

# Statistical methods for linguistic research: Foundational Ideas – Part II

Bruno Nicenboim

University of Potsdam, Potsdam, Germany

Shravan Vasishth

University of Potsdam, Potsdam, Germany and

CEREMADE (Centre de Recherche en Mathématiques de la Décision), Université Paris

Dauphine, Paris France

July 31, 2016

## Abstract

We provide an introductory review of Bayesian data analytical methods, with a focus on applications for linguistics, psychology, psycholinguistics, and cognitive science. The empirically oriented researcher will benefit from making Bayesian methods part of their statistical toolkit due to the many advantages of this framework, among them easier interpretation of results relative to research hypotheses, and flexible model specification. We present an informal introduction to the foundational ideas behind Bayesian data analysis, using, as an example, a linear mixed models analysis of data from a typical psycholinguistics experiment. We discuss hypothesis testing using the Bayes factor, and model selection using cross-validation. We close with some examples illustrating the flexibility of model specification in the Bayesian framework. Suggestions for further reading are also provided.

*Keywords:* Bayesian data analysis, linguistics, linear mixed models, Bayes factor, Model selection

## Introduction

In Part I of this review, we presented the main foundational ideas of frequentist statistics, with a focus on their applications in linguistics and related areas in cognitive science. Our main goal there was to try to clarify the underlying ideas in frequentist methodology. We discussed the meaning of the p-value, and of Type I and II errors and Type S and M errors. We also discussed some common misinterpretations associated with hypothesis testing in this framework.

There is no question that frequentist data-analytic methods such as the t-test, ANOVA, etc. are and will remain an important part of the toolkit of the experimentally inclined linguist. There is, however, another framework that is not yet part of the standard statistics curriculum in linguistics, but can be of great value: Bayesian data analysis. Learning to use this framework for data analysis is easier than it seems; in fact, most of us already think like Bayesians when we carry out our frequentist analyses. One example of this is the (incorrect) interpretation of the p-value as the probability of the null hypothesis being true. If we are faced with a p-value of 0.06 (i.e., a value just above the threshold of statistical significance), we often resort to expressions of gradedness, saying that “the result was marginally significant”, or “the results did not reach the conventional level of significance”, or even “a non-significant trend towards significance.”<sup>1</sup> This desire to take the non-significant result as meaningful comes about because of an irresistible temptation to (erroneously) give a Bayesian interpretation to the p-value as the probability of the null hypothesis being true. As discussed in Part I, the p-value is the probability that the statistic is at least as extreme as the one observed given that the null is true. When we misinterpret the p-value in terms of the probability of the null being true, a p-value of 0.06 doesn't seem much different from 0.04. Several studies have shown that such interpretations

---

<sup>1</sup>For a list of over 500 remarkably imaginative variants on this phrasing, see <https://mchankins.wordpress.com/2013/04/21/still-not-significant-2/>.

of the p-value are not uncommon (among many others: Haller & Krauss, 2002; Lecoutre, Poitevineau, & Lecoutre, 2003). The interpretation of frequentist confidence intervals also suffers from similar problems (Hoekstra, Morey, Rouder, & Wagenmakers, 2014). Given that the Bayesian interpretation is the more natural one that we converge to anyway, why not simply do a Bayesian analysis? One reason that Bayesian methods have not become mainstream may be that, until recently, it was quite difficult to carry out a Bayesian analysis except in very limited situations. With the increase in computing power, and the arrival of several probabilistic programming languages, it has become quite easy to carry out relatively complicated analyses.

The objective of this paper is to provide a non-technical but practically oriented review of some of the tools currently available for Bayesian data analysis. Since the linear mixed model (LMM) is so important for linguistics and related areas, we focus on this model and mention some extensions of LMMs. We assume here that the reader has fitted frequentist linear mixed models (Bates, Maechler, Bolker, & Walker, 2015). The review could also be used to achieve a better understanding of papers that use Bayesian methods for statistical inference.

We start the review by outlining what we consider to be the main advantages of adopting Bayesian data-analytic methods. Then we informally outline the basic idea behind Bayesian statistics: the use of Bayes' theorem to incorporate prior information to our results. Next, we review several ways to verify the plausibility of a research hypothesis with a simple example from psycholinguistics using Bayesian linear mixed models. In the second part of the paper, we discuss some example applications involving standard and less standard (but very useful) models. In the final section, we suggest some further readings that provide a more detailed presentation.

### **Why bother to learn Bayesian data analysis?**

Statisticians have been arguing for decades about the relative merits of the frequentist vs Bayesian statistical methods. We feel that both approaches, used as intended, have their

merits, but that the importance of Bayesian approaches remains greatly underappreciated in linguistics and related disciplines.

We see two main advantages to using Bayesian methods for data analysis. First, Bayesian methods allow us to directly answer the question we are interested in: How plausible is our hypothesis given the data? We can answer this question by quantifying our uncertainty about the parameters of interest. Second, and perhaps more importantly, it is easier to flexibly define hierarchical models (also known as mixed effects or multilevel models) in the Bayesian framework than in the frequentist framework. Hierarchical models, whether frequentist or Bayesian, are highly relevant for the repeated measures designs used in linguistics and psycholinguistics, because they take both between- and within-group variances into account, and because they pool information via “shrinkage” (see Gelman, Hill, & Yajima, 2012). These properties have the desirable effects that we avoid overfitting the data, and we avoid averaging and losing valuable information about group-level variability (Gelman & Hill, 2007 provide more details). For example, both subjects and items contribute independent sources of variance in a standard linguistics repeated measures design. In a hierarchical model, both these sources of variance can be included simultaneously. By contrast, in repeated measures ANOVA, one has to aggregate by items (subjects), which artificially eliminates the variability between items (subjects). This aggregation leads to artificially small standard errors of the effect of interest, which leads to an increase in Type I error.

The frequentist linear mixed model standardly used in psycholinguistics is generally fit with the `lme4` package (Bates, Maechler, et al., 2015) in R. However, if we want to include the maximal random effects structure justified by the design (Barr, Levy, Scheepers, & Tily, 2013; Schielzeth & Forstmeier, 2009), these models tend to not converge or to give unrealistic estimates of the correlations between random effects (Bates, Kliegl, Vasishth, & Baayen, 2015). In contrast, the maximal random effects structure can be fit without problems using Bayesian methods, as discussed later in this review (also see Bates, Kliegl, et al., 2015; Chung, Gelman, Rabe-Hesketh, Liu, & Dorie, 2013; Sorensen, Hohenstein, & Vasishth,

2015). In addition, Bayesian methods can allow us to hierarchically extend virtually any model: non-linear models (which are not generalized linear models) and even the highly complex models of cognitive processes (Lee, 2011; Shiffrin, Lee, Kim, & Wagenmakers, 2008). See also the discussion about hierarchical models in the section entitled Examples of applications of Bayesian methods.

### Bayesian data analysis: An informal introduction

In linguistics, we are usually interested in determining whether there is an effect of a particular factor on some dependent variable; an example from psycholinguistics is the difference in processing difficulty between subject and object relative clauses as measured by reading times. In the frequentist paradigm, we assume that there is some unknown point value  $\mu$  that represents the difference in reading time between the two relative clause types; the goal is to reject the null hypothesis that this true  $\mu$  has the value 0. In the Bayesian framework, our goal is to obtain an estimate of  $\mu$  given the data along with an uncertainty estimate (such as a credible interval, discussed in detail below) that gives us a range over which we can be reasonably sure that the true parameter value lies. We obtain these estimates given the data and given our prior knowledge/information about plausible values of  $\mu$ . We elaborate on this idea in the next section when we present a practical example, but the essential point is that the distribution of  $\mu$  (called the posterior distribution) can be expressed in terms of the prior and likelihood:<sup>2</sup>

$$\text{Posterior} \propto \text{Prior} \times \text{Likelihood} \quad (1)$$

To repeat, given some prior information about the parameter of interest, and the likelihood, we can compute the posterior distribution of the parameter. The focus is not

---

<sup>2</sup>The term likelihood may be unfamiliar. For example, if we have  $n$  independent data points,  $x_1, \dots, x_n$  which are assumed to be generated from a Normal distribution with parameters  $\mu$  and  $\sigma$  and a probability density function  $f(\cdot)$ , the joint probability of these data points is the product  $f(x_1) \times f(x_2) \times \dots \times f(x_n)$ . The value of this product is a function of different values of  $\mu$  and  $\sigma$ , and it is common to call this product the Likelihood function, and it is often written  $L(x_1, \dots, x_n; \mu, \sigma^2)$ . The essential idea here is that the likelihood tells us the joint probability of the data for different values of the parameters.

on rejecting a null hypothesis but on what the posterior distribution tells us about the plausible values of the parameter of interest. Recall that frequentist significance testing is focused on calculating the p-value  $P(\textit{statistic} \mid \mu = 0)$ , that is, the probability of observing a test statistic (such as a t-value) at least as extreme as the one we observed given that the null hypothesis that  $\mu = 0$  is true. By contrast, Bayesian statistics allows us to talk about plausible values of the parameter  $\mu$  given the data, through the posterior distribution of the parameter. Another important point is that the posterior is essentially a weighted mean of the prior and the likelihood. The implications of this statement are made clear graphically in Figure 1. Here, we see the effect of two types of prior distributions on binomial data given 10 or 100 observations. Two points are worth noting. First, when the prior is spread out over a wide range and assign equal probability to all possible values, the posterior distribution ends up closer to the likelihood; this alignment to the likelihood is more pronounced when we have more data (larger sample size). Second, when we have weakly informative priors, with sparse data, the posterior is closer to the prior, but with more data, the posterior is again closer to the likelihood. What this implies is that when we have little data, it is worth investing time in developing priors informed by prior knowledge; but when we have a lot of data, the likelihood will dominate in determining the posterior (we return to this point later, with an example). Indeed, in large-sample situations, we will usually find that the Bayesian posterior and the frequentist mean, along with their uncertainty estimates, are nearly identical or very similar (even though their meaning is quite different—see the discussion below on Bayesian credible intervals).

