Assignment 1: Eyetracking

Deadline for hand-in: 26/02-20
Github: https://github.com/askesvane/4-1-portfolio/blob/master/Data_visualizations.md

Foraging Experiment

Introduction

According to the eye-mind hypothesis (Just & Carpenter, n.d.), eye-movements are linked to the processing of the mind basically meaning that is it possible to infer cognitive processing in the brain from the movements of the eyes. In this experiment we investigated to what extent top-down constraints affect eye movements. The experiment is based upon a former experiment conducted by (Rhodes et al., 2014) that examined the effect of different instructions on eye movements. To examine the question we hypothesized that the length of the saccades will be different in search conditions compared to count conditions due to different degrees of top-down processing. In the visual count task, the stimuli to which the eyes have been presented will dictate the eye-movements. In the visual search tasks, the properties of the picture will leave no information about where to find the star.

H1: The top-down constraint of a search task will elicit significantly different saccade amplitudes than the top-down constraint of a count task.

Methods & Materials

Participants

In the experiment there were a total number of 6 participants (4 females, 2 males). They were all undergraduate students from Cognitive Science at Aarhus University.

Task

Two different tasks called ‘count’ and ‘search’ were presented to each participant. In the condition ‘count’, they had to count a given kind of stimuli in a presented picture. This could for instance be the numbers of e.g. birds, sheeps, or water drops. In the ‘search’ task, the participants had to look after a semi-transparent star hidden within the picture. A pool of 10 pictures were included in the experiment - each condition contained 5 trials randomized between the trials. The order of the pictures and which condition they were included in, was randomized across participants.

Procedure

The Eye Link 1000 was used to track the participants eye movements. 3 out of 4 females and 1 male had a left dominant eye, where the rest was right eyed. Data was recorded with a sampling rate of 500 Hz with an accuracy of 0.15º-0.5º. The the IViewX ninepoint automated calibration procedure was used and was repeated until the calibration was fulfilled.

Event estimation from Eye Link 1000 was used to define fixations and saccades. All data points that fell outside the screen coordinates (1680, 1050) were removed.

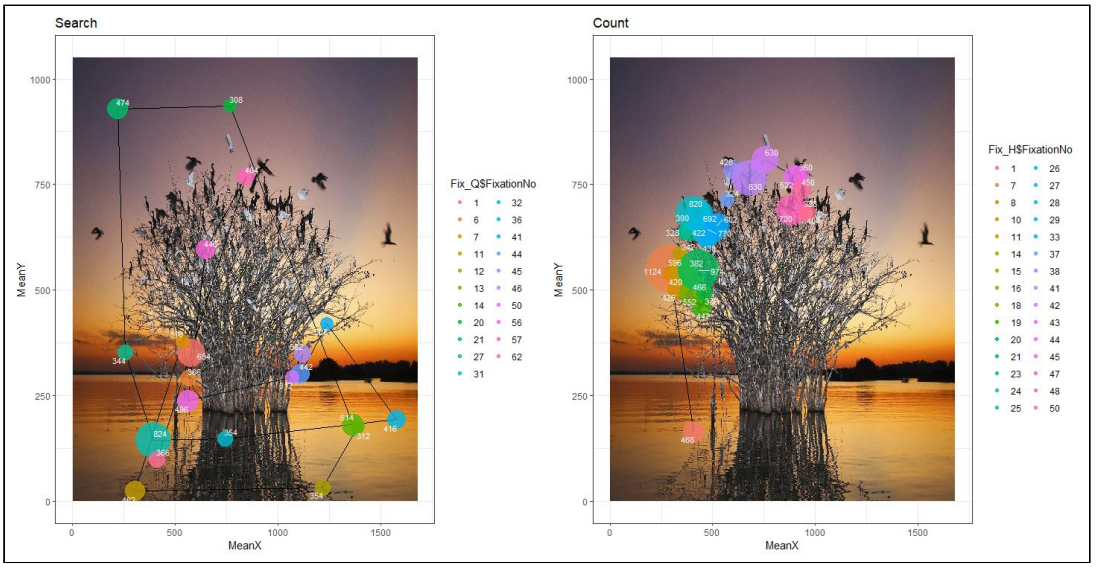


Figure 1: Scanpath for two different participants looking at the same stimuli but in the two different conditions

Figure 1 shows examples of scanpaths in the two different conditions after preprocessing of the data. Based on the plots we assume that the preprocessing and the event estimates of EyeLink 1000 was fair. Moreover, it indicates a difference in foraging patterns across conditions.
The effect of condition on eye movement (in this case length of saccades) can also be seen in figure 2 underlying the above-mentioned quality.

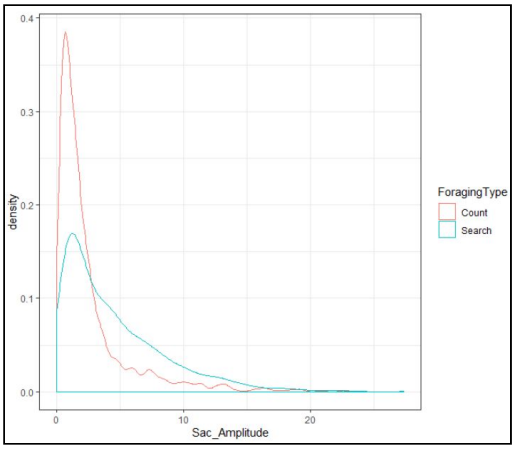


Figure 2: Density of saccade amplitude for each condition

We predicted length of saccades (length is measured as amplitude in degrees of the visual field) from condition (either 'search' or 'count'), as we have hypothesized that the eye movements will be different in the two conditions. As it was a repeated measures design, where all participants participated in both conditions, we have added varying effects. The model included varying effects for participant ID, with a varying slope for condition, and for stimulus (the different pictures shown), also with a varying slope for condition.
The model was a generalized linear model (GLM) with a lognormal distribution. We decided to use a lognormal distribution based on the assumption that saccade amplitude only deals with positive values. Furthermore, running both a model with gaussian and lognormal distributions revealed that the lognormal distribution fitted the data better. The analysis was conducted in R Studio (R Core Team, 2020) using the LME4 package (Bates et al., 2015)

Amplitude of saccades ~ 1 + Condition + (1+ Condition | ID) + (1 + Condition | Stimulus)

After obtaining the beta value, it was recalculated into a scale that is not lognormal in order to make the beta interpretable. Furthermore, we predicted the data only from our model and the recalculations made the predictions comparable to the actual data. This also enabled us to do a sanity check to look at residuals using the DHARMA package.

Results

The difference in saccade length from 'count' to 'search' condition was found to be significant on a lognormal scale ($b = 0.53735$, $SE = 0.07907$, p-value $< .05$). After calculating the beta value to be on a normal scale, the beta is 1.8. As the output variable is in 'degrees of the visual field', the mean difference in saccade amplitude in the condition 'search' is 1.8 degrees longer than in the condition 'count' - when taking varying effects into account.

From examining the spaghetti plot (figure 3), it is apparent that the length of saccades increases from 'count' condition to 'search' condition. However, differently for participant to participant, which is the reason for the varying effect.

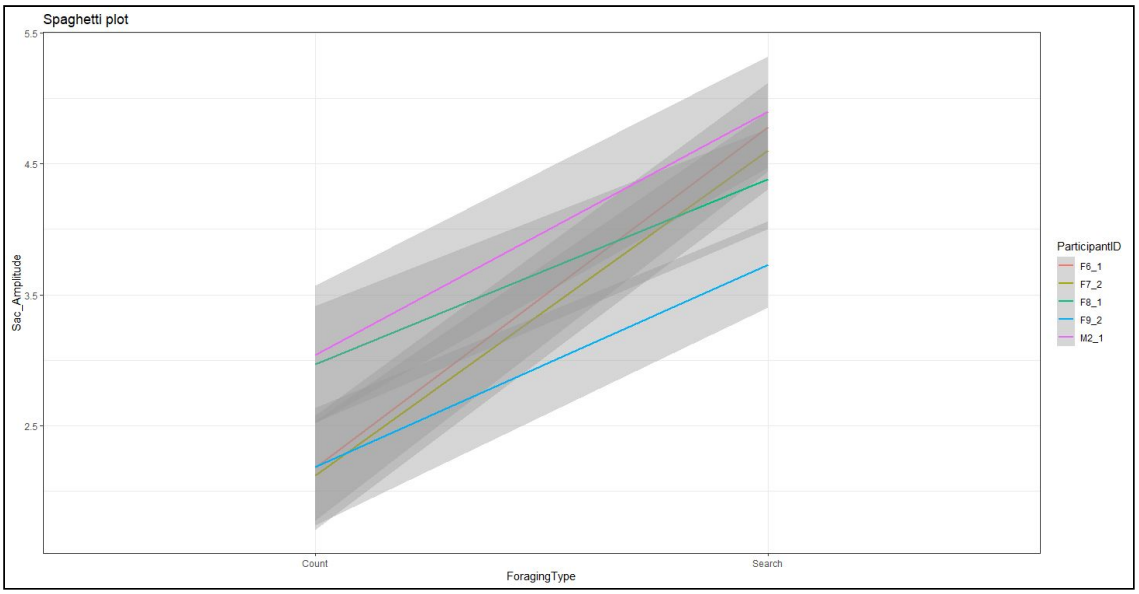


Figure 3: Spaghetti plot of saccade amplitude, between conditions for each participant.

However, condition only explains a (very) small proportion of the variance in saccade length, R^2 = 0.0269386. It roughly explains 3% of the variance in saccade length.

It is also worth noting that the residuals between predicted values and actual values, are of different magnitude.

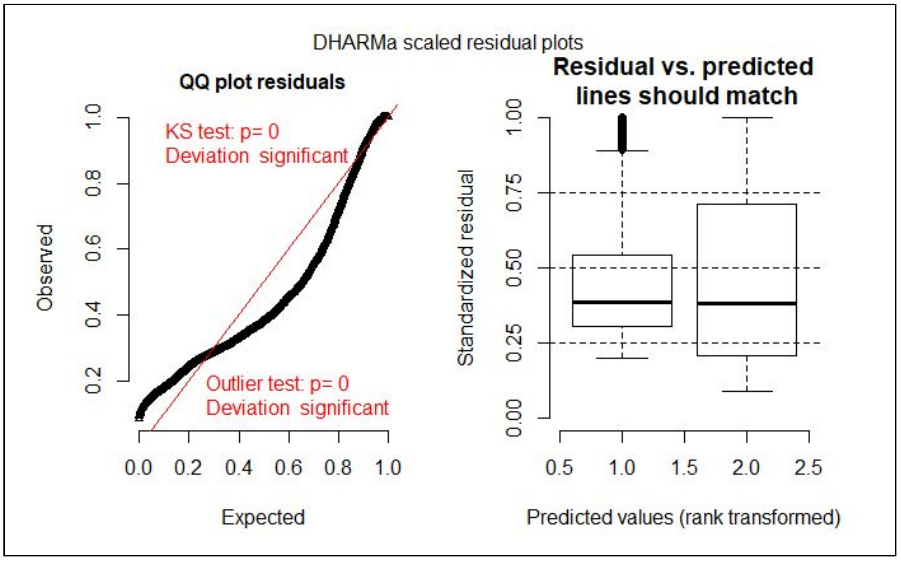


Figure 4: Residual plot for model (Harting, 2020)

For lower observed values the model predicts too low saccade amplitude values. However, at middle observed values the model predicts too high a saccade amplitude value (See figure 4, left). This could be an indication that the saccade amplitude distribution is perhaps not lognormally distributed.

In the right plot in fig. 4, it is apparent that - on the whole - the model undershoots for both condition ‘count’ and ‘condition’ search. Moreover there seems to be heteroscedasticity of residuals as the spread of residuals seems to be larger for count condition, than for the search condition.

Discussion

We hypothesized that a search condition would elicit greater mean saccade amplitudes than a count condition due to two different underlying top-down processes. Our results support the hypothesis even though the effect seems vanishingly small. Together with our limitations in terms of few participants lead us to the conclusion that more research within the field needs to be done.

