Assignment 4: The role of priors and cumulative science

Deadline for hand-in: 30/4-2020

Github: https://github.com/saraoe/assignment-4-priors-and-cumulative-science

Introduction

This assignment will examine the role of priors, when making a model and what impact this may have on cumulative science. Firstly, we will make a meta-analysis of 41 studies. Then, we will create 3 different Bayesian models to analyze two new studies - two of these models using a skeptical prior and the last using informed priors based on the meta-analysis. One the basis of these models, we will discuss the role of priors and what this would imply for cumulative science.

Data Collection

Meta-analysis

This meta-analysis is based on previous studies investigating pitch variability in typically developing (TD) children and children with autism spectrum disorder (ASD). The studies assessed vary in publication year, geographic placement etc. (see table 1 for details in demographics).

Studies (n)	Population (n)	Year of publication(range)	Total ASD (n)	Total TD (n)	Age of ASD in months (mean, sd)	Age of TD in months (mean, sd)	Languages assessed (n)	Language of interest (dk)	Language of interest (en)
41	36	1982-2018	847	807	171.43 (42.70)	166.69 (30.66)	9	5	26

Table 1: demographics of meta-analysis."Population" refers to the population of participants, and does not overlap completely with studies, given that different studies sometimes used the same participants.

New studies

We include two new studies. One study with Danish speaking participants and one study with American speaking participants. Each ASD participant had an associated TD participant (table 2). An elaborate illustration of the data distribution can be seen in figure 1.

Number of participants by diagnosis (TD)	Gender males (TD)	Mean age in months for ASD (sd)	Mean age in months for TD (sd)	Number of Danish patients (TD)	Number of American patients (TD)
513 (561)	459 (442)	138.1 (25.35)	139.76 (27.57)	335 (427)	178 (134)

Table 2: demographics of the 2 new studies

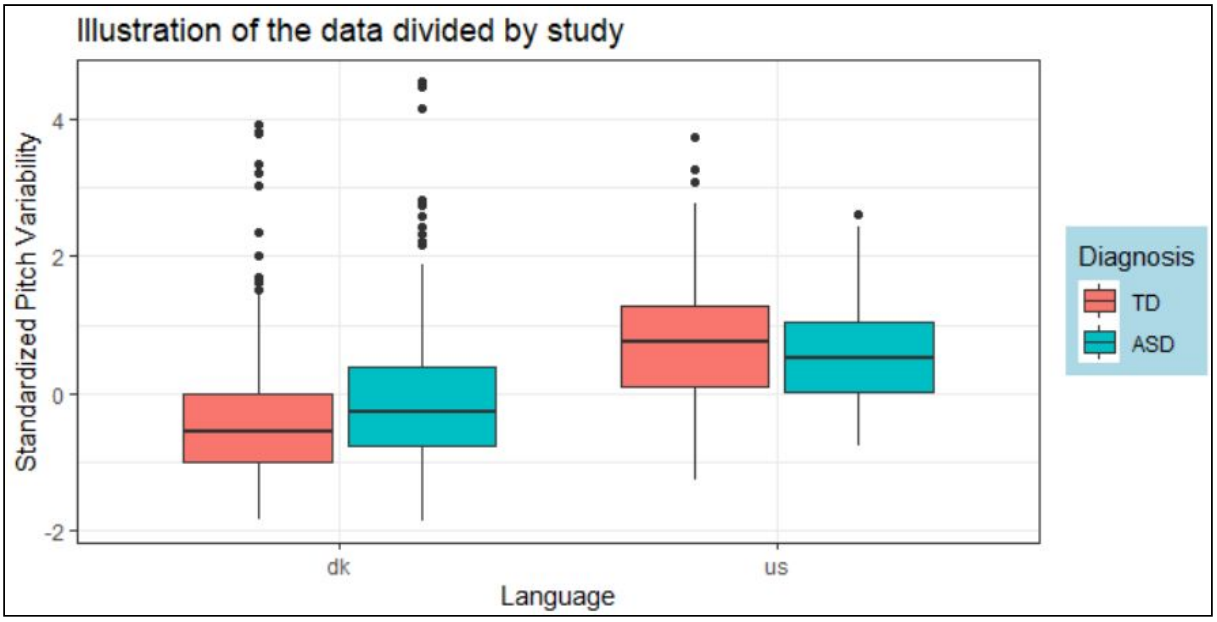


Figure 1: Illustration of data from the two new studies.

Analysis

Throughout the entire analysis of the data, we have used R (R Core Team, 2020). In order to extract key measures from the meta analysis, we used the package Metafor (Viechtbauer 2010). We operated in the package BRMS (Bürkner, 2017; Bürkner, 2018) for our Bayesian analysis.