Bayes' theorem is just a mathematical rule that allows us to calculate any posterior distribution. In practice, however, this is true for a very limited number of cases, and, in fact, the posterior of many of the models that we are interested in cannot be derived analytically. Fortunately, the posterior distribution can be approximated with numerical techniques such as Markov Chain Monte Carlo (MCMC). Many of the probabilistic programming languages freely available today (see the final section for a listing) allow us define our models without having to acquire expert knowledge about the relevant numerical techniques.

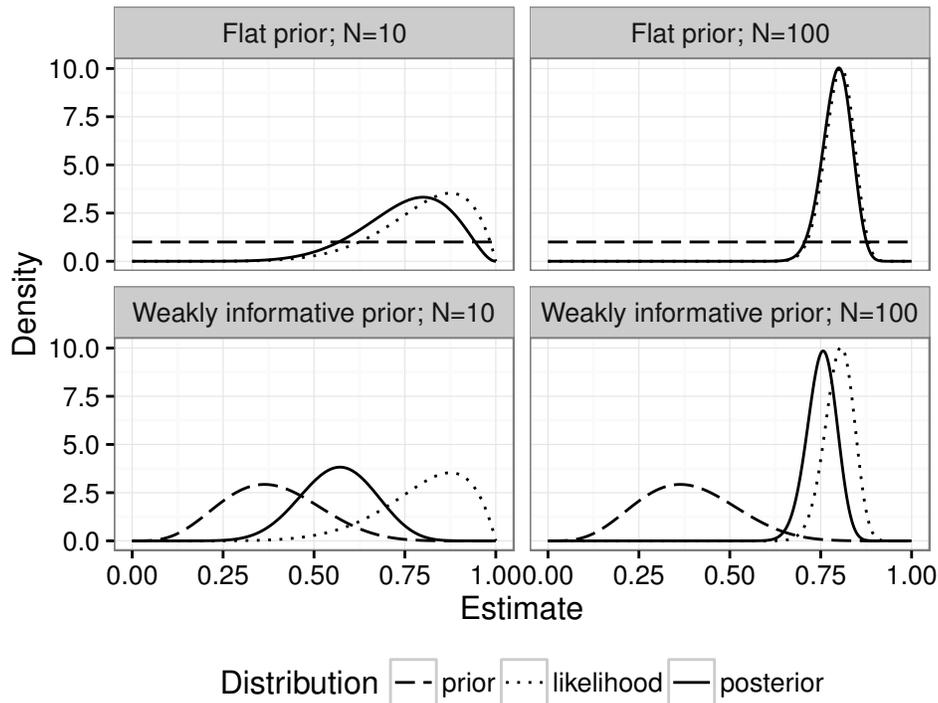


Figure 1. Posterior distributions given different likelihoods and priors for binomial data.

It is all very well to talk about Bayesian methods in the abstract, but how can they be used by linguists and psycholinguists? To answer this question, consider a concrete example from psycholinguistics. We feel that it is easier to show an example that the reader can relate to in order to convey a feeling for how a Bayesian analysis would work; for further study, excellent introductory textbooks are available (we give some suggestions in the final section).

### An example of statistical inference using Bayesian methods

In contrast to significance testing in frequentist statistics, Bayesian inference is not necessarily about dichotomous decisions (reject null or fail to reject null), but rather about the evidence for and against different hypotheses. We illustrate one way to carry out statistical inference using a Bayesian linear mixed model with the data from Gibson and Wu (2013), which has a simple two-condition repeated measures design. Gibson and Wu (2013)

compared reading times at the head noun for subject and object relative clauses and argued for facilitation in the case of object relative clauses (providing a counter-example to the cross-linguistic generalization that subject relative clauses are easier to process than object relative clauses; also see Hsiao & Gibson, 2003). We will assume, as we have done elsewhere (Sorensen et al., 2015), that the dependent variable, reading times, has a lognormal distribution, and thus we will use log reading times as the dependent variable.<sup>3</sup> We fit the linear mixed model with the full random effects structure justified by the design (Barr et al., 2013), and we code object relative clauses with 1 and subject relative clauses with  $-1$ . With this sum contrast coding, Gibson and Wu’s prediction that object relative clauses are easier than subject relative clauses in Chinese would, if correct, result in an effect with a negative sign (i.e. shorter reading times for the condition coded as 1). In the Gibson and Wu dataset, we have 37 participants and 15 items; due to some missing data, we have a total of 547 data points. The conditions were presented to participants in a counterbalanced manner using a Latin square.

The researcher familiar with `lme4` (for details see Bates & Sarkar, 2007 and `lme4` documentation: Bates, Maechler, et al., 2015) will not find the transition to the Bayesian approach difficult. In `lme4` syntax, the model we would fit would be

```
lmer(log(rt) ~ cond + (cond|subj)+(cond|item))
```

We can fit an analogous Bayesian linear mixed model with the `stan_lmer` function from the `rstanarm` package (Gabry & Goodrich, 2016); see the code in Listing 1. The main novelty in the syntax is the specification of the priors for each parameter. Some other details need to be specified, such as the desired number of chains and iterations that the MCMC algorithm requires to converge to the posterior distribution of the parameter of interest. To speed up computation, the number of processors (cores) available in their computer can also be specified. In the example shown in Listing 1, we have left other parameters of the MCMC algorithm at the default values, but they may need some fine tuning in case of

---

<sup>3</sup>Notice, however, that this is not necessarily the best characterization of latencies; see Nicenboim, Logačev, Gattei, & Vasishth, 2015; Ratcliff, 1993; Rouder, 2005.

non-convergence (this is announced by a warning message).

A comparison of the estimates from the `lme4` “maximal” LMM and the analogous Bayesian LMM are shown in Table 1. An important point to notice is that correlations between the varying intercepts and slopes in the fitted model using the `lmer` function are on the boundary; although this does not register a convergence failure warning in `lmer`, such boundary values constitute a failure to estimate the correlations (Bates, Kliegl, et al., 2015). This failure to estimate the correlations is due to the sparsity of data: we have only 37 subjects and even fewer items (15) and are asking too much of the `lmer` function. In the Bayesian LMM, the prior specified on the correlation matrix of the random effects ensures that if there is insufficient data, the posterior will have a mean correlation near 0 with a wide uncertainty associated with it (this is the case in the items intercept-slope correlation), and if there is more data, the posterior will be a compromise between the prior and the likelihood, although the uncertainty associated with this estimate may still be high (this is the case in the subjects intercept-slope correlation). See Sorensen et al. (2015) for more discussion on this point.

### Prior specification

Since priors are an important part of the model, it is worth saying a few words about them. In order to fit a Bayesian model, we need to specify a prior distribution on each parameter; these priors express our initial state of knowledge about the possible values that the parameter can have. It is possible to specify completely *uninformative priors*, such as flat priors, but these are far from ideal since they concentrate too much probability mass outside of any reasonable posterior values. This can have the consequence that without enough data, this prior will dominate in determining the posterior mean and the uncertainty associated with it (Gelman, 2006). Priors that give some minimal amount of information improve inference and are called *regularizing* or *weakly informative priors* (see also Chung et al., 2013; Gelman, Jakulin, Pittau, & Su, 2008).

In the typical psycholinguistic experiment, different weakly informative priors gener-

```

1 library(rstanarm)
2 dgw <- read.table("gibsonwu2012data.txt")
3 dgw_hn <- subset(dgw, subset = region == "headnoun")
4 dgw_hn$cond <- ifelse(dgw_hn$type == "obj-ext", 1, -1)
5 m1 <- stan_lmer(formula = log(rt) ~ cond + (cond | subj) + (cond | item),
6                 prior_intercept = normal(0, 10),
7                 prior = normal(0, 1),
8                 prior_covariance = decov(regularization = 2),
9                 data = dgw_hn,
10                chains = 4,
11                iter = 2000,
12                cores = 4)
13 #summary(m1) # Very long summary with all the parameters in the model

```

Listing 1: Code for fitting a linear mixed model with `stan_lmer`. A major difference from `lme4` syntax is that priors are specified for (a) the intercept, (b) the slope, and (c) the variance-covariance matrices the random effects for subject and item. The other specifications, regarding chains and iterations, relate to the how samples are taken from the posterior distribution, and specifying the number of cores can speed up computation.

Table 1

*Comparison of the frequentist and Bayesian model estimates for the Gibson and Wu dataset. The estimates for both the coefficients and the variance components are comparable, but notice that the correlations between varying intercepts and slopes are quite different in the frequentist and Bayesian models. The `stan_lmer` packages provides the median and the standard deviation of the median absolute difference (MAD) for the fixed effects, but one could equally well compute the mean and standard error, to mirror the `lme4` convention*

<b>lmer</b>				<b>stan_lmer</b>	
<u>Random effects</u>				<u>Random effects</u>	
Groups	Name	Std.Dev.	Corr	Std.Dev.	Corr
subj	(Intercept)	0.2448		0.2425	
	cond	0.0595	-1.00	0.0762	-0.521
item	(Intercept)	0.1820		0.1829	
	cond	0.0002	1.00	0.0475	0.012
Residual		0.5143		0.5131	
<u>Fixed effects</u>				<u>Fixed effects</u>	
	Estimate	Std. Error	t value	Median	MAD-SD
(Intercept)	6.06180	0.0657	92.24	6.0641	0.0658
cond	-0.03625	0.0242	-1.50	-0.0364	0.0301

ally don't have much of an effect on the posterior, but it is a good idea to do a *sensitivity analysis* by evaluating the effect of different priors on the posterior; see Levshina (2016); Vasishth, Chen, Li, and Guo (2013) for examples from linguistics. Another use of priors is to include valuable prior knowledge about the parameters into our model; these *informative priors* could come from actual prior experiments (the posterior from previous experiments) or from meta-analyses (Vasishth et al., 2013), or from expert judgements (O'Hagan et al., 2006; Vasishth, 2015). Such a use of priors is not widespread even among Bayesian statisticians, but could be a powerful tool for incorporating prior knowledge into a new data analysis. For example, starting with relatively informative priors could be a huge improvement over starting every new study on relative clauses with the assumption that we know nothing about the topic. In addition there might be computational reasons to choose a particular prior, that is, some specific distributions are chosen as priors because they can make sampling from the posterior more efficient (see, for example, Ghosh, Li, & Mitra, 2015).

