Bayesian data analysis in the phonetic sciences: A tutorial introduction

Shravan Vasishth[a,*], Mary E. Beckman[b,*], Bruno Nicenboim[a], 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 uses a case study involving voice onset time (VOT) data from Dongbei (Northeastern) Mandarin and North American English to demonstrate how Bayesian linear mixed models can be fit using the programming language Stan. Through the case study, we demonstrate some of the advantages of the Bayesian framework: (i) the researcher can flexibly define the underlying process that they believe to have generated the data; (ii) one can obtain direct information regarding the uncertainty about the parameter that relates to the theoretical question being studied; (iii) because the posterior distributions of the parameters depend only on the priors and the data at hand, there is no reliance on assumed properties of the data generative process under hypothetical repeated sampling, leading to more realistic expectations about what we have learnt from the data. However, 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 a detailed case study, 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.

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

The last 30 years have seen several rapid developments in probabilistic programming languages for carrying out Bayesian inference. After the BUGS (Bayesian Inference Using Gibbs Sampling) project began in 1989 at the MRC Biostatistics Unit at Cambridge, several software packages, such as WinBUGS (Lunn et al., 2000), and JAGS (Plummer, 2016) have made Bayesian methods accessible for non-statistician end-users. As a consequence of the wide availability of such software, even relative newcomers to Bayesian methods can deploy these approaches in their research quite easily. The latest move in the direction of flexible and user-friendly software is Stan (Stan Development Team, 2017). Developed by a group of statisticians led by Andrew Gelman at Columbia University, Stan represents a powerful new addition to applied statistics.

Stan is potentially of great value to the phonetic sciences, and therefore we feel that it is worth putting in effort to understand its capabilities. In this paper, do not aim to convince the reader that it is worth learning to fit Bayesian models; for such an introduction, other sources, such as Kruschke (2014) and McElreath (2016) are recommended. Here, we simply assume that the reader is interested in a quick, practical, and non-technical introduction to Bayesian modeling.

From our perspective, the main attraction of Bayesian data analysis is that it allows for flexible model specification (Lee, 2011; Nicenboim and Vasishth, 2016). Bayesian methods directly answer the research question, instead of focusing on rejecting a strawman null hypothesis and then indirectly drawing inferences about the research hypothesis.

Bayesian analysis is in principle easy and intuitive. We set up a probability model, where our data y are assumed to be generated from some probability density function that takes an unobserved vector of parameters $\boldsymbol{\theta}$: $\boldsymbol{y} \sim f(\boldsymbol{\theta})$. For example, a common assumption is to treat voice onset time (VOT) values as being generated from a normal distribution with some unknown mean μ and standard deviation σ (thus, the vector of parameters is $\boldsymbol{\theta} =< \mu, \sigma >$). We can write the assumed generative process as $VOT \sim Normal(\mu, \sigma)$. Crucially in the Bayesian framework, parameters of the vector $\boldsymbol{\theta}$ have a prior distribution defined over them (this is discussed in more detail below). Having specified the priors and the probability model, using the observed data we derive (using software packages such as Stan) the conditional probability density function of the parameter(s) given the data; this is written $f(\boldsymbol{\theta} \mid \boldsymbol{y})$. The inferential goal of modeling is to obtain the posterior distributions of the parameters, and to draw conclusions from these posterior distributions.

Although the logic behind Bayesian data analysis is easy, a price to be paid is that fitting and evaluating a model takes effort, time, and thought. Many decisions must

be made along the way. Prior distributions of parameters must be defined, model convergence assessed, and (depending on the goal of the modeling) the predictive performance of the model evaluated.

In this tutorial, we will provide an informal introduction to Bayesian data analysis in Section 2, and then present in more detail a case study geared for the phonetic sciences. This case study is intended to provide a practical first entry to Bayesian data analysis in an informal manner. We do not pretend to cover all aspects of Bayesian modeling here; but suggestions for further reading are provided at the end. In our case study, we will focus on (generalized) linear mixed models (Pinheiro and Bates, 2000; Baayen et al., 2008; Bates et al., 2015), also known as hierarchical models (Gelman and Hill, 2007), because these are the standard tool today in experimental research in linguistics and the psychological sciences. We assume in this paper that the reader has fit linear mixed models using the R package lme4 (Bates et al., 2015).

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 English /t/ and /k/ produced by male and female participants. Participants in each gender category (sum-contrast coded as +1 for female and −1 for male) are asked to produce multiple stop-initial words, resulting in multiple measurements of VOT for each participant. For $i = 1, \ldots, I$ participants and $j = 1, \ldots, J$ items, we could fit a varying intercepts linear mixed model of the following type:

$$VOT_{ij} = \beta_1 + u_i + w_j + \beta_2 \times \text{gender}_i + \epsilon_{ij} \tag{1}$$

where u_i and w_j are, respectively, the by-participant and by-item adjustments to the intercept coefficient β_1; in other words, the distribution of u_i is assumed to be $Normal(0, \sigma_u)$; similarly, $w_j \sim Normal(0, \sigma_w)$. The residual error, ϵ, is assumed to have a distribution $Normal(0, \sigma_e)$. Thus, this model requires the estimation of the following five parameters: $\beta_1, \beta_2, \sigma_u, \sigma_w, \sigma_e$. The intercept β_1 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. Gender is, however, a within-items factor, so varying slopes could be fit by items.

An alternative way to write the model is in terms of the assumed generative distribution:

$$VOT_{ij} \sim Normal(\beta_1 + u_i + w_j + \beta_2 \times \text{gender}_{ij}, \sigma_e) \tag{2}$$

The standard frequentist (more precisely, Neyman-Pearson) approach would proceed as follows. The four parameters are assumed to have unknown point values in nature, and we use maximum likelihood estimation (or restricted maximum likelihood) to obtain estimates of these parameters. For the estimates of the $\boldsymbol{\beta}$ parameters, we also obtain standard errors (SEs) and this allows us to carry out inference, especially on β_2, the effect of gender. We posit a null hypothesis, e.g., $H_0 : \beta_2 = 0$, and then evaluate whether the observed estimate $\hat{\beta}_2$ from the data is unexpected under this hypothesis. If it is unexpected, it is usual to argue that we found evidence for the effect; here, the effect of gender. If we fail to show that the data are unexpected under the null, it is common to accept the null as true; notice that arguing for the null is not valid under the frequentist framework; we can only reject the null hypothesis or fail to do so.[1] Absence of evidence is generally not evidence of absence.

"Unexpected" has a specific meaning here: the absolute value of the observed statistic $t_{obs} = (\hat{\beta}_2 - 0)/SE$ has to be greater than the approximate critical value 2 in order to reject the null hypothesis with Type I error 0.05.

The Bayesian approach proceeds differently. First, the parameters are assumed to have distributions; they are not just unknown point values. To make this concrete, consider the following prior distributions for the four parameters ($Normal_+$ indicates a half normal distribution including only positive values):

1. $\beta_1 \sim Normal(0, 200)$
2. $\beta_2 \sim Normal(0, 50)$
3. $\sigma_e \sim Normal_+(0, 100)$.
4. $\sigma_u \sim Normal_+(0, 100)$.
5. $\sigma_w \sim Normal_+(0, 100)$.

The distributions for the vectors of $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}$ parameters are visualized in Figure 1. They express the prior belief about the possible values of the parameters; for example, the β_1 parameter is assumed to lie within -400 and 400 with 95% probability.[2] Why these priors and not others? They express a belief about plausible values of the

[1] When we have high statistical power, it can be valid to accept the null (Hoenig and Heisey, 2001). But that is a very rare situation in linguistics and psychology, where power can be as low as 6-20% (Jäger et al., 2017).

[2] It may seem odd to use the prior that we have for the intercept β_1 given that VOT can never be negative in a voiceless stop. In a full Bayesian analysis, the effect of different priors on the posterior would be considered in a so-called sensitivity analysis. An example is presented in Vasishth et al. (2013). In the present case, we don't explore the effect of different mildly informative priors because these will not have much effect on the posterior, which will be largely determined by the data. If there had been much less data, the effect of the priors would be more noticeable.

parameters before we have seen the data. Such priors, which allow a broad range of values, are called vague or mildly informative priors (Gelman et al., 2017). For most applications of Bayesian modeling, it is standard to use mildly informative priors, but informative priors based on expert opinion or prior knowledge could be considered as well.

