Bayesian data analysis in the phonetic sciences: A tutorial introduction

Shravan Vasishth[a,*], Bruno Nicenboim[a], Mary E. Beckman[b,*], Fangfang Li[c], Eun Jong Kong[d]

[a] *Department of Linguistics, University of Potsdam*
[b] *Department of Linguistics, Ohio State University*
[c] *Department of Psychology, University of Lethbridge*
[d] *Department of English, Korea Aerospace University*

Abstract

This tutorial analyzes voice onset time (VOT) data from Dongbei (Northeastern) Mandarin Chinese and North American English to demonstrate how Bayesian linear mixed models can be fit using the programming language Stan via the R package `brms`. Through the case study, we demonstrate some of the advantages of the Bayesian framework: researchers can (i) flexibly define the underlying process that they believe to have generated the data; (ii) obtain direct information regarding the uncertainty about the parameter that relates the data to the theoretical question being studied; and (iii) incorporate prior knowledge into the analysis. Getting started with Bayesian modeling can be challenging, especially when one is trying to model one's own (often unique) data. It is difficult to see how one can apply general principles described in textbooks to one's own specific research problem. We address this barrier to using Bayesian methods by providing three detailed case studies, with source code to allow easy reproducibility. The examples presented are intended to give the reader a flavor of the process of model-fitting; suggestions for further study are also provided. All data and code are available from: https://github.com/vasishth/jopbayes.

Keywords: Bayesian data analysis, Linear mixed models, voice onset time, gender effects, vowel duration

[*] Corresponding authors.

Email addresses: `vasishth@uni-potsdam.de` (Shravan Vasishth), `beckman.2@osu.edu` (Mary E. Beckman)

1. Introduction

In phonetics and other related areas of the language sciences, the vast majority of studies are designed to elicit several data points from each participant for each level of the linguistic variable of interest. This design poses difficulties for classic ANOVA models, which can accommodate only one random effect at a time, so that either the sets of datapoints for each participant or the sets of datapoints for each item must be replaced with the mean values (Clark, 1973). Over the last two decades, phoneticians have addressed these difficulties by turning to other methods, and linear mixed models—sometimes referred to as multilevel or hierarchical linear models—have become a standard tool, perhaps *the* standard tool for analyzing repeated measures data. The `lme4` package (Pinheiro and Bates, 2000; Baayen et al., 2008; Bates et al., 2015b) in R has greatly simplified model specification and data analysis for repeated measures designs. Even more recently, a Bayesian alternative to frequentist linear mixed models has become available, largely due to the emergence of a new programming language, Stan (version 1.17.3) (Stan Development Team, 2017b). In this article, we provide a tutorial introduction to fitting Bayesian linear mixed models. In order to make it easy for the newcomer to Bayesian data analysis to fit models, we use the popular and powerful R package `brms`, version 2.1.9 (Bürkner, 2016), which uses `lme4` syntax that researchers in linguistics and psychology are familiar with.

Fitting Bayesian models takes more time and effort than their frequentist analogues. Why bother to learn this relatively complex approach? We feel that there are several important advantages to fitting Bayesian models. Perhaps the most important one is that it gives us a degree of flexibility in defining models that is difficult to match with frequentist tools (Lee, 2011; Nicenboim and Vasishth, 2016). Examples are linear mixed models with a “maximal” random-effects structure (Barr et al., 2013),[1] and models incorporating non-standard assumptions (we discuss such a case below). A second advantage of Bayesian modeling is that we can focus our attention on quantifying our uncertainty about the magnitude of an effect. Instead of drawing a conclusion like “gender affects voice onset time”, using the Bayesian framework we can identify a credible interval of plausible values representing the effect. In other words, we can present a probability distribution of plausible values, instead of focusing on whether a particular confidence interval does or does not contain the value 0. Such quantitative summaries of an effect tell us much more about the research question than binary statements like “effect present” or “effect absent.” A third advantage

[1] Of course, one has to exercise judgement when deciding to fit such overparameterized models. See the discussions in Bates et al. (2015a); Baayen et al. (2017); Matuschek et al. (2017).

of Bayesian data analysis is that we can incorporate prior knowledge or beliefs in the model in an explicit way with the use of so-called informative priors. Such a use of priors is not widespread, but could be a powerful tool for building on what we already know about a research question.

In this tutorial, we will provide an informal introduction to Bayesian data analysis, and then present three case studies involving retrospective measurements of productions in a large cross-linguistic phonetic corpus. These examples are intended to provide a practical first entry to Bayesian data analysis. We do not aim to cover all aspects of Bayesian modeling here, but suggestions for further reading are provided at the end. In our case studies, we will focus on (generalized) linear mixed models (Pinheiro and Bates, 2000; Baayen et al., 2008; Bates et al., 2015b), because they are a standard tool today in experimental research in linguistics and the psychological sciences. We assume in this paper that the reader knows how to fit linear mixed models using the R package lme4 (Bates et al., 2015b).

All data and code are available from https://github.com/vasishth/jopbayes.

2. An informal introduction to Bayesian data analysis

Consider a simple case where we carry out an experiment in which we measure voice onset time in milliseconds in recordings of word-initial stops such as Mandarin $/t^h/$ and $/k^h/$ produced by male and female participants. Participants in each gender category are asked to produce multiple stop-initial words, resulting in repeated measurements of VOT from each participant. An example data-frame is shown in Listing 1.

For $i = 1, \ldots, I$ participants and $j = 1, \ldots, J$ items, we could fit a so-called varying intercepts and varying slopes linear mixed model of the type specified in (1) – the equation for a frequentist linear mixed model for the effect of gender on VOT. A notational convention we use here: a varying intercept always has index 0, and a varying slope has index 1. Thus, a varying intercept for item j is written $w_{0,j}$ and a varying slope is written $w_{1,j}$. Fixed intercepts and slopes also have the same numerical subscript convention of 0 for intercepts, and 1 for the slope (with increasing numbers in the case of multiple predictors).

Using these notational conventions, a frequentist linear mixed model for the effect of gender on VOT could be specified as follows:

$$VOT_{ij} = \beta_0 + u_{0,i} + w_{0,j} + (\beta_1 + w_{1,j}) \times \text{gender}_{ij} + \epsilon_{ij} \tag{1}$$

where $u_{0,i}$ and $w_{0,j}$ are, respectively, the by-participant and by-item adjustments to the intercept coefficient β_0, and $w_{1,j}$ is the by-item adjustment to the slope term for

```
subject item     gender VOT
F01     kh^.l&r    0.5  105
F01     kh^.tIxN   0.5  120
F01     khA9       0.5  104
F01     khE.tS&p   0.5  127
F01     khek       0.5  141
F01     khev       0.5  106
...
M20     thu.n&     -0.5 101
M20     thub       -0.5  66
M20     thuT       -0.5  67
M20     twhI.stIxd -0.5  69
M20     twhi.z&rz  -0.5  93
M20     twhIn      -0.5  85
```

Listing 1: Example data-set from English.

gender, β_1. The varying intercepts for subjects, $u_{0,i}$, are assumed to be distributed as $Normal(0, \sigma_{u0})$; similarly, the varying intercepts for items $w_{0,j}$ have the distribution $Normal(0, \sigma_{w0})$, and the varying slopes for item by gender, $w_{1,j}$ have the distribution $Normal(0, \sigma_{w1})$. (Note that we follow the convention here that a Normal distribution is written in terms of its mean and standard deviation, not variance; in other words, we write $Normal(\mu, \sigma)$). The residual error, ϵ, is assumed to have the distribution $Normal(0, \sigma_e)$. Finally, the varying intercepts and slopes for item, $w_{0,j}, w_{1,j}$ are assumed to have correlation ρ_w. In `lme4` syntax, the above model corresponds to the following (`datE_stops` refers to the data frame):

```
lmer(VOT ~ 1 + gender + (1 | subject) + (1 + gender | item),
     dat = datE_stops)
```

Because `lme4` assumes an intercept term, the `1 +` can be omitted, as in:

```
lmer(VOT ~ gender + (1 | subject) + (gender | item), dat = datE_stops)
```

The above model requires the estimation of the parameters listed in 2. (Note that in Bayesian linear mixed models, u_{0i}, w_{0j}, w_{1j} are also parameters; but these are not of primary interest in studies such as this example which address questions only about group effects rather than about patterns of differences across individuals or across items.)

$$\beta_0, \beta_1, \sigma_{u0}, \sigma_{w0}, \sigma_{w1}, \rho_w, \sigma_e \tag{2}$$

The intercept β_0 represents the grand mean VOT. Note that it does not make sense to fit varying slopes for gender by participants in this model because gender is a between-participants factor (i.e., we can't investigate the effect of gender on the participants). Gender is, however, a within-items factor, so varying slopes for gender can be fit by items (i.e., we *can* investigate the effect of gender on the items).

In the frequentist framework, we would just need to run the `lmer` function as shown above. However, in the Bayesian linear mixed model, some more work is needed before we can run the corresponding function in the package `brms`. While it is possible to fit the previous function with `brms` without the specifications described below, this would make use of default priors, which may not be adequate for every data-set.

The very first step is to define prior distributions for each of the parameters in the model. Once the priors are defined, the model is fit, as we will show below. The end-product of a Bayesian analysis is a so-called joint posterior distribution of all the parameters. These posterior distributions show the probability distributions of plausible values of the parameters, given the data and the model. These posteriors are then used for statistical inference.