Returning to the practical issue of prior specification in R functions like, `stan_lmer`, it is a good idea to specify the priors explicitly; the default priors assumed by the function may not be appropriate for your specific data. For example, in our example above, if we set the prior for the intercept as a normal distribution with a mean of zero and a standard deviation of ten,  $\text{Normal}(0,10)$ ,<sup>4</sup> we are assuming that we are 68% sure that the grand mean in log-scale will be between  $-10$  (very near zero milliseconds) and  $10$  ( $\approx 22$  seconds). This prior is extremely vague, and can be changed but it won't affect the results much. However, notice that had we not log-transformed the dependent variable, we would be assuming that we are 68% sure that the grand mean is between  $-10$  and  $10$  milliseconds, and 95% sure that it is between  $-20$  and  $20$  milliseconds! We are guaranteed to not get sensible estimates if we do that.

We start the analysis by setting  $\text{Normal}(0,1)$  as the prior for the effect of subject vs. object relative clauses. This means that we assume that we are 68% certain that the

---

<sup>4</sup>This happens to be the default value in this version of `rstanarm`, but this might change in future versions.

difference between the conditions should be less than 1300 ms.<sup>5</sup> In reality, the range is likely to be much smaller, but we will start with this prior. We will exemplify the effect of the priors by changing the prior of the estimate for the effect of the experimental condition in the next section. The reader may notice that there is also a prior for the covariance matrix of the random effects (Chung et al., 2013); the regularization parameter in this prior specification to a value larger than one can help to get conservative estimates of the intercept-slope correlations when we don't have enough data; recall the discussion regarding Table 1 above. For further examples, see Bates, Kliegl, et al. (2015) and the vignettes in the `RePsychLing` package (<https://github.com/dmbates/RePsychLing>). A more in-depth discussion is beyond the scope of this paper, and the reader is referred to the tutorial by Sorensen et al. (2015) and a more general discussion by Chung et al. (2013).<sup>6</sup>

### The posterior and statistical inference

As mentioned earlier, the result of a Bayesian analysis is a posterior distribution, that is, a distribution showing the relative plausibilities of each possible value of the parameter of interest, conditional on the data, the priors, and the model. Every parameter of the model (fixed effects, random effects, shape of the distribution) will have a posterior distribution. Typically, software such as the `rstanarm` package in R will deliver samples from the posterior that we can use for inference. In our running example, we will focus on the posterior distribution of the effect of object relative clauses in comparison with subject relative clauses; see Figure 2.

To communicate our results, we need to summarize and interpret the posterior distribution. We can report point estimates of the posterior probability such as the mean or the median (in some cases also the mode of the distribution, known as the *maximum a posteriori* or MAP, is also reported). When the posterior distribution is symmetrical and

<sup>5</sup>The mean reading time at the critical region is 550ms, and since we assumed a lognormal distribution, to calculate the ms, we need to find out  $\exp(\log(550) + 1) - \exp(\log(550) - 1)$ , which is 1300; the 68% is the probability mass between [-1,1] in the standard normal distribution.

<sup>6</sup>We left the priors of other parameters such as the standard deviation of the distribution, or residuals in `lme4` terms, and the scale of the random effects at their default values. These priors shouldn't be ignored when doing a real analysis! The user should verify that they make sense.

```

14 samples_m1 <- as.data.frame(m1) # It saves all the samples from the model.
15 posterior_condition <- samples_m1$cond
16 options(digits = 4)
17 mean(posterior_condition)
## [1] -0.0358
19 median(posterior_condition)
## [1] -0.03573

```

Listing 2: Code for summarizing point estimates

```

21 mean(posterior_condition < 0) # Posterior probability that lies below zero.
## [1] 0.89

```

Listing 3: Code for finding the mass of the posterior probability that lies below zero

approximately normal in shape, the mean and median almost converge to the same point (see Figure 2). In this case it won't matter which one we report, since they only differ in the fourth decimal digit; see Listing 2.

It is also important to summarize the amount of posterior probability that lies below or above some parameter value. In Gibson and Wu's data, since the research question amounts to whether the parameter is positive or negative, and Gibson and Wu predict that it will be negative, we can compute the posterior probability that the difference between object and subject relative clauses is less than zero,  $P(\hat{\beta} < 0)$ . This probability is 0.89; see Listing 3 and also Figure 3(a).

There is nothing special about zero; and since the difference between English object and subject relative clauses is, in general, quite large, and Gibson and Wu predict that in Chinese the same processes as in English give an advantage to object relatives over subject relative clauses, we could be interested, instead, in knowing the probability that the advantage of object relative clauses is at least 20 ms. This advantage can be translated

into approximately  $-0.02$  from the grand mean in log-scale; we can inspect the posterior distribution and find out that this,  $P(\hat{\beta} < -0.02)$ , is 0.67; see also Figure 3(b).

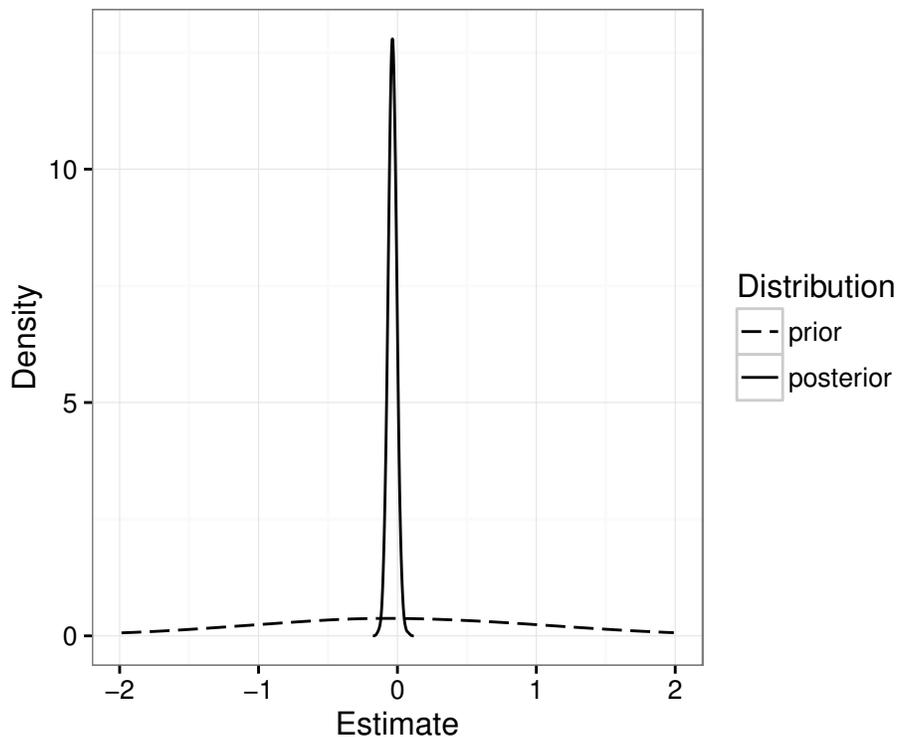


Figure 2. Posterior distribution of the difference between object and subject relative clauses given a prior distribution  $\text{Normal}(0,1)$ .

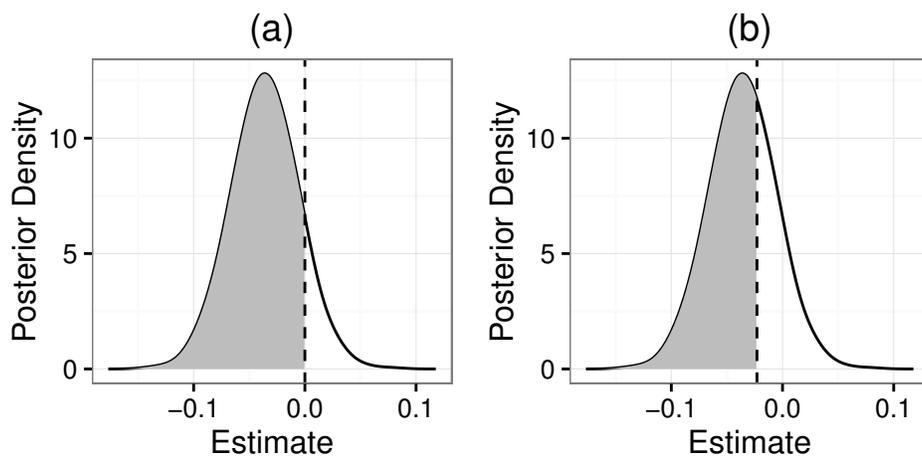


Figure 3. Posterior probability that the difference between object and subject relative clauses is less than zero (a), and less than  $-0.02$  (b).

### The 95% credible interval

It is possible (and desirable) to report an interval of posterior probability, that is, two parameter values that contain between them a specified amount of posterior probability. This type of interval is also known as *credible interval*. A credible interval demarcates the range within which we can be certain with a certain probability that the “true value” of a parameter lies. It is the true value not out there “in nature”, but true in the model’s logical world (see also the interesting distinction between *small* and *large worlds* in McElreath, 2015).

The Bayesian credible interval is different from the frequentist confidence interval because the credible interval can be interpreted with the data at hand, while the frequentist counterpart is a property of the statistical procedure. The statistical procedure only indicates that frequentist confidence intervals across a series of hypothetical data sets produced by the same underlying process will contain the true parameter value in a certain proportion of the cases (Hoekstra et al., 2014; Morey, Hoekstra, Rouder, Lee, & Wagenmakers, 2015).