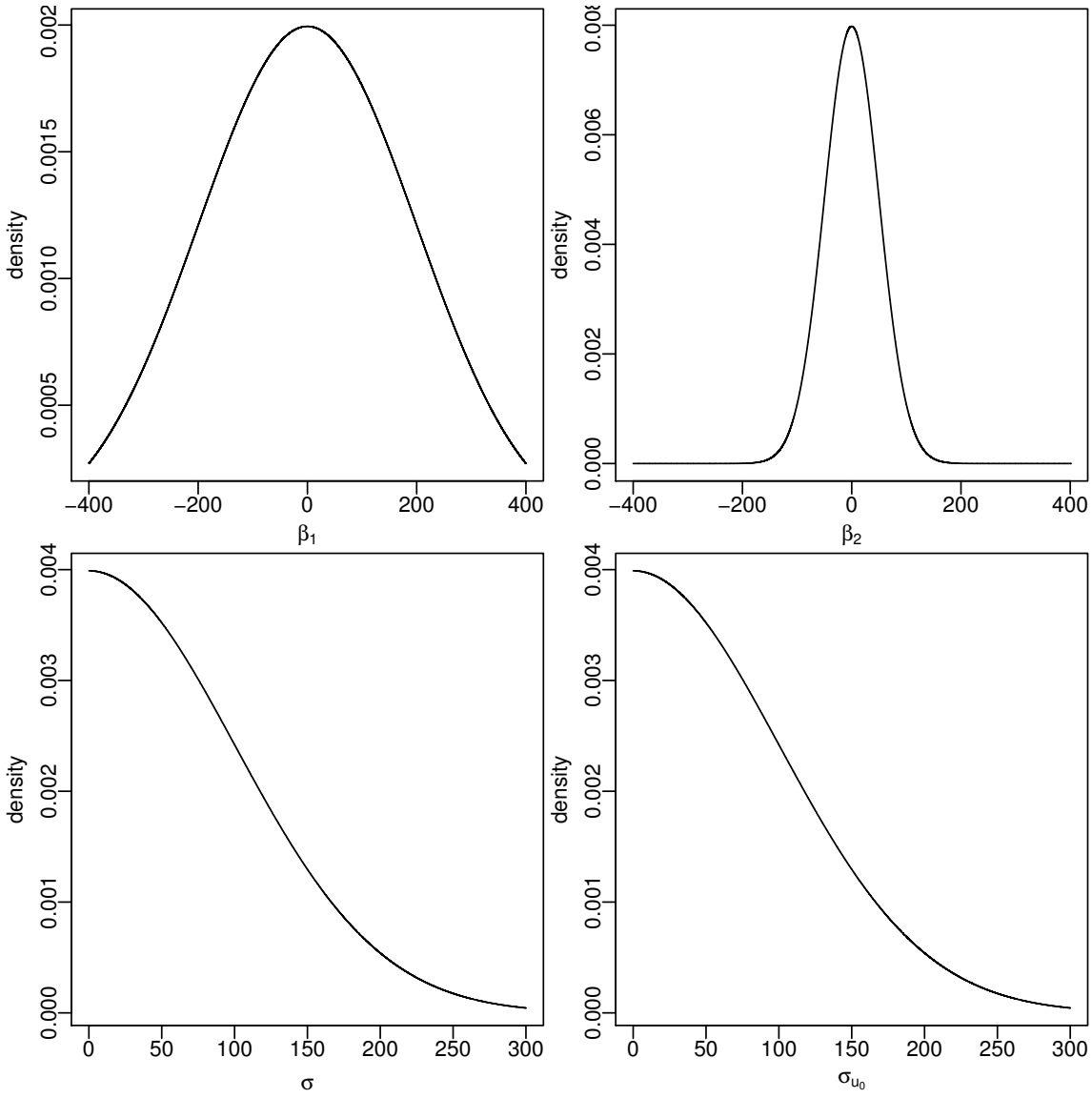

Figure 1: Prior distributions for the parameters of the varying intercepts hierarchical linear model.

Having specified vague or mildly informative prior distributions for the parameters, the data are then used to derive a posterior distribution for each of the parameters. The posterior distribution of each parameter can be seen as a weighted mean of the prior and the data: if a great deal of data are available, the posterior will be determined largely by the data, but if little data are available, the posterior will be largely influenced by the prior. This relationship between the posterior, the data, and the priors arises from Bayes' rule, which can be informally stated as follows:

$$\text{Posterior} \propto \text{Data} \times \text{Prior} \tag{3}$$

The end-product of a Bayesian analysis is always a posterior distribution for each parameter; the posterior tells us what the plausible values of the parameters are given the data and the priors. Once we have the posterior, we can (a) compute summary statistics such as the mean, median, and 95% credible intervals, (b) compute the probability that the parameter is positive or negative given the data, and (c) carry out hypothesis tests using Bayes factors or model comparison. More details are provided in Nicenboim and Vasishth (2016).

What are the advantages of carrying out a Bayesian analysis? We mention here the ones that we consider to be the most important.

First, models can be flexibly adapted to model the underlying generative process assumed. For example, a hierarchical mixture model may be a better way to model a particular data-set; this is straightforward in a Bayesian setting (Nicenboim and Vasishth, 2017; Vasishth et al., 2017). By contrast, the frequentist linear mixed model generally ties the researcher down to a fixed class of model.

Second, the posterior distribution places the focus squarely on the research question of interest; we can directly answer the question: what is our best estimate of the parameter of interest? By contrast, the frequentist approach involves a very indirect process of reasoning: we must look for evidence against a straw-man null hypothesis in order to argue for the specific research hypothesis. Despite the obvious problem that evidence against the null does not entail that there is evidence in favor of the specific alternative hypothesis of interest.

Third, the posterior distributions derived in the Bayesian approach depend only on the priors and the data we actually have at hand. By contrast, the frequentist approach appeals to hypothetical properties of data we *didn't* collect; often, these imagined properties are just a fantasy, as the recent controversy around replication failures suggests (Open Science Collaboration, 2015). The Bayesian approach leads to more realistic expectations about (non-)replicability.

In sum, the Bayesian approach avoids the common confusions that arise from trying to interpret hypothetical replications, and the associated Type I and II error

properties of an experiment. Instead of using the statistical significance filter to decide whether a result is "reliable", the researcher can focus on achieving higher precision of the parameter estimates that are of direct interest, and actually attempting to replicate their findings (Vasishth and Gelman, 2017).

What are the disadvantages of fitting a Bayesian model? The principal disadvantage is that it takes time to develop, evaluate, and understand a model. If the researcher is used to quickly fitting a frequentist linear mixed model to check whether an effect is significant or not, they may be surprised at the amount of effort that must be invested into working in the Bayesian framework. If one is used to simply obtaining a binary decision (accept or reject the null) from a data-set, Bayesian modeling will look like a nuisance. The value of Bayesian modeling is appreciated only once one realizes that the goal is to model the underlying process that generated the data, and to quantify our uncertainty about the parameter(s) of interest.

As mentioned above, the goal of this article is not to discuss all aspects of Bayesian data analysis, but rather to provide an entry point to researchers in the phonetic sciences. Towards this end, we limit ourselves to discussing a case study that unpacks the process of fitting a Bayesian hierarchical model. All data and code are available from the website https://github.com/vasishth/jopbayes.

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 (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.

4. Question 1: The effect of gender on VOT in long-lag stops

In order to address question 1, we begin by plotting positive voice onset time values (in milliseconds) by gender in each language; see Figure 2. In the current paper, we focus only on aspirated stops in Mandarin and voiceless stops in English, as these are the two categories that have shown relatively consistent gender related differences in both languages. Also, focusing on just voiceless stops in English lets us avoid for now the vexing question of how to evaluate distributions that involve negative VOT values (which have been treated in many different ways in the previous literature).

Using lme4 (Bates et al., 2015) syntax, a frequentist hierarchical linear model for each language, with gender as a predictor and varying intercepts for subjects and items, would have the following general form:

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

The variable item refers here to word type; each word type was elicited from each participant.

We now define a Bayesian hierarchical model with varying intercepts for participants and items. This is exactly the same model as the lmer syntax above implies.

