selker, R., Gronau, Q. F., Šmíra, M., Epskamp, S. et al. (2018), 'Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications', *Psychonomic Bulletin & Review* **25**(1), 35–57.

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.

Winter, B. (2019), *Statistics for Linguists: An Introduction Using R*, Routledge.