The two most common types of Bayesian credible intervals are the percentile interval and highest posterior density interval (HPDI; Box & Tiao, 1992). In the first one, we assign equal probability mass to each tail. This is the most common way to report credible intervals, because non-Bayesian intervals are usually percentile intervals. As with frequentist confidence intervals, it is common to report 95% intervals (see Figure 4). The second option is to report the HDPI, that is the narrowest interval containing the specified probability mass. This interval will show the parameter values most consistent with the data, but it can be noisy and depends on the sampling process (Y. Liu, Gelman, & Zheng, 2013). When the posterior is symmetrical and normal looking, it is very similar to the percentile interval: In the case of Gibson and Wu’s data, the difference between them is in the second or third decimal digit; see Listing 4.

```
23 options(digits = 4)
24 posterior_interval(m1, par = "cond", prob = 0.95) # 95% Percentile Interval
##           2.5%   97.5%
## cond -0.09553 0.02427
27 library(SPIin) # For calculating the HPDI
28 bootSPIin(posterior_condition)$spin # 95% HPDI
## [1] -0.09563 0.02355
```

Listing 4: Code for 95% Percentile Interval and HPDI

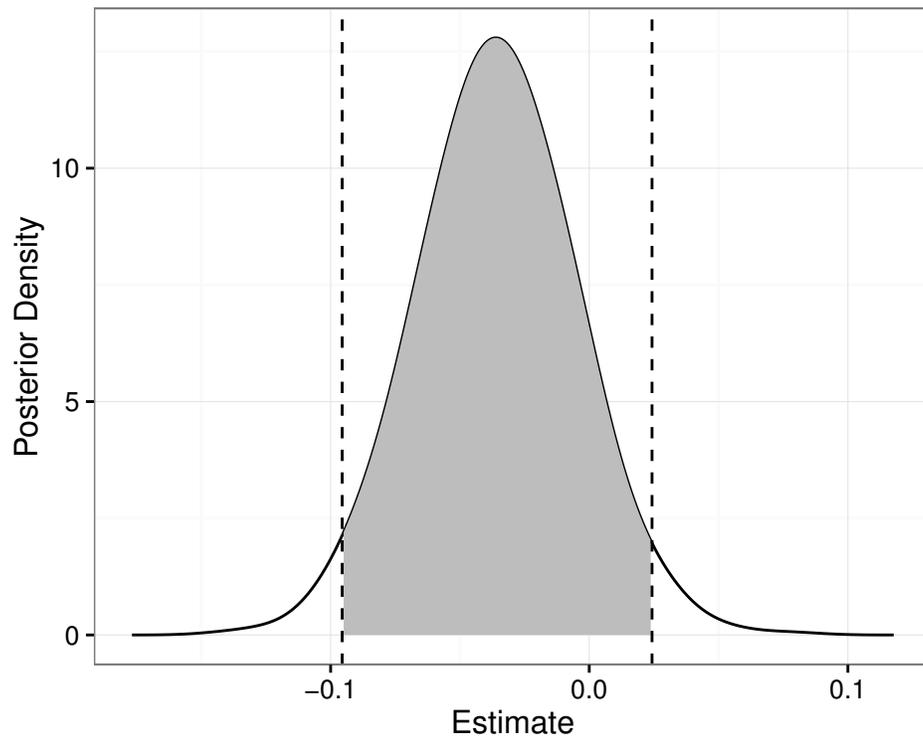


Figure 4. 95% Credible Interval.

### Investigating the effect of prior specification on posteriors

So far we dealt with a very weakly informative prior; what would happen with more informative ones? Let us start by choosing a reasonable alternative prior, such as  $Normal(0, 0.21)$  or  $Normal(0, 0.11)$ ; these assume that the difference between conditions will be around 200 or 100 ms, and can be positive or negative; see the top row of Figure 5. Alternatively, we could have chosen unrealistically tightly constrained priors, such as (a)  $Normal(0.02, 0.02)$ , which assumes a difference four times larger between subject and object relative clauses than what the data shows; (b)  $Normal(0.05, 0.02)$ , which assumes that object relative clauses *are slower* than subject relative clauses also in Chinese; or (c)  $Normal(-0.05, 0.02)$ , which assumes an unusually precise prior information about the effect. With such unreasonable priors, we will get unreasonable posteriors; see the bottom row of Figure 5. This is because as we increase the precision of the priors we have more influence on the posterior distribution. Figure 5 and Table 2 illustrate how the prior influences the posterior. It is worth noticing that in all these examples we are assuming a normal distribution, in some cases, however, it may be necessary to choose a distribution that is less informative and more robust against outliers, such as Student- $t$  (Ghosh et al., 2015).

At this point, the reader may well ask: why do I have to decide on a “reasonable” prior? What is reasonable anyway? Isn’t this injecting an uncomfortable level of subjectivity into the analysis? Here, one should consider that the way we actually reason about research follows this methodology, albeit informally. When we review the literature on a particular topic, we report some pattern of results, often classifying them as “significant” and “non-significant” effects. For example, if we are reviewing the literature on English relative clauses, we might conclude that most studies have shown a subject relative advantage. However, if we stop to consider what the average magnitude of the reported effects is, we already have much more information than the binary classification of significant or not significant. For the relative clause example, in self-paced reading studies, at the critical region (which is the relative clause verb in English), we see 67 milliseconds (SE approximately 20)

(Grodner & Gibson, 2005); 450 ms, 250 ms, 500 ms, and 200 ms (approximate SE 50 ms) in experiments 1-4 respectively of Gordon, Hendrick, and Johnson (2001); 20 ms in King and Just (1991) (their figure 6). In eye-tracking studies reporting first-pass reading time during reading, we see 48 ms (no information provided to derive standard error) in Staub (2010); and 12 ms (no SE provided) in Traxler, Morris, and Seely (2002). Normally we pay no attention to this information when conducting a new analysis; but using this information is precisely what the Bayesian framework allows us to do. In effect, it allows us to formally build on what we already know.

Table 2

*Summary of posterior distributions of the coefficients for the object relative advantage in the Gibson and Wu data, assuming different priors. The first three priors can be considered weakly informative and reasonable, but the last three are overly constrained and we can see that as a consequence they dominate the posterior, in the sense that the posterior is largely determined by the prior.*

Prior	95% CrI		$P(\hat{\beta} < 0)$	$\hat{\beta}$
<i>Normal</i> (0, 1)	-0.1	0.02	0.88	-0.04
<i>Normal</i> (0, 0.21)	-0.09	0.02	0.88	-0.03
<i>Normal</i> (0, 0.11)	-0.08	0.02	0.86	-0.03
<i>Normal</i> (-0.18, 0.02)	-0.2	-0.15	1	-0.17
<i>Normal</i> (0.05, 0.02)	0.01	0.06	0	0.04
<i>Normal</i> (-0.05, 0.02)	-0.07	-0.02	1	-0.04

### Inference using the credible interval

So we have checked that the posterior is not too sensitive to different weakly informative priors. What inferences can we draw from the model? Are object relative clauses easier than subject relative clauses in Chinese? We do have some evidence, although it is rather weak. Although we do not need to make an accept/reject decision, for situations where we really want to make a decision, Kruschke, Aguinis, and Joo (2012) suggest that, since the 95% credible intervals created using HDPI include the most credible values of the parameter, they can be used as a decision tool: One simple decision rule is that any value outside the 95% HDPI is rejected (also see Dienes, 2011). Note also that in symmetric posterior distributions, the percentile interval will have a range similar to the HPDI and

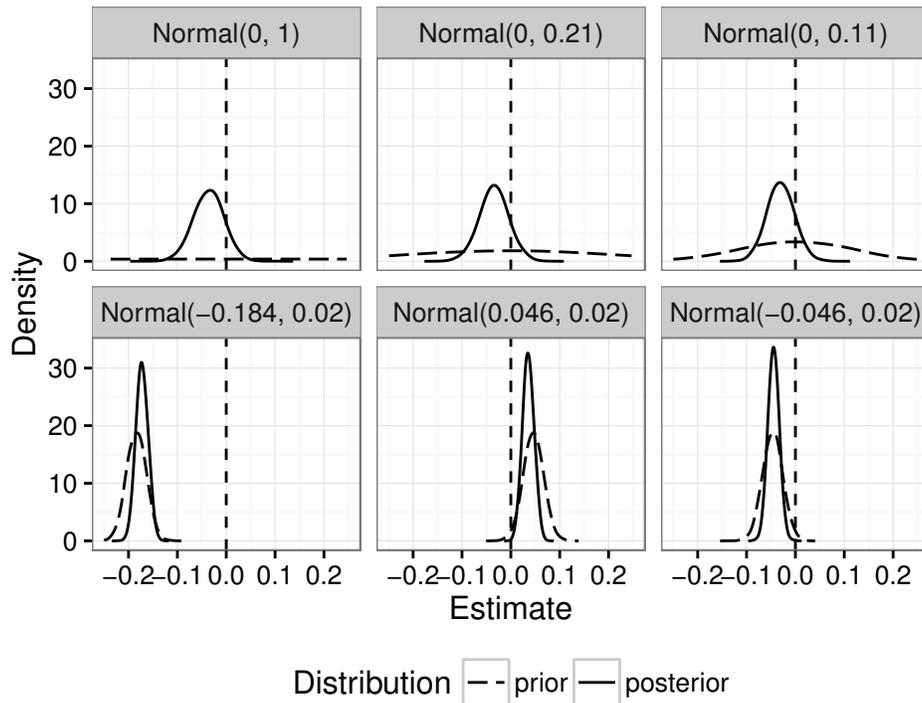


Figure 5. Posterior distributions given different type of priors. The first row shows different weakly informative priors, while the second row shows unreasonably constrained priors.

can equally well be used.