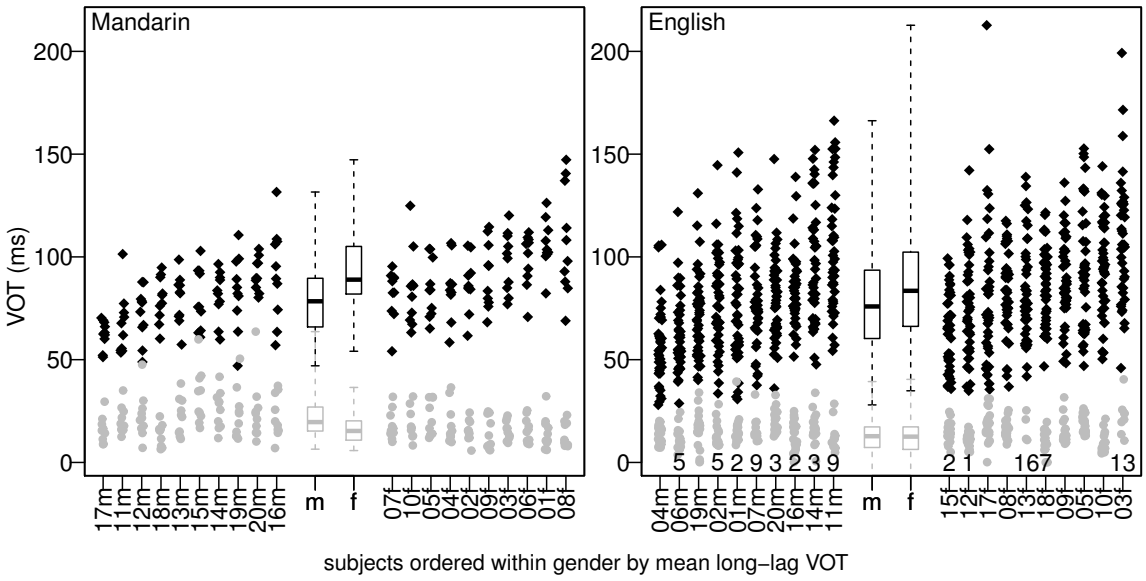

Figure 2: Plots showing the Mandarin and English VOT data by gender and aspiration/voicing in long-lag stops. The grey shaded dots are for unaspirated/voiced stops that are not of interest for the current paper but are included nonetheless for comparison purposes.

For $i = 1, \ldots, I$ participants and $j = 1, \ldots, J$ items, we could fit a varying intercepts linear mixed model of the following type:

$$VOT_{ij} = \beta_1 + u_i + w_j + \beta_2 \times \text{gender}_i + \epsilon_{ij} \tag{4}$$

where u_i and w_j are, respectively, the by-participant and by-item adjustments to the intercept coefficient β_1; that is, u_i is assumed to be distributed as $Normal(0, \sigma_u)$ and w_j is assumed to be distributed as $Normal(0, \sigma_w)$. The residual error, ϵ, is assumed to have a distribution $Normal(0, \sigma_e)$. Thus, this model requires the estimation of the following five parameters: $\beta_1, \beta_2, \sigma_u, \sigma_w, \sigma_e$.

Defining the model in Stan requires us to specify the priors for each of the

parameters, and the underlying probability model for the data:

$$\begin{aligned}
\beta_1 \sim & Normal(0, 200) \\
\beta_2 \sim & Normal(0, 50) \\
\sigma_e \sim & Normal_+(0, 100) \\
\sigma_u \sim & Normal_+(0, 100) \\
\sigma_w \sim & Normal_+(0, 100) \\
VOT_{ij} \sim & Normal(\beta_1 + u_i + w_j + \beta_2 \times \text{gender}_{ij}, \sigma_e)
\end{aligned} \tag{5}$$

Here, it is sufficient that the priors allow a range of values, perhaps even implausibly large ones; it is easily established through a sensitivity analysis that different mildly informative priors will not change the posteriors much.

Once the model is defined (see Listing 1 in the Appendix), fitting the model involves running several (typically four) parallel chains to sample from the posterior distributions of each parameter. The number of iterations for each chain is typically 2000, and the first half of the iterations are discarded because in the early stages the sampler may not sample from the posterior distribution; the first half is usually referred to as warm-up or burn-in. For reproducible code, please see the github site accompanying this paper.

The end-product of the model fits for Mandarin and English are the posterior distributions of the parameters. One can summarize the posterior distributions in terms of the 95% credible intervals, as in Table 1. The parameter of interest for evaluating the effect of gender is β_2. In the Mandarin and the English data, the credible intervals suggest a positive value for the parameter given the data; for Mandarin the interval is $[2, 11]$ with $P(\beta_2 > 0) \approx 1$ and for English $[-3, 10]$ with $P(\beta_2 > 0) = 0.88$. This suggests that, given these data, compared to males, female speakers have longer voice onset times in both languages. Note, however, that 0 is a plausible value for English (the credible interval contains 0), but not for Mandarin.

4.1. Interpreting the results of a Bayesian analysis

How to interpret these results? If we had carried out a frequentist analysis using the lme4 package, we would have found a "significant" effect of gender on VOTs in Mandarin but "no significant effect" of gender in English. Should the conclusion be that Mandarin shows effects of gender but English does not? These kinds of binary decisions (based on p-values or any other statistic such as credible intervals) are highly misleading (Wasserstein and Lazar, 2016). 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 investigated, using for example a random-effects meta-analysis (Jäger et al., 2017). From these data alone, we would conclude that there is suggestive evidence for the effect of gender on VOT in the two languages. One great advantage of the Bayesian approach is that we can talk explicitly about what we consider plausible values of the effect to be in each language; this is a big step beyond simplistic binary statements like "effect present" and "effect absent."

Thus, for standard hierarchical models such as these, the model fitting process is not challenging. Notice that we are not rejecting any null hypothesis, and 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_2). If necessary, one can use this posterior to carry out hypothesis testing (using Bayes factors); see Nicenboim and Vasishth (2016) and Paape et al. (2017) for examples from linguistics.

4.2. Assessing model convergence

	Mandarin					English				
	mean	2.5%	97.5%	n_eff	Rhat	mean	2.5%	97.5%	n_eff	Rhat
beta[1]	86	79	92	793	1.01	82	74	90	495	1.00
beta[2]	7	2	11	1032	1.00	4	-3	10	535	1.01
sigma_e	14	13	16	4000	1.00	17	16	18	4000	1.00
sigma_u	9	6	14	2110	1.00	13	9	20	2651	1.00
sigma_w	7	3	13	2179	1.00	17	14	22	4000	1.00

Table 1: Posterior means and 95% credible intervals of the parameters for the model evaluating the effect of gender in Mandarin and English. The column Rhat refers to a convergence diagnostic of the Markov Chain Monte Carlo samples, with a value near 1 indicating convergence; the column n_eff is a measure of effective sample size and should be higher than approximately 10% of the number of samples after warm-up or burn-in (here 10% of 4000).

In Bayesian modeling, it is important to check whether the model has converged. One metric for convergence is the so-called Rhat statistic; 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. When n_eff is below 10%, a simple reparameterization of the model can be implemented; this point is covered in the Stan documentation (Stan Development Team, 2017).

Another indication of successful convergence is that, when the chains are plotted, they look like "fat hairy caterpillars". This is a visual check that confirms that the chains are mixing well. Figures 3 and 4 show trace plots for the Mandarin and English models. These plots suggest the model has converged.

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

Although not discussed in the present paper, a complete Bayesian analysis would also assess how well the model's predictions of future data match up with the observed data. These so-called posterior predictive checks can easily be generated from the Stan model. Essentially, we generate many instances of new data after computing the posterior distributions of the parameters and compare them to observed data. A more sophisticated alternative is to hold out some data and then compare the model's predictions to that held-out data (k-fold cross validation). A fully worked example of this approach (with reproducible code and data) is in Nicenboim and Vasishth (2017), and a detailed discussion is in Nicenboim and Vasishth (2016).

5. Question 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. Throughout this paper, when we talk about vowel duration as a predictor, we mean scaled, centered vowel duration on a millisecond scale. This means that we subtract from the estimated vowel duration of each participant the mean vowel duration of all participants and we divide this by the standard deviation of the vowel durations; this has the effect that when the scaled, centered vowel duration is 0, this represents the mean of all participants. Centering generally reduces the correlation between predictors and makes the intercept and the effect of interest easier to interpret (Gelman and Hill, 2007), and scaling improves convergence.