Meta-analysis

Before commencing with our analysis, we started off by standardizing the pitch variability to have the different studies on the same scale. This enabled us to get the effect sizes in a measure of Cohen's D, as well as getting the Standard Error (uncertainty in Cohen's d).
We ran the following model using a Bayesian framework:

EffectSize | se(StandardError) ~ 1 + (1 | Population)

We defined sceptical priors for the model that assumed no effect. The prior for the effect size was defined as normally distributed, with a mean of 0 and a sd of 0.1, and for the difference related to Study as normally distributed, with a mean of 0 and a sd of 0.3.
The summary statistics revealed a mean pitch variability from the meta analysis on 0.43, with a SE of 0.09. Furthermore, based on the statistics we expect the heterogeneity to be 0.32 referring to the expected error the model will do when we look at different (new) studies. We here refrained from taking publication bias, or other biases into account.

The Bayesian analysis of 2 new studies

In the second part of the analysis, we analysed pitch variability (in terms of standardized interquartile range) in groups of ASD patients as well as in associated control groups.
First, we assessed the distribution of standardized pitch variability in order to select an appropriate distribution for the models. Off hand, the data seems log-normally distributed. The values which we

have extracted from the meta-analysis are on a 'Cohen's d' scale. A 'Cohen's d' scale has both negative and positive values distributed around a mean of 0. Though, a log-normal distribution does not account for negative values as opposed to a Gaussian distribution. In order to make the two new studies compatible with the meta-analytic values, we assumed a Gaussian distribution and standardized pitch variability (See figure 2).

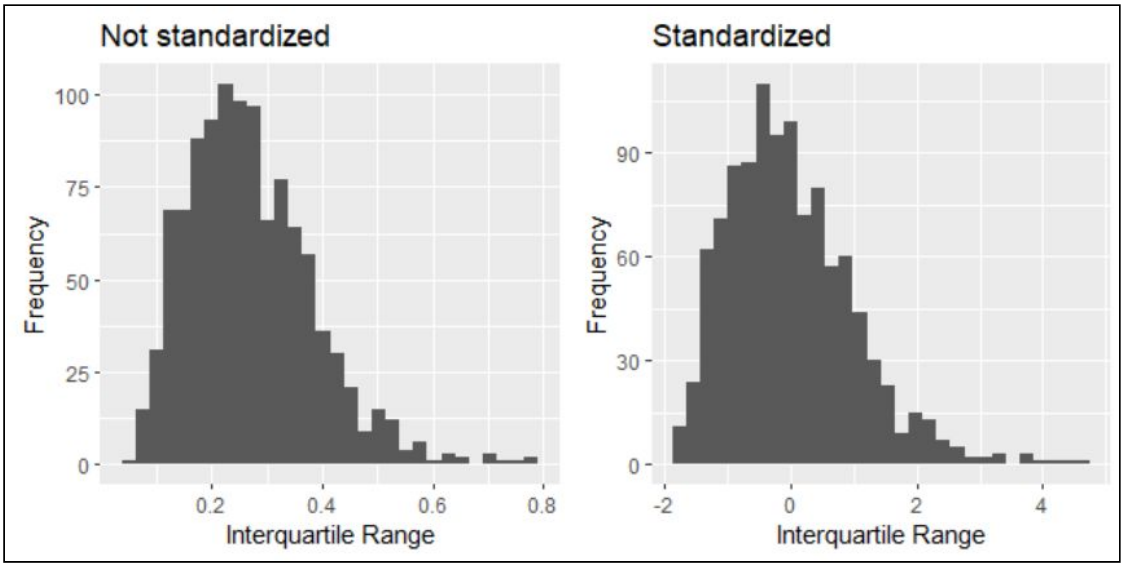


Figure 2: Left: Histogram illustrating the frequencies of different IQR values. Right: Histogram of the same values, but standardized.

Model 1

We first built a regression model to predict Pitch variability from Diagnosis alone, not taking language into account. We used the following formula:

PitchVariability ~ 1 + Diagnosis + (1|ID)

For this, we assumed a Gaussian distribution of the outcome variable. We defined sceptical un-informed priors for the model that assumed no effect (diagnosis TD and ASD have the same IQR pitch). The prior for pitch variability in TDs (intercept) was defined as normally distributed, with a mean of 0 and a sd of 0.3, the difference between pitch variability of TD's and ASD's (the slope) defined as normally distributed, with a mean of 0 and a sd of 0.1. We also assumed that pitch variability would vary slightly by participant (varying/random intercept) so it was set with the same values. For the overall variance of the model (sigma) the prior was defined as normally distributed, with a mean of 0.5 and a sd of 0.3. The reason why we chose such priors was primarily based on the fact that we did not have any prior knowledge. Furthermore, we wanted the evidence to *overcome* the assumption of no difference between ASD and TD instead of assuming a difference already before assessing the data.
We did sanity checks on our models, using prior and posterior predictive checks - all of which can be seen in the Appendix 1.