Social Engagement experiment

Github: https://github.com/askesvane/4-1-portfolio/blob/master/social_engagement.md

Introduction

All of the studies of social cognition have been focusing most on observing social interactions, but what happens if we make the participants engage in a task? The cognitive processes of observing social behaviour might not be the same as the cognitive processes of actually engaging in social interactions. In this experiment we investigated the impact of social engagement based on a former study by Tylén et al. (2012). More specifically it was investigated how social engagement through eye contact influence pupil size, under the assumption that pupil size is a measurement for emotional arousal.

H: We hypothesize that interactively engagement increases physiological arousal that will lead to greater pupil dilation.

Methods & Materials

Participants

In the experiment there were a total number of 6 participants (5 females, 1 male). They were all undergraduate students from Cognitive Science at Aarhus University.

Task

In this experiment the participant was shown stimuli of a person doing an action.
The experiment included 8 different video clips that were shown to all participants in a random order. The videos showed both a female and a male actor to correct for gender. More than gender, the videos varied over two variables: ostensive/non-ostensive and direct/averted (see figure 1). This makes it a two by two factorial design with within participant measures. The participants were told that they would have to answer a question about either the person in the video or the object the person were holding. They were given no other instructions.

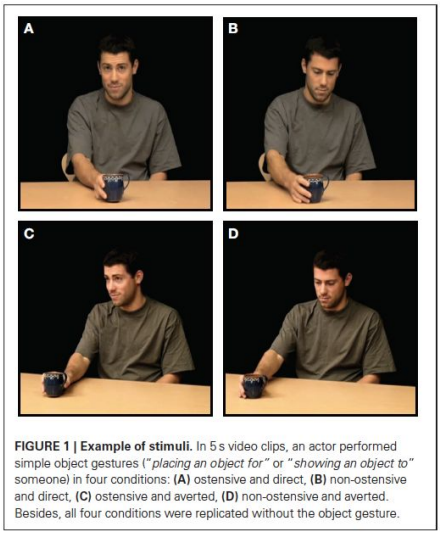


FIGURE 1 | Example of stimuli. In 5 s video clips, an actor performed simple object gestures ("*placing an object for*" or "*showing an object to*" someone) in four conditions: **(A)** ostensive and direct, **(B)** non-ostensive and direct, **(C)** ostensive and averted, **(D)** non-ostensive and averted. Besides, all four conditions were replicated without the object gesture.

Figure 1: Example of stimuli (Tylén et al., 2012)

Procedure

The Eye Link 1000 was used to track the participants eye movements. 2 out of 5 females and the male had a left dominant eye. Data was recorded with a sampling rate of 500 Hz with an accuracy of 0.15º-0.5º. The the IViewX ninepoint automated calibration procedure was used and was repeated until the calibration was fulfilled.
Event estimation from Eye Link 1000 was used to define fixations and saccades. All data points that fell outside the screen coordinates (1680, 1050) were removed.
Across participants the relative pupil size varied across a scale defined by the software (see table 1).

Mean pupil size	**Min. pupil size**	**Max pupil size**	**Standard deviation**	**Participants**
6523.7	1815	8488	835.8	6

Table 1: overview of mean pupil size, minimum pupil size, maximum pupil size, and participant number

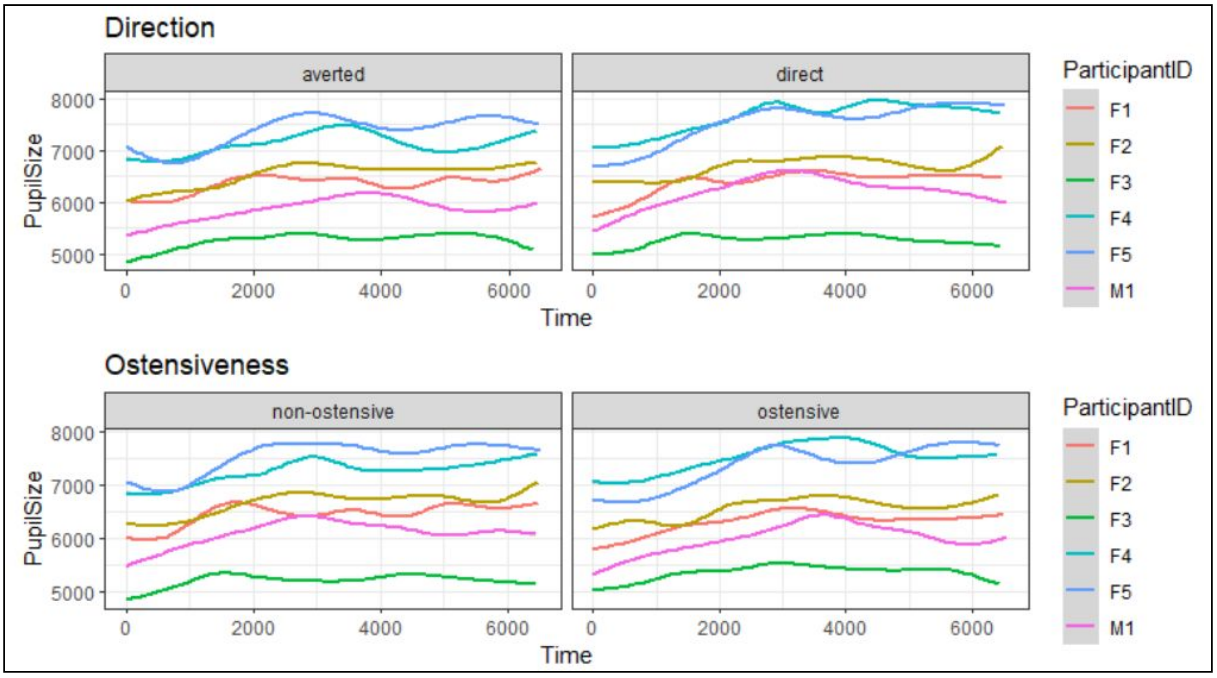


Figure 2: density plot of pupil size for each participant in direct/averted and ostensive/non-ostensive respectively

Figure 2 shows pupil size over time for each participant in terms of ostensiveness and direction. Based on this, we assume that the preprocessing and the pupil size estimates of EyeLink 1000 was fair.

We predicted pupil size from both direct/averted and ostensive/non-ostensive. As it is a two by two factorial design we included an interaction between the two. As it was a repeated measures design, where all participants saw all stimuli, we have added varying effects. The model included varying effects for participant ID, with a varying slope for the interaction effect between our two predictor variables, as we believe that the baseline of pupil size direction and ostensive cue will vary across participants. We also hypothesized (for the interaction effect) that the direction would moderate the effect ostensiveness has on pupil size. We modelled pupil size using generalized linear models (GLM). We used two models, one assuming a lognormal distribution, and one assuming gaussian distribution. We then predicted values from our two models, to see which model was best. The analysis was conducted in R Studio (R Core Team, 2020) using the LME4 package (Bates et al., 2015)

Pupil size ~ Direct * Ostensive + (1 + Direct * Ostensive | ID)

Alternatively, we also ran a model without the interaction effect in case the interaction turned out insignificant.

Pupil size ~ Direct + Ostensive + (1 + Direct + Ostensive | ID)

Results

From the two models (one assuming a lognormal distribution, and one assuming gaussian distribution) we compared predictions with actual values for both, to see which model was best. From these results we chose lognormal distribution (see fig. 3).

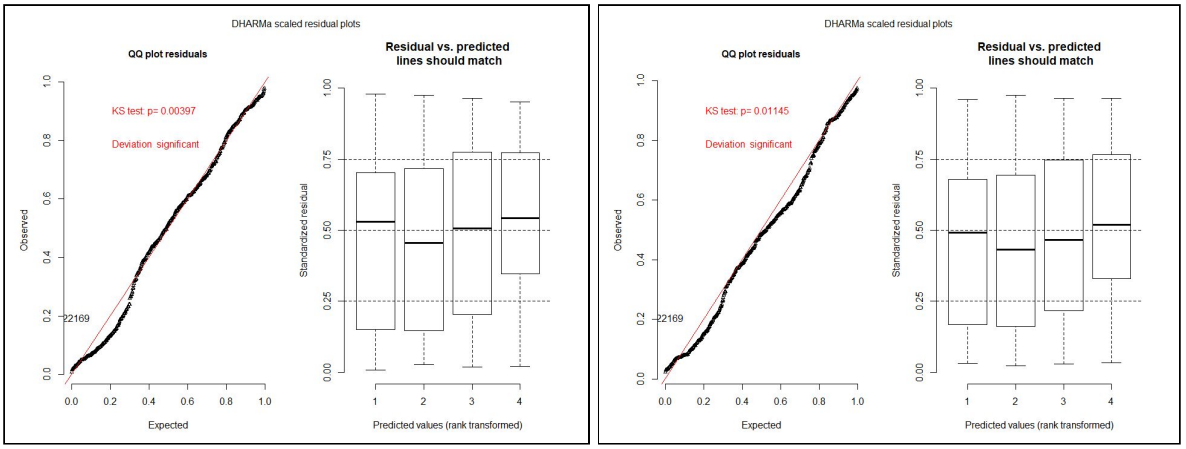


Figure 3: DHARMA plot for our model predictions using a lognormal distribution (left) and gaussian distribution (right)

No significant results were found in the model with interaction. Therefore, the model without interaction was used. In the condition of direction, when going from averted to direct the pupil size increases with significance on a lognormal scale (b = 0.03, SE = 0.006, p > 0.05). When the beta value is converted to the scale used by the equipment, it translates to 193.4.

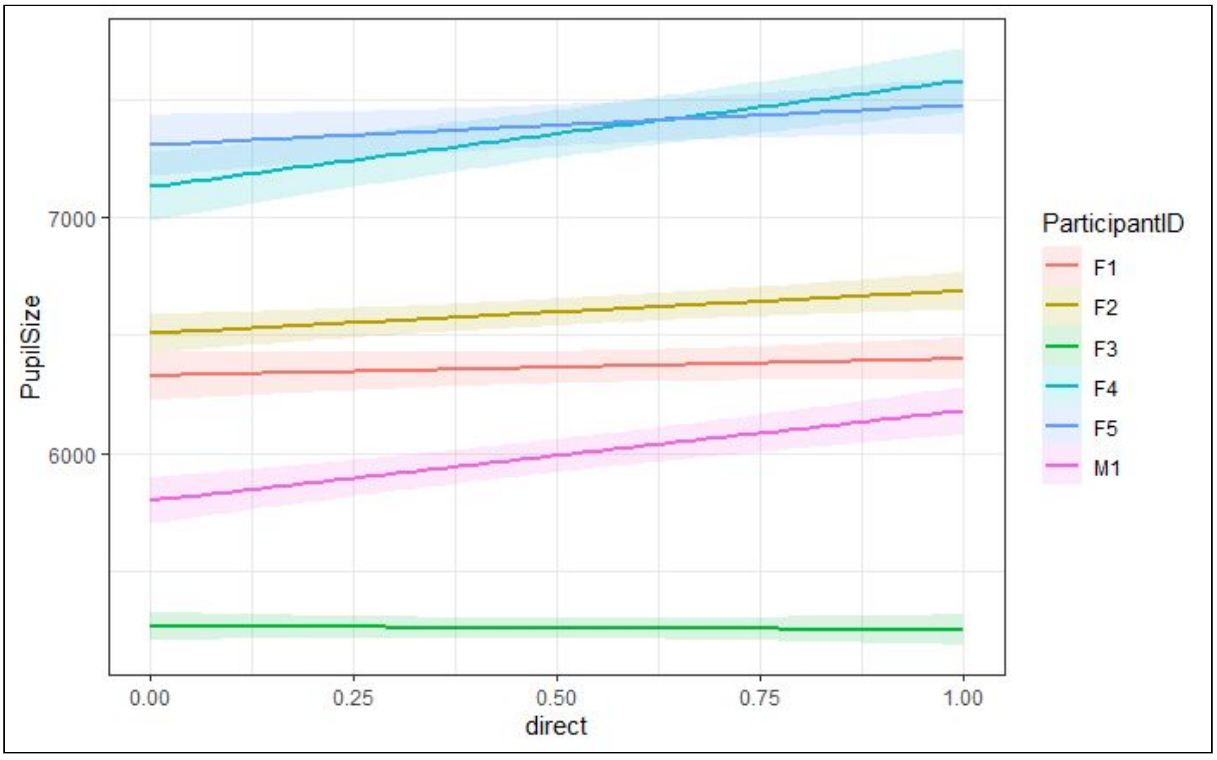


Figure 4: Spaghetti plot. The effect of direction on pupil size.