2.1. Defining priors

We start by choosing the following prior distributions for the parameters in the model. The notation "$\sim$" should be read as "is distributed as." $Normal_+(0, 100)$ is a short-hand for a truncated (or half) normal distribution with mean 0 and standard deviation 100, which includes only positive values. The choice of a truncated normal distribution instead of a normal distribution is necessary for the priors on standard deviations, because standard deviation cannot be less than 0.

1. $\beta_0 \sim Normal(0, 200)$
2. $\beta_1 \sim Normal(0, 50)$
3. $\sigma_e \sim Normal_+(0, 100)$
4. $\sigma_{u0} \sim Normal_+(0, 100)$
5. $\sigma_{w0}, \sigma_{w1} \sim Normal_+(0, 100)$
6. $\rho_w \sim LKJ(2)$

When defining priors, it is a good idea to visualize them so that the researcher can decide whether these are reasonable. The priors chosen here are visualized in Figure 1.

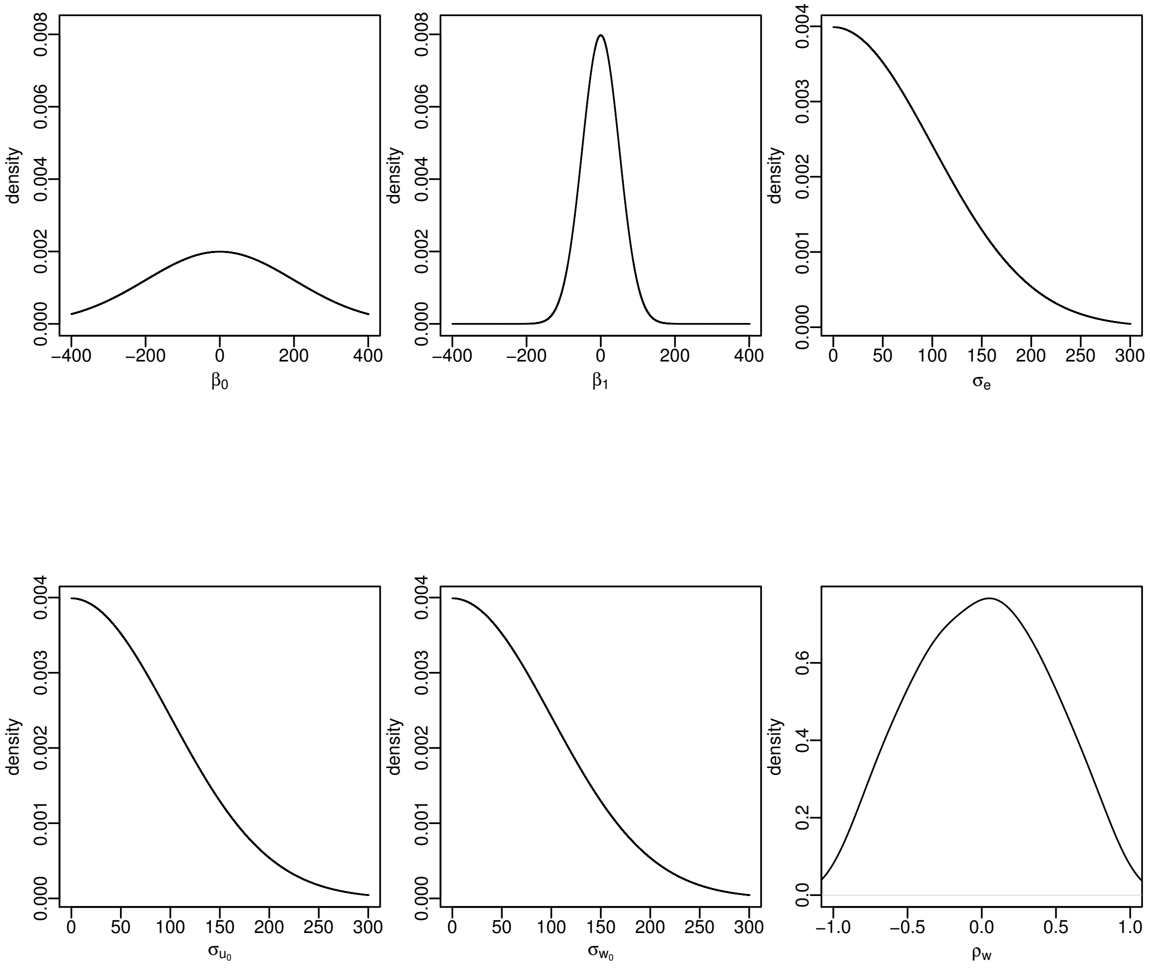

Figure 1: Prior distributions for the parameters of the varying intercepts and varying slopes linear mixed model.

Priors express beliefs about the plausible values of the parameters; these beliefs can be based on expert or domain knowledge, or could be based on already-available data. For example, the theoretically interesting parameter for us is the effect of gender, β_1. This is assumed here to have a prior distribution Normal(0, 50), which implies that the parameter is believed to lie between -100 and $+100$ ms with probability 95%. This range arises because 95% of the probability in a Normal distribution with mean μ and standard deviation σ is contained within the approximate range $\mu \pm 2 \times \sigma$. This prior for β_1 assumes quite a wide range of possible values; it could easily be much more constrained. For example, if we know from previous research that gender effects on VOT are unlikely to be larger than 40 ms, the prior Normal(0, 20) could be quite reasonable. As we will show below, a so-called sensitivity analysis (which is standard practice in medicine, economics, and other fields) can be useful to check whether the posterior distribution of the parameter is affected by the prior specification.

The priors on the standard deviations ($Normal_{+}(0, 100)$) express the fact that a value below 0 is impossible, and that larger values (larger than 200) are unlikely. Why did we assume that values larger than 200 are unlikely? This decision should ideally come from knowledge about and experience with VOT data. However, nothing hinges on this particular choice of prior; we could have chosen a prior that is even more spread out (has an even larger standard deviation) without any substantial change in the outcome.

An important point to notice in the prior specification for the intercept β_0 is that we allow negative values in the prior distribution. Since we assume in this paper that VOT can only have values greater than 0, we could in principle constrain the prior to allow only positive values. However, because there is sufficient data in the present examples, these decisions about the prior will not have a major impact on the posterior distribution. If we had very little data to work with, the prior would be highly influential in determining the outcome.

The correlation parameter ρ_w uses the LKJ-correlation prior which is based on a method for generating random correlation matrices developed by Lewandowski et al. (2009). This takes a numerical parameter that determines the shape of the distribution. A standard choice is to choose the LKJ(2) prior, because it assumes that extreme values (±1) are highly unlikely. This prior, which is currently only available in Stan (Stan Development Team, 2017b) (and of course in `brms`), can be used for essentially arbitrarily large correlation matrices of random effects.

The priors we have chosen here allow a broad range of values for the parameters, and are called regularizing, weakly informative priors (Gelman et al., 2017). "Regularizing" here means that extreme values are disallowed or downweighted; for example, a prior on a correlation parameter would be regularizing if it disallows or downweights extreme values such as -1 or $+1$, which are quite unlikely in data.

Weakly informative priors give some minimal amount of information and have the objective of yielding stable inferences (see also Chung et al., 2013; Gelman et al., 2008). For most applications of Bayesian modeling, it is standard to use regularizing, weakly informative priors, but informative priors based on expert opinion or prior knowledge can and should be considered as well; we illustrate this with an example below.

```
library(brms)

priors <- c(set_prior("normal(0, 200)", class = "Intercept"),
                      set_prior("normal(0, 50)", class = "b",
                                coef = "gender"),
                      set_prior("normal(0, 100)", class = "sd"),
                      set_prior("normal(0, 100)", class = "sigma"),
                      set_prior("lkj(2)", class = "cor"))
```

Listing 2: Example of prior specification in `brms`.

2.2. Specifying priors in `brms`

The prior specification in Listing 2 defines different priors for each `class` of parameter. Class `Intercept` is the intercept parameter (i.e., β_0), class `b` are all the slopes in a model (in this case it indicates the β_1 parameter), i.e., the slope for gender; the parameter for gender can be marked by writing the name of the predictor variable in the data-frame (here, female is coded as $+1/2$ and male as $-1/2$). The parameters of class `sd` are the standard deviation parameters for the random effects (in this case, σ_{u0}, σ_{w0}, and σ_{w1}) and the class `sigma` is the standard deviation of the residual error ϵ (i.e., σ_e). The parameters of class `sd` and `sigma` are automatically constrained by `brms` to not have values lower than 0. Finally, the parameter of class `cor` is the correlation parameter, and can be used to define LKJ priors for correlations in an essentially arbitrarily complex random effects structure. It is this LKJ prior that ensures that the correlation parameter(s) can generally be estimated, even when data are relatively sparse. Note, however, that when data are too sparse to estimate such parameters, the uncertainty of the estimate will be high—one would learn nothing new (beyond what is specified through the prior distribution) about that parameter from the data.