Hypothesis testing
We assessed the following hypothesis for model 1:

M1H1: ASD's have a higher Pitch Variability than TD

Model 2

In the second Bayesian regression model, we told the model that we have data from two different languages, Danish and English:

PitchVariability ~ 0 + Language + Language:Diagnosis + (1|ID)

We hypothesised that the pitch variability would differ between the studies, as the phonetic structures of the two languages are different to begin with. Therefore, we defined 'Language' as a main effect. We have incorporated an interaction effect between 'Language' and 'Diagnosis', as the difference between the groups (ASD and TD) might differ across languages.
In this model, we also used sceptical un-informed priors. For the intercept of the languages (DK and US) the priors were both defined as normally distributed, with a mean of 0 and sd of 0.3. For the two slopes (between TD and ASD for each language) priors were defined equivalently to the sceptical priors of model 1 (with a mean of 0 and sd of 0.1). The prior for the varying effect of participants were also defined equivalent to model 1 (with a mean of 0 and sd of 0.1) and for the overall variance of the model the prior was defined with a mean of 0.5 and sd of 0.3. We chose these priors for the same reason as in model 1.
We did sanity checks on our models, using prior and posterior predictive checks - all of which can be seen in the Appendix 1.

Model comparison and hypothesis testing
We addressed the qualities of the first and second model. The model comparison was operationalized through weights and the IC criterion LOO. Afterwards, we assessed the following hypothesis for model 2:

M2H1: ASD's have a higher Pitch Variability than TD, for Danish speaking

M2H2: ASD's have a higher Pitch Variability than TD, for American speaking

M2H3: There is a larger difference in Pitch variability between ASD's and TD's in Danish speaking compared to American speaking

Model 3

We re-ran the best of the two models above (the one including language as a fixed effect) with informed priors. The formula was identical to the previous:

PitchVariability ~ 0 + Language + Language:Diagnosis + (1|ID)

For this model, the priors were made based on the results of the meta-analysis. The priors for both intercepts of the model (one for each language) were defined as normally distributed, with a mean of -0.215 and sd of 0.3, based on the mean of the effect size of the meta-analysis divided by two. The priors for the two slopes were defined as normally distributed, with means of 0.43 and sd of 0.09, based on the effect and standard deviation of the meta-analytical model. The prior for the varying

effect of participants were defined equivalent to model 1 and 2 (normally distributed, with a mean of 0 and sd of 0.1) and the prior for sigma was defined based on the heterogeneity of the meta-analysis, which was found to be 0.32 with an estimated error of 0.1, also normally distributed.
We did sanity checks on our models, using prior and posterior predictive checks - all of which can be seen in the Appendix 1.

Model comparison and hypothesis testing
Lastly, we compared the two models which included language. As one had an uninformed prior and the other had an informed prior (from the meta analysis), it enabled us in examining the role of an informed prior.
We compared the posterior distributions of the two models, together with a model comparison using the information criterion; LOO.

We assessed the following hypothesis for model 3:

M3H1: ASD's have a higher Pitch Variability than TD for danish speaking

M3H2: ASD's have a higher Pitch Variability than TD, for American speaking

M3H3: There is a larger difference in Pitch Variability between ASD's and TD's in Danish speaking compared to American speaking

All three models were run on two Monte Carlo Markov Chains. Plots of the chains can be seen in the Appendix 2.

Results

Model 1:

Model 1 showed a small positive effect on pitch variability (b = 0.08, CIs = -0.06 - 0.23), although quite uncertain when going from TD to ASD.
M1H1:
The evidence ratio for the first hypothesis was 6.41, meaning (according to model 1) there is 6.41 times more evidence for a positive effect on pitch variability.

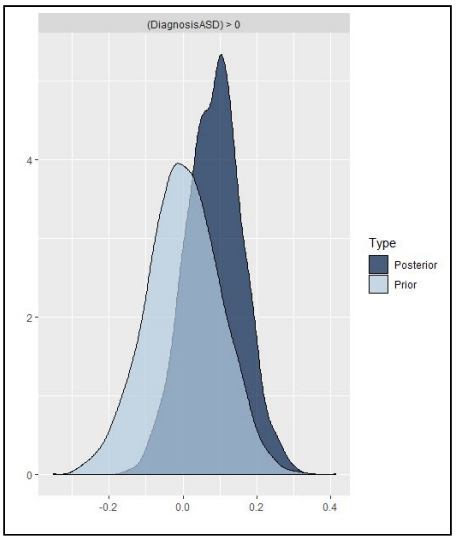


Figure 3: The posterior predictions for betas, with TD as intercept

Model 2

Model 2, likewise, found a small positive effect on pitch variability (b = 0.09, CIs = -0.08 - 0.25) in the study with Danish-speaking participants. However, the opposite effect was found in the study with american-speaking participants (b = -0.02, CIs = -0.18 - 0.15). The effect of language on pitch variability can also be seen in figure 4.