However, the direction of the actor (direct/averted) even though significant, didn't explain much of the variance (R^2 = 0.0139). This can also be seen in figure 4 - that the effect of direction on pupil size is very small.

Discussion

The non-significant results when going from ostensive to non-ostensive and the small effect of direction might be due to the fact that the eye contact is not with an actual person but a video. One might speculate that doing the same experiment with an actual person as stimuli would elicit different results. Further research should therefore be performed in real life.

References

Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.

Florian Hartig (2020). DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models. R package version 0.2.7. https://CRAN.R-project.org/package=DHARMa

K. Tylén, M. Allen, B.K. Hunter, A. Roepstorff (2012). Interaction vs. observation: distinctive modes of social cognition in human brain and behavior? A combined fMRI and eye-tracking study. Retrieved from https://www.frontiersin.org/articles/10.3389/fnhum.2012.00331/full

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/.

Theo Rhodes, Christopher T. Kello & Bryan Kerster (2014) Intrinsic and extrinsic contributions to heavy tails in visual foraging, Visual Cognition, 22:6, 809-842, DOI: 10.1080/13506285.2014.918070

Assignment 2: CogSci Knowledge

Deadline for hand-in: 11/3-2020
Github: https://github.com/saraoe/assignment-2-4

Part 1

1. What's Riccardo's estimated knowledge of CogSci? What is the probability he knows more than chance (0.5)

- According to the posterior distribution of Riccardo's (RF's) knowledge of CogSci, it is most probable that Riccardo's knowledge of CogSci is 50% (the peak probability is at 0.5)
- Extracted from the distribution as well, there is a 50 % chance that he knows more than chance (because the area under the curve above 0.5 on the x-axis is 0.5).

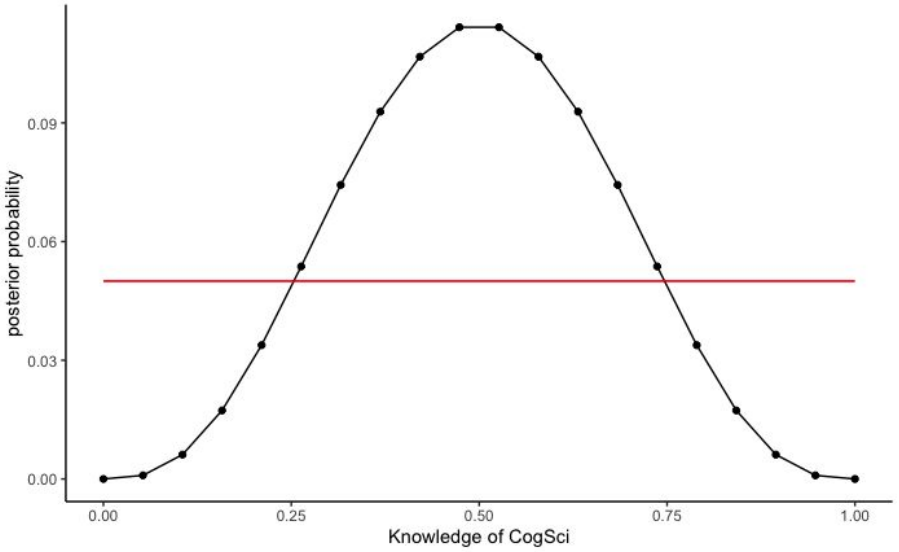


Figure 1: Prior and posterior distribution of Riccardo's knowledge of CogSci

2. Estimate all the teachers' knowledge of CogSci. Who's best? Use grid approximation. Comment on the posteriors of Riccardo and Mikkel.

- KT has the highest peak at 1 on the posterior probability, meaning he has the highest probability of answering all future questions correctly. JS has a slightly lower peak at .81 knowledge of CogSci, however, with a much higher probability (expressed on y-axis in figure 2). Also, there is approx. 0% probability that JS answers below change, while there is approx. 12% probability that KT does (table 1). From sampling the posterior distribution, we have also found that JS has a 52.4 percentage chance of having more knowledge than KT (table 2).
- MW and RF both answered half of their questions correct. However, even though having the same prior, the posterior distribution of the two differs. This is due to MW answering more questions, and the distribution is therefore more narrow around 50%. In other words, fewer

data points means the prior weighs more on the posterior (relative to the data). From sampling the posterior distribution, we have also found that RF has a 49.6 percentage chance of having more knowledge than MW (table 2).

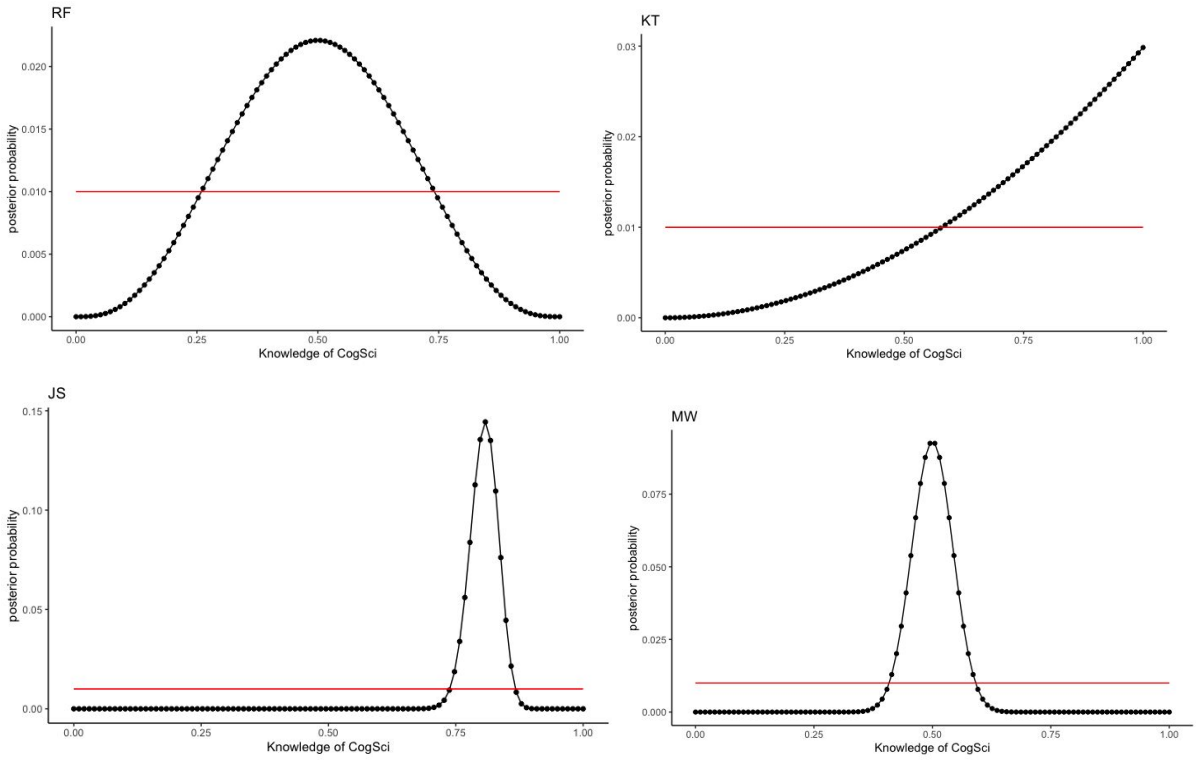


Figure 2: Prior and posterior distribution of all teacher's knowledge of CogSci with uniform prior

Teacher	Knowledge at MAP	Probability of knowledge above chance
RF	0.5	0.5
KT	1	0.8769
JS	0.81	Approx. 1
MW	0.5	0.5

Table 1: Teacher's knowledge of CogSci from posterior calculated from a flat prior

	KT>	JS>	RF>	MW>
KT	na	52.4	16.52	13.06
JS	47.57	na	3.28	Approx. 0
RF	83.48	96.7	na	50.37
MW	86.93	100	49.62	na

Table 2: Matrix of the comparison of teacher's knowledge given the flat prior (e.g. JS has a 52.4 percentage chance of having more knowledge than KT)

3. Change the prior. Given your teachers have all CogSci jobs, you should start with a higher appreciation of their knowledge: the prior is a normal distribution with a mean of 0.8 and a standard deviation of 0.2. Do the results change (and if so how)?

- The posterior distributions have changed as we have changed the prior probability distribution. By changing the prior, the posterior changes according to the new prior.
- This is especially clear when looking at Mikkel Wallentin and Riccardo. Mikkel is hardly influenced (the peak on the posterior distribution is still around .5), whereas Riccardo is highly influenced with a posterior distribution skewed towards .8 (and therefore looks closer to the prior).
- To also compare the results regarding who has the highest probability of being the smartest teacher, we reran the previous comparisons with the new prior.
  In table 4, we can see the probability of the different teachers having a true knowledge of CogSci higher than the others. KT and JS both have very high probabilities of being the smartest. An example of change is that the probability of KT being smarter than JS is now >50, whereas it was <50, with the flat prior.

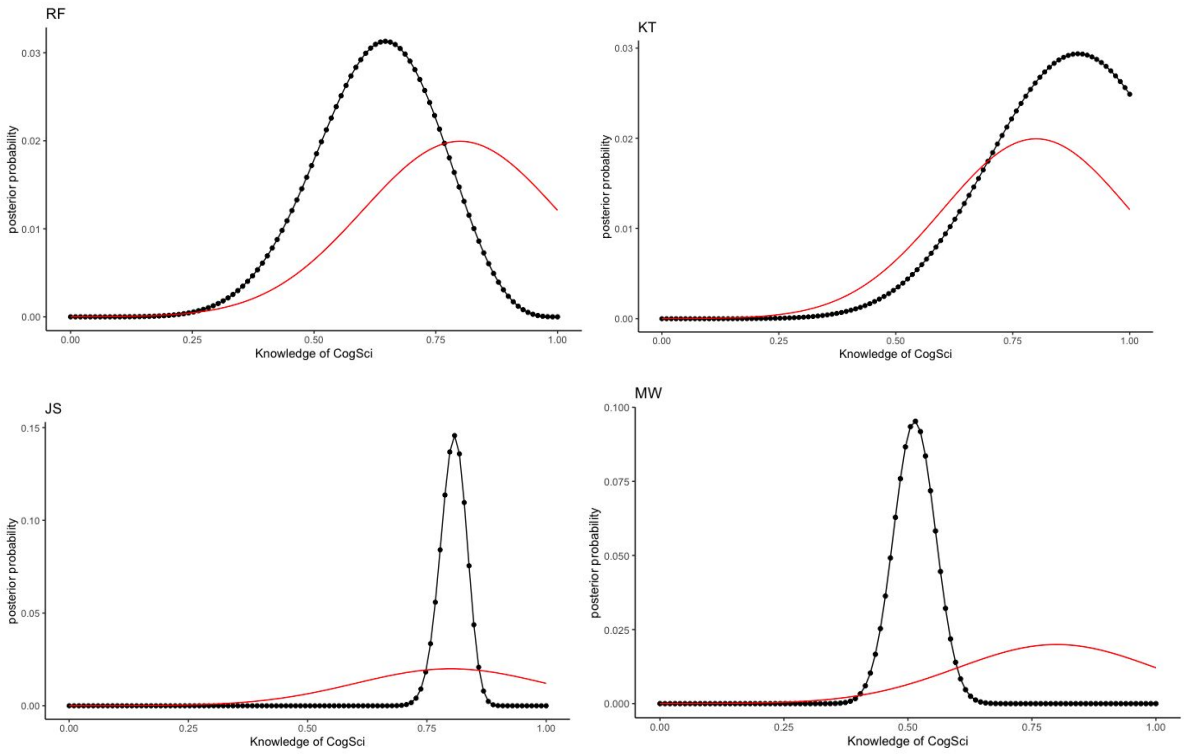


Figure 3: Prior and posterior distribution of all teacher's knowledge of CogSci with normally distributed prior

Teacher	Knowledge at MAP	Probability of knowledge above chance
RF	0.65	0.8418
KT	0.89	0.9761
JS	0.81	Approx. 1
MW	0.52	0.62

Table 3: Teacher's knowledge of CogSci from posterior calculated from a normally distributed prior

	KT>	**JS>**	**RF>**	**MW>**
KT	na	45.4	17.11	3.2
JS	54.38	na	7.84	Approx. 0
RF	82.74	91.66	na	20
MW	96.6	100	80.2	na

Table 4: Matrix of the comparison of teacher's knowledge given the new prior (0.8 mean and 0.2 SD) e.g. JS has a 45.4 percentage chance of having more knowledge than KT

4. You go back to your teachers and collect more data (multiply the previous numbers by 100). Calculate their knowledge with both a uniform prior and a normal prior with a mean of 0.8 and a standard deviation of 0.2. Do you still see a difference between the results? Why?