A more sophisticated decision rule according to Kruschke et al. (2012) also allows us to accept a null result. A region of practical equivalence (ROPE) around the null value can be established; we assume that values in that interval, for example  $[-0.005, 0.005]$ , are practically zero. We would reject the null value if the 95% HPDI falls completely outside the ROPE (because none of the most credible values is practically equivalent to the null value). In addition, we would accept the null value if the 95% HPDI is completely inside the ROPE, because the most credible values are practically equivalent to the null value. The crucial thing is that 95% HPDI gets narrower as the sample size gets larger.

### Reporting the results of the Gibson and Wu analysis

So how can we report the analysis of the Gibson and Wu experiment, and what can we conclude from it? After providing all the relevant details about the linear mixed model,

including the priors, we would report the mean of the estimate of the effect, its credible interval, and the probability of a negative effect.

In the present case, we would report that (i) the prior for the intercept is  $Normal(\mu = 0, \sigma = 10)$ , (ii) the prior for the effect of interest (the object-subject difference) is  $Normal(0, 1)$ , and (iii) the regularization on the covariance matrix of random effects is 2. We would also report that four chains were run for 2000 iterations each. We would also mention that a sensitivity analysis using weakly informative priors showed that the posterior is not overly influenced by the prior specification.

The results of the `stan_lmer` based analysis are repeated below for convenience (see Table 1 for a comparison with `lmer` output). If the variance components and correlations of the random effects are also of theoretical interest (this could be the case if individual differences are relevant theoretically), then the credible intervals for these can also be reported.

Table 3

*Summary of the Bayesian linear mixed model estimates for the Gibson and Wu data-set. The `stan_lmer` packages provides the median and the standard deviation of the median absolute difference (MAD) of the fixed effects, but one could equally well compute the mean and standard error, to mirror the `lme4` convention.*

<u>Random effects</u>			
Groups	Name	Std.Dev.	Corr
subj	(Intercept)	0.2425	
	cond	0.0762	-0.521
item	(Intercept)	0.1829	
	cond	0.0475	0.012
Residual		0.5131	
<u>Fixed effects</u>			
	Median	MAD-SD	
(Intercept)	6.0641	0.0658	
cond	-0.0364	0.0301	

The effect of interest is the difference between the object and subject relative clause reading times, and this can be summarized in terms of the estimated mean of the posterior, and the credible intervals: ( $\hat{\beta} = -0.04$ , 95% *CrI* =  $[-0.1, 0.02]$ ). The posterior probability of this effect being less than zero,  $P(\hat{\beta} < 0)$ , is 0.89. It has the correct sign following Gibson

and Wu’s prediction, but the evidence that it is negative is not very strong.

It would also be very helpful to the reader of a published result to have access to the original data and the code that led to the analysis; this allows the researcher to independently check the analyses himself or herself, possibly with different prior specifications, and to build on the published work by using the information gained from the published result. Trying out different priors would be specially helpful to the expert researcher who has a different opinion (based on their own knowledge about the topic) on what the true effect might be.

### Hypothesis testing using the Bayes factor

The Bayes factor (BF) provides a way to quantify the evidence for the model under which the observed data are most likely relative to another model. This is accomplished by computing the ratio of the *marginal likelihoods* of two models  $M_0$  and  $M_1$ , which correspond to research hypotheses  $H_0$  and  $H_1$  (we will use the words model and hypothesis interchangeably below):

$$BF_{01} = \frac{p(D|M_0)}{p(D|M_1)} \quad (2)$$

$BF_{01}$  then indicates the extent to which the data supports  $M_0$  over  $M_1$ . The marginal likelihood of a model,  $p(D|M)$ , is the probability of the data  $D$  given the model  $M$ . For example, if we toss a coin five times and get four heads, we can compute the probability of getting four heads by using the probability mass function for the binomial distribution:

$$\binom{n}{k} p^k (1-p)^{(n-k)} \quad (3)$$

Here, we have five trials ( $n=5$ ), four heads ( $k=4$ ), and some probability  $p$  of getting a heads. If the parameter  $p$  of the binomial is believed to be 0.5, we can compute the the probability of getting exactly four heads:

$$\binom{5}{4} 0.5^4 (1 - 0.5)^{(5-4)} = 0.16 \quad (4)$$

This is the marginal likelihood under a particular model (the assumption that the parameter  $p = 0.5$ ).

As mentioned above, the Bayes factor is a ratio: if we want to compare the null hypothesis,  $H_0$ , with a specific alternative hypothesis,  $H_1$ , the ratio we would compute is  $BF_{01} = p(D|H_0)/p(D|H_1)$ . An outcome smaller than one will mean more evidence for  $H_1$  than for  $H_0$ . Crucially, the Bayes factor can provide evidence in favor of the null hypothesis: This is the case when the outcome of the calculation is bigger than one (see Gallistel, 2009). For example, suppose our null hypothesis is that our coin is fair, i.e.,  $p = 0.5$ , and the alternative is that  $p = 0.8$ . For the five coin tosses with four successes, the marginal likelihoods under the two hypotheses are 0.16 (for  $p=0.5$ ), and 0.41 (for  $p=0.8$ ). If we take the ratio of these two values, then we have a Bayes factor of 0.38. This is weak evidence in favor of the alternative hypothesis that  $p = 0.8$ . Alternatively, if we had two heads in five tosses, then the situation would have been different: the marginal likelihoods under the two hypotheses would then be 0.31 (for  $p=0.5$ ), and 0.05 (for  $p=0.8$ ), and the Bayes factor would be 6.1, which is weak evidence in favor of the null hypothesis. A scale has been proposed to interpret Bayes factors according to the strength of evidence in favor of one model (corresponding to some hypothesis) over another (see Lee & Wagenmakers, 2014, citing Jeffreys, 1961). On this scale, a Bayes factor of 10-30 would constitute strong evidence in favor of Model 1 over Model 2; larger values than 10 are very strong evidence, and smaller values constitute weaker evidence. Obviously, values smaller than 1 would then favor Model 2.

These examples with coin tosses are simple, but the approach can be scaled up to the situation where we have a prior defined for our parameter(s); in this case, the marginal likelihood would be computed by taking a sum over the likelihoods, weighted by the probability assigned to each possible value of the parameter. To take a simplified example, assume that

our prior is in the five-trial coin-toss example above is that  $p = 0.1$  with probability 0.4 and  $p = 0.8$  with probability 0.6, then the marginal likelihood when we have four heads is:

$$0.4 \times \binom{5}{4} 0.1^4 (1 - 0.1)^{(5-4)} + 0.6 \times \binom{5}{4} 0.8^4 (1 - 0.8)^{(5-4)} \quad (5)$$

As discussed earlier, in reasonably large samples, the posterior distribution is not overly influenced by weakly informative priors. In contrast, the Bayes factor *is* sensitive to the priors (C. C. Liu & Aitkin, 2008). When priors are defined to allow a broad range of values, the result will be a lower marginal likelihood (which in turns influences the Bayes factor, as we saw in the examples above). This sensitivity of the Bayes factor to priors can be considered a liability (Lee & Wagenmakers, 2014, chapter 7.5). However, the dependency on the prior can be studied explicitly with a sensitivity analysis, in which one varies the prior and studies the fluctuations of the Bayes factor.

A challenge with the Bayes factor is that, when sample sizes are moderate or the models are relatively complicated, the marginal likelihood is often quite difficult to estimate using sampling (Carlin & Louis, 2008, 196). Nevertheless, there are tools for computing the Bayes factor corresponding to t-tests (Rouder, Speckman, Sun, Morey, & Iverson, 2009), and the R package `BayesFactor` and the JASP software package also provide functions for computing Bayes factor for repeated measures ANOVA designs. There is also another method, called the Savage-Dickey density ratio method that can be used directly with Bayesian linear mixed models. We present below a practical example of computing Bayes factor using this method.

### **An example: Computing Bayes Factor in the Gibson and Wu data**

The Savage–Dickey density ratio method (Dickey, Lientz, et al., 1970) is a straightforward way to compute the Bayes factor for nested models. The method consists of dividing the height of the posterior for a certain parameter by the height of the prior of the same parameter, at the point of interest (see Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010 for a complete tutorial and the mathematical proof). Critically, we can use the height

```

30 library(polyspline)
31 fit_posterior <- logspline(posterior_condition)
32 posterior <- dlogspline(0, fit_posterior) # Height of the posterior at 0
33 prior      <- dnorm(0, 0, 1) # Height of the prior at 0
34 (BF01 <- posterior/prior) #BF01 shows clear support for H0
## [1] 15.67

```

Listing 5: Code for calculating the Bayes Factor for the Gibson and Wu data.

of an approximation of the posterior distribution from the samples obtained from the numerical method employed (such as MCMC). In our case, we could calculate the evidence in favor or against our predictor (condition) being zero. The model with the experimental condition and the null model will have several parameters in common that are not of interest (such as the intercept of the fixed effects and random effects, the standard deviation, etc.), but these parameters won't influence the calculation of the Savage–Dickey density ratio (Wagenmakers et al., 2010).

Listing 5 illustrates how to perform the calculation for the Gibson and Wu example. We see that the comparison clearly favors the null hypothesis: it is showing 15.67 times more evidence for the null than for any other value. However, this might be because when priors allow a broad range of values, and thus are too uninformative, the alternative hypothesis to the null,  $H_1$ , (that the effect is different from zero) is penalized for assigning too much prior mass to values that are too unlikely (while all the prior mass of the null hypothesis,  $H_0$ , is concentrated in zero). Without a proper specification of priors,  $H_0$  would always be more likely than  $H_1$ .

Table 4 shows the Bayes factor under different weakly informative priors: the first column represents the numerator of a Bayes factor, while the second column the denominator. The priors in Table 4 represent our prior beliefs on the plausibility of different values of the effect of object vs. subject relative clauses. The table shows that as we provide tighter

and more realistic priors, the evidence in favor of  $H_0$  decreases, which means that we don't have enough evidence to accept the null hypothesis. But notice that this doesn't mean that we can accept  $H_1$  either.