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

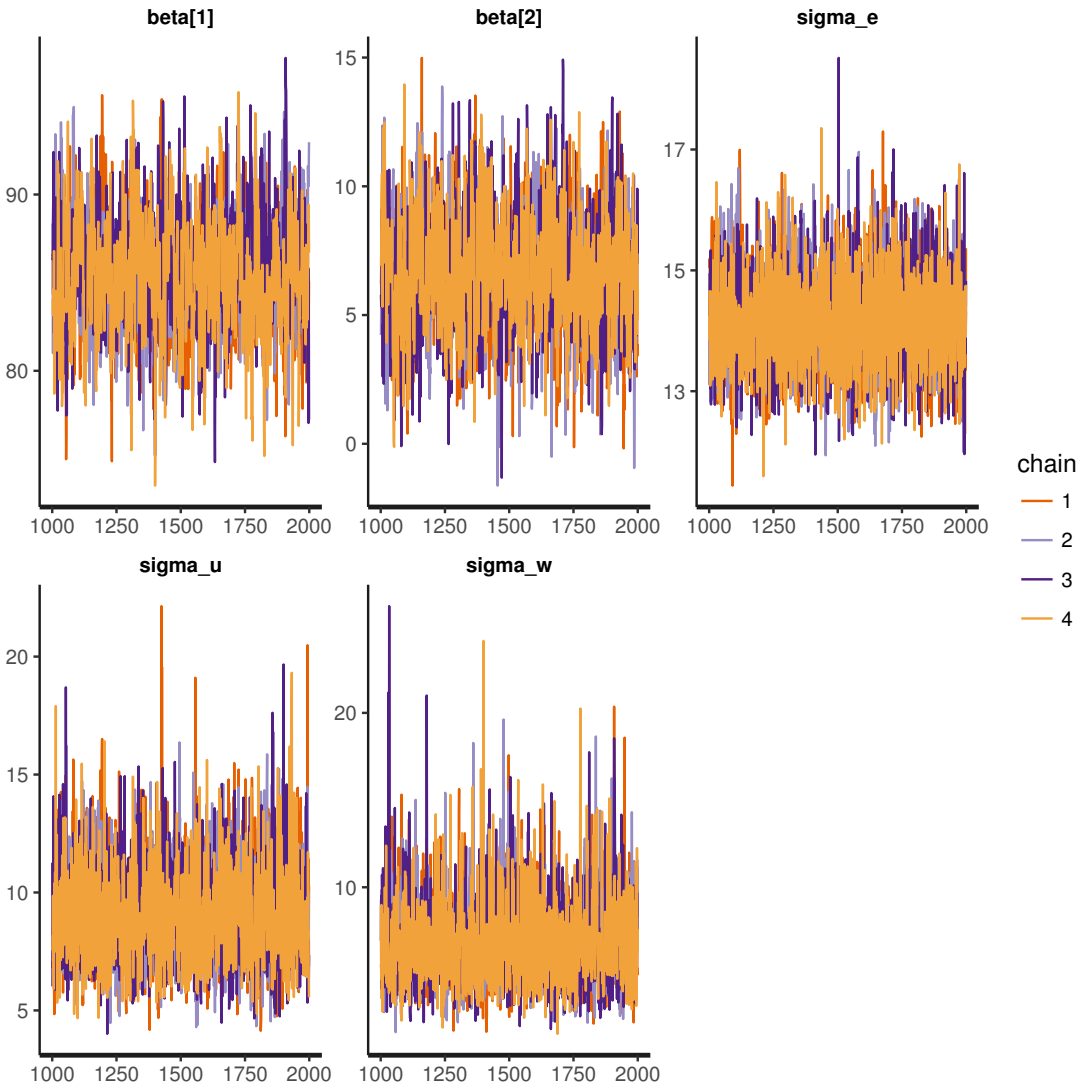

Figure 3: Trace plots for the effect of gender on VOT in Mandarin.

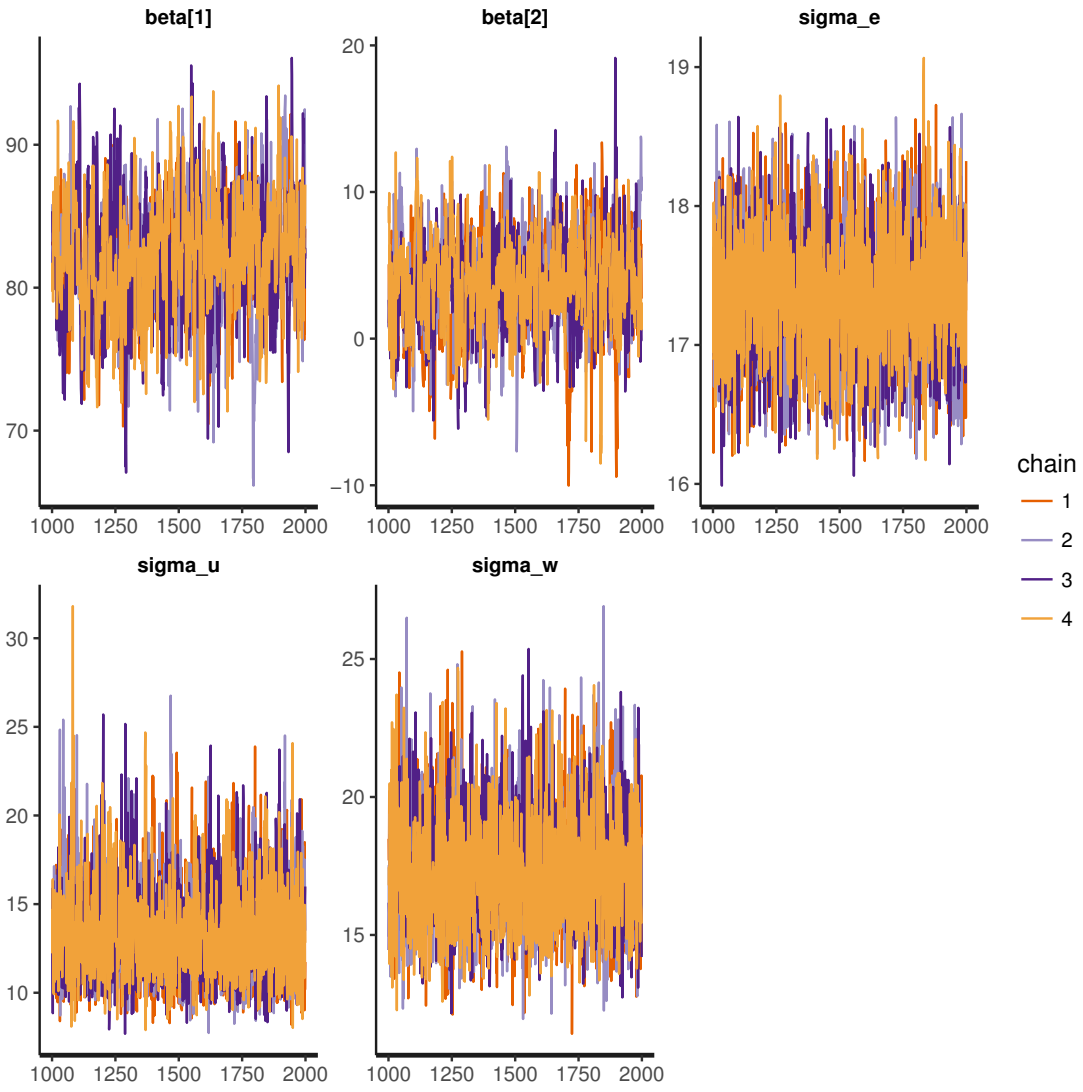

Figure 4: Trace plots for the effect of gender on VOT in English.

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 5 shows, for the two languages, the relationship between the mean VOTs and mean vowel duration, along with the standard deviations 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 5, it seems clear that there is a linear relationship between mean vowel duration and mean VOT in both languages. Indeed, a linear model fit to the data does seem to show clear evidence that this is the case; see Tables 2 and 3.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.9618	25.6796	0.31	0.7601
meanvdur	1.2967	0.1542	8.41	0.0000

Table 2: Linear model showing the effect of mean vowel duration on mean VOT in Mandarin.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	35.8649	40.1668	0.89	0.3837
meanvdur	1.1048	0.2241	4.93	0.0001

Table 3: Linear model showing the effect of mean vowel duration on mean VOT in English.

In both languages, a unit increase in mean vowel duration is predicted to increase mean VOT by approximately 1.3, SE 0.15 in Mandarin, and by 1.1, SE 0.22 in English (the units are ms). 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.[3]

In our model, we will take this measurement error of the vowel duration estimate into account. Thus, if the vowel duration estimated for one participant i is vdur_i, we can also record the standard error of that estimate, *and take that uncertainty into*

[3]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 participant or by item and then analysed using repeated measures ANOVA.