- In the results above we found that differences in the prior affected the posterior distribution heavily. However, after adding extra data (by multiplying the previous numbers by 100) the differences between the posterior distributions with different priors decreases as the data increases.

Uniform prior data:

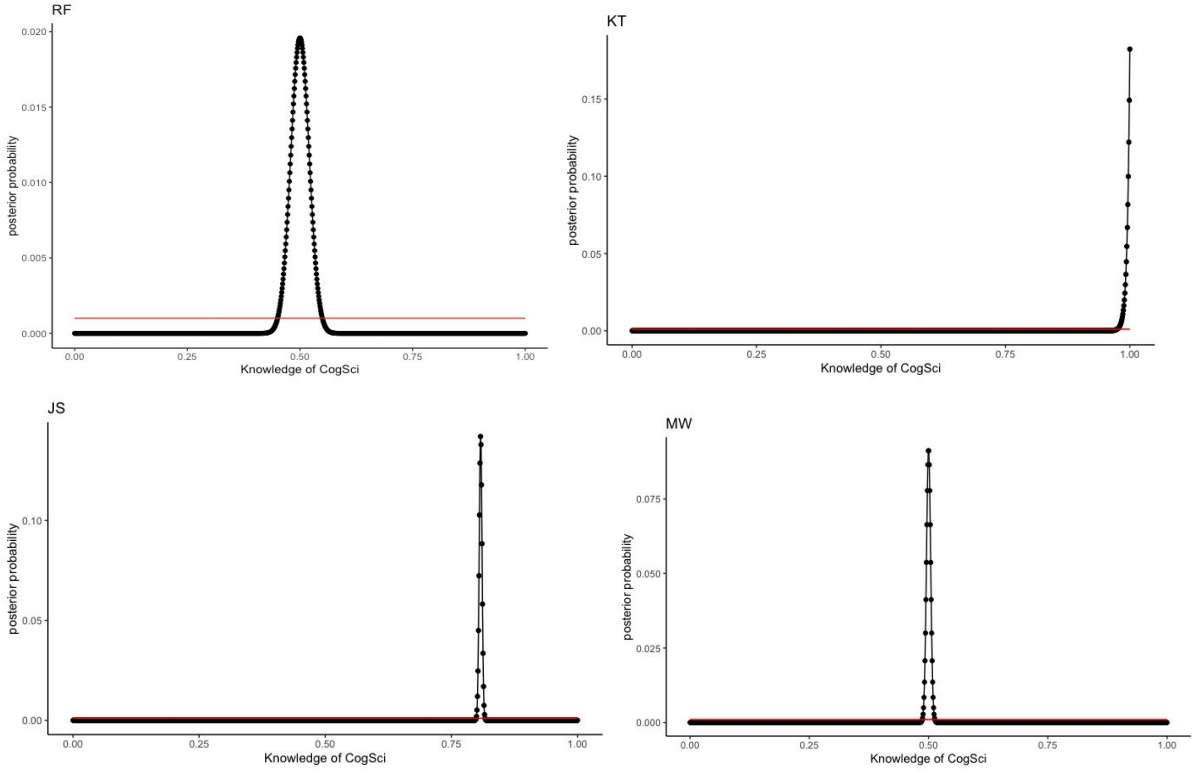


Figure 4: Prior and posterior distribution of all teacher's knowledge of CogSci with uniform prior and new data

Teacher	**Knowledge at MAP**	**Probability of knowledge above chance**
RF	0.5	0.5

KT	1	Approx. 1
JS	0.808	Approx. 1
MW	0.5	0.5

Table 4: Teacher's knowledge of CogSci from posterior calculated from a flat prior with the new data

Normal prior data:

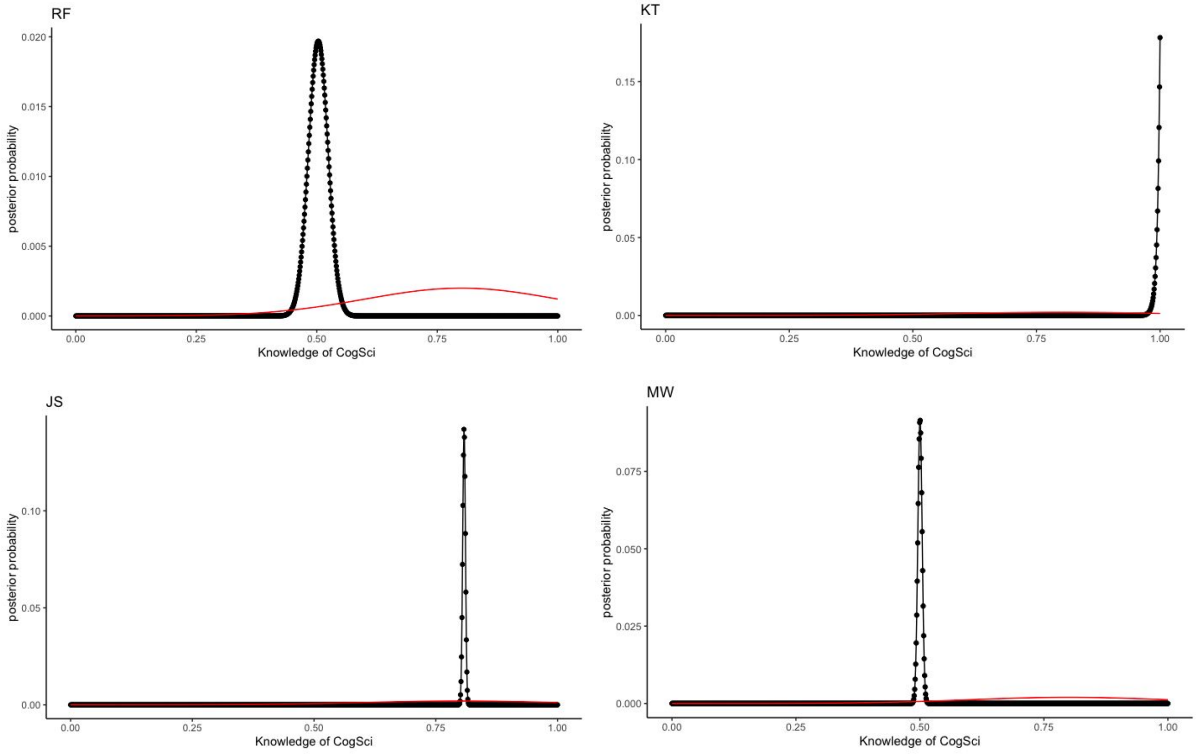


Figure 5: Prior and posterior distribution of all teacher's knowledge of CogSci with a normally distributed prior and the new data

Teacher	**Knowledge at MAP**	**Probability of knowledge above chance**
RF	0.504	0.5604
KT	1	Approx. 1
JS	0.808	Approx. 1
MW	0.501	0.5130

Table 5: Teacher's knowledge of CogSci from posterior calculated from a normally distributed prior with the new data

5. Imagine you're a skeptic and think your teachers do not know anything about CogSci, given the content of their classes. How would you operationalize that belief?

- Since it does not change the facts about the data we have to change the prior (our own assumption)

- As we are skeptical, we believe that the mean will be .5 (chance). Moreover, we have made a small SD (0.05), as we are fairly certain that all teachers will answer close to half of the questions correct by change.

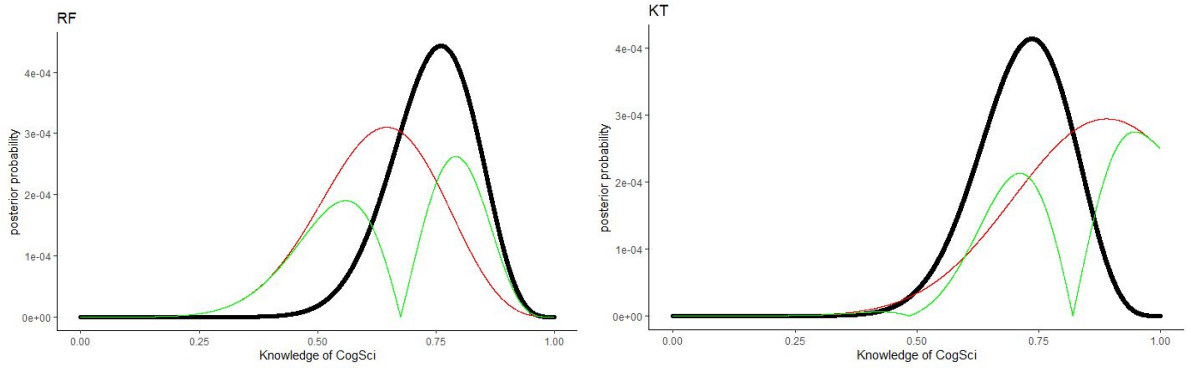
Part 2 - Focusing on predictions

1. Write a paragraph discussing how assessment of prediction performance is different in Bayesian vs. frequentist models

In a frequentist framework, we make a null hypothesis to test our assumptions. Our null hypothesis would be that there is no difference in the teacher's knowledge of CogSci from the first to the second year. We make a t-test and investigate this hypothesis, and if the p-value is below 0.05 we can reject the null hypothesis.
This type of model can lead to overconfident estimations of what is right and wrong, because the conclusion is binomial - either we reject the null hypothesis or we don't.

In a Bayesian framework, we would not test the difference on an arbitrary cutoff. Rather, we would just explain the differences in the prior, and the posterior. This we could do in a number of different ways, e.g. plotting prior and posterior, point estimate (e.g. MAP), intervals (e.g. PI or HPDI) or subtracting the prior from the posterior.