Table 4

*Bayes factor under different weakly informative priors*

H0	H1	Prior
3.63	1	Normal(0, 0.21)
2.14	1	Normal(0, 0.11)

In sum, the Bayes Factor can be a useful tool, but it should be borne in mind that it will always be affected by the prior, so a sensitivity analysis is a good idea when reporting Bayes Factors. According to Dienes (2011), its calculation depends on answering a question about which there may be disagreement among researchers: “What way of assigning probability distributions of effect sizes as predicted by theories would be accepted by protagonists on all sides of a debate?” One of the clearest advantages of the Bayes Factor is that once the minimal magnitude of an expected effect is agreed upon, evidence can be gathered in favor of the null hypothesis.

### Model selection using cross-validation

Another way to make a decision about the hypothesis that object relative clauses are easier than subject relative clauses in Chinese is to treat the hypothesis as a model that can be compared with other models such as the null. We will focus on cross-validation; for other approaches, see Shiffrin et al. (2008).

The question whether object relative clauses are easier than subject relative clauses in Chinese can be also be phrased in terms of evaluating the model on its ability to make predictions about future or unseen observations, in comparison with, for example, a null model (or another model). However, it may not be the most suitable way to compare nested linear mixed models when the effects being investigated are small (Gelman, Hwang, & Vehtari, 2014; Wang & Gelman, 2014).

This approach to model selection is based on finding the most “useful model” for

characterizing future data, and not necessarily the true model: the true model is not guaranteed to produce the best predictions, and a false model is not guaranteed to produce poor predictions (Wang & Gelman, 2014). The ideal measure of a model’s fit would be its (out-of-sample) predictive performance for new observations that are produced by the same data-generating process. When the future observations are not available the predictive performance can be estimated by calculating the *expected predictive performance* (Gelman, Hwang, & Vehtari, 2014; Vehtari & Ojanen, 2012).

The cross-validation techniques that we review below are based on comparing the expected predictive performance of a model with its actual performance (but see Vehtari & Ojanen, 2012 and Piironen & Vehtari, 2015 for a more complete review). We will focus on Bayesian leave-one-out cross-validation (LOO-CV; Geisser & Eddy, 1979) and three approximations: (a) k-fold-cross-validation (k-fold-CV; Vehtari & Ojanen, 2012), (b) Pareto smoothed importance sampling (PSIS-LOO; Vehtari & Gelman, 2015), and (c) the widely applicable information criterion (or Watanabe-Akaike information criterion: WAIC; Watanabe, 2009, 2010). The latter two are implemented in the R package *loo* (Vehtari, Gelman, & Gabry, 2015b), but they should be used with care, since they are affected by highly influential observations. When highly influential observations are present, k-fold-CV is recommended (the code for implementing k-fold-CV in Stan is available in Vehtari, Gelman, & Gabry, 2015a).

The basic idea of cross-validation is to split the data such that each subset is used as a validation set, while the the remaining sets (the training set) are used for estimating the parameters. LOO-CV method depicts the case when the training set only excludes one observation. The main advantage of this methods is its robustness, since the training set is as similar as possible to the real data, while the same observations are never used simultaneously for training and evaluating the predictions. A major disadvantage is the computational burden (Vehtari & Ojanen, 2012), since we need to fit a model as many times as the number of observations.

The k-fold-CV (Vehtari & Ojanen, 2012) can be used to reduce the computation time

by reducing the number of models we need to fit. In the k-fold-CV approach, the data are split into k subsets (or folds), where k is generally around ten. Each subset is in turn used as the validation set, while the remaining data are used for parameter estimation. A further reduction in computation time can be achieved with PSIS-LOO (Vehtari & Gelman, 2015), which is faster compared to LOO-CV, and does not require fitting the model multiple times (Vehtari & Gelman, 2015; Vehtari et al., 2015a).

Information criteria are commonly used for selecting Bayesian models, since they are directly related to assessing the predictive performance of the models. In addition, WAIC is asymptotically equal to LOO (Watanabe, 2010). The distinguishing feature of WAIC in comparison with AIC (Akaike Information Criterion; Akaike, 1974), DIC (Deviance Information Criterion; Spiegelhalter, Best, Carlin, & Van Der Linde, 2002), and BIC (Bayesian Information Criterion; Schwarz, 1978, which also has a different goal than the other measures discussed here), is that WAIC is point-wise: the uncertainty is calculated point-by-point in the data over the entire posterior distribution (Gelman, Hwang, & Vehtari, 2014). This is important because some observations are harder to predict than others. In addition, AIC does not work well with strong priors, and while DIC can take into account informative priors, and it is the measure of choice in many Bayesian applications, it may give unexpected results when the posterior distribution is not well summarized by its mean (for example, if the posterior is substantially skewed). For a complete comparison between AIC, DIC, and WAIC, see Gelman, Hwang, and Vehtari (2014).

Cross validation techniques are ideally suited for comparing highly different models, and may be a fully Bayesian replacement for AIC or DIC. However, even with moderate sample size, it can be difficult to compare nested hierarchical models (such as linear mixed models) based on predictive accuracy (Wang & Gelman, 2014). An experimental manipulation can produce a tiny change in predictive accuracy, which can be nearly indistinguishable from noise, but it can be still useful for evaluating a psycholinguistic theory. This means that unless we are dealing with huge effects or with a very large sample size, the null model would always be almost as good as the model with the predictor of interest, as far as predic-

tive accuracy is concerned, while the complexity of the model with the predictor of interest is penalized (Gelman, Hwang, & Vehtari, 2014; Wang & Gelman, 2014).

**Some closing remarks on inference.** In the previous sections, we have presented Bayesian methods as a useful and important alternative to null hypothesis significance testing (NHST) for statistical inference. In contrast to NHST, where a sharp binary decision is made between rejecting or failing to reject the null hypothesis, we can now directly talk about the strength of the evidence for a certain effect. We pointed out that the 95% Bayesian credible interval as a possible way to summarize the evidence. The credible interval has a very intuitive interpretation (which researchers often ascribe mistakenly to frequentist confidence intervals): it gives the range over which we can be 95% certain that the true value of the effect lies, given of course the data, the priors, and the model. We also pointed out that the mass of the probability of the posterior distribution below (or above) zero can give valuable information about the plausibility of a negative (or positive) effect.

As a rule of thumb, we can interpret the evidence as strong if zero lies outside the 95% credible interval (Kruschke et al., 2012). If zero is included within the interval, there might still be weak evidence for an effect, if the probability of the estimate being less than (or greater than) zero is large enough. Our interpretation of the evidence should also take into account that the range of possible magnitudes of the effect makes sense theoretically; for example, if we find an effect of less than one millisecond, we shouldn't interpret it as strong evidence, just because its credible interval doesn't include zero. A difference of less than one millisecond would likely have no theoretical relevance in psycholinguistics or linguistics.

Given the results from the Bayesian analysis of the Gibson and Wu data, we could claim that there is some weak evidence for the claim that object relative clauses are easier than subject relative clauses in Chinese, depending on what we make of the magnitude of the effect on this experiment in comparison with other similar experiments. Importantly, we wouldn't be able to claim that we have evidence for no effect. This is a common problem in the way that null hypothesis significance testing is used in linguistics and psycholinguistics; a failure to find an effect is presented as evidence for the null hypothesis that the parameter

is 0 (also see Part I of this review for more discussion). The effect is also considered to be zero even if repeated experiments consistently show, say, a negative sign of the effect that do not reach statistical significance. In such a situation, if theory suggests “no effect”, we could (and should) establish an interval around zero that would be practically equivalent to “no effect”, the region of practical equivalence or ROPE, and find that the 95% credible interval of the estimate of the effect falls completely inside it. Alternatively, one could identify the smallest effect we would expect, and then use Bayes factors.

It should be stressed that Bayesian analysis is not immune to the dangers of over-interpreting noise, if several different models are tried or if models are fitted in several correlated regions (as in self-paced reading) or measures (as in eye-tracking), it is likely that some model will yield some effect that looks as if it were robust (see section 6 of Part I). As we stressed in Part I of the review, there’s nothing as convincing as a replication to assess the robustness of an effect.

In the final section below, we review some examples that use Bayesian tools, and mention some of the possibilities for fitting more complex and interesting models. Any of these example applications can serve as a starting point for the researcher.

### **Examples of applications of Bayesian Methods**

It has become relatively straightforward to fit complex Bayesian models due to the increase in computing power and the appearance of probabilistic programming languages, such as WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000), JAGS (Plummer, 2012), and Stan (Stan Development Team, 2015). Even though these statistical packages allow the user to define models without having to deal with the complexities of the sampling process, some background statistical knowledge is needed before one can define the models.

There are some alternatives that allow Bayesian inference in R without having to fully specify the model “by hand”. The packages *rstanarm* (Gabry & Goodrich, 2016) and *brms* (Buerkner, 2015) emulates many popular R model-fitting functions, such as (g)lmer, using Stan for the back-end estimation and sampling, and can be useful for a smooth transition

between frequentist linear mixed models and Bayesian ones.<sup>7</sup> In addition, the *BayesFactor* (Morey & Rouder, 2015) package emulates other standard frequentist tests (t-test, ANOVA, linear models, etc.), and provides the Bayes Factor given some pre-specified priors. For a simpler option, JASP (Love et al., 2015) provides a graphical user interface, and is an alternative to SPSS.