2.3. Specifying the linear mixed model in `brms`

After we have defined the priors as shown above, we define the linear mixed model using `lme4` syntax, as shown in Listing 3. The `brms` code has some differences from `lme4`:

1. The term `family = gaussian()` makes explicit the underlying likelihood function that is implicit in `lme4`. Other linking functions are possible, exactly as in the `glmer` function in `lme4`.
2. The term `prior` takes as argument the list of priors we defined in Listing 1. Although this specification of priors is optional, we suggest that the researcher explicitly specify each prior. Otherwise, `brms` will define a prior by default, which may or may not be appropriate for the research area. We return to this point below.
3. The term `iter` refers to the number of iterations that the sampler makes to sample from the posterior distribution of each parameter (by default 2000).
4. The term `warmup` refers to the number of iterations from the start of sampling that are eventually discarded (by default half of `iter`).
5. The term `chains` refers to the number of independent runs for sampling (by default four).
6. The term `control` refers to some optional control parameters for the sampler, such as `adapt_delta`, `max_treedepth`, and so forth.

At this beginning stage, it is not important to understand the details behind iterations, warmup, chains, and the control structure. The values used in Listing 3 will suffice for typical data-sets in phonetics. Most of these are the default values for those terms, and in case they lead to warnings, Stan and `brms` will print out detailed suggestions on how to proceed; the researcher should follow the instructions in the warning mesages, and consult guide to Stan's warnings (`mc-stan.org/misc/warnings.html`). For example, when we used the default model in Listing 3 with the Mandarin data frame, the model did not converge and it was recommended to increase the default value (.8) of `adapt_delta` (see section 3.1.5).

```
m1M <- brm(formula = VOT ~ gender + (1 | subject) + (gender | item),
           data = datM_stops, family = gaussian(), prior = priors,
           iter = 2000, chains = 4, control = list(adapt_delta = 0.99))
```

Listing 3: Example of model specification in `brms`.

After the Bayesian linear mixed model has been fit by running the code shown in Listing 3, the next question is: how to draw inferences from the model output, and how to summarize the results? In the next section, we work through some examples illustrating the steps. In our first example below, in order to illustrate the advantages of using Stan and `brms`, we also compare the performance of the Bayesian model with the frequentist estimates.

3. Research questions

In our case study, we use published voice onset time (VOT) data measured in milliseconds for word-initial stops elicited from 10 adult female and 10 adult male speakers that use differences in VOT in some way to contrast at least two series of stops. We use data from 20 speakers of Dongbei (Northeastern) Mandarin Chinese (Li, 2013) and 20 speakers of North American English (Kong et al., 2012). The target stop productions were elicited in the same way across the two languages, using a picture-prompted word-repetition task that was developed to elicit word productions from young children, (Edwards and Beckman, 2008). Because the VOT measurements were made using the same criteria by the same group of researchers and their research assistants, they are amenable to evaluating the following questions:

1. Does VOT in the long-lag stops (aspirated stops in Mandarin and voiceless stops in English) differ by gender in each language?
   Li (2013), Peng et al. (2014), and Ma et al. (2017) show that in three different varieties of Mandarin, women tend to produce aspirated stops with longer VOT values relative to men. In motivating her study, Li reviews many previous studies showing that in both North American English and British English, women tend to produce voiceless stops with longer VOT values relative to men. These studies include Morris et al. (2008); Robb et al. (2005); Ryalls et al. (1997); Swartz (1992).
2. Is VOT in the long-lag stops predicted by speaker's typical vowel duration (as a proxy for speech rate)?
   A number of studies reviewed in Simpson (2012) suggest that cross-linguistically, women tend to speak more slowly and clearly. For example, Byrd (1994) measured longer utterance durations in female speakers of North American English and found that they tend to use less vowel reduction. Similarly, Hillenbrand et al. (1995) and many others have shown that female speakers produce longer stressed vowels than men. Building on this work as well as on work such as Kessinger and Blumstein (1997) and Pind (1995) showing that VOT is correlated with speaking rate, Li (2013) suggests that it is important to test for

effects of inter-speaker rate differences when examining apparent gender effects on VOT values.

3. Are there cross-linguistic differences between English and Mandarin for questions 1 and 2?

We investigate each of these questions next.

3.1. The effect of gender on VOT in long-lag stops

In order to address question 1, we begin by plotting voice onset time values for each participant in each language, grouping the participants by gender: see Figure 2. As the figure shows, the female speakers on average have longer VOT values than male speakers in the stops in the long-lag category (i.e., the voiceless stops of English and the aspirated stops of Mandarin). In contrast, in the other stop type, the effect of gender is either non-existent (in the voiced stops of English) or in the opposite direction (in the unaspirated stops of Mandarin). Li (2013) interpreted the interaction between gender and stop type for Mandarin as evidence that the effect of gender is an indirect result of a gender effect on speech clarity, with male speakers tending to reduce the contrast between unaspirated and aspirated stops and female speakers tending to enhance it. Evaluating the evidence for this interpretation of the Mandarin interaction is complicated by the fact that the distributions of the VOT values for the speakers with larger mean values cover a wider range and are more skewed relative to the distributions for the speakers with smaller mean values. This pattern suggests that a log transform is in order (Gelman and Hill, 2007, 59-65). However, the fact that most of the English speakers produced at least some tokens of voiced stops with voicing lead precludes a simple application of the transform to address the question of the gender effect in both languages using the same model. In keeping with the purpose of this tutorial introduction, therefore, we will defer the problem of how the gender effect depends upon the stop type for a future paper, and here include just the VOT values for the stop types with long-lag VOT.

We fit the Bayesian linear mixed model shown in Listing 3 for Mandarin and for English repeated here in Listing 4 for convenience. Along the way, we also fit the corresponding `lme4` models.

```
m1E <- brm(formula = VOT ~ gender + (1 | subject) + (gender | item),
           data = datE_stops, family = gaussian(), prior = priors,
           iter = 2000, chains = 4, control = list(adapt_delta = 0.99))
```

Listing 4: Model specification in `brms` to address Question 1.

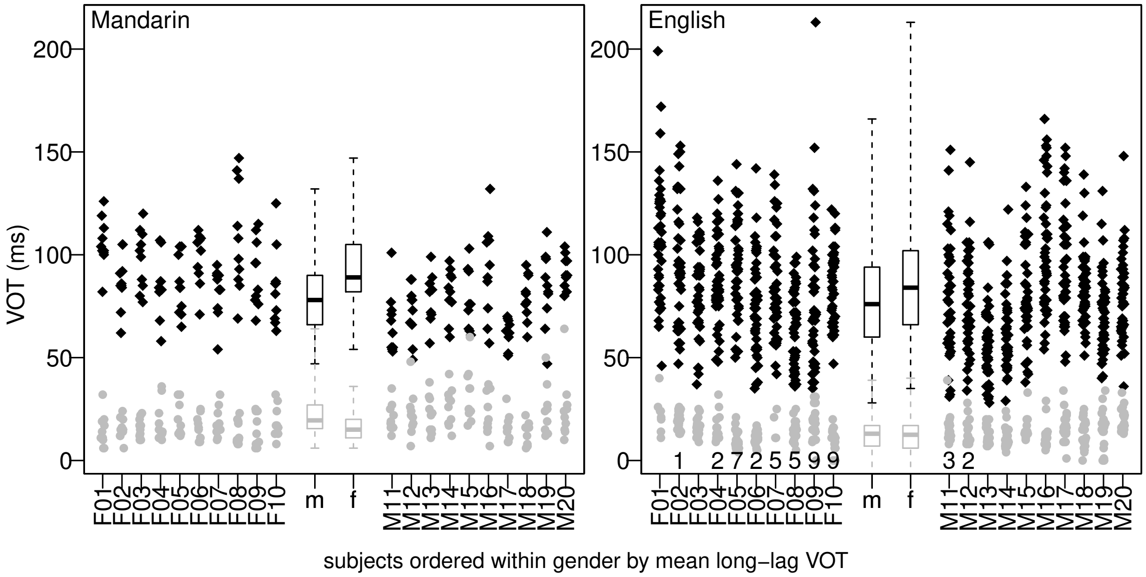

Figure 2: Stripcharts showing the distribution of VOT values for each of the participants, grouped by language (with Mandarin speakers in the left panel and English speakers in the right panel) and by gender within each language. Boxplots in the center of each panel show the median, inter-quartile range, and range for each gender and stop type. (The whiskers for the minimum VOT for English voiced stops are well below the bottom of the plot, due to the 77 tokens with voicing lead.) The grey shaded dots are for unaspirated/voiced stops that are not of interest for the current paper but are included nonetheless for comparison purposes.

3.1.1. Summarizing the results of a Bayesian analysis

We can summarize the posterior distributions of the parameters of a model graphically using the function `stanplot`, that calls the package `bayesplot` (Gabry and Mahr, 2017). In Listing 5, we plot the posterior distributions of the model `m1E` using histograms.

```
stanplot(m1E, type="hist")
```

Listing 5: Code for plotting posterior distributions of the model `m` using histograms.

The plots produced by `bayesplot` are based on the popular package `ggplot2` (Wickham, 2009) and thus plots made with `bayesplot` can be modified with `ggplot2` syntax. In Figures 3 and 4, for example, we modified the plots to show 95% Bayesian credible intervals, as well as the frequentist `lme4` estimates and 95% credible intervals, for Mandarin and English respectively. Note that by default `lme4` only outputs point value estimates for the standard deviation and correlation parameters; the Bayesian model will always deliver a posterior distribution. The figures show the posterior distributions as well as 95% credible intervals: these are the range over which we can be 95% certain that the true values of the parameter lie, given these particular data and the model. Credible intervals have a different meaning than frequentist confidence intervals; the latter refer to intervals that would contain the true unknown point mean value if the experiment were hypothetically be repeated multiple times. Thus, a single confidence interval technically does not tell us about the uncertainty about our estimate of the parameter (although researchers often treat frequentist confidence intervals as Bayesian credible intervals). For more on confidence intervals versus credible intervals, see Morey et al. (2015).