2. Provide at least one plot and one written line discussing prediction errors for each of the teachers.



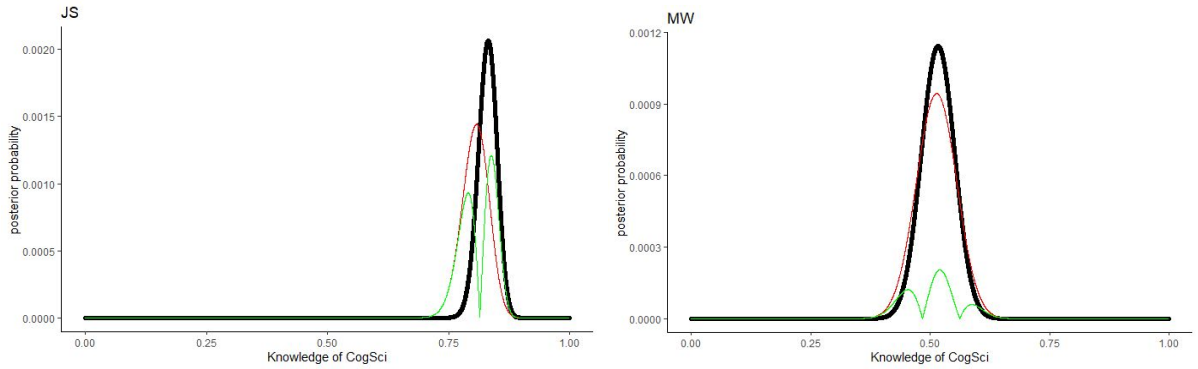


Fig. 6: Plot of prior (red), posterior (black) and the difference between the two in absolute numbers (green)

We have calculated a new posterior from the new data, using the posterior from the first year as the prior. The distributions are shown in figure 6.
The results from our model on the data from the first year seems to predict the data from the second year fairly well. The predictions are especially good for MW, where the difference between the posterior from the first year and the posterior from the second is very small (see green line on fig. 6). The overlap of the old posterior (now prior) and the posterior distributions are, however, quite large for all teachers, indicating that it makes relatively good predictions about the teachers knowledge of CogSci.

We have also calculated the difference in maximum a posteriori (MAP) for the two posterior distributions (old data and new data) of each teacher (table 5). This too shows that the prediction is good for MW. However, there is a somewhat large difference in MAP for KT and RF. Though, when examining the plots of the distributions (fig. 6) there is a great overlap.

teacher	MAP_posterior	MAP_prior	MAP_diff
JS	0.8321	0.8079	0.0242
KT	0.7364	0.8899	0.1535
MW	0.5166	0.5136	0.0030
RF	0.7611	0.6464	0.1147

Table 5: MAP for both posterior (map_posterior) and prior (map_prior) distributions for each teacher and the difference between those (map_diff)

	Prior HPDI (50%)	**Posterior HPDI (50%)**
RF	0.5658566 - 0.7353735	0.6969697 - 0.8187819
KT	0.8137814 - 0.9838984	0.6679668 - 0.7981798
JS	0.7868787 - 0.8243824	0.8192819 - 0.8449845
MW	0.4826483 - 0.5391539	0.4944494 - 0.5417542

Table 7: HPDI intervals for each individual, prior (which was the posterior for the old data) and posterior for the new data

Now, we assess how the model we have generated from the old data will predict the new data. We sample 10.000 times from the posterior distribution of each teacher and asses the change of obtaining the specific number of correct answers given the model we have from the old data (+/- 3%).

JS	RF	MW	KT
28.24%	9.6%	24.11%	10.72%

Table 8: Prediction of new data given old model with +/- 3% interval from actual data

The old model predicts the new data for JS and MW well - above 20% of the times the model would estimate a guess within the 3% interval of the actual guess in the new data (table 8). However, the old model does not predict RF and KT as well. This can also be seen in the plot of prediction errors (figure 7), where the errors of JS and MW have a higher density around zero (no prediction error), while this is not the case for RF and KT.

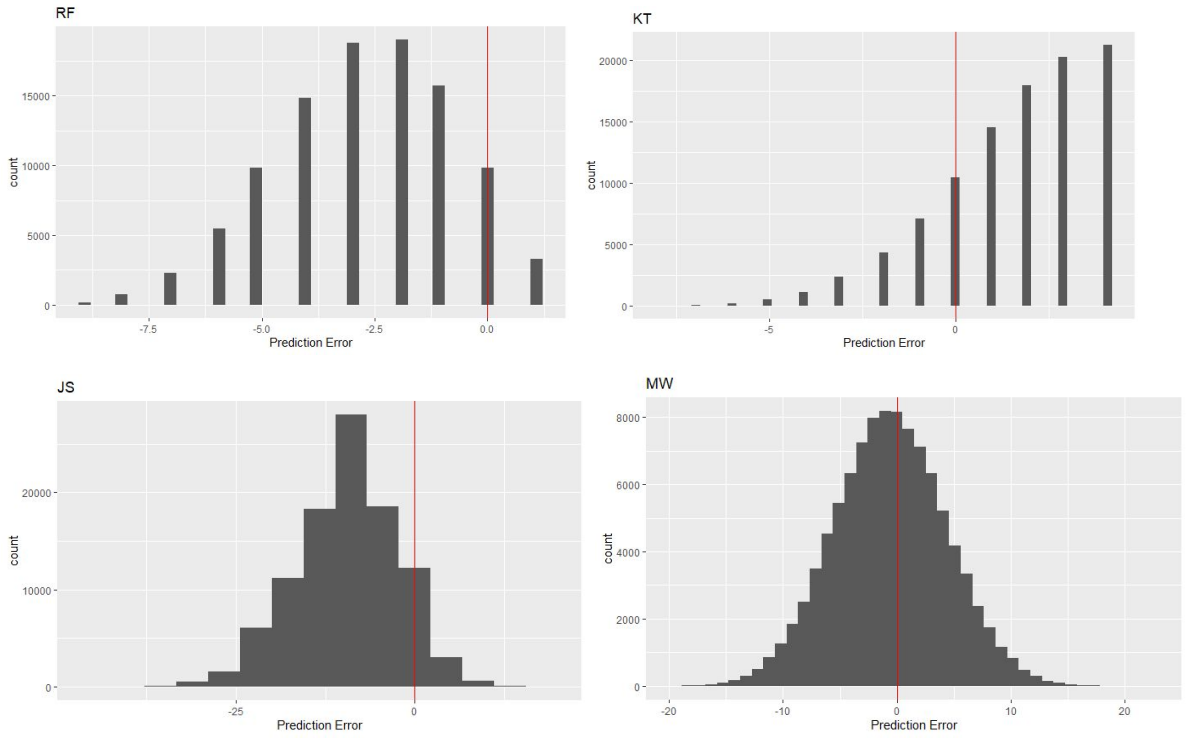


Figure 7: Prediction errors for each teacher, when using the old posterior to predict the new data. The red line is zero prediction error.

Assignment 3: Causal Inference

Deadline for hand-in: 26/03-2019
Github: https://github.com/saraoe/assignment_3_causal_inference

First part

Q1.1) Test if schizophrenia involve higher altercentric intrusion
To test if people with schizophrenia have higher altercentric intrusion, we have made a model that predicts the score for altercentric intrusion from the diagnosis (control or schizophrenia)

Altercentric Intrusion ~ 0 + Diagnosis

We define the same priors for both estimates - for schizophrenics and controls - which is a normal distribution, as altercentric intrusion is a continuous variable, with a mean of 4 and a SD of 1. By having the same prior, our prior belief is that there is no difference between the two groups. The prior for sigma we defined as a normal distribution with a mean of 1 and a SD of 2.

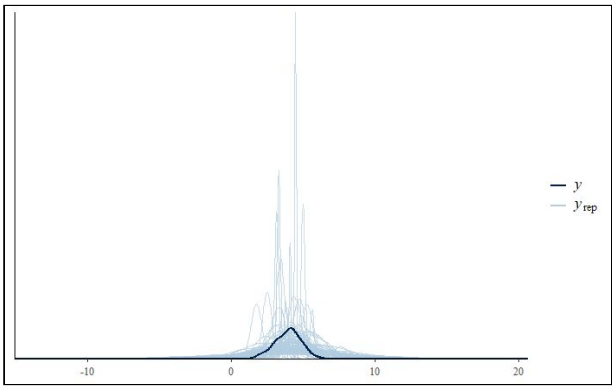


Figure 1: Prior predictive check

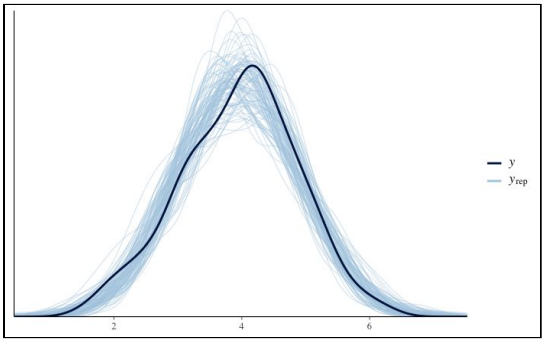


Figure 2: Posterior predictive check

The model indicates a credible difference in altercentric intrusion in the two groups supporting our hypothesis (b = 0.36, CIs = 0.16, 0.57, ER = 1332). Controls showed on average an altercentric intrusion effect of 3.86 (CIs 3.74, 3.98), and schizophrenia of 4.22 (CIs = 4.01, 4.43). The difference between controls and schizophrenics can also be seen visually in figure 3 and 4.

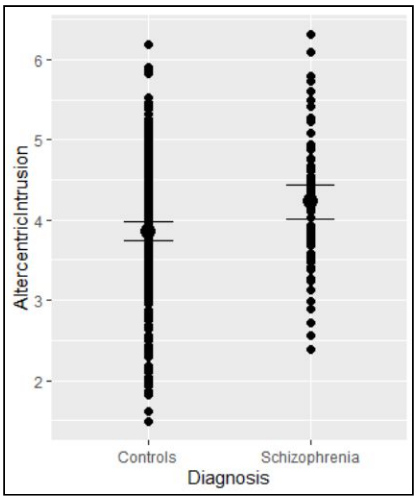


Figure 3: Plot of actual data

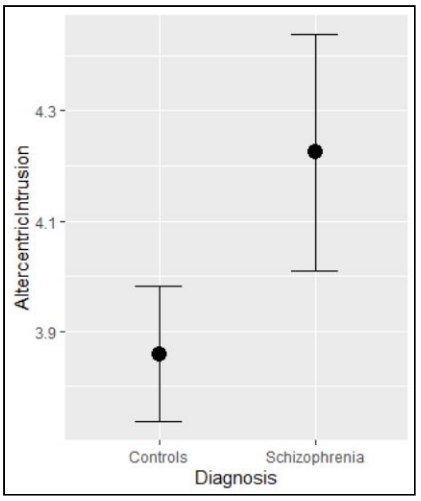


Figure 4: Difference between controls and Schizophrenics

Q1.2) Is altercentric intrusion related to specific symptoms in the patients?

To test if altercentric intrusion is related to Voice Hearing, Mind reading and Apathy we made the following models:

$$Altercentric\ Intrusion \sim 1 + Voice\ Hearing$$
$$Altercentric\ Intrusion \sim 1 + Mind\ Reading$$
$$Altercentric\ Intrusion \sim 1 + Apathy$$

We have mean centered all predictor variables for better interpretation of the intercept of the models. For all models we have a prior posterior which is normally distributed with a mean of 4 and a SD of 1 and a sigma, which is also normally distributed, with a mean of 1 and a SD of 2.
For the voice hearing and mind reading models, we have made a normally distributed prior for the beta, which has a mean of 0.5 (as we expect a positive slope) and a SD of 1. For the apathy model, we have also made a normally distributed prior for the beta, with a mean of - 0.5 (as we expect a negative slope) and a SD of 1.

The model for voice hearing showed a small positive effect of voice hearing on altercentric intrusion (b = 0.08, CIs = -0.19, 0.35). The credibility interval crosses zero. Therefore, there is a probability that voice hearing actually has no effect on altercentric intrusion.