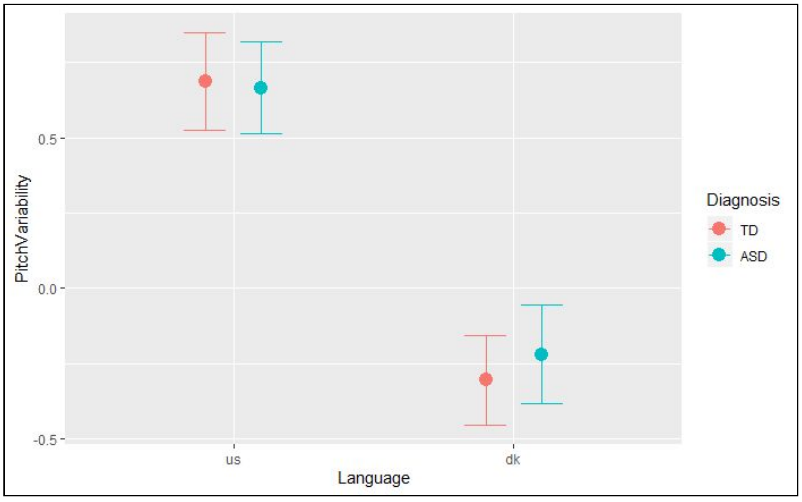


Figure 4: Effects of Model 2

When testing the three hypotheses related to the second model, we found the following.

M2H1: The evidence ratio for the first hypothesis was 5.64, meaning (according to model 2) there is 5.64 times more evidence for a positive effect on pitch variability for Danish-speakers, than for a negative or no effect.

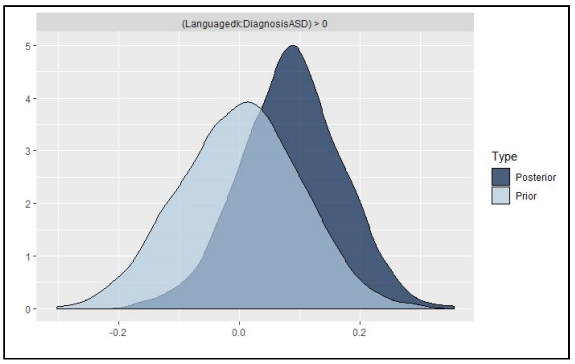


Figure 5: The posterior predictions for betas, with TD as intercept, and only for danish speaking.

M2H2: The evidence ratio for the second hypothesis was 0.65, meaning (according to model 2) there is 0.65 times more evidence for a positive effect on pitch variability for US-speakers, than for a negative or no effect. Thereby, the model indicates that no effect or a negative effect is more likely.

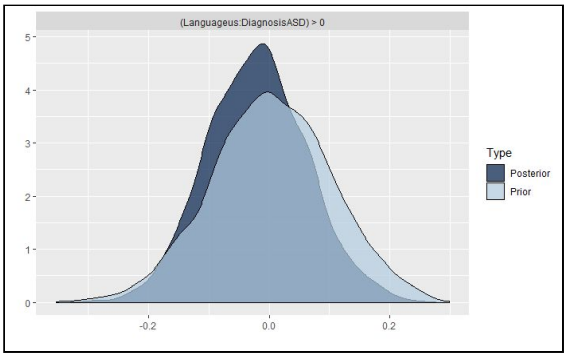


Figure 6: The posterior predictions for betas, with TD as intercept, and only for English American speaking

M2H3: The evidence ratio for the third hypothesis was 4.21, meaning (according to model 2) there is 4.21 more evidence for a larger effect on pitch variability for Danish-speakers compared to american-speakers.

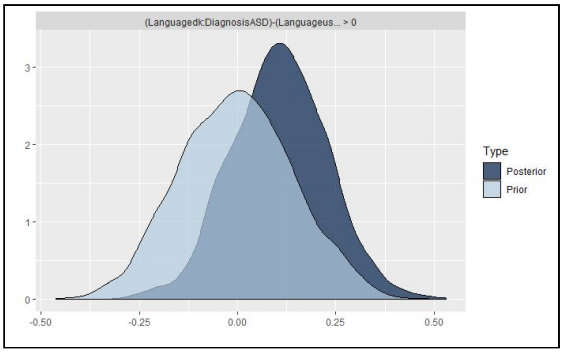


Figure 7: The posterior predictions for betas, with TD as intercept, when we subtract the beta in Danish speaking with the beta in English American speaking

Model 1 and 2 - Comparison

When comparing model 1 and model 2 using the LOO information criterion, the second model seemed natably better. Model weights revealed the same indicating that the quality of model 2 is higher (see table 3).

	Weight

Model 1	0.313
Model 2	0.687

Table 3: LOO comparison of model 1 and model 2

Model 3

Model 3, likewise, found a small positive effect on pitch variability (b = 0.38, CIs = 0.26 - 0.5) in the study with Danish-speaking participants, when going from TD to ASD.
A similar effect was found for American participants, when going from TD to ASD (b = 0.28, CIs = 0.16 - 0.41) (see figure 6).
Moreover, model 3 showed a small positive effect on pitch variability (b = 0.09, CIs = -0.09 - 0.27), when having TD as intercept, and subtracting the beta in the Danish speaking study with the beta for American speaking study.