The posteriors for the Mandarin data show that female speakers have an increased VOT (over the grand mean), with an estimated mean 13 ms, and a 95% credible interval 3, 23 ms. Comparing the parameter estimates of `lme4` vs. Stan, we see the effect of regularizing priors in the correlation parameter. The `lme4` estimate for the correlation is on the boundary (-1), indicating a failure to estimate the parameter (Bates et al., 2015a); and the standard deviation of varying slope for gender by item is much smaller in `lme4`. In `lme4`, a correlation estimate near $+1$ or -1 suggests that there is insufficient data to estimate this parameter, and a simpler model without the correlation parameter should be fit (Bates et al., 2015a).

Consider now the Bayesian estimate for the correlation. Like all posterior distributions in a Bayesian analysis, this is a compromise (analogous to a weighted mean) between the prior and the data: when data are sparse, the prior will dominate in

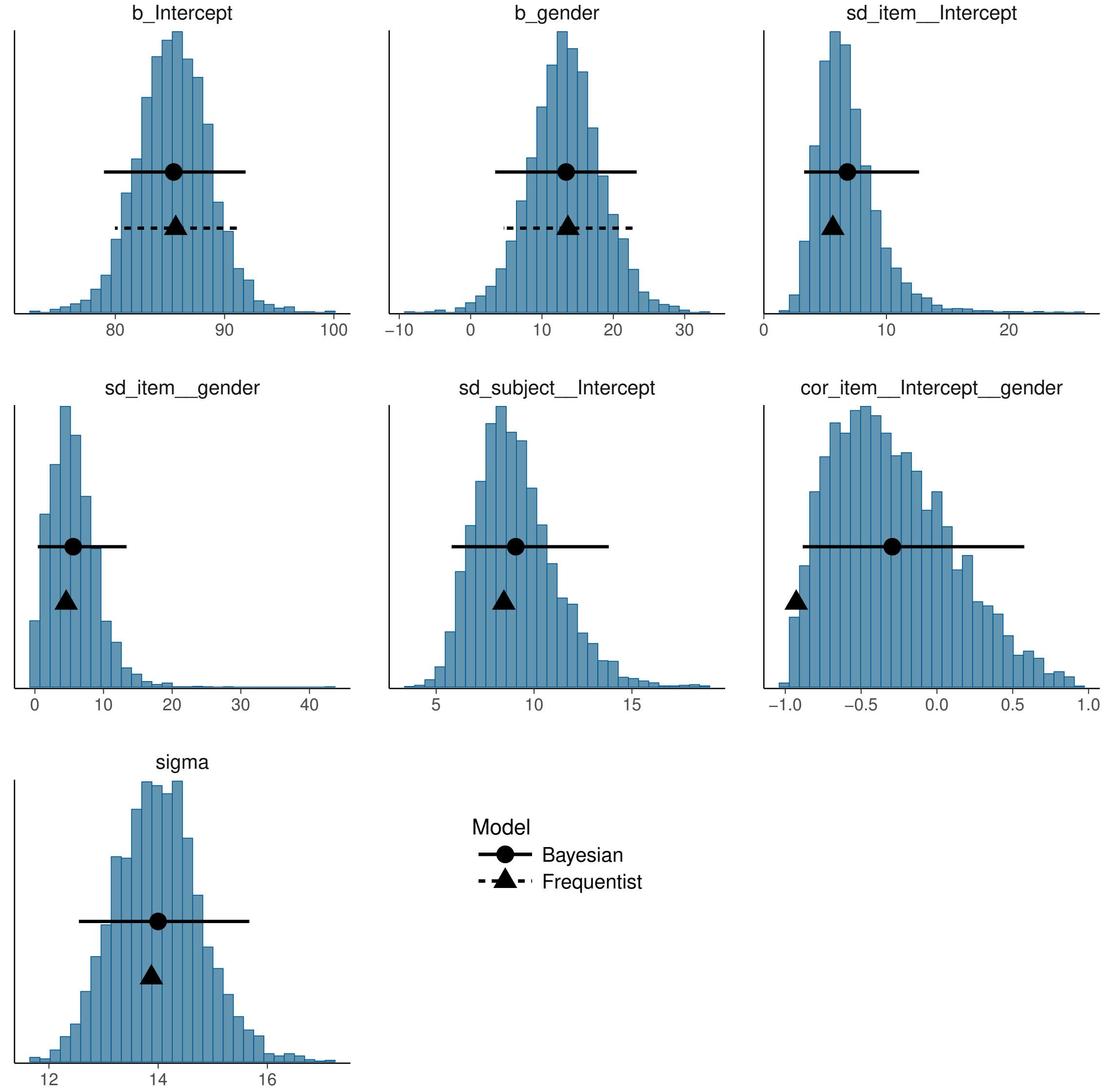

Figure 3: Posterior distributions of the parameters for the Mandarin linear mixed effects model investigating the effect of gender on VOT. The circles and the solid lines represent the mean of the posterior and the 95% Bayesian credible intervals respectively; the triangles and the dashed lines represent the frequentist (`lme4`) estimates and the 95% frequentist confidence intervals respectively.

determining the posterior, but when there are sufficient data-points, the data will largely determine the posterior and lead to estimates similar to lme4's. Notice, for example, that the correlation's posterior distribution is widely spread out between -1 and 1; the estimate is near zero but it has very high uncertainty. This wide distribution of the correlation is due to the regularizing effect of the LKJ(2) specification. Note that with brms, we have succeeded in estimating the posterior distribution of the parameter in the sense that we will not have a convergence failure. But we haven't learnt much about plausible values of the correlation parameter. If we had much more data from Mandarin, we could in principle get very accurate estimates of the correlation (which may or may not be 0). But as things stand, all that the inclusion of the correlation parameter in the model achieves is that it incorporates this source of uncertainty in the model. Also note that, in this particular data-set for Mandarin, the correlation parameter is not going to affect our posterior distribution of the effect of gender (β_1); without the correlation parameter, the posterior distribution has the same mean and credible interval (14, 95% credible interval 3, 23 ms). Such a no-correlation model can be fit in brms using the double-vertical-bar syntax of lme4:

```
VOT ~ gender + (1 | subject) + (1 + gender || item)
```

This is equivalent to the following:

```
VOT ~ gender + (1 | subject) + (0 + gender | item) + (1 | item)
```

The posteriors from the English data show that the effect of gender has an estimated mean 6 ms, 95% credible interval -5, 20 ms. We discuss the interpretation of this and the Mandarin result in section 3.1.3.

3.1.2. Interpreting the results of a Bayesian analysis

Having fit the models for Mandarin and English, we now discuss different ways of drawing inferences from the posterior distributions. Our favored approach is to display the posterior distribution of the parameter of interest, because we believe that the researcher's focus should be on the estimate of the parameter and our uncertainty of that estimate. However, other approaches exist; we discuss hypothesis testing using Bayes factors, and predictive evaluation using an approximation of leave-one-out cross validation.

3.1.3. Interpreting the posterior distribution

How to interpret these posterior distributions of the gender effect? If we had carried out a frequentist analysis using the lme4 package or the like, we would have