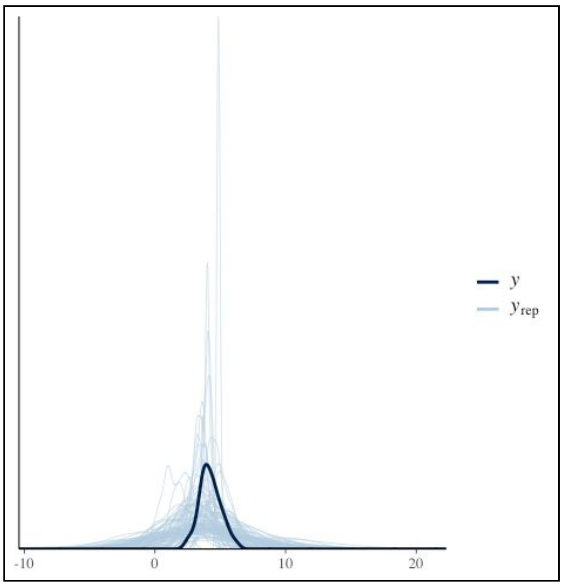


Figure 5: Prior predictive check for model with Voice Reading

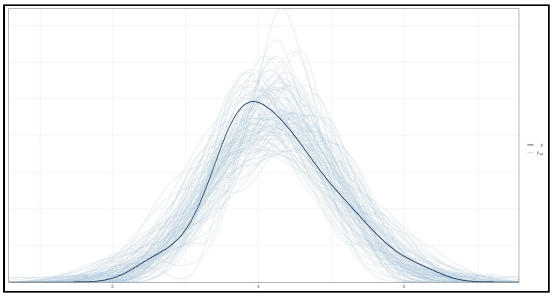
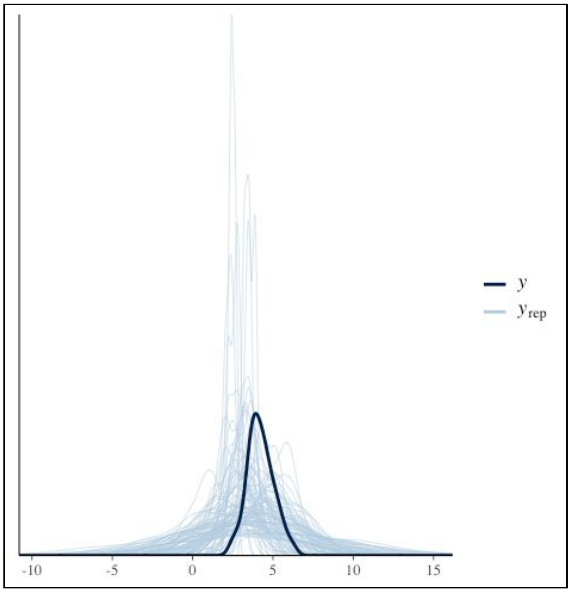


Figure 6: Posterior predictive check for model with Voice Reading

The model for mind reading showed a small positive effect of mind reading on altercentric intrusion (b = 0.09, CIs = -0.14, 0.30). The credibility interval crosses zero, there is a probability that mind reading actually has no effect on altercentric intrusion.



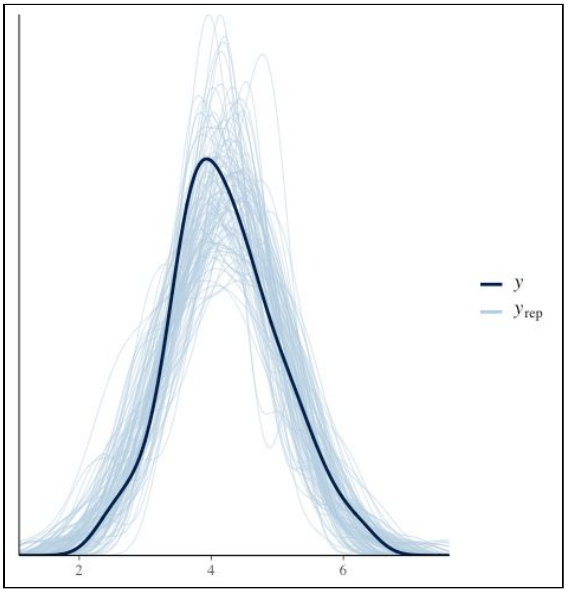


Figure 7 (left): Prior predictive check for the model with Mind Reading
Figure 8 (right): Posterior predictive check for the model with Mind Reading

The model for apathy showed a small negative effect of apathy on Altercentric Intrusion (b = - 0.23, CI-95% = -0.49-0.03). The credibility interval crosses zero, there is a probability that apathy actually has no effect on altercentric intrusion.

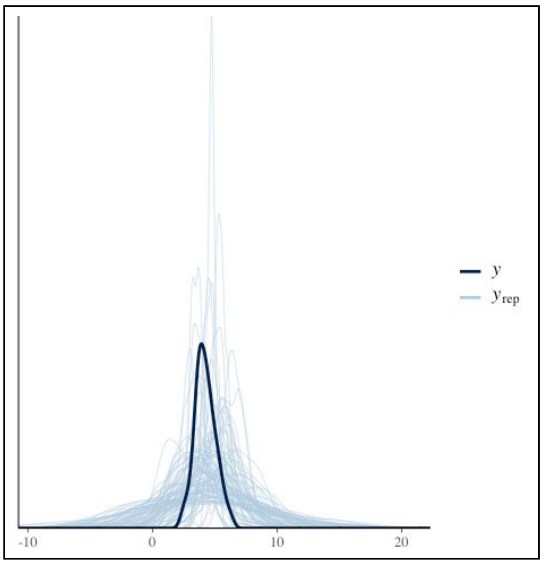


Figure 7: Prior predictive check for the model with Apathy

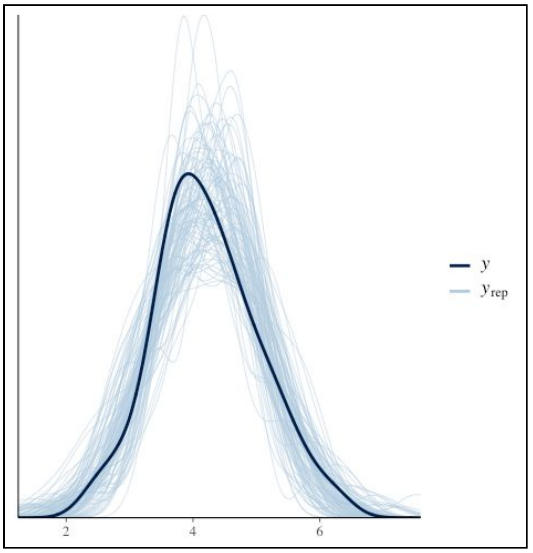


Figure 8: Posterior predictive check for the model with Apathy

Multiple Regression Model

As none of the symptoms seem to predict altercentric intrusion especially well on their own, we suspect that there might be a masking effect of the other variables. To unmask this effect, we have chosen to include all three symptoms in our final multiple regression model.

Altercentric Intrusion ~ 1 + Voice Hearing + Mind Reading + Apathy

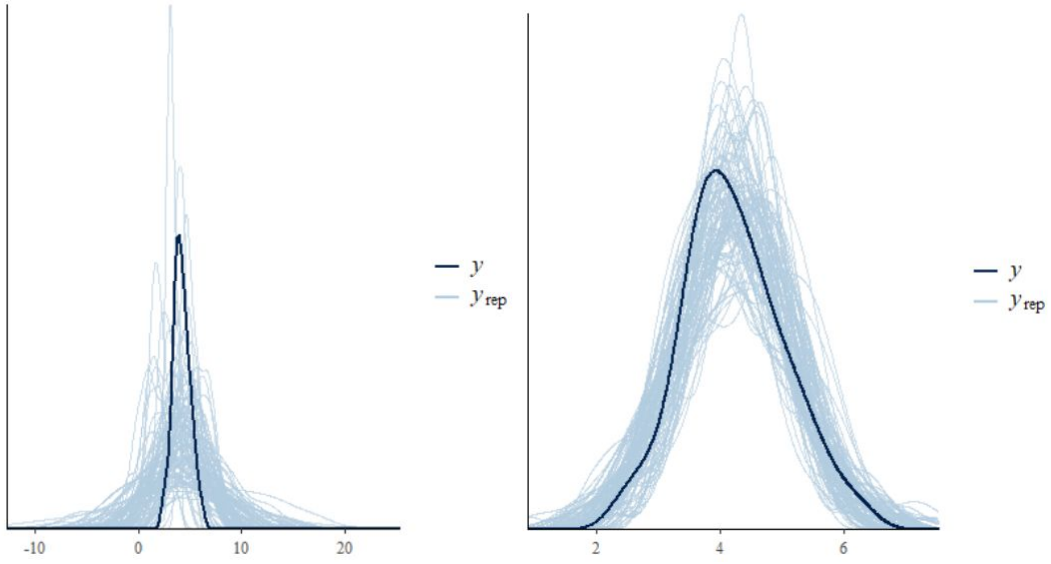


Figure 9 (left): Prior predictive check for our multiple regression model
Figure 10 (right): Posterior predictive check our multiple regression model

The effects for the multiple regression model is very similar to the effects of the symptoms in the models above: There is a small positive effect for both voice hearing (b = 0.05, CIs = -0.26, 0.36) and mind reading (b = 0.03, CIs = -0.25, 0.30) on altercentric intrusion, while there is a small negative effect of apathy (b = -0.21, CIs = -0.53, 0.11). However, all with credibility intervals crossing zero, making the estimated effects unlikely (see figure 12).

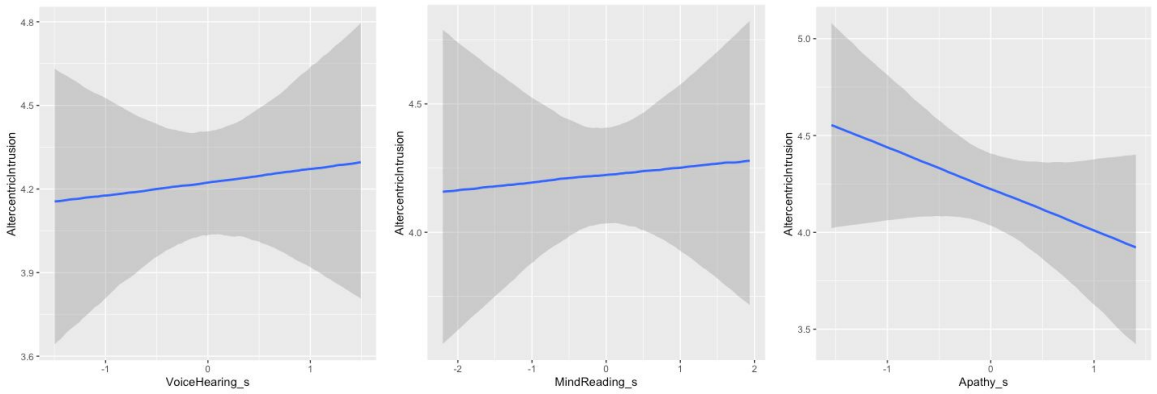


Figure 11: Plotting the effect of the three predictors (left = Voice Hearing, middle = Mind Reading, right = Apathy), CI = 95%.

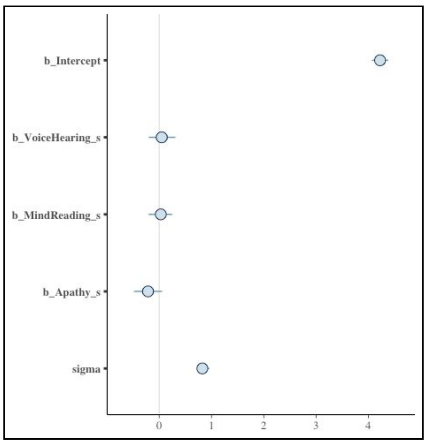


Figure 12: coefficient plot, all parameters overlap with 0, CI = 95%. As such, the plot shows how bad the model is.