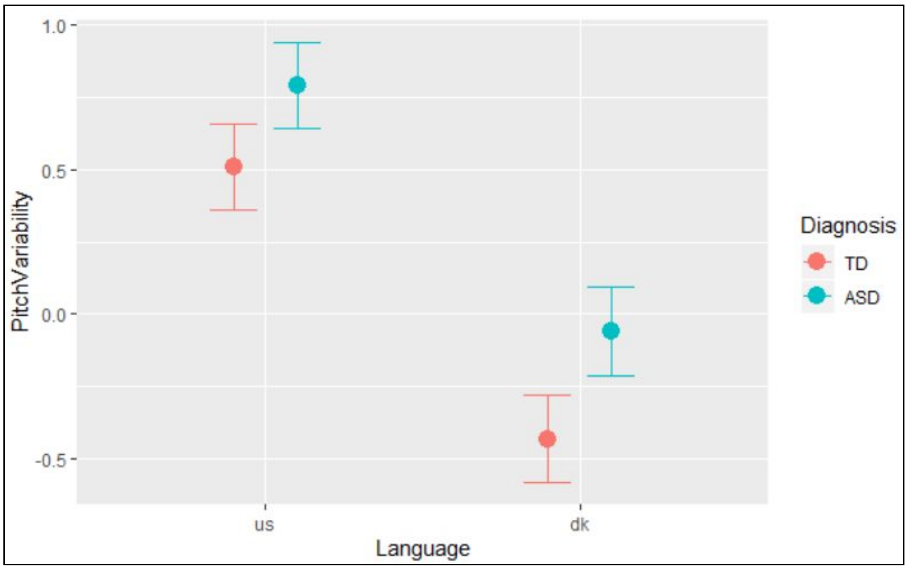


Figure 8: Effects of Model 3

M3H1: The evidence ratio for the first hypothesis was > 2000, meaning (according to model 3) it is very likely that ASD's have higher pitch variability than TD's, for the Danish speaking study

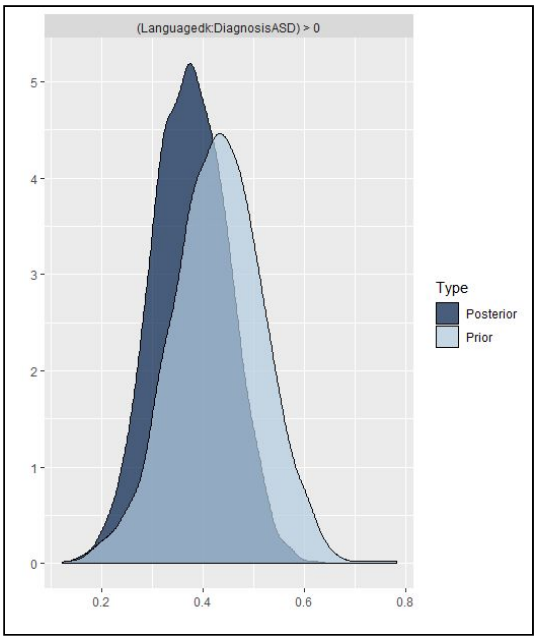


Figure 9: The posterior predictions for betas, with TD as intercept, and only for danish speaking.

M3H2:

The evidence ratio for the second hypothesis was > 2000, meaning (according to model 3) it is very likely that ASD's have higher pitch variability than TD's, for American speaking study

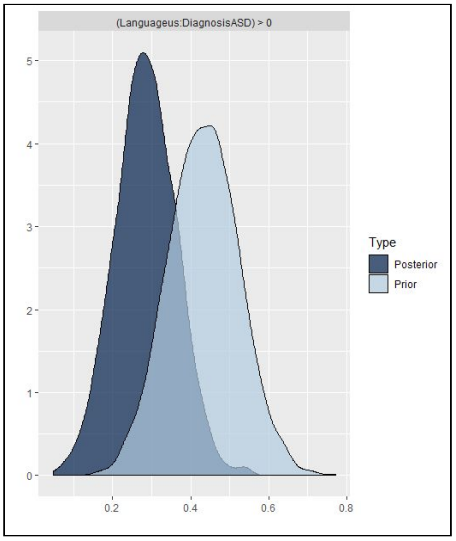


Figure 10: The posterior predictions for betas, with TD as intercept, and only for English American speaking

M3H3:

The evidence ratio for the third hypothesis was = 3.68, meaning (according to model 3) it is 3.68 times more likely that the difference between ASD's and TD's is higher for the Danish speaking study than for the American speaking study.