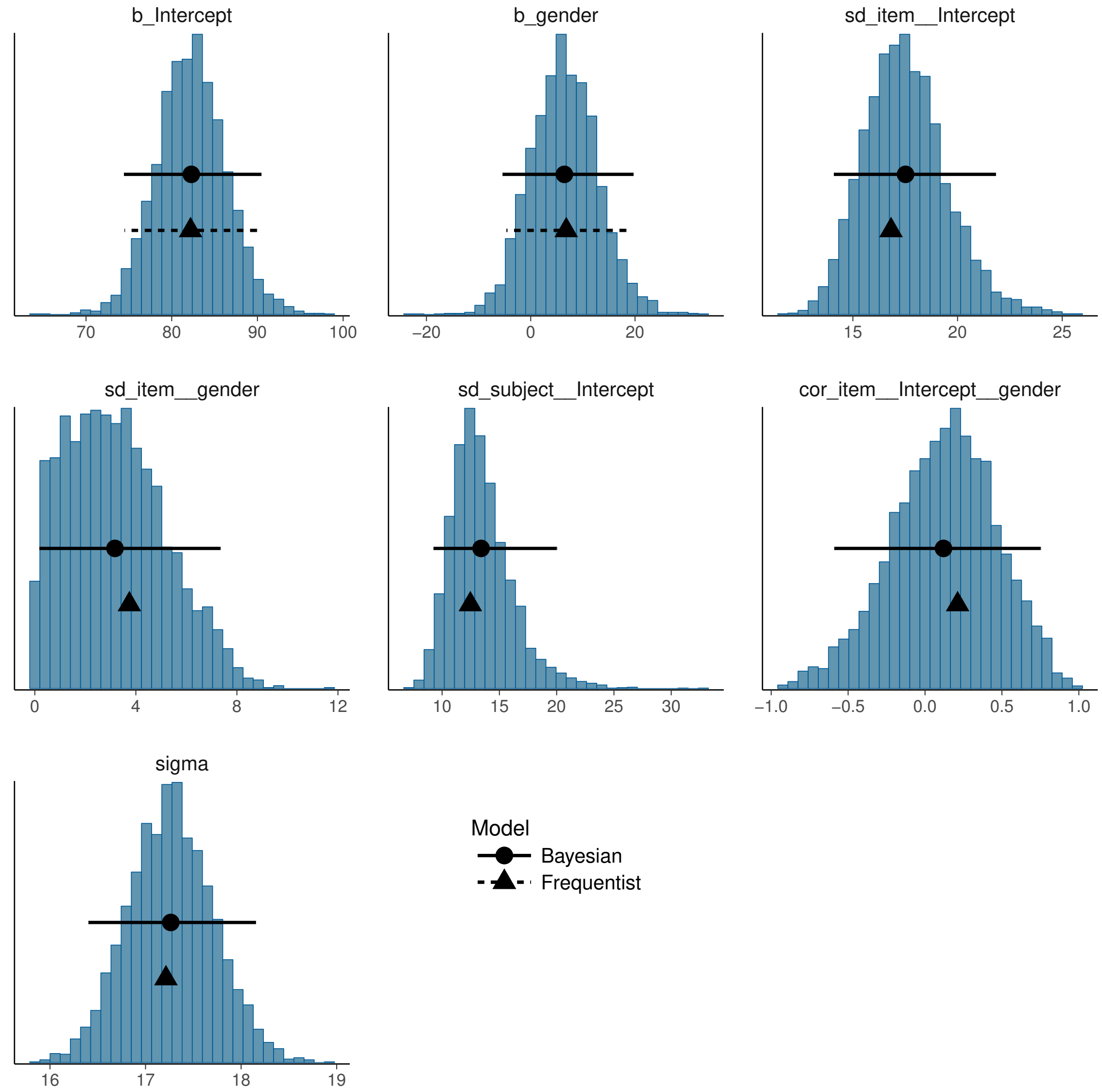

Figure 4: Posterior distributions of the English data investigating the effect of gender on VOT. The circles and the solid lines represent the mean of the posterior and the 95% Bayesian credible intervals respectively; the triangles and the horizontal dashed lines represent the frequentist (`lme4`) estimates and the 95% frequentist confidence intervals respectively.

found a “significant” effect of gender on VOTs in Mandarin but “no significant effect” of gender in English. We can see this by just examining the frequentist confidence intervals for the effect of gender in Figures 3 and 4: if the error bar in the frequentist estimate for the gender effect spans zero on the x-axis, the effect would not be significant at Type I error probability of 0.05.

Should the conclusion be that Mandarin shows effects of gender but English does not? As statisticians have repeatedly pointed out (Wasserstein and Lazar, 2016), these kinds of binary decisions (based on p-values or any other statistic such as credible intervals) are common but highly misleading. On the one hand, when power is low, if an effect comes out significant it is guaranteed to be an overestimate (Vasishth et al., 2018). On the other hand, when power is low and an effect is found to be non-significant, this often misleads researchers into the invalid belief that they have shown that the null hypothesis is true. An example from psycholinguistics is Phillips et al. (2011), where the absence of interference effects are presented as evidence of absence. We suggest that the focus should instead be on obtaining the estimates and our uncertainty of these estimates. Furthermore, in order to interpret the effect of gender in these languages, the totality of the evidence available in the literature for these languages should be quantitatively investigated, using, for example, a meta-analysis (Jäger et al., 2017; Nicenboim et al., 2018). Evidence synthesis has been taken to a new level through the MetaLab project at Stanford (http://metalab.stanford.edu). Phonetics can also benefit greatly by quantifying what we have learnt from previous studies, instead of classifying the literature on a phenomenon into two bins, significant and non-significant results.

From the Mandarin and English data (taken out of the context of previous work on this topic), we would conclude that there is some evidence for the effect of gender on VOT in the two languages.

Notice that we are not rejecting any null hypothesis here, and we are computing no p-value. The most useful information we can obtain from a Bayesian model is the posterior distribution of the parameter of interest (here, β_1). However, if necessary, one can use this posterior to carry out hypothesis testing using Bayes factors (Lee and Wagenmakers, 2014).

3.1.4. Using Bayes factor for hypothesis testing

Simplifying somewhat, the Bayes factor is the ratio of the likelihoods of the two models under comparison. For example, if we want to carry out a Bayes factor analysis of the effect of gender on VOT in Mandarin, we could compare the following two models `m1M` and `m0M`:

```
## m1M: more complex model
```

```
VOT ~ gender + (1 | subject) + (gender | item)
## m0M: simpler model
VOT ~ 1 + (1 | subject) + (gender | item)
```

We will follow a convention of indicating with a subscript the order in which the models are being compared. For example, the Bayes factor, BF_{10}, indicates the extent to which the data supports m1M over m0M; BF_{01} would indicate the support for m0M over m1M. A convention widely followed in Bayesian statistics is that a Bayes factor value of 10-30 would constitute strong evidence in favor of the more complex model, and smaller values, say 3-10, constitute weaker evidence. Values close to one indicate no meaningful evidence for one model or the other.

Example code for computing the Bayes factor is shown in Listing 6. The full model m1M has the same syntax as shown earlier, except that one term must be added: save_all_pars = TRUE; this specification is needed to save the samples for all parameters as the samples are needed for the Bayes factor calculation. After fitting the full model m1M, one fits the reduced model, and then one can compare the two models using the function bayes_factor in brms. Note the order of the models m1M and m0M in the function; this order matters in interpreting the Bayes factor because it is a ratio of likelihoods of two alternative models. A ratio like 3 computed using the function bayes_factor(m1M, m0M) states that data is three times more likely to have occurred under m1M than m0M. If the function call had been bayes_factor(m0M, m1M), the output would be 1/3.

When computing Bayes factors, it is important to check the sensitivity of the Bayes factor to the prior for the parameter we are interested in testing. This is because the Bayes factor is very sensitive to the choice of the prior, even in cases where the posterior is not (or barely) affected. For example, in the Mandarin case, we computed Bayes factors under three different priors for β_1, the effect of gender. Table 1 shows that when the prior on β_1 is very constrained (Normal(0, 20)), the evidence is in favor of an effect of gender. With increasingly diffuse priors, the evidence for the effect of gender becomes progressively weaker. Table 1 also shows the mean and 95% credible interval for the gender parameter; this remains largely unchanged despite the different priors used (with more diffuse priors, the estimate of the mean increases slightly). Thus, the Bayes factor is being affected by the prior. The point to take away here is that Bayes factors can be a useful tool, but one should think carefully about the prior, and one should consider the Bayes factor under several different priors, including informative ones. In the present case, for example, the more diffuse priors might be quite unrealistic. Researchers in psycholinguistics and phonetics are not used to thinking about what constitutes a reasonable prior; but this is not

```
# We use the same priors as before:
prior_summary(m1M,all=F)
##                   prior     class   coef group resp dpar nlpar bound
## 2         normal(0, 50)         b gender
## 3        normal(0, 200) Intercept
## 4  lkj_corr_cholesky(2)         L
## 6        normal(0, 100)        sd
## 12       normal(0, 100)     sigma

m1M <- brm(formula = VOT  ~ gender + (1 | subject) + (gender | item),
            data = datM_stops,
            family = gaussian(), prior = priors,
            save_all_pars = TRUE,
            iter = 2000, chains = 4,
            control = list(adapt_delta = 0.99))

m0M <- update(m1M, formula = ~ .-gender)

BF10 <- bayes_factor(m1M, m0M)
```

Listing 6: Example code showing how Bayes factor can be computed in `brms`.

unusual in areas like medicine, where expert opinion often needs to be incorporated into the data analysis (O'Hagan et al., 2006).

Prior on β_1	BF_{10}	Posterior of β_1		
		mean	lower	upper
Normal(0,20)	6.19	12.71	2.97	22.45
Normal(0,50)	2.85	13.53	3.51	23.42
Normal(0,70)	2.21	13.48	3.44	23.52

Table 1: The influence of the prior on Bayes factor; an example from the Mandarin data. The models being compared are linear mixed effects models with and without the gender factor.

3.1.5. Assessing model convergence

In Bayesian modeling, it is important to check whether the model has converged. One metric for convergence is the so-called $\hat{R}$ statistic (`Rhat` in the model output); this is the ratio of the between to within chain variance. When each of the chains is sampling from the posterior, the end-result is that the amount of between-chain variability is approximately the same as within-chain variability, so that the ratio of these variances is approximately 1. Thus, an Rhat of approximately 1 for each parameter is one indication that the model has converged. In addition, one should check the effective sample size (`n_eff`). This is an estimate of the number of independent draws from the posterior distribution. Since the samples are not independent, `n_eff` will generally be smaller than the total number of samples. How large `n_eff` should be depends on the summary statistics that we want to use. But as a rule of thumb, `n_eff` should be larger than 10% of the total number of samples. Thus, in our case, the number of samples is 4000 (1000 from each of the four chains, having discarded the first 1000 as warm-up), so `n_eff` should ideally be larger than 400.

The fitted models above provide information about convergence diagnostics. As shown in Listing 7, the `summary` function in `brms` provides information regarding the `Rhat` and n_eff diagnostics.

Apart from `Rhat` values and the number of effective samples, another indication of successful convergence is that, when the chains are plotted, they overlap. This is a visual check that confirms that the chains are mixing well. Figure 5 shows an example trace plot for the Mandarin model's fixed effects parameters (the intercept and slope). This plot is generated using a built-in function call:

```
stanplot(m1M, pars = c("^b"))
```

```
summary(m1E)

 Family: gaussian
  Links: mu = identity; sigma = identity
Formula: VOT ~ gender + (1 | subject) + (gender | item)
   Data: datE_stops (Number of observations: 836)
...
Population-Level Effects:
          Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
Intercept    82.29      4.05    74.43    90.43        475 1.00
gender        6.46      6.42    -5.39    19.65        611 1.00
...
```

Listing 7: An extract from the summary output from a fitted brms model.