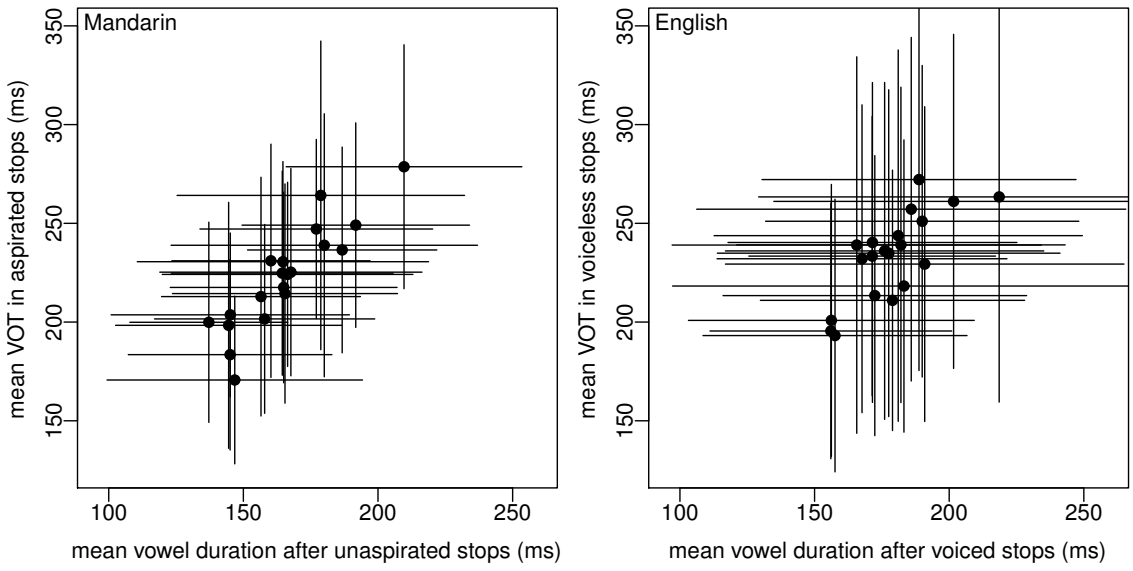

Figure 5: Mean VOT values are shown against mean vowel duration in the two languages; the error bars represent standard deviations of each measure. There is a clear positive relationship between the means, but that does not take the uncertainty of these estimates into account.

account in our hierarchical model. Such a hierarchical measurement error model is straightforward to implement in Stan.

We assume that each participant i has an estimated vowel duration vdur_i, which has standard error SE_i, and that each participant has a true but unknown underlying vowel duration true_vdur_i. Thus, the estimated vowel durations are assumed to be generated from the following distribution: $\text{vdur}_i \sim Normal(\text{true\_vdur}_i, SE_i)$. The variable true_vdur_i is a latent variable much like the varying intercepts for participants and for items. In the hierarchical model specification, instead of using the estimated vowel duration as a predictor, we use the latent variable true_vdur_i. What such a model achieves is that it takes into account the uncertainty of the estimate for the vowel duration when estimating parameters.

$$
\begin{aligned}
&\text{Priors:} \\
&\beta_1 \sim Normal(0, 200) \\
&\beta_2 \sim Normal(0, 50) \\
&\sigma_e \sim Normal_+(0, 100) \\
&\sigma_{u_0} \sim Normal_+(0, 100) \\
&\sigma_{w_0} \sim Normal_+(0, 100) \\
&\text{true\_vdur}_i \sim Normal(0, 200) \\
&\text{Measurement error:} \\
&\text{vdur}_i \sim Normal(\text{true\_vdur}_i, SE_i) \\
&\text{Probability model for data:} \\
&VOT_{ij} \sim Normal(\beta_1 + u_i + w_j + \beta_2 \times \text{true\_vdur}_i, \sigma_e)
\end{aligned}
\tag{6}
$$

The Stan code corresponding to this model is shown in Listing 2. The posterior distributions of the parameters for Mandarin and English are summarized in Table 4, and the trace plots are shown in Figures 6 and 7. The trace plots show no indication of convergence problems.

5.1. Interpreting the results

The models for both languages suggest that there is no evidence that vowel duration affects VOT: in both languages, the posterior distribution for the parameter β_2 is centered near 0 with 95% credible intervals that span a range of approximately 5 ms in either direction. This could mean that vowel duration is not a good measure of speech rate (which might yet have an effect on VOT; cf. Kessinger and Blumstein, 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 to these 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. An example of how evidence can be synthesized is discussed in Jäger et al. (2017).

Note that if we had fit a hierarchical model without the measurement error, the 95% credible intervals for the effects involving vowel duration would be tighter—the measurement error takes into account the uncertainty of the measurement error on vowel durations.

Our purpose here in introducing the measurement error model is to illustrate the flexibility of Bayesian modeling. Once we have decided on the underlying variance components, expressing the model in Stan is relatively easy. A further extension of the model could take the uncertainty in the dependent variable, VOT values, into account as well.

	Mandarin					English				
	mean	2.5%	97.5%	n_eff	Rhat	mean	2.5%	97.5%	n_eff	Rhat
beta[1]	85	78	93	898	1.00	82	74	90	190	1.02
beta[2]	0	-6	5	558	1.00	-1	-7	5	319	1.02
sigma_e	14	13	16	4000	1.00	17	17	18	4000	1.00
sigma_u	12	8	17	3299	1.00	14	9	20	4000	1.00
sigma_w	7	3	12	2385	1.00	17	14	22	4000	1.00

Table 4: Posterior means and 95% credible intervals of the parameters for the measurement error model evaluating the vowel duration in Mandarin and English. The column Rhat refers to a convergence diagnostic of the Markov Chain Monte Carlo samples with a value near 1 indicating convergence; the column n_eff is a measure of effective sample size and should be higher than approximately 10% of the number of samples after warm-up or burn-in (here 10% of 4000).

6. Question 3: 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 (using measurement error) and language.

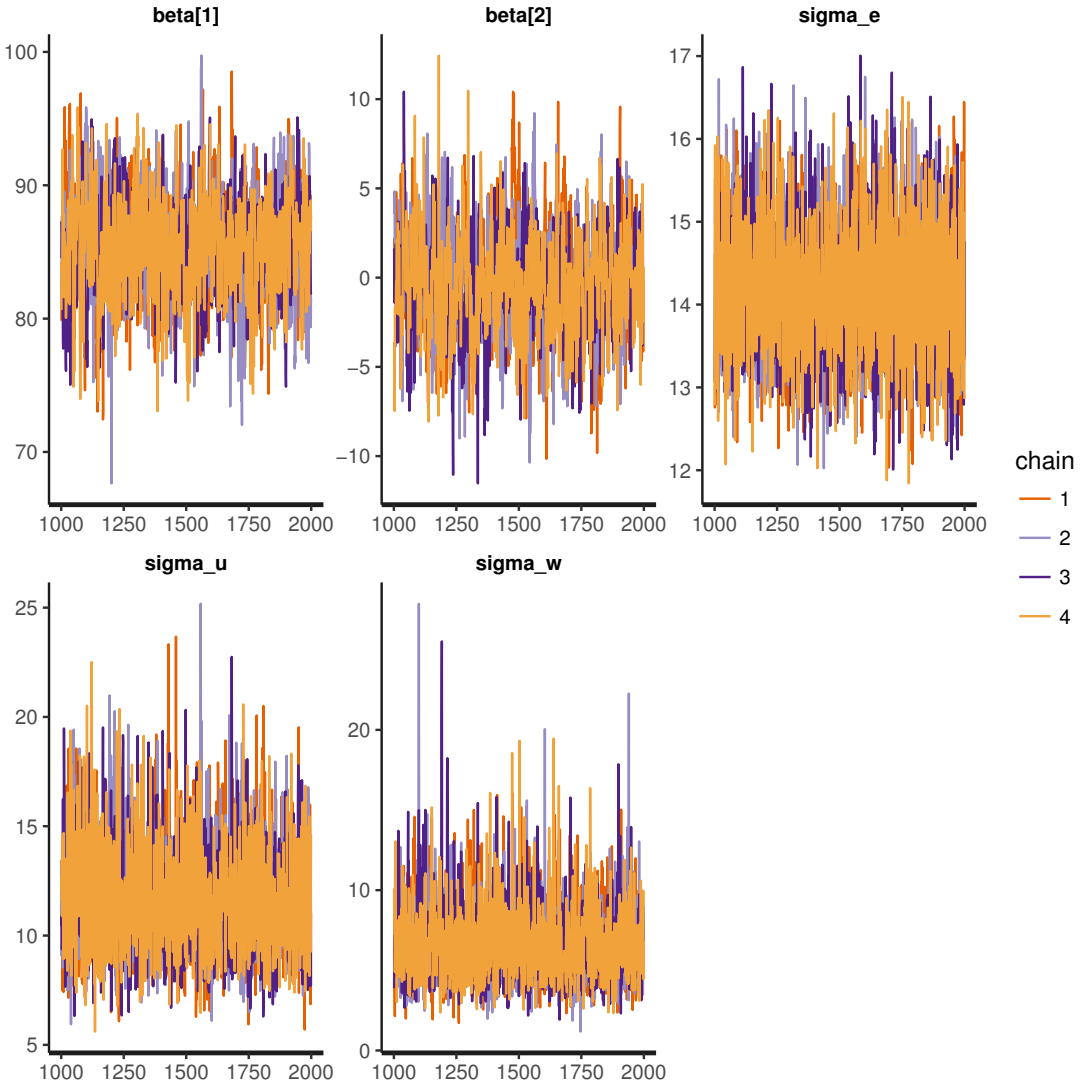

Figure 6: Trace plots for the Mandarin measurement error model addressing question 2: the effect vowel duration (beta[2]) on VOT, taking the uncertainty in the estimation of the vowel duration into account.