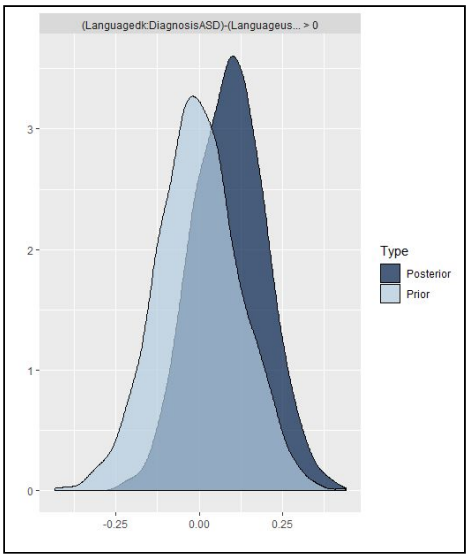


Figure 11: The posterior predictions for betas, with TD as intercept, when we subtract the beta in Danish speaking with the beta in English American speaking

Model 2 and 3 - Comparison

Comparing model 2 (with sceptical priors) and model 3 (with informed priors) as well based on the IC criterion 'LOO' also revealed quality differences. Model 2 seems slightly better, with 56.9% of the weight, however model 3 still has 43.1% of the weight, so we can't rule out that model 3 is in fact the best model (see table 4).

	Weight
Model 2	0.569
Model 3	0.431

Table 4: LOO comparison of model 2 and model 3

Discussion

In this assignment we have been testing two different priors: a conservative and meta-analytic prior. From our findings, the sceptical prior is slightly better but we cannot rule out the conservative prior. Assessing both priors, reveals the advantages and disadvantages of using a meta-analytic prior.

When looking at the effects estimates for the two models, there is a large difference as the direction has changed for the US beta value. Model 2, with the skeptical prior, pitch variability decreases, when going from TD to ASD, however, the opposite was found for the model with meta-analytic priors (model 3). This difference in estimate can also be seen in the plots of the posterior distribution of betas for US-speaking (figure 6 and 10). This indicates that the meta-analytic prior drags the estimate in a positive direction. One could speculate that dividing the studies from the meta-analysis into English and Danish speaking studies - and making different prior for the two beta values - could have affected the direction of the estimates. Moreover, the meta-analytic prior was made as a normal distribution with the estimates from the meta-analysis. This prior allowed almost no probability for an

estimate below zero. One could have made the distribution a student's t-distribution instead, which has fatter tails, so that we would allow for more probability of an estimate of zero and below. This would perhaps have been optimal.
The two models have almost the same weights, when comparing them using the LOO information criterion. Even though the model with conservative prior (model 2) has more weight in the small world, it is not enough to rule out that the model with the meta-analytic prior is the best model in the small world.

In assessing the quality of our models we took different approaches for different parts of quality assessment. The priors were assessed using prior predictive checks, which all looked good, except from the slight skew to the right of the data which the model did not capture, indicating that a gaussian distribution of all priors might not be the best fit (log normal might be a better choice). To assess the model fit further, we used trace plots to get an ocular overview of the models, looking for stationarity and good mixing, which all looked good (see Appendix 2). Rhat for all models was 1 indicating good convergence of the Markov chains. The effective sample size for the models also gave sufficient samples to rely on the mean calculations, but were slightly below the threshold of 2000 for relying on 99% CI (tail ESS = 1638), which could indicate that we might not have used quite enough samples.

Meta-analytic priors should have a role in scientific practice. We should always think about how we can include prior knowledge, as that is what the bayesian framework allows for. The more prior knowledge the better - the reason for it is quite simple - it will be more likely that studies represent the big world. This is only a given, if the general research field is not biased. If we have bad science in, we get bad science out, and in such a situation it is of course not good to have meta-analytic priors. If a field is very biased it might lead to more biased results, where we otherwise would have gotten good results.

Informed priors might overrule the evidence of a new study that has captured a true effect that might not have been discovered earlier (e.g. if the research field suffers from publication bias). This is a huge drawback of the method. It might be somewhat circumvented though. Informed meta-analytic priors would not suffer from the drawer problem to the same extent, if science developed into being more "Openly Scientific" - meaning that we pre-registered our studies to a larger extent. This might be useful in a combination with the Bayes framework of informed priors.

Seen in a bigger perspective, meta-analytic priors are a way of practicing the idea of cumulative science where we stand on the shoulders of giants. We use the information we have obtained from others' work balanced with thought through methods. In most fields of research, letting conservative approaches complement the more cumulative nature of using informed priors could be beneficial. The advantages of using informed priors are overwhelming but by investigating the field of interest with a sceptical approach as well, one is more likely not to be misled by general biases within the field. Our two models had quite similar stacking weights. We could have used an ensemble of the two models in order to not throw out any of the models.

References

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

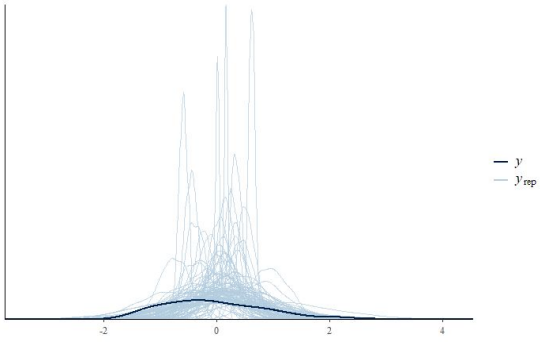
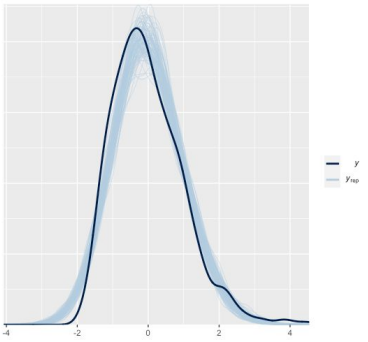
Paul-Christian Bürkner (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. Journal of Statistical Software, 80(1), 1-28. doi:10.18637/jss.v080.i01