If the chains had not converged, one would see the trajectories of the chains going in different directions.

A more detailed investigation of convergence can also be achieved using the shinystan package (Stan Development Team, 2017a). This package provides a self-contained graphical user interface for interactively exploring the posterior of a Bayesian model, including help and glossaries. For example, we can explore the model m1M in the following way:

```
library(shinystan)
shiny_m1M <- launch_shinystan(m1M)
```

While convergence problems may seem daunting at first, suggestions regarding how to fix them appear on brms output when warnings are printed. Moreover, in many cases the convergence problems appear due to an incorrect model specification (e.g., having varying slopes for gender by participants in the previous model: ...+ (gender | subject)), or in the selection of priors (e.g., using a uniform prior when there is not enough data), and thus they can be easily fixed. In some specialized models (not discussed here), convergence problems are due to the geometry of the posterior distribution and this may require reparameterization by modifying the Stan code generated by brms (using make_stancode()); see the Reparameterization section of the Optimizing Stan Code chapter in the Stan documentation (Stan Development Team, 2017b). Discussion of this topic is beyond the scope of this introductory tutorial.

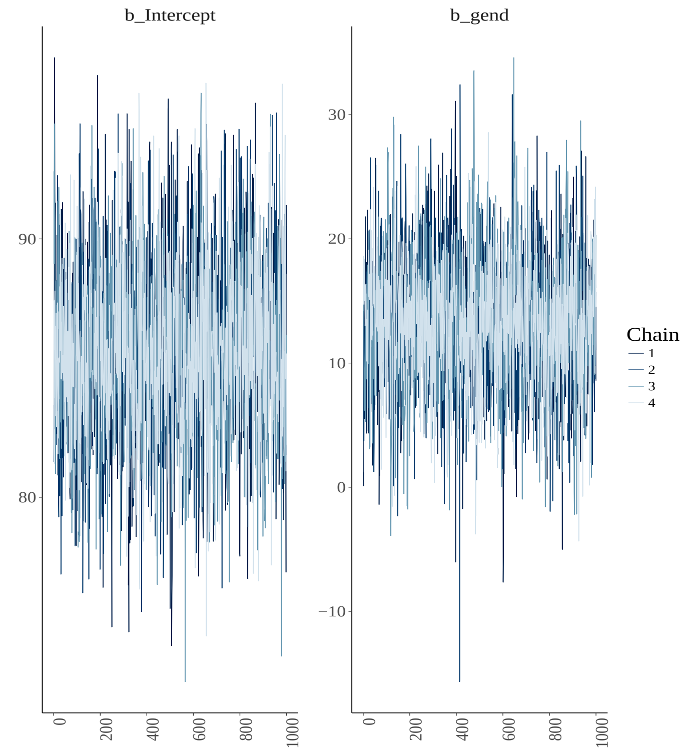

Figure 5: Trace plots for the fixed effects parameters in the Mandarin data.

3.1.6. Assessing model fit, sensitivity analysis, and model comparison

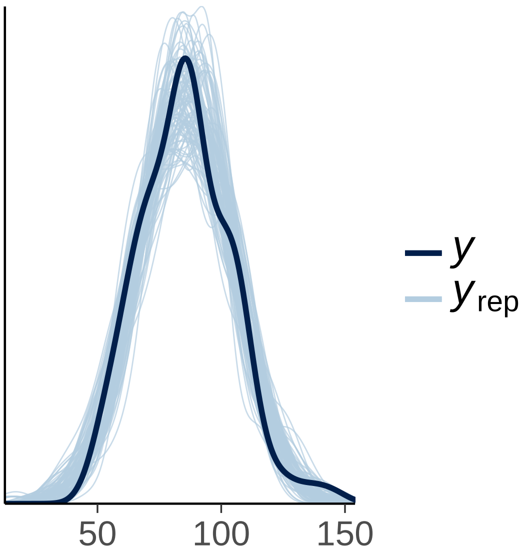

Figure 6: Posterior predictive checks for the Mandarin data. The lines marked y_{rep} refer to the posterior predictive values generated by the model, and the black solid line are the observed data.

One commonly used method for assessing how well the model matches up with the observed data is to use so-called posterior predictive checks. Essentially, we generate many instances of new data after computing the posterior distributions of the parameters and compare them to observed data.

Posterior predictive samples can easily be generated from model fit using `brms`. An example is shown in Figure 6. This figure was generated by typing the following command, which compares the data with 100 predicted samples:

```
pp_check(m1M, nsamples = 100)
```

Here, the observed data are plotted alongside the predicted data generated by the model. If the predicted and observed data have similar distributions, we can conclude that the model has a reasonable fit.

There is an obvious drawback to this approach: one is evaluating the model against the very data that was used to estimate the parameters. It should not be surprising that the model predicts data that were used to fit the parameters! However, when model assumptions are grossly violated, even this easy test will fail. For example, if there are some (say, 5%) 0 ms VOTs in a data-set (e.g., due to data loss or some other reason), or if there is a mixture of distributions generating the data, and the model assumes a Gaussian likelihood, the posterior predictive distributions and the distribution of the data will not line up. For a real-life example of such a situation, see Vasishth et al. (2017). In Figure 7, we use the Mandarin data to simulate such a situation by randomly replacing 5% of the data with 0 ms values. The mismatch between the data and the posterior predictive values is clear visually.

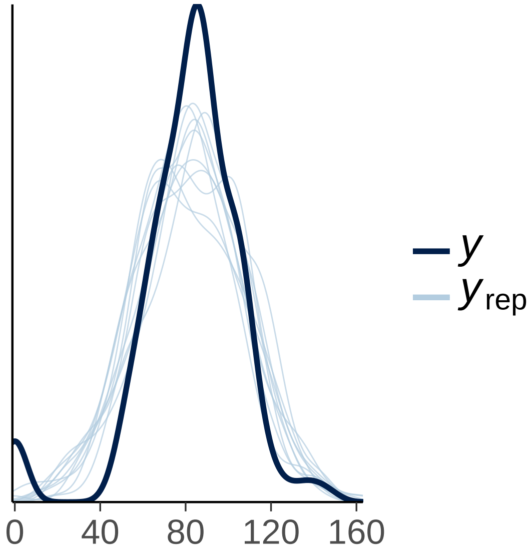

Figure 7: An artificial example of how a mismatch between the model assumptions and the data can lead to poor posterior predictive fits. Here, 5% of the Mandarin VOT values were randomly replaced with 0 ms values and the same model as the one for research question 1 was fit. Now the posterior predictive check shows that the lack of fit between the data and the predicted values.

A better approach for evaluating the predictive performance of a model may be to test the model's predictions against new data, or against held out subsets of data (Vehtari and Ojanen, 2012; Gelman et al., 2014; Piironen and Vehtari, 2015). This procedure is called k-fold cross validation. Another variant is called leave-one-out (LOO) cross-validation; in LOO, we leave one data point out and fit the model, and then predict the held-out data (Vehtari et al., 2015a). The distance between the predicted and observed data can then be used to quantify the relative predictive error when comparing competing models. The `brms` package provides tools for doing these kinds of model evaluations. In `brms` one can do an approximation of LOO cross-validation using the built-in function `loo` (Vehtari et al., 2015b). In the above example for LOO, the function call is simple: `loo(m1M, m0M)`. The output of this command in the present case would be a quantity called the LOO Information Criterion (LOOIC)

```
loo(m1M, m0M)
##                LOOIC    SE
## m1M          1653.25 22.32
## m0M          1654.69 22.59
## m1M - m0M      -1.44  2.17
```

Listing 8: Model comparison using PSIS-LOO. A smaller value of LOOIC indicates a model with better predictions.

for each model; this quantifies the estimated predictive error, and displayed alongside it is its standard error. The function `loo` also computes the difference in estimated predictive error between the two models, along with a standard error of the difference. These two quantities can then be used to compare the two models: we compute the difference in LOOIC values of the two models ($\Delta LOOIC$), and then use the standard error to determine whether the difference in LOOIC includes 0 as a value by computing $\Delta LOOIC \pm 2 \times SE$.

Listing 8 shows the output of the model comparison using LOO. Here, the difference in predictive error, the final line in the output, is $\Delta LOOIC \pm 2 \times SE = -1.44 \pm 2.17 = -5.78, 2.91$. The Bayes factor based hypothesis test showed some weak evidence in favor of `m1M`, LOO shows virtually no difference between the models. This is because the experimental manipulation produced a very small change in the predictive performance of `m1M` in comparison with `m0M`. This absence of a difference in predictive performance does *not* mean that small effects are not important for evaluating a phonetic theory. In general, even with moderate sample size, it can be difficult to compare nested hierarchical models (such as linear mixed models) based on predictive performance (Wang and Gelman, 2014).

This method of model selection is useful, however, when one is interested in comparing the predictive performance of very different competing models. For fully worked examples of this approach (with reproducible code and data) in the context of cognitive modeling in psycholinguistics, see Nicenboim and Vasishth (2018); Vasishth et al. (2017).