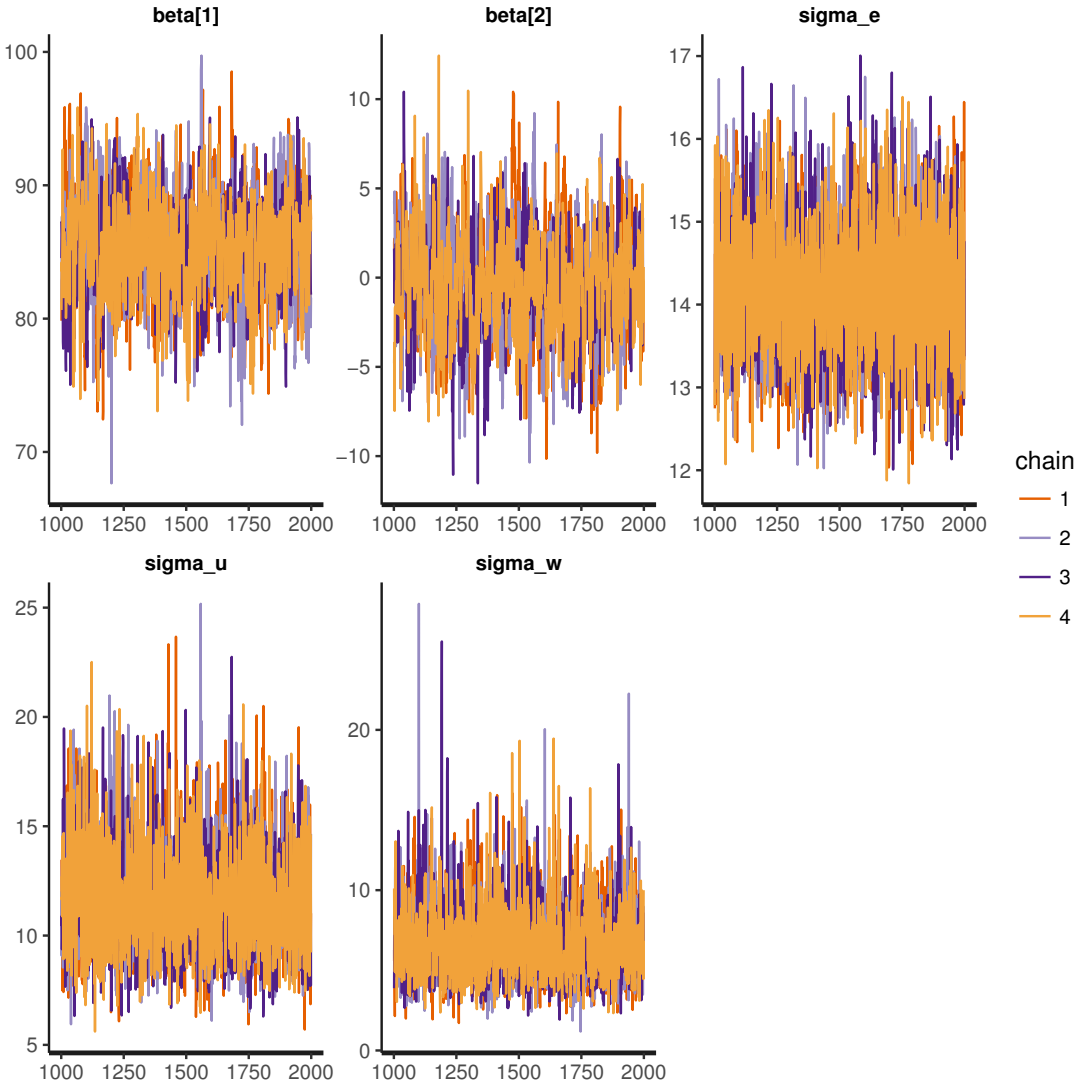

Figure 7: Trace plots for the English measurement error model addressing question 2: the effect vowel duration (beta[2]) on VOT, taking the uncertainty in the estimation of the vowel duration into account.

Model (a) is shown below, along with the priors assumed:

$$
\begin{aligned}
\beta_1 &\sim Normal(0, 200) \\
\beta_2, \beta_3, \beta_4 &\sim Normal(0, 50) \\
\sigma_e &\sim Normal_+(0, 100) \\
\sigma_{u_0} &\sim Normal_+(0, 100) \\
\sigma_{w_0} &\sim Normal_+(0, 100) \\
VOT_{ij} &\sim Normal(\beta_1 + u_i + w_j + \beta_2 \times \text{gender}_i + \\
&\qquad \beta_3 \times \text{language}_i + \beta_4 \times \text{gender}_i \times \text{language}_i, \sigma_e)
\end{aligned}
\tag{7}
$$

Model (b) has the same priors as (a) and in the model for question 2, and has a measurement error component:

Measurement error:

$$\text{vdur}_i \sim Normal(\text{true\_vdur}_i, SE_i)$$

Probability model for data:

$$
\begin{aligned}
VOT_{ij} \sim Normal(&\beta_1 + u_i + w_j + \beta_2 \times \text{true\_vdur}_i + \\
&\beta_3 \times \text{language}_i + \beta_4 \times \text{language}_i \times \text{true\_vdur}_i, \sigma_e)
\end{aligned}
\tag{8}
$$

The corresponding Stan models for (a) and (b) are shown in Listings 3 and 4.

The posterior distributions are summarized in Tables 5 and 6, and the trace plots are shown in Figures 8 and 9. The trace plots show that convergence was achieved: the chains are mixing well. In Table 6, the n_eff for the beta[2] parameter (the effect of vowel duration) and the beta[4] parameter (the interaction of language and vowel duration) have somewhat low values—we had said earlier that 400 would be a reasonable lower bound. The n_eff could be improved through reparameterization, as discussed in the Stan documentation. However, even though the value is low, the chains are mixing well, and so it would be acceptable to leave the model unmodified.

6.1. Interpreting the results

From Table 6, it seems clear that the gender effect holds once we pool the data from the two languages: females have a higher VOT than males. Females have an estimated VOT 5 ms above the grand mean estimate (the estimate beta[1]) of 84 ms, and males have an estimated VOT 5 ms below the grand mean. The 95% credible