Paul-Christian Bürkner (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. The R Journal, 10(1), 395-411. doi:10.32614/RJ-2018-017

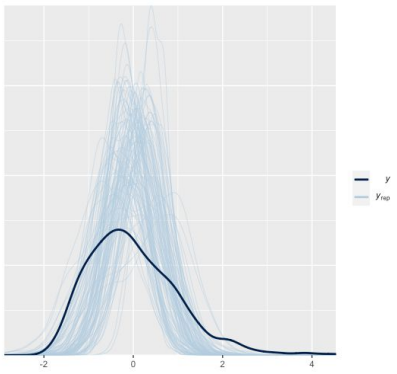
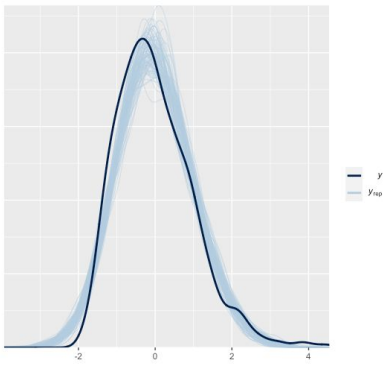
Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. Journal of Statistical Software, 36(3), 1-48. URL: http://www.jstatsoft.org/v36/i03/

Appendix 1

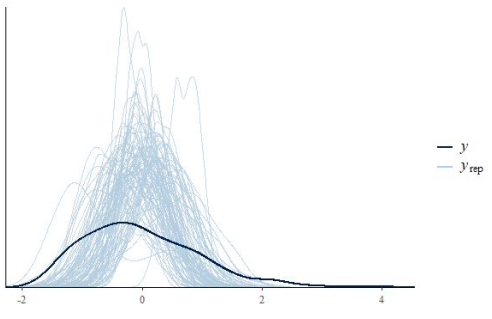
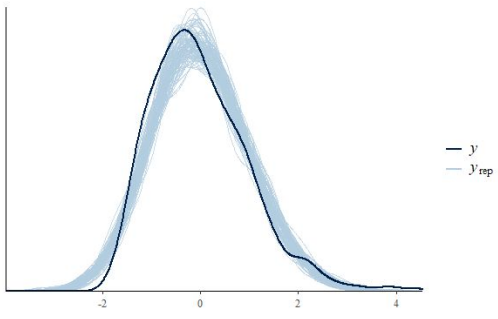
Model 1

	
Prior predictive check for first model 1	*Posterior predictive check for first model 1*