Thus, our analysis demonstrates how a complete data analysis can be carried out in the Bayesian framework. To recapitulate the steps:

1. Explore the data using graphical tools; visualize the relationships between variables of interest.
2. Define model(s) and priors.
3. Fit model(s) to data.

4. Check for convergence (Rhat, n_eff, trace plots).
5. Carry out inference by
   (a) summarizing and displaying posterior distributions, or
   (b) Bayes factors, or
   (c) evaluating predictive performance of competing models using k-fold cross-validation or approximations of leave-one-out cross-validation.

A Bayesian analysis is clearly more involved than a frequentist one and requires some thought and judgment when defining priors. The models can take a long time to compile, which can be quite frustrating. However, the reward is substantial: one can turn the focus to quantitative estimates of effect sizes. This is much more informative than the significant/not-significant distinction, as discussed earlier. When quantitative models exist, these empirical estimates can be compared against model predictions. In cases where quantitative models don't yet exist, empirical estimates provide the basis for developing such models.

In the remainder of the paper, we address the other two questions we posed, and in doing so, demonstrate the flexibility of the Bayesian framework.

3.2. The effect of typical vowel duration on VOT

In order to investigate the effect of vowel duration on VOT, we can use the vowel duration for each participant as a predictor to a model in the same way we used gender before; the only difference is that the vowel duration is a continuous measure, whereas gender was a categorical variable that we coded using sum contrasts.

One question that arises here is: How do we estimate the vowel duration for each speaker? One possibility is to take the mean vowel durations from the same long-lag stop trials that provide the VOT values and use those as a predictor; another is to take the mean vowel durations from the unaspirated (Mandarin) and voiced (English) trials. We take the second alternative in this paper in order to avoid using information from the long-lag trials twice in the same model.

Figure 8 shows, for the two languages, the relationship between the mean VOTs and mean vowel duration, along with the standard errors of each mean (the error bars). This uncertainty expressed by the standard deviation arises because we measure each participant's vowel duration and VOT values multiple times, and these measurements will naturally have some error about the (unknown) true value for that participant.

Looking at Figure 8, it seems that there is a linear relationship between mean vowel duration and mean VOT in both languages. A linear model fit to the data yields the following estimates for Mandarin: mean 0.26, 95% confidence interval -0.02, 0.53. For English, the estimates are: 0.2, 95% confidence interval -0.18, 0.59.

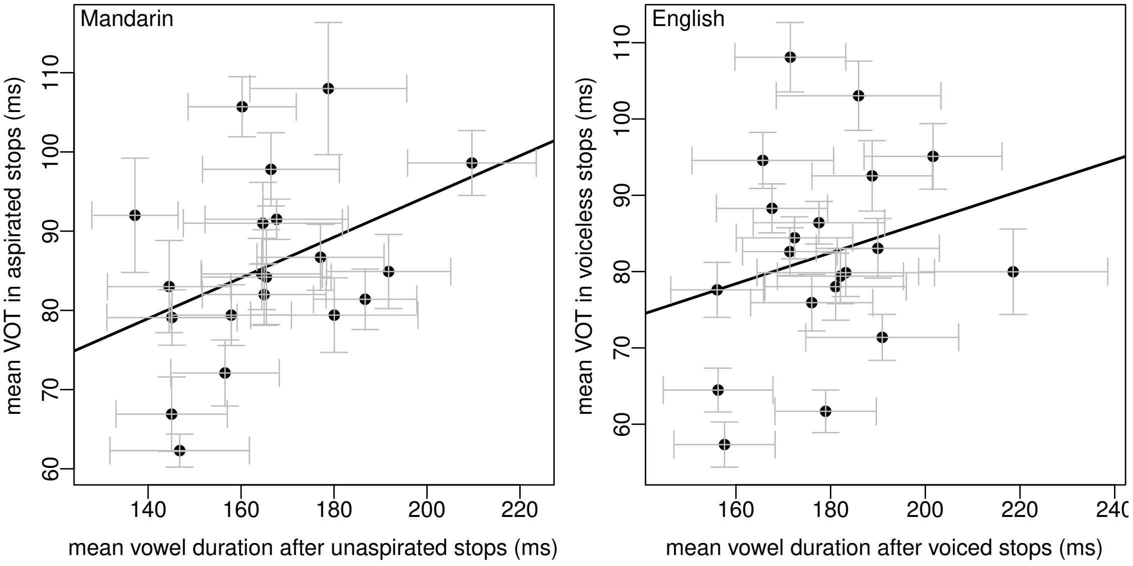

Figure 8: Mean VOT values are shown against mean vowel duration in the two languages; the error bars represent standard errors of each measure. There seems to be a positive relationship between the means; we see this from the linear regression line fitted to the means. But this linear fit does not take the uncertainty of these estimated means into account.

However, what these linear models do not take into account is the uncertainty of each of the estimated mean values. It is well-known that aggregating data in this way can lead to correlations arising from ignoring the relevant variance components.[2]

In our models, we will take the measurement errors of the mean VOT and mean vowel duration estimates into account. Thus, if the VOT and vowel duration estimated for one participant i is VOT_i and vdur_i, we can also record the standard errors of these estimates, *and take that uncertainty into account in our model.* Such a measurement error model is straightforward to implement in `brms`, and is shown in Listing 9. Here, we define priors for the intercept and slope fixed effect, and for the standard deviation of the residuals. The predictor includes not only vowel duration but also the corresponding standard error; this is written `me(c_meanvdur, sevdur)`, where `c_meanvdur` is the centered mean vowel duration, and `sevdur` is the standard error of the mean vowel duration. For defining the prior for this predictor, we use the concatenation of the string `me(c_meanvdur, sevdur)`, with the brackets and commas stripped out: `mec_meanvdursevdur`. This is just how `brms` deals with this parameter name. We chose a Cauchy(0, 5) prior for the predictor; this prior assumes "fat tails", which means that values far from 0 are considered to be possible. As an exercise, the reader may wish to change this prior to Normal(0, 10) to see whether the posterior changes substantially (it should not). The dependent variable, meanVOT, also has a standard error associated with it, and this is expressed in `brms` by writing `meanVOT | se(seVOT)`. An important detail in `brms` syntax when fitting a measurement error model on the dependent variable is that residual error is estimated by adding the term `(1 | subject)`.

3.3. Interpreting the results

The estimates from the two measurement error models (for Mandarin and English) are shown in Tables 2 and 3. Such a tabular summary is an alternative way to summarize the posterior distributions of interest. The 95% credible intervals show that for both languages, there isn't any evidence for an effect of vowel duration on VOT: the estimates are near 0 now. This is because the standard errors of the estimates for VOT and for vowel duration add uncertainty to the estimates.

These results could mean that mean vowel duration is not a good measure of speech rate (which might yet have an effect on VOT; cf. Kessinger and Blumstein,

[2] An analogous problem arises when we use repeated measures ANOVA; an effect that is "not significant" using linear mixed models can become "significant" once the data are aggregated by averaging over sets of items (or over groups of participants) and then analyzed using repeated measures ANOVA.

```
## data frame used:
head(meansM)
## # A tibble: 6 x 5
##   subject meanVOT    seVOT c_meanvdur   sevdur
##     <chr>   <dbl>    <dbl>      <dbl>    <dbl>
## 1     F01   105.7 3.791951 -5.3116317 11.63752
## 2     F02    86.7 4.158125 11.5364481 13.66959
## 3     F03    97.8 4.623130  0.8793482 14.72106
## 4     F04    84.9 4.677250 26.1910909 13.35989
## 5     F05    84.6 4.492463 -1.1088362 13.03397
## 6     F06    98.6 4.099322 44.1128520 13.84189
priors_cauchy <- c(set_prior("normal(0, 200)", class = "Intercept"),
            set_prior("cauchy(0, 5)", class = "b",
                      coef = "mec_meanvdursevdur"),
            set_prior("normal(0, 20)", class = "sdme"),
            set_prior("normal(0, 20)", class = "sd"))

m2M_error <- brm(formula = meanVOT | se(seVOT)  ~ me(c_meanvdur, sevdur) +
            (1 | subject),
            data = meansM, family = gaussian(), prior = priors_cauchy,
            iter = 2000, chains = 4,
            control = list(adapt_delta = 0.999,
                          max_treedepth=15))
```

Listing 9: Measurement error model, investigating the effect of vowel duration on VOT in Mandarin.

	Estimate	lower	upper
Intercept	85.92	79.36	95.1
mean vowel duration	0.65	-0.32	2.61

Table 2: Estimates from the measurement error model for Mandarin.

	Estimate	lower	upper
Intercept	86.55	49.26	124.42
mean vowel duration	1.6	-8.14	10.12

Table 3: Estimates from the measurement error model for English.

1997 and Pind, 1995). Another possibility is that the effect is very small, and that we do not have enough data to draw any conclusions. As always, an important question to ask is, how do the present data relate to existing work on this topic? A quantitative evaluation of the current data in the context of existing estimates is a very important but underappreciated tool. If we had a systematic way to summarize our prior knowledge on this question, we could have incorporated this knowledge by using informative priors in the analysis.

3.4. Cross-linguistic differences between Mandarin and English for questions 1 and 2

We can address this question by fitting two separate hierarchical models: (a) the main effects and interaction of language and gender, (b) the main effects and interaction of mean vowel duration and language. For simplicity, we ignore measurement error on the mean vowel duration, but this can be easily added to the model if necessary.