Second part

Q2.1) Draw a causal graph (Directed Acyclical Graph) of the variables. Discuss which biases you might have introduced

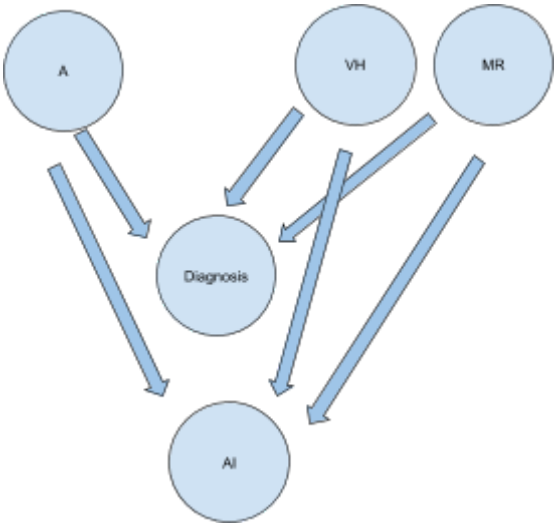


Figure 13: DAG plot showing causal inferences. AI = Altercentric Intrusion, VH = Voice Hearing, MR = Mind Reading, A = Apathy

The multi regression model we made above was made on a subset of the data, where we only included the schizophrenics. However, subsetting the data like this introduced a bias in our model (see figure 13). Both altercentric intrusion and the symptoms that we used as predictors cause schizophrenia - i.e. the higher score one has in the different symptoms, the more likely the person is to have schizophrenia, and also the higher score one have in altercentric intrusion, the more likely the person is to have schizophrenia.

Q2.2.) Redesign your analysis following the graph and report how the results change
To redesign the study, we refrain from conditioning on diagnosis. In other words, we run the same model, but on data for both controls and people with schizophrenia. Thereby, we have closed the backdoor so the same variance will not flow to altercentric intrusion more than once.

By running the model on the complete dataset the coefficients change. There is a positive effect for both voice hearing (b = 0.15, CIs = 0.05, 0.25) and mind reading (b = 0.16, CIs = 0.05, 0.27), and the credible intervals no longer overlap with zero, which indicates a small probability that the effect is zero - i.e. no effect (see figure 14). However, even though we hypothesized that apathy would have a negative impact on altercentric intrusion - it most likely has not. There is a small positive effect of apathy (b = 0.01, CIs = -0.10, 0.13), though, very uncertain. Furthermore, we made a model without apathy predicting altercentric intrusion from voice hearing and mind reading alone. This is illustrated with the pseudo code underneath. The results can be seen in the right plot of figure 14. The estimates of the other predictors do not change markedly.

Altercentric Intrusion ~ 1 + Voice Hearing + Mind Reading

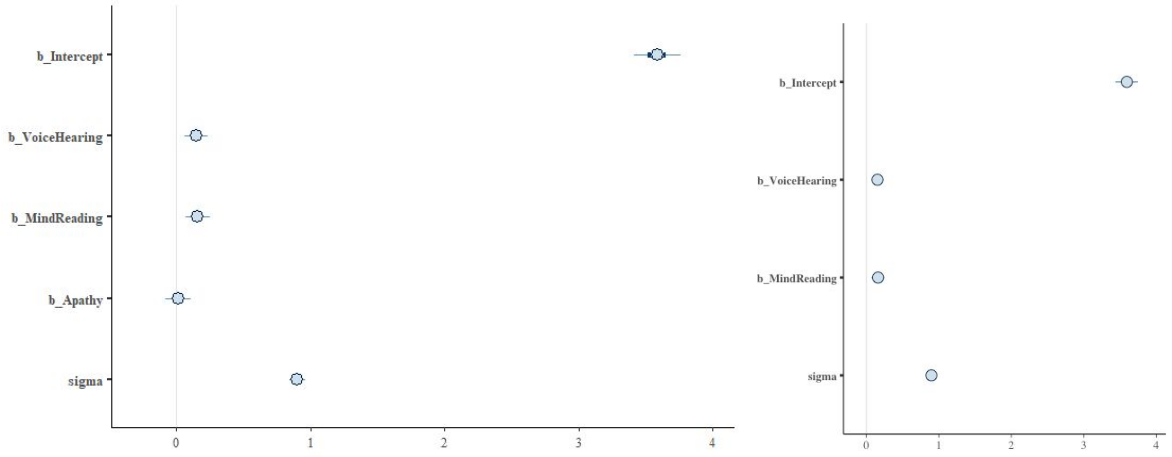


Figure 14: Coefficient plots using the entire dataset (left = multi regression model including Apathy, right = multi regression model excluding Apathy), CI = 95%

Third part

Q3.1) Look through the code and identify whether the results you have match the underlying truth. (or if the direction at least is the same). Discuss what you have learned.
It turns out that apathy did not predict any variance in altercentric intrusion, as we thought it would have in our model. The true model is shown in figure 15 and only differs by the missing causal link between apathy and altercentric intrusion, that we anticipated (see figure 13).

When conditioning on diagnosis, we did not find the effects of mind reading and voice hearing. When we ran the same model on the entire data set with both schizophrenics and controls, we found a positive effect of both mind reading and voice hearing (as one would according to the true model) and no effect of apathy, even though we added this as a predictor.
The beta estimates were not precisely 0.2. They were 0.15 and 0.16 so very close to the true effect of the symptoms. This could be due to the priors of the betas. We defined the mean of the prior as 0.1 for both betas and this might have 'drawn' the estimate away from 0.2 and closer to 0.1, leaving out estimates a bit low compared to the 'true' estimates. However, the true effect lies within the CIs and one would always expect some noise.

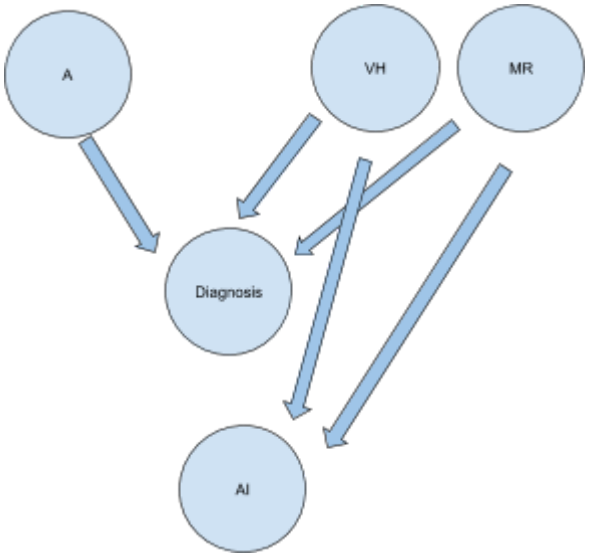


Figure 15: The "true" model

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

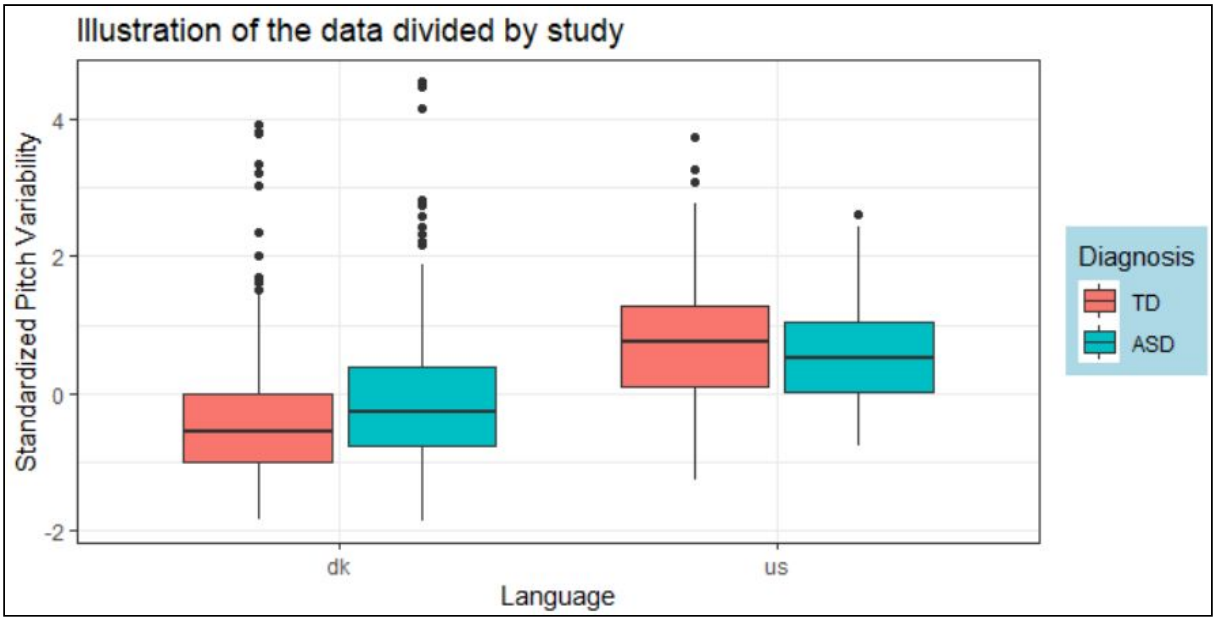


Figure 1: Illustration of data from the two new studies.

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).

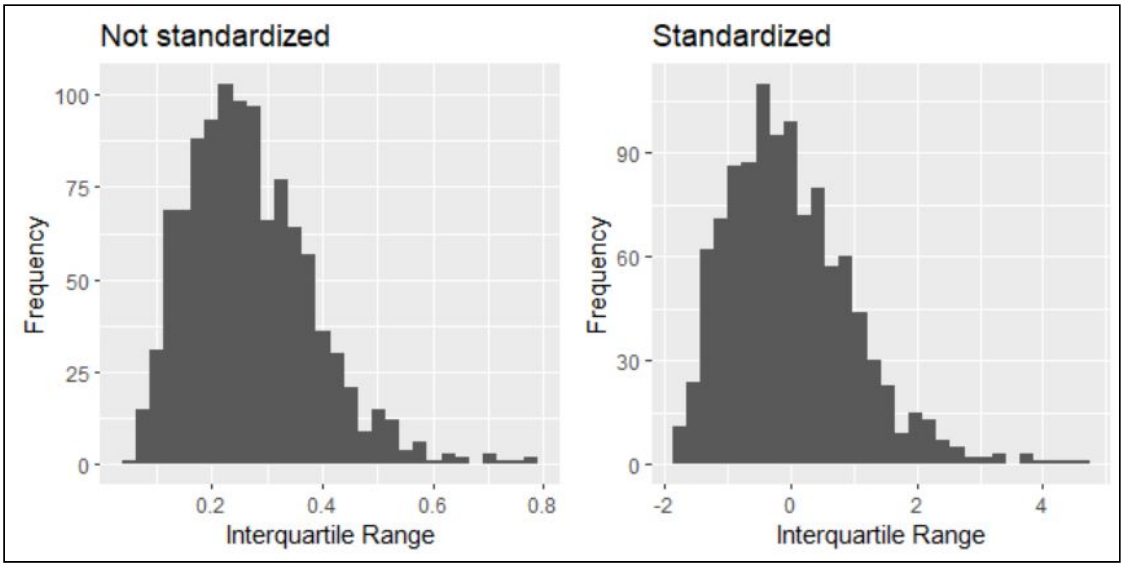


Figure 2: Left: Histogram illustrating the frequencies of different IQR values. Right: Histogram of the same values, but standardized.

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.

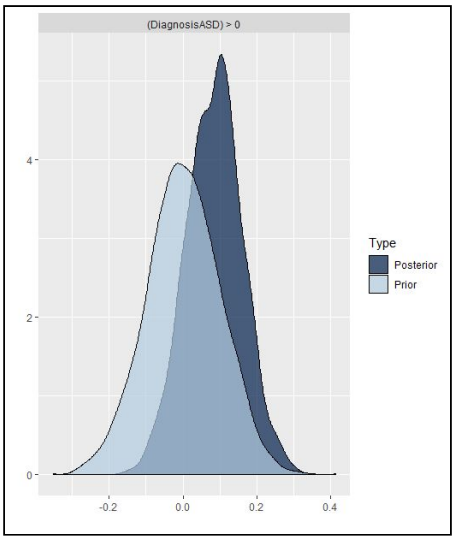


Figure 3: The posterior predictions for betas, with TD as intercept

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.