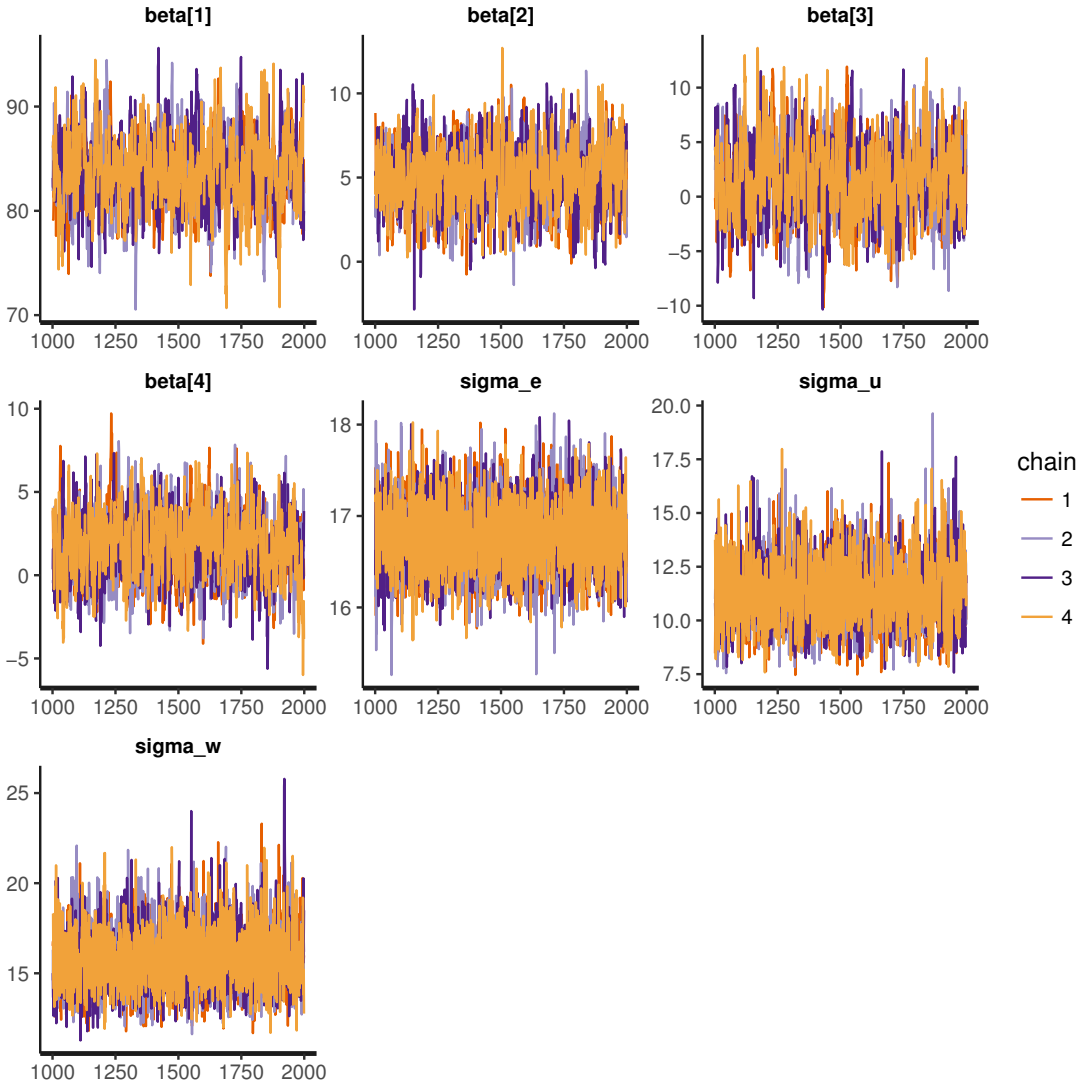

Figure 8: Trace plots for the model evaluating the effect of gender (beta[2]), language (beta[3]), and their interaction (beta[4]).

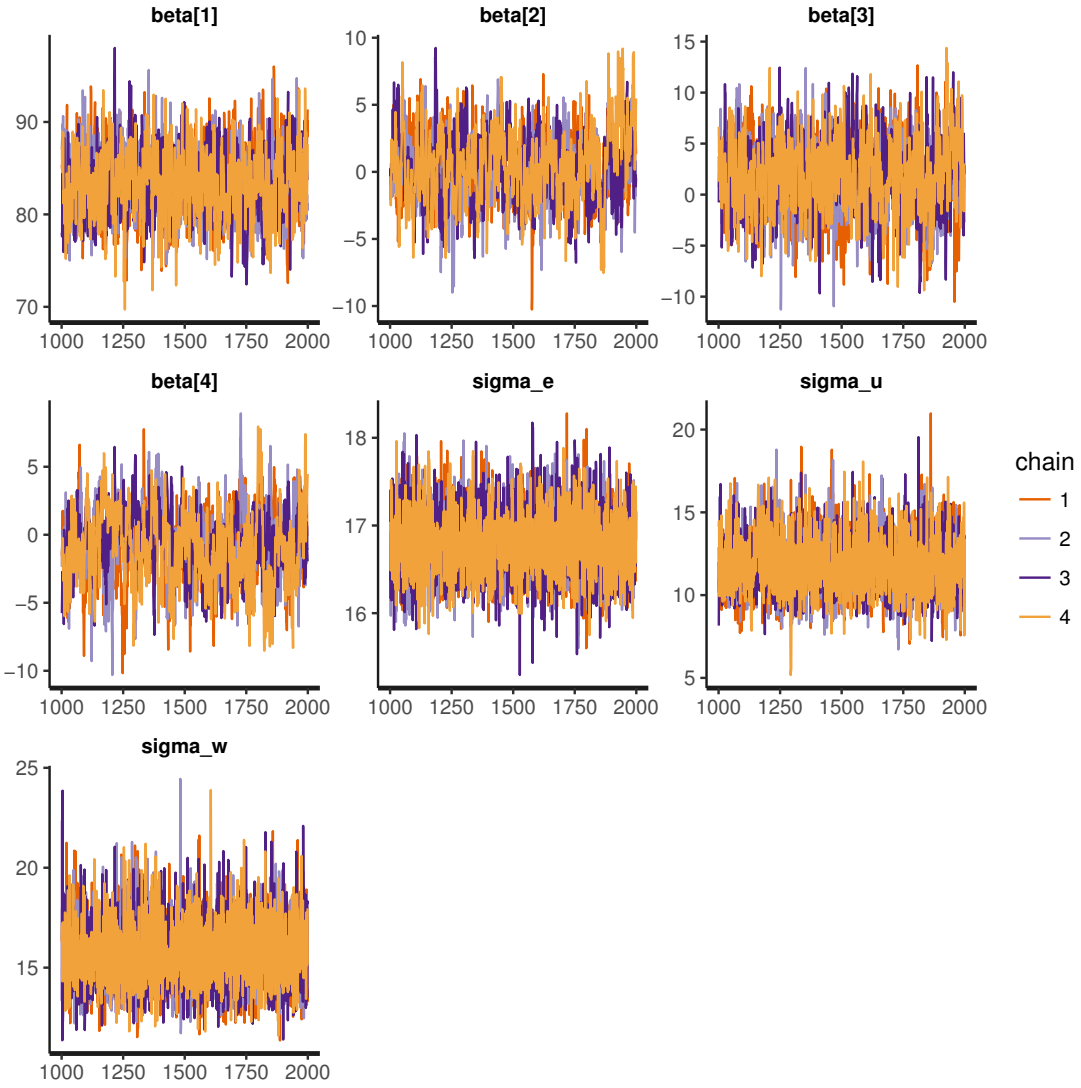

Figure 9: Trace plots for the measurement error model evaluating the effect of vowel duration (beta[2]), language (beta[3]), and their interaction (beta[4]).

	mean	2.5%	97.5%	n_eff	Rhat
beta[1]	84	77	91	539	1.00
beta[2]	5	1	9	769	1.00
beta[3]	2	-5	8	596	1.01
beta[4]	2	-2	6	650	1.00
sigma_e	17	16	18	4000	1.00
sigma_u	11	9	15	1981	1.00
sigma_w	16	13	20	2669	1.00

Table 5: Posterior means and 95% credible intervals of the parameters for the model evaluating the effect of gender (beta[2]), language (beta[3]), and their interaction (beta[4]). The column Rhat refers to a convergence diagnostic of the Markov Chain Monte Carlo samples with a value near 1 indicating convergence; the column n_eff is a measure of effective sample size and should be higher than 10% of the number of samples after warm-up or burn-in (here 10% of 4000).

interval of $1-9$ ms tells us what the plausible range of values is given these particular data. Table 6 shows that there isn't much evidence for a difference due to language or a gender $\times$ language interaction. A larger data-set may well give us tighter estimates, so one cannot draw any strong conclusions about language differences and about a potential interaction.

The posterior distributions in Table 6 for the parameters of interest are all centered around zero, leaving us with no reason to believe (given these data) that vowel duration or language have any effect on VOT values.

7. Concluding remarks

We attempted to provide a practical entry point into Bayesian modeling using hierarchical models and the probabilistic programming language Stan. We have therefore only covered some of the critical steps that should be taken in carrying out the analysis. Other important issues relating specifically to hierarchical models are discussed in Nicenboim and Vasishth (2016) and Sorensen et al. (2016). For accessible book-length treatments intended for the empirical sciences, we suggest Kruschke (2014) and McElreath (2016).

	mean	2.5%	97.5%	n_eff	Rhat
beta[1]	84	77	90	662	1.01
beta[2]	0	-5	5	371	1.01
beta[3]	2	-5	9	648	1.00
beta[4]	-1	-6	4	334	1.02
sigma_e	17	16	18	4000	1.00
sigma_u	12	9	15	4000	1.00
sigma_w	16	13	19	4000	1.00

Table 6: Posterior means and 95% credible intervals of the parameters for the measurement error model evaluating the effect of vowel duration (beta[2]), language (beta[2]), and their interaction (beta[3]). The column Rhat refers to a convergence diagnostic of the Markov Chain Monte Carlo samples with a value near 1 indicating convergence; the column n_eff is a measure of effective sample size and should be higher than 10% of the number of samples after warm-up or burn-in (here 10% of 4000).