Model 2

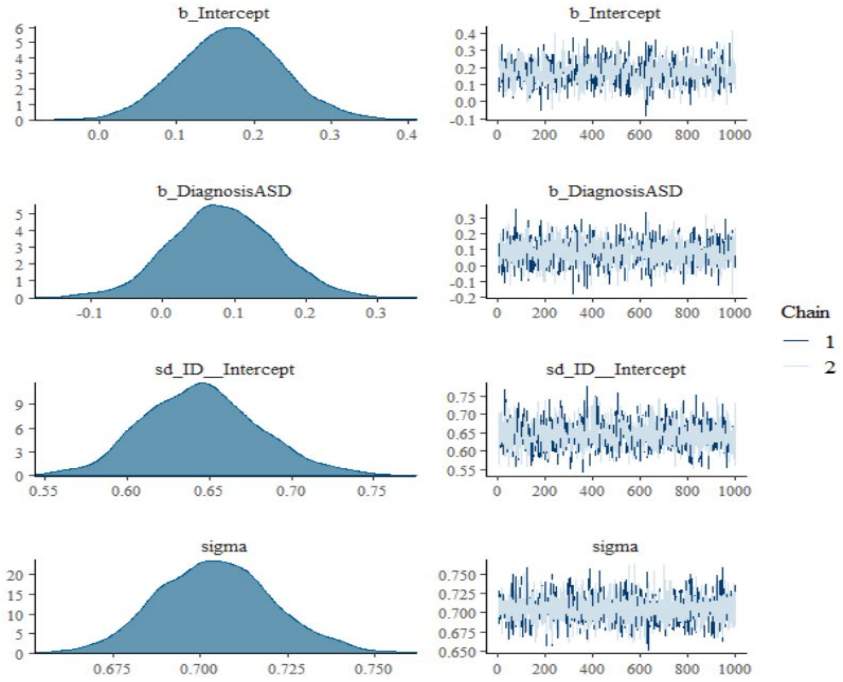
	
Prior predictive plot for model 2	*Posterior predictive check for model 2*

Model 3

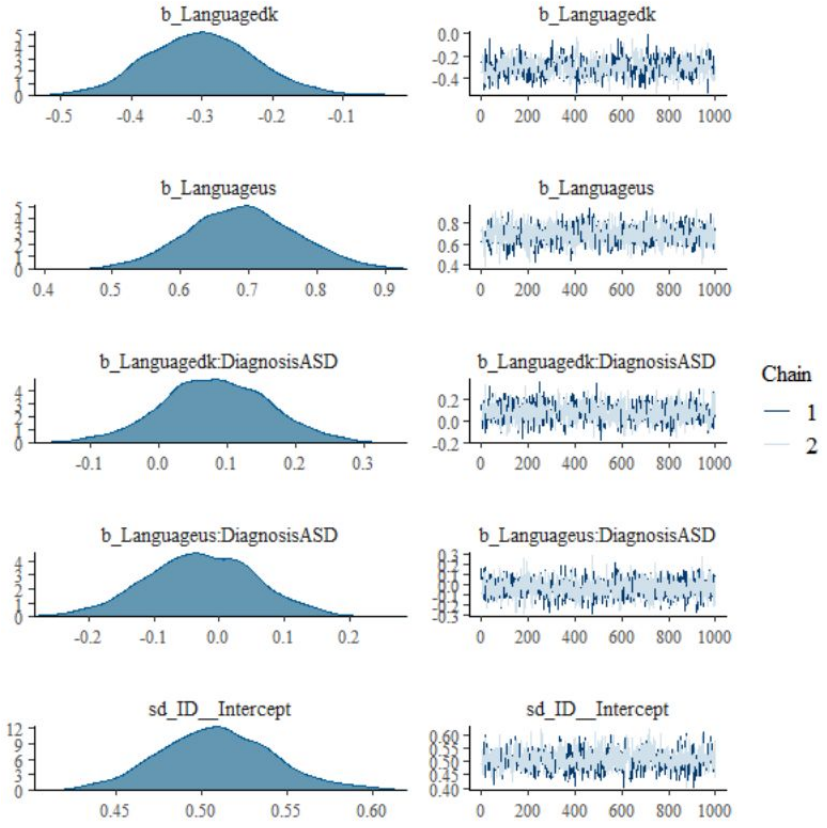
	
Prior predictive plot for model 3	*Posterior predictive plot for model 3*

Appendix 2

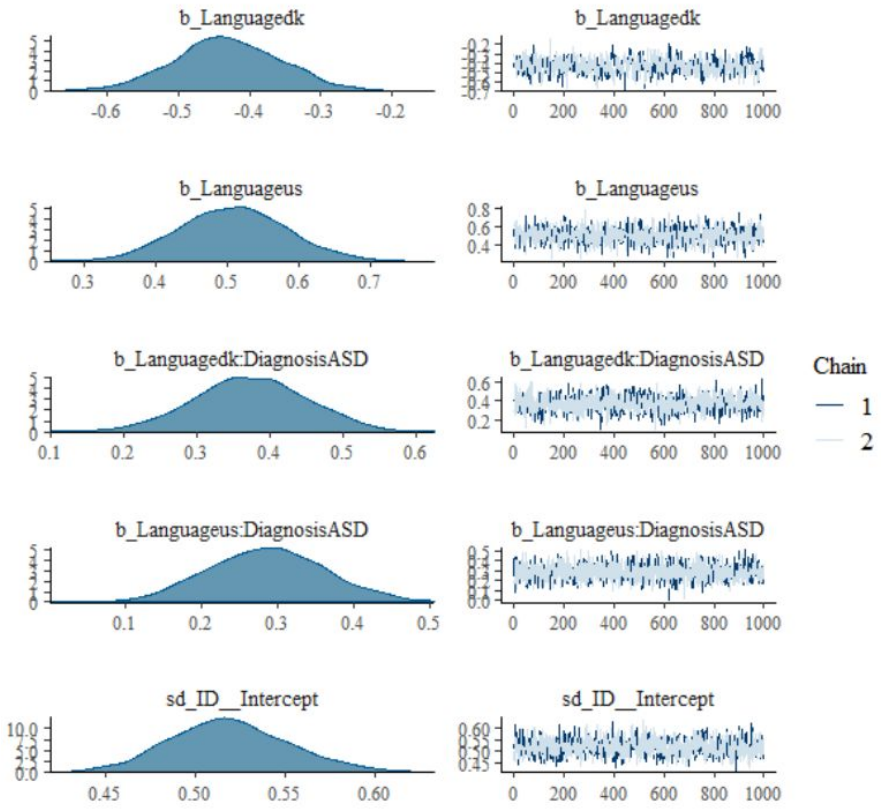
Model 1



Model 2



Model 3


b_Languagedk
b_Languagedk
b_Languageus
b_Languageus
b_Languagedk:DiagnosisASD
b_Languagedk:DiagnosisASD
Chain
1
2
b_Languageus:DiagnosisASD
b_Languageus:DiagnosisASD
sd_ID__Intercept
sd_ID__Intercept