For linear mixed models, one strength of Bayesian methods is that we can fit models with a full random structure that would not converge with frequentist methods or would yield overestimates of correlations between the random effects (Bates, Kliegl, et al., 2015). This can be achieved by using appropriate weakly informative priors for the correlation matrices (so-called LKJ priors, and see also Sorensen et al., 2015 for a tutorial).<sup>8</sup> Some examples of papers using Bayesian linear mixed models in psycholinguistics are Frank, Trompenaars, and Vasishth (2015); Hofmeister and Vasishth (2014); Husain, Vasishth, and Srinivasan (2014). However, the major advantage of Bayesian methods lies in the possibility of moving beyond linear models. These become relevant for modeling distributions of reaction and reading times (RTs), which are limited on the left by some amount of time (i.e., the shift of the distribution), and are highly right skewed. RTs can be reciprocal- or log-transformed to incorporate them in linear mixed models, but these transformations still assume a shift of 0 ms. Rouder (2005) suggests the shifted log-normal hierarchical model as a suitable model for RTs. This type of model is not linear but can be fit straightforwardly, and can be used for inferences in experiments with self-paced reading tasks (Nicenboim et al., 2015). Another potential use of non-linear hierarchical models that to our knowledge has not been applied in psycholinguistics or linguistics is ordered probit hierarchical models for acceptability judgments or any type of rating task that uses a scale (Kruschke, 2015, Chapter 23). A further interesting application is that one can synthesize evidence from

---

<sup>7</sup>Regarding `rstanarm`, our experience has been that one can fit hierarchical linear models much faster if one specifies the model by hand in an efficient matrix format; see the `RePsychLing` package (<https://github.com/dmbates/RePsychLing>) for example code.

<sup>8</sup>An issue that should be taken into account when fitting hierarchical models by manually specifying the full model is that the geometry of the posterior distributions can be very complex, and with very large data sets, sampling can be excruciating slow. One way to speed up such models is through reparametrization (see for example page 211 in Stan Development Team, 2016 and Papaspiliopoulos, Roberts, & Sköld, 2007).

existing studies by carrying out a meta-analysis (Vasishth et al., 2013). Meta-analysis does face the potential problem that researchers often do not provide estimates of standard errors, which are needed for the analysis. However, our experience has been that even when estimates of standard errors are not reported, these can usually be derived from the reported statistics. This is how we carried out a Bayesian random-effects (hierarchical) meta-analysis in Engelmann, Jäger, and Vasishth (2016). Meta-analysis is not widely used in linguistics and psycholinguistics, but it can play a very important role in literature reviews.

Bayesian cognitive modeling is another extremely fruitful use of Bayesian methods, and some of the methods discussed in Lee (2011) and Lee and Wagenmakers (2014) could easily be adapted for psycholinguistics. It is important to note the distinction between using Bayesian methods for modeling cognitive processes, assuming that (some aspect of) the mind is Bayesian, and using Bayesian methods for modeling cognitive processes without necessarily assuming a Bayesian mind. Some examples of the former category are Bayesian/noisy channels approaches to parsing (for a review see Traxler, 2014) or to word learning (see, for example, Xu & Tenenbaum, 2007); and the belief update models presented by Myslín and Levy (2016) and Kleinschmidt, Fine, and Jaeger (2012). Indeed, even though Kleinschmidt et al. (2012) model adaptation as a Bayesian belief update, the model itself was fit using frequentist methods. An example of the second category, i.e., using Bayesian methods for modeling without assuming that the mind is Bayesian, is Logačev and Vasishth (2016). Here, the focus is on evaluating different models of parsing in the face of ambiguities.

In addition, there is a class of models that is mostly used in two-forced choice tasks and that have the strength of integrating accuracy and reaction times instead of wrongly treating them as independent outcomes. This class of models is based on the idea that response selection can be modeled by a process that accumulates evidence until a threshold is reached. This could be applied in deciding whether a string of letters is a word or a non-word, whether a word is the right completion of a sentence, whether a sentence is grammatical or not, and so forth. One of the most widely applied evidence accumulation models is the Ratcliff diffusion model (see Ratcliff & Rouder, 1998), but several other models

based on the similar ideas exist, such as the Ballistic and linear Ballistic accumulator (Brown & Heathcote, 2005, 2008). The implementation of these models with frequentist methods is notoriously complicated, and in order to fit all their parameters they require a large number of trials. Bayesian methods allow extending these models hierarchically, with all the benefits that this implies, that is, the ability to take into account within- and between-subjects and between-items variability, and to do partial pooling. Vandekerckhove, Tuerlinckx, and Lee (2011), for example, provide an implementation in WinBUGS of a hierarchical diffusion model. The linear ballistic accumulator has also been implemented in WinBUGS (Donkin, Averell, Brown, & Heathcote, 2009), and it could be extended in the same way as the hierarchical diffusion model was. Another Bayesian model based on the accumulation of evidence is the lognormal race model (Rouder, Province, Morey, Gomez, & Heathcote, 2014), not as feature rich as the diffusion model and the linear ballistic accumulator, but its approach can generalize to any number of choices (including just one choice).

Finally, another advantage that Bayesian methods can provide is the use of informative priors in situations where data are scarce but we have previous knowledge about the effects. This is the idea behind the Small N Acceptability Paradigm for Linguistic Acceptability Judgments (SNAP Judgments; Mahowald, Graff, Hartman, & Gibson, 2015), which allows us to obtain quantitative and statistically valid data in syntax and semantics research in situations where it would be difficult to consult with many native speakers. And finally, the use of informative priors could be a significant advantage in studies with impaired participants (such as aphasics), where it is difficult to have a large sample size.

### Concluding remarks

Carrying out Bayesian data analysis clearly requires thought and effort; even if one uses convenient packages like `rstanarm`, several decisions have to be made: we have to define priors, carry out sensitivity analyses, and decide how to interpret the results. By comparison, fitting a linear mixed model using `lme4` is much easier: just write a single line of code and extract the t-value(s) or the like. To add insult to injury, the overhead in terms

of time and effort of fitting a Bayesian model seems unjustified given that, for large sample sizes, the estimates for the fixed effects from a Bayesian model and the corresponding `lme4` model will be quite similar (if not identical), especially with weakly informative priors (for examples from psycholinguistics, see Bates, Kliegl, et al., 2015). Why bother to use Bayesian methods then? One compelling reason is that although p-values answer *a* question, they answer the wrong question. Once one realizes that the p-value doesn't provide any direct evidence for the research question, the motivation to compute it fades. Another reason is that since we already tend to interpret the result of frequentist analyses in a Bayesian manner, we might as well carry out a Bayesian analysis. Finally, as discussed earlier, Bayesian probabilistic programming languages provide a degree of flexibility in defining models that is difficult to match with frequentist tools.

### Further reading

For a first introduction to Bayesian methods, we suggest McElreath (2015) and Kruschke (2015). Lynch (2007) is also excellent but assumes some calculus. For a more advanced treatment of the topic, see Gelman, Carlin, et al. (2014). Linear mixed models are covered from both the frequentist and Bayesian perspective by Gelman and Hill (2007). For an accessible introduction of Bayesian methods for cognitive modeling, see Lee and Wagenmakers (2014).

### Acknowledgment

Thanks to Lena Jäger, Dario Paape, and Daniela Mertzen for helpful comments on previous versions of this review.

### References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–723.

- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.
- Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). *Parsimonious mixed models*. Retrieved from <http://arxiv.org/abs/1506.04967> (ArXiv e-print)
- Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*. (In Press)
- Bates, D., & Sarkar, D. (2007). lme4: Linear mixed-effects models using S4 classes [Computer software manual]. (R package version 0.9975-11)
- Box, G. E., & Tiao, G. C. (1992). *Bayesian inference in statistical analysis* (First ed.). John Wiley & Sons.
- Brown, S. D., & Heathcote, A. (2005). A ballistic model of choice response time. *Psychological review*, *112*(1), 117.
- Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive psychology*, *57*(3), 153–178.
- Buerkner, P.-C. (2015). brms: Bayesian regression models using Stan [Computer software manual]. Retrieved from <http://CRAN.R-project.org/package=brms> (R package version 0.6.0)
- Carlin, B. P., & Louis, T. A. (2008). *Bayesian methods for data analysis*. CRC Press.
- Chung, Y., Gelman, A., Rabe-Hesketh, S., Liu, J., & Dorie, V. (2013). Weakly informative prior for point estimation of covariance matrices in hierarchical models. *Manuscript submitted for publication*.
- Dickey, J. M., Lientz, B., et al. (1970). The weighted likelihood ratio, sharp hypotheses about chances, the order of a markov chain. *The Annals of Mathematical Statistics*, *41*(1), 214–226.
- Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, *6*(3), 274–290.
- Donkin, C., Averell, L., Brown, S., & Heathcote, A. (2009). Getting more from accuracy

- and response time data: Methods for fitting the linear ballistic accumulator. *Behavior Research Methods*, 41(4), 1095–1110.
- Engelmann, F., Jäger, L. A., & Vasishth, S. (2016). *The determinants of retrieval interference in dependency resolution: Review and computational modeling*. (Manuscript submitted)
- Frank, S. L., Trompenaars, T., & Vasishth, S. (2015). Cross-linguistic differences in processing double-embedded relative clauses: Working-memory constraints or language statistics? *Cognitive Science*, n/a.
- Gabry, J., & Goodrich, B. (2016). rstanarm: Bayesian applied regression modeling via stan [Computer software manual]. Retrieved from <http://CRAN.R-project.org/package=rstanarm> (R package version 2.9.0-1)
- Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review*, 116(2), 439–453. Retrieved from <http://dx.doi.org/10.1037/a0015251> doi: 10.1037/a0015251
- Geisser, S., & Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, 74(365), 153–160.
- Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian analysis*, 1(3), 515–534.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (Third ed.). Chapman and Hall/CRC.
- Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge, UK: Cambridge University Press.
- Gelman, A., Hill, J., & Yajima, M. (2012, Apr). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, 5(2), 189–211. Retrieved from <http://dx.doi.org/10.1080/19345747.2011.618213> doi: 10.1080/19345747.2011.618213
- Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24(6), 997–1016.

- Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 1360–1383.
- Ghosh, J., Li, Y., & Mitra, R. (2015, July). On the Use of Cauchy Prior Distributions for Bayesian Logistic Regression. *ArXiv e-prints*.
- Gibson, E., & Wu, H.-H. I. (2013). Processing Chinese relative clauses in context. *Language and Cognitive Processes*, 28(1-2), 125–155.
- Gordon, P. C., Hendrick, R., & Johnson, M. (2001). Memory interference during language processing. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 27(6), 1411–1423.
- Grodner, D., & Gibson, E. (2005). Consequences of the serial nature of linguistic input. *Cognitive Science*, 29, 261–290.
- Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers. *Methods of Psychological Research*, 7(1), 1–20.
- Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014, Jan). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, 21(5), 1157–1164. Retrieved from <http://dx.doi.org/10.3758/s13423-013-0572-3> doi: 10.3758/s13423-013-0572-3
- Hofmeister, P., & Vasishth, S. (2014). Distinctiveness and encoding effects in online sentence comprehension. *Frontiers in Psychology*, 5, 1–13. (Article 1237) doi: doi:10.3389/fpsyg.2014.01237
- Hsiao, F. P.-F., & Gibson, E. (2003). Processing relative clauses in Chinese. *Cognition*, 90, 3–27.
- Husain, S., Vasishth, S., & Srinivasan, N. (2014). Strong expectations cancel locality effects: Evidence from Hindi. *PLoS ONE*, 9(7), 1–14.
- Jeffreys, H. (1961). *Theory of probability*. Oxford: Clarendon Press.
- King, J., & Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of memory and language*, 30(5), 580–602.

- Kleinschmidt, D., Fine, A. B., & Jaeger, T. F. (2012). A belief-updating model of adaptation and cue combination in syntactic comprehension. In *Proceedings of the 34rd Annual Meeting of the Cognitive Science Society (CogSci12)* (pp. 605–10).
- Kruschke, J. K. (2015). *Doing Bayesian data analysis* (Second ed.). Boston: Academic Press. doi: <http://dx.doi.org/10.1016/B978-0-12-405888-0.09999-2>
- Kruschke, J. K., Aguinis, H., & Joo, H. (2012, Sep). The time has come: Bayesian methods for data analysis in the organizational sciences. *Organizational Research Methods*, *15*(4), 722–752. Retrieved from <http://dx.doi.org/10.1177/1094428112457829>  
doi: 10.1177/1094428112457829
- Lecoutre, M.-P., Poitevineau, J., & Lecoutre, B. (2003). Even statisticians are not immune to misinterpretations of null hypothesis significance tests. *International Journal of Psychology*, *38*(1), 37–45.
- Lee, M. D. (2011, Feb). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology*, *55*(1), 1–7. Retrieved from <http://dx.doi.org/10.1016/j.jmp.2010.08.013> doi: 10.1016/j.jmp.2010.08.013
- Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.
- Levshina, N. (2016). A Bayesian mixed-effect multinomial model of English permissive constructions reveals a remarkable alignment of linguistic, cognitive, social and colostruational distances. *Cognitive Linguistics*. (In Press)
- Liu, C. C., & Aitkin, M. (2008, Dec). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology*, *52*(6), 362–375. Retrieved from <http://dx.doi.org/10.1016/j.jmp.2008.03.002> doi: 10.1016/j.jmp.2008.03.002
- Liu, Y., Gelman, A., & Zheng, T. (2013). Simulation-efficient shortest probability intervals. *arXiv preprint arXiv:1302.2142*.
- Logačev, P., & Vasishth, S. (2016). Understanding underspecification: A comparison of two computational implementations. *Quarterly Journal of Experimental Psychology*. (Accepted)

- Love, J., Selker, R., Marsman, M., Jamil, T., Dropmann, D., Verhagen, A. J., . . . Wagenmakers, E. J. (2015). *JASP (version 0.7)[computer software]*.
- Lunn, D., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS-A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and computing*, *10*(4), 325–337.
- Lynch, S. M. (2007). *Introduction to applied Bayesian statistics and estimation for social scientists*. Springer.
- Mahowald, K., Graff, P., Hartman, J., & Gibson, E. (2015). *Snap judgments: A small n acceptability paradigm (snap) for linguistic acceptability judgments*.
- McElreath, R. (2015). *Statistical rethinking: A bayesian course with r examples*. Chapman and Hall/CRC.
- Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2015, Oct). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*. Retrieved from <http://dx.doi.org/10.3758/s13423-015-0947-8> doi: 10.3758/s13423-015-0947-8
- Morey, R. D., & Rouder, J. (2015). *Bayesfactor: An r package for bayesian analysis in common research designs*. Retrieved from <https://cran.r-project.org/web/packages/BayesFactor/index.html>
- Myslín, M., & Levy, R. (2016). Comprehension priming as rational expectation for repetition: Evidence from syntactic processing. *Cognition*, *147*, 29–56.
- Nicenboim, B., Logačev, P., Gattei, C., & Vasishth, S. (2015). *When high-capacity readers slow down and low-capacity readers speed up: Working memory differences in unbounded dependencies*. (Resubmitted)
- O’Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., . . . Rakow, T. (2006). *Uncertain judgements: Eliciting experts’ probabilities*. John Wiley & Sons.
- Papaspiliopoulos, O., Roberts, G. O., & Sköld, M. (2007). A general framework for the parametrization of hierarchical models. *Statistical Science*, 59–73.

- Piironen, J., & Vehtari, A. (2015). Comparison of Bayesian predictive methods for model selection. *arXiv preprint arXiv:1503.08650*.
- Plummer, M. (2012). JAGS version 3.3.0 manual. *International Agency for Research on Cancer. Lyon, France*.
- Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin*, *114*(3), 510.
- Ratcliff, R., & Rouder, J. N. (1998, Sep). Modeling Response Times for Two-Choice Decisions. *Psychological Science*, *9*(5), 347–356. Retrieved from <http://dx.doi.org/10.1111/1467-9280.00067> doi: 10.1111/1467-9280.00067
- Rouder, J. N. (2005, Jun). Are unshifted distributional models appropriate for response time? *Psychometrika*, *70*(2), 377–381. Retrieved from <http://dx.doi.org/10.1007/s11336-005-1297-7> doi: 10.1007/s11336-005-1297-7
- Rouder, J. N., Province, J. M., Morey, R. D., Gomez, P., & Heathcote, A. (2014). The lognormal race: A cognitive-process model of choice and latency with desirable psychometric properties. *Psychometrika*, 1–23.
- Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009, Apr). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237. Retrieved from <http://dx.doi.org/10.3758/PBR.16.2.225> doi: 10.3758/pbr.16.2.225
- Schielezeth, H., & Forstmeier, W. (2009). Conclusions beyond support: Overconfident estimates in mixed models. *Behavioral Ecology*, *20*(2), 416–420. doi: 10.1093/beheco/arn145
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.
- Shiffrin, R., Lee, M., Kim, W., & Wagenmakers, E.-J. (2008, Dec). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *HCOG*, *32*(8), 1248–1284. Retrieved from <http://dx.doi.org/10.1080/03640210802414826> doi: 10.1080/03640210802414826

- Sorensen, T., Hohenstein, S., & Vasishth, S. (2015). *Bayesian linear mixed models using Stan: A tutorial for psychologists, linguists, and cognitive scientists*. Retrieved from <http://arxiv.org/abs/1506.06201> (ArXiv e-print)
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4), 583–639.
- Stan Development Team. (2015). *Stan: A C++ library for probability and sampling, version 2.7.0*. Retrieved from <http://mc-stan.org/>
- Stan Development Team. (2016). Stan modeling language users guide and reference manual, version 2.9 [Computer software manual]. Retrieved from <http://mc-stan.org/>
- Staub, A. (2010). Eye movements and processing difficulty in object relative clauses. *Cognition*, 116(1), 71–86.
- Traxler, M. J. (2014). Trends in syntactic parsing: anticipation, bayesian estimation, and good-enough parsing. *Trends in cognitive sciences*, 18(11), 605–611.
- Traxler, M. J., Morris, R. K., & Seely, R. E. (2002). Processing subject and object relative clauses: Evidence from eye movements. *Journal of Memory and Language*, 47(1).
- Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2011). Hierarchical diffusion models for two-choice response times. *Psychological methods*, 16(1), 44.
- Vasishth, S. (2015). *A meta-analysis of relative clause processing in Mandarin Chinese using bias modelling*. Sheffield, UK. Retrieved from <http://www.ling.uni-potsdam.de/~vasishth/pdfs/VasishthMScStatistics.pdf>
- Vasishth, S., Chen, Z., Li, Q., & Guo, G. (2013, 10). Processing Chinese relative clauses: Evidence for the subject-relative advantage. *PLoS ONE*, 8(10), 1–14.
- Vehtari, A., & Gelman, A. (2015). Pareto smoothed importance sampling. *arXiv preprint arXiv:1507.02646*.
- Vehtari, A., Gelman, A., & Gabry, J. (2015a). Efficient implementation of leave-one-out cross-validation and waic for evaluating fitted bayesian models. *arXiv preprint arXiv:1507.04544v2*.

- Vehtari, A., Gelman, A., & Gabry, J. (2015b). loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models [Computer software manual]. Retrieved from <https://github.com/jgabry/loo> (R package version 0.1.3)
- Vehtari, A., & Ojanen, J. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statist. Surv.*, *6*(0), 142–228. Retrieved from <http://dx.doi.org/10.1214/12-SS102> doi: 10.1214/12-ss102
- Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010, May). Bayesian hypothesis testing for psychologists: A tutorial on the savage–dickey method. *Cognitive Psychology*, *60*(3), 158–189. Retrieved from <http://dx.doi.org/10.1016/j.cogpsych.2009.12.001> doi: 10.1016/j.cogpsych.2009.12.001
- Wang, W., & Gelman, A. (2014). Difficulty of selecting among multilevel models using predictive accuracy. *Statistics at its Interface*, *7*, 1–8.
- Watanabe, S. (2009). *Algebraic geometry and statistical learning theory* (Vol. 25). Cambridge University Press.
- Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *The Journal of Machine Learning Research*, *11*, 3571–3594.
- Xu, F., & Tenenbaum, J. B. (2007). Sensitivity to sampling in Bayesian word learning. *Developmental science*, *10*(3), 288–297.