AppendixA. Stan model code

The R package RStan must be installed before Stan code can be run. For installation instructions, see the Quick Start guide on the Stan website mc-stan.org. All the Stan code presented below is available under https://github.com/vasishth/jopbayes.

```
data {
  int<lower = 1> N;                  //number of data points
  vector[N] y;                      //dependent variable vot
  vector<lower = -1, upper = 1>[N] gend;    //predictor
  int<lower = 1> J;                  //number of subjects
  int<lower = 1> K;                  //number of items
  int<lower = 1, upper = J> subj[N];   //subject id
  int<lower = 1, upper = K> item[N];   //item id
}

parameters {
  vector[2] beta;            //fixed intercept and slope
  vector[J] u;               //subject intercepts
  vector[K] w;               //item intercepts
  real<lower = 0> sigma_e;     //error sd
  real<lower = 0> sigma_u;     //subj sd
  real<lower = 0> sigma_w;     //item sd
}

model {
  vector[N] mu;
  //priors
  beta[1] ~ normal(0, 200);
  beta[2] ~ normal(0, 50);
  sigma_e ~ normal(0, 100);
  sigma_u ~ normal(0, 100);
  sigma_w ~ normal(0, 100);
  u ~ normal(0,sigma_u);    //subj random effects
  w ~ normal(0,sigma_w);    //item random effects
  // likelihood
  mu = beta[1] + u[subj] + w[item] + beta[2]*gend;
  y ~ normal(mu,sigma_e);

}
```

Listing 1: Stan code for addressing question 1.

```
data {
  int<lower = 1> N;                  //number of data points
  vector[N] y;                       //dep variable vot
  int<lower = 1> J;                  //number of subjects
  int<lower = 1> K;                  //number of items
  vector[J] meanvdur;                // noisy centered mean vdur
  vector<lower = 0>[J] se;             // se mean vdur
  int<lower = 1, upper = J> subj[N];   //subject id
  int<lower = 1, upper = K> item[N];   //item id
}

parameters {
  vector[2] beta;              //fixed intercept and slopes
  vector[J] true_mvdur; // true unknown value mvdur
  vector[J] u;                 //subject intercepts
  vector[K] w;                 //item intercepts
  real<lower=0> sigma_e;       //error sd
  real<lower=0> sigma_u;       //subj sd
  real<lower=0> sigma_w;       //item sd
}

model {
  vector[N] mu;
  //priors
  true_mvdur ~ normal(0, 200);
  meanvdur ~ normal(true_mvdur, se); // measurement model
  beta[1] ~ normal(0, 200);
  beta[2] ~ normal(0, 50);
  sigma_e ~ normal(0, 100);
  sigma_u ~ normal(0, 100);
  sigma_w ~ normal(0, 100);
  u ~ normal(0, sigma_u);    //subj random effects
  w ~ normal(0, sigma_w);    //item random effects
  // likelihood
  mu = beta[1] + u[subj] + w[item] + beta[2] * true_mvdur[subj];
  y ~ normal(mu, sigma_e);
}
```

Listing 2: Stan code for addressing question 2.

```
data {
  int<lower = 1> N;                  //number of data points
  vector[N] y;                      //dependent variable vot
  vector<lower = -1, upper = 1>[N] gend;    //predictor
  vector<lower = -1, upper = 1>[N] lang;    //predictor
  int<lower = 1> J;                  //number of subjects
  int<lower = 1> K;                  //number of items
  int<lower = 1, upper = J> subj[N];   //subject id
  int<lower = 1, upper = K> item[N];   //item id
}

parameters {
  vector[4] beta;            //fixed intercept and slope
  vector[J] u;               //subject intercepts
  vector[K] w;               //item intercepts
  real<lower = 0> sigma_e;     //error sd
  real<lower = 0> sigma_u;     //subj sd
  real<lower = 0> sigma_w;     //item sd
}

model {
  vector[N] mu;
  //priors
  beta[1] ~ normal(0, 200);
  beta[2:4] ~ normal(0, 50);
  sigma_e ~ normal(0, 100);
  sigma_u ~ normal(0, 100);
  sigma_w ~ normal(0, 100);
  u ~ normal(0, sigma_u);    //subj random effects
  w ~ normal(0, sigma_w);    //item random effects
  // likelihood
  mu = beta[1] + u[subj] + w[item] + beta[2]*gend
     + beta[3] * lang + beta[4] * gend .* lang;
  y ~ normal(mu,sigma_e);

}
```

Listing 3: Stan code for addressing question 3(a), the effect of gender, language and their interaction.

```
data {
  int<lower = 1> N;                  //number of data points
  vector[N] y;                       //dep variable vot
  int<lower = 1> J;                  //number of subjects
  int<lower = 1> K;                  //number of items
  vector<lower = -1, upper = 1>[N] lang;    //predictor
  vector[J] meanvdur;                // noisy centered mean vdur
  vector<lower = 0>[J] se;             // se mean vdur
  int<lower = 1, upper = J> subj[N];   //subject id
  int<lower = 1, upper = K> item[N];   //item id
}

parameters {
  vector[4] beta;              //fixed intercept and slopes
  vector[J] true_mvdur; // true unknown value mvdur
  vector[J] u;                 //subject intercepts
  vector[K] w;                 //item intercepts
  real<lower=0> sigma_e;       //error sd
  real<lower=0> sigma_u;       //subj sd
  real<lower=0> sigma_w;       //item sd
}

model {
  vector[N] mu;
  //priors
  true_mvdur ~ normal(0, 200);
  meanvdur ~ normal(true_mvdur, se); // measurement model
  beta[1] ~ normal(0, 200);
  beta[2:4] ~ normal(0, 50);
  sigma_e ~ normal(0, 100);
  sigma_u ~ normal(0, 100);
  sigma_w ~ normal(0, 100);
  u ~ normal(0, sigma_u);    //subj random effects
  w ~ normal(0, sigma_w);    //item random effects
  // likelihood
  mu = beta[1] + u[subj] + w[item] + beta[2] * true_mvdur[subj]
     + beta[3] * lang + beta[4] * true_mvdur[subj] .* lang;
  y ~ normal(mu, sigma_e);
}
```


Listing 4: Stan code for addressing question 3(b), the effect of vowel duration, language and their interaction.

Acknowledgments

The recordings and measurements described in this paper were made with the support of NIH grant DC02932 to Jan Edwards, of an Ohio State University Department of Linguistics Targeted Investment Award to Fangfang Li and Eun Jong Kong, and of 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.

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.

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

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

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

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

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

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.

Hoenig, J. M., Heisey, D. M., 2001. The abuse of power: The pervasive fallacy of power calculations for data analysis. The American Statistician 55 (1), 19–24.

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., Feb 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

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.

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.

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.

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

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., 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., 2017. Models of retrieval in sentence comprehension: A computational evaluation using Bayesian hierarchical modeling. Journal of Memory and LanguageAccepted.
URL https://arxiv.org/abs/1612.04174

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

Paape, D., Nicenboim, B., Vasishth, S., 2017. Does antecedent complexity affect ellipsis processing? An empirical investigation. Glossa 2.
URL https://www.glossa-journal.org/articles/abstract/10.5334/gjgl.290/

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.

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.

Plummer, M., 2016. JAGS version 4.2.0 user manual.

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.

Sorensen, T., Hohenstein, S., Vasishth, S., 2016. Bayesian linear mixed models using Stan: A tutorial for psychologists, linguists, and cognitive scientists. Quantitative

Methods for Psychology 12 (3), 175–200.
URL http://www.ling.uni-potsdam.de/~vasishth/statistics/BayesLMMs.html

Stan Development Team, 2017. 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., Chen, Z., Li, Q., Guo, G., 10 2013. Processing Chinese relative clauses: Evidence for the subject-relative advantage. PLoS ONE 8 (10), 1–14.

Vasishth, S., Gelman, A., 2017. The statistical significance filter leads to overconfident expectations of replicability. In: Proceedings of Cognitive Science Conference. London, UK.
URL http://arxiv.org/abs/1702.00556

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/

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