	Estimate	lower	upper
Intercept	83.84	76.79	90.63
gender	10.03	2.56	17.38
lang	3.06	-9.81	16.10
gender:lang	6.93	-8.18	22.59

Table 4: The main effects of gender and language, and their interaction.

As Tables 4 and 5 show, we see some evidence for gender (mean 10 ms, 95% credible interval 3, 17); and some weak evidence for vowel duration affecting VOT (mean 4 ms, 95% credible interval 0, 8). All other effects have wide uncertainty and have means not far from 0.

	Estimate	lower	upper
Intercept	83.52	76.85	90.44
cmeanvdur	3.83	-0.34	8.08
lang	3.13	-10.54	16.90
cmeanvdur:lang	1.75	-6.11	9.83

Table 5: The main effects of centered and scaled vowel duration and language, and their interaction.

4. Concluding remarks

We have attempted to provide a practical entry point into Bayesian modeling using the package `brms`, which serves as a convenient and easy-to-use front-end to the probabilistic programming language Stan. Other ways to use Stan are through the front-end `rstanarm` (Gabry and Goodrich, 2016), and the R package `rstan` (Guo et al., 2016). The package `rstanarm` has fewer customizations possible compared to `brms`, but has precompiled code for commonly used models, which leads to faster data analysis. Other versions also exist for python (pystan), Matlab, Mathematica, Julia, Stata; see `mc-stan.org` for more detail on these alternatives.

In order to develop a better understanding of this approach to analyzing data, it is of course important to acquire experience and further exposure to fitting and interpreting models. Several useful books have recently appeared that are intended for a general audience. Two important recent ones are Kruschke (2014) and McElreath (2016); these provide a complete introduction to different aspects of Bayesian modeling.

Bayesian methods also find application in cognitive modeling; two useful introductory books are Lee and Wagenmakers (2014) and Farrell and Lewandowsky (2018).

Acknowledgments

The recordings and measurements described in this paper were made with the support of NIH grant DC02932 to Jan Edwards, an Ohio State University Department of Linguistics Targeted Investment Award to Fangfang Li and Eun Jong Kong, and University of Lethbridge start-up funds to Fangfang Li. For partial support of this research, we thank the Volkswagen Foundation (through grant 89 953) and the Deutsche Forschungsgemeinschaft (through grant VA 482/8-1) to Shravan Vasishth, which funded Bruno Nicenboim. We are very grateful to the reviewers for helpful comments, and especially to Paul-Christian Bürkner for developing `brms`, for helpful comments on the `brms` code appearing in this paper, and for real-time support.

References

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.

Baayen, R. H., Vasishth, S., Kliegl, R., Bates, D., 2017. The cave of shadows: Addressing the human factor with generalized additive mixed models. Journal of Memory and Language, 206–234.

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., 2015a. Parsimonious mixed models, arXiv e-print.
URL http://arxiv.org/abs/1506.04967

Bates, D., Maechler, M., Bolker, B., Walker, S., 2015b. Fitting linear mixed-effects models using lme4. Journal of Statistical Software 67, in Press.

Bürkner, P.-C., 2016. brms: An R package for Bayesian multilevel models using Stan. Journal of Statistical Software 80 (1), 1–28.

Byrd, D., October 1994. Relations of sex and dialect to reduction. Speech Communication 15 (1–2), 39–54.

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.

Clark, H. 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), 335–359.

Edwards, J., Beckman, M. E., 2008. Methodological questions in studying consonant acquisition. Clinical Linguistics and Phonetics 22 (12), 937–956.

Farrell, S., Lewandowsky, S., 2018. Computational Modeling of Cognition and Behavior. Cambridge University Press.

Gabry, J., Goodrich, B., 2016. rstanarm: Bayesian applied regression modeling via stan. R package version 2 (1).

Gabry, J., Mahr, T., 2017. bayesplot: Plotting for Bayesian Models. R package version 1.4.0.
URL https://CRAN.R-project.org/package=bayesplot

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

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.

Gelman, A., Simpson, D., Betancourt, M., Aug. 2017. The prior can generally only be understood in the context of the likelihood. ArXiv e-prints.

Guo, J., Lee, D., Sakrejda, K., Gabry, J., Goodrich, B., De Guzman, J., Niebler, E., Heller, T., Fletcher, J., 2016. rstan: R Interface to Stan. R 534, 0–3.

Hillenbrand, J., Getty, L. A., Clark, M. J., Wheeler, K., May 1995. Acoustic characteristics of American English vowels. Journal of the Acoustical Society of America 97 (5), 3099–3111.

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.

Kessinger, R. H., Blumstein, S. E., 1997. Effects of speaking rate on voice-onset time in thai, french, and english. Journal of Phonetics 25 (2), 143–168.

Kong, E. J., Beckman, M. E., Edwards, J. R., November 2012. Voice onset time is necessary but not always sufficient to describe acquisition of voiced stops: The cases of Greek and Japanese. Journal of Phonetics 40 (6), 725–744.

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

Lee, M. D., 2011. How cognitive modeling can benefit from hierarchical Bayesian models. Journal of Mathematical Psychology 55 (1), 1–7.
URL http://dx.doi.org/10.1016/j.jmp.2010.08.013

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

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

Li, F., 2013. The effect of speakers' sex on voice onset time in Mandarin stops. The Journal of the Acoustical Society of America 133 (2), EL142–EL147.

Ma, J., Chen, X., Wu, Y., Zhang, L., 2017. Effects of age and sex on voice onset time: Evidence from Mandarin voiceless stops. Logopedics Phoniatrics Vocology, 1–7.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, R. H., Bates, D., 2017. Balancing Type I Error and Power in Linear Mixed Models. Journal of Memory and Language 94, 305–315.

McElreath, R., 2016. Statistical rethinking: A Bayesian course with examples in R and Stan. Vol. 122. CRC Press.

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., Wagenmakers, E.-J., Oct 2015. The fallacy of placing confidence in confidence intervals. Psychonomic Bulletin & Review.
URL http://dx.doi.org/10.3758/s13423-015-0947-8

Morris, R. J., McCrea, C. R., Herring, K. D., 2008. Voice onset time differences between adult males and females: Isolated syllables. Journal of Phonetics 36 (2), 308–317.

Nicenboim, B., Roettger, T., Vasishth, S., 2018. Using meta-analysis for evidence synthesis: The case of incomplete neutralization in German, under review, Journal of Phonetics.

Nicenboim, B., Vasishth, S., 2016. Statistical methods for linguistic research: Foundational Ideas - Part II. Language and Linguistics Compass 10 (11), 591–613.
URL http://dx.doi.org/10.1111/lnc3.12207

Nicenboim, B., Vasishth, S., 2018. Models of retrieval in sentence comprehension: A computational evaluation using Bayesian hierarchical modeling. Journal of Memory and Language.

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.

Peng, J.-F., Chen, L.-M., Lee, C.-C., 2014. Voice onset time of initial stops in Mandarin and Hakka: Effect of gender. Taiwan Journal of Linguistics 12 (1), 63–79.

Phillips, C., Wagers, M. W., Lau, E. F., 2011. Grammatical illusions and selective fallibility in real-time language comprehension. Experiments at the Interfaces 37, 147–180.

Piironen, J., Vehtari, A., 2015. Comparison of Bayesian predictive methods for model selection. arXiv preprint arXiv:1503.08650.

Pind, J., 1995. Speaking rate, voice-onset time, and quantity: The search for higher-order invariants for two Icelandic speech cues. Attention, Perception, & Psychophysics 57 (3), 291–304.

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

Robb, M., Gilbert, H., Lerman, J., 2005. Influence of gender and environmental setting on voice onset time. Folia Phoniatrica et Logopaedica 57 (3), 125–133.

Ryalls, J., Zipprer, A., Baldauff, P., 1997. A preliminary investigation of the effects of gender and race on voice onset time. Journal of Speech, Language, and Hearing Research 40 (3), 642–645.

Simpson, A. P., March 2012. Phonetic differences between male and female speech. Language and Linguistics Compass 3 (2), 621–640.

Stan Development Team, 2017a. shinystan: Interactive visual and numerical diagnostics and posterior analysis for Bayesian models. R package version 2.4.0.
URL http://mc-stan.org/

Stan Development Team, 2017b. Stan: A C++ library for probability and sampling, version 2.15.0.
URL http://mc-stan.org/

Swartz, B. L., 1992. Gender difference in voice onset time. Perceptual and motor skills 75 (3), 983–992.

Vasishth, S., Mertzen, D., Jäger, L. A., Gelman, A., 2018. The statistical significance filter leads to overoptimistic expectations of replicability.
URL https://psyarxiv.com/hbqcw

Vasishth, S., Nicenboim, B., Chopin, N., Ryder, R., 2017. Bayesian hierarchical finite mixture models of reading times: A case study, submitted to Psychological Review.
URL https://osf.io/fwx3s/

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. R package version 0.1.3.
URL https://github.com/jgabry/loo

Vehtari, A., Ojanen, J., 2012. A survey of Bayesian predictive methods for model assessment, selection and comparison. Statist. Surv. 6 (0), 142–228.
URL http://dx.doi.org/10.1214/12-SS102

Wang, W., Gelman, A., 2014. Difficulty of selecting among multilevel models using predictive accuracy. Statistics at its Interface 7, 1–8.

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.
URL http://dx.doi.org/10.1080/00031305.2016.1154108

Wickham, H., 2009. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.
URL http://ggplot2.org