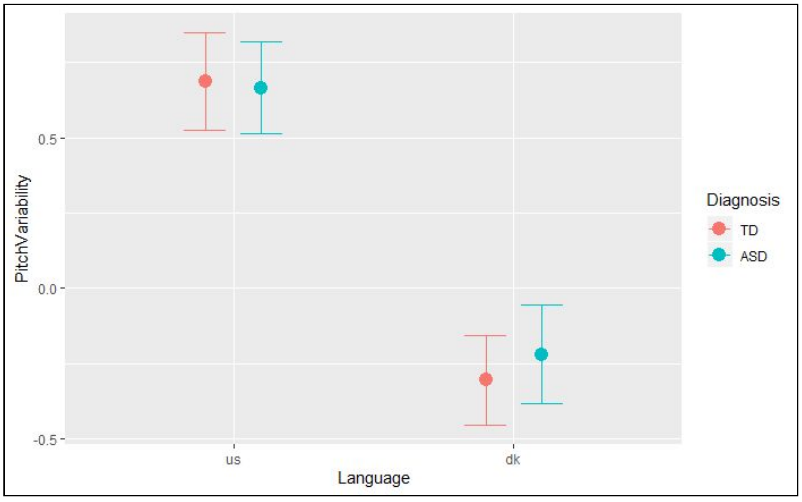


Figure 4: Effects of Model 2

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.

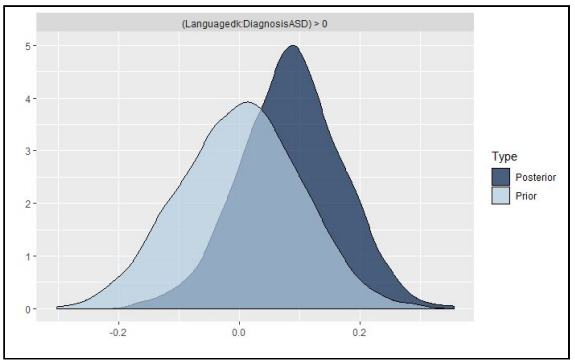


Figure 5: The posterior predictions for betas, with TD as intercept, and only for danish speaking.

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.

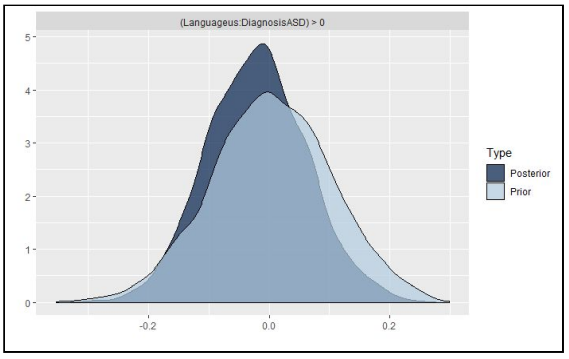


Figure 6: The posterior predictions for betas, with TD as intercept, and only for English American speaking

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.

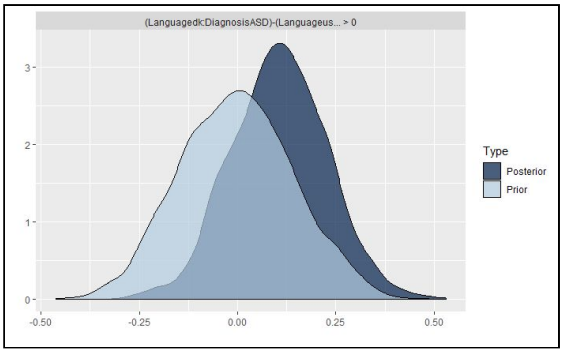


Figure 7: The posterior predictions for betas, with TD as intercept, when we subtract the beta in Danish speaking with the beta in English American speaking

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.

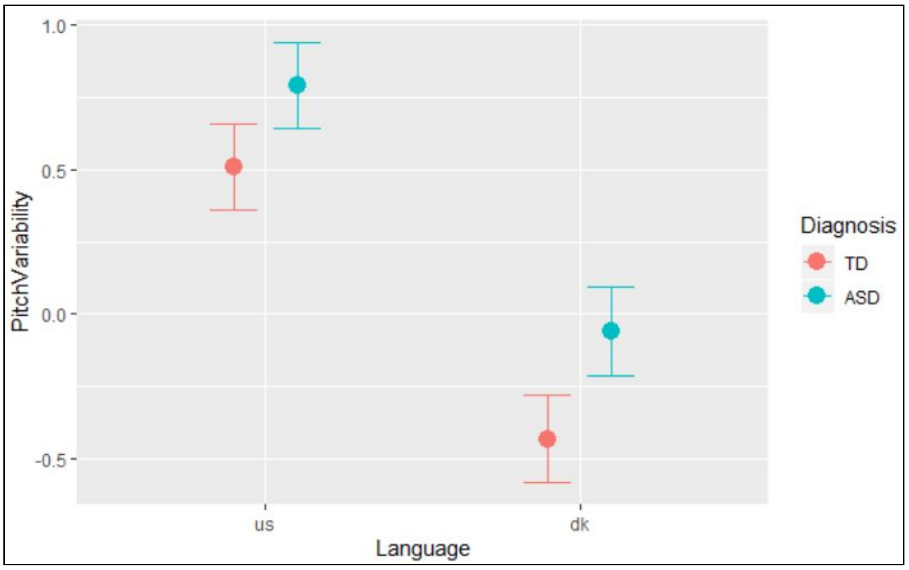


Figure 8: Effects of Model 3

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

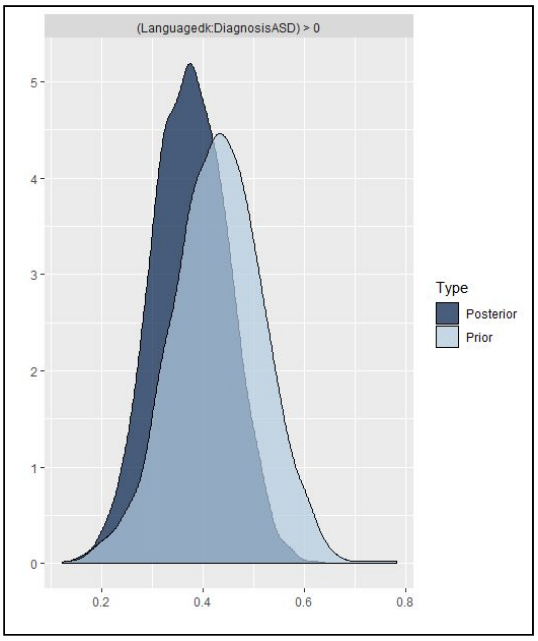


Figure 9: The posterior predictions for betas, with TD as intercept, and only for danish speaking.

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

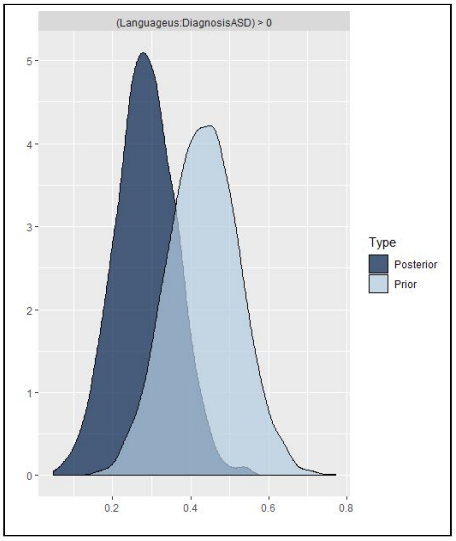


Figure 10: The posterior predictions for betas, with TD as intercept, and only for English American speaking

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.

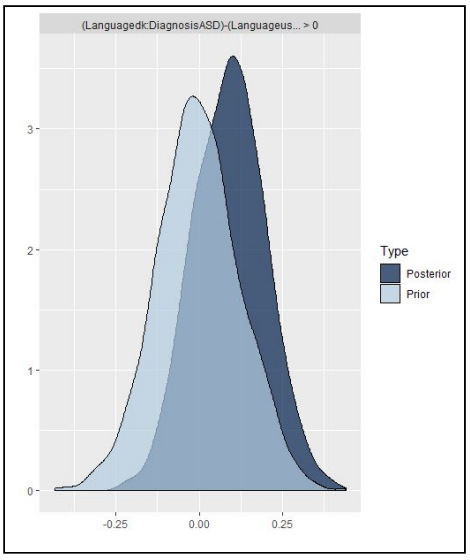


Figure 11: The posterior predictions for betas, with TD as intercept, when we subtract the beta in Danish speaking with the beta in English American speaking

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/.

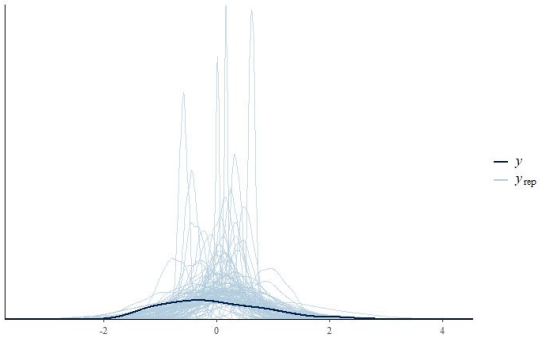
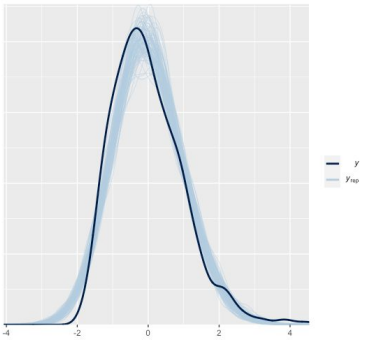
Paul-Christian Bürkner (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. Journal of Statistical Software, 80(1), 1-28. doi:10.18637/jss.v080.i01

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

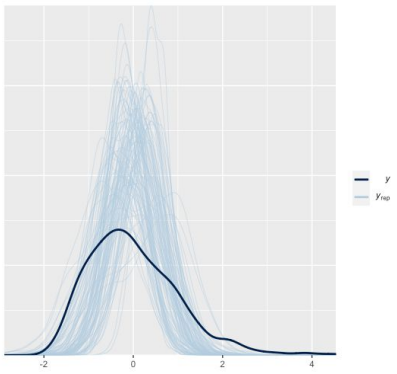
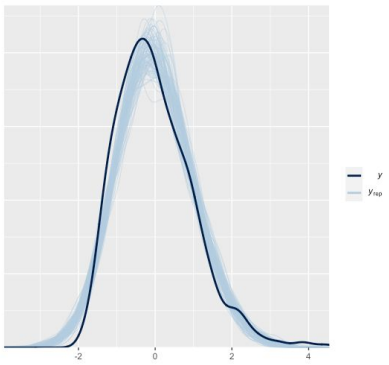
Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. Journal of Statistical Software, 36(3), 1-48. URL: http://www.jstatsoft.org/v36/i03/

Appendix 1

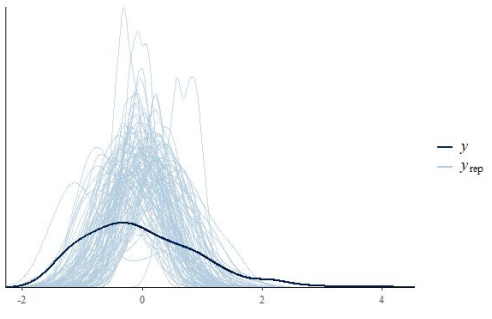
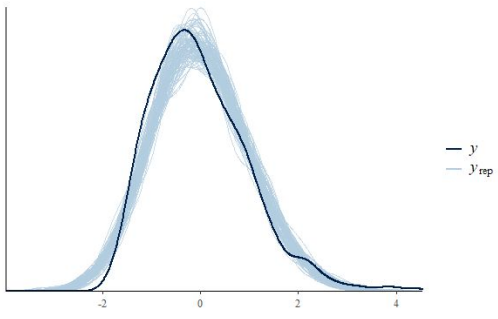
Model 1

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

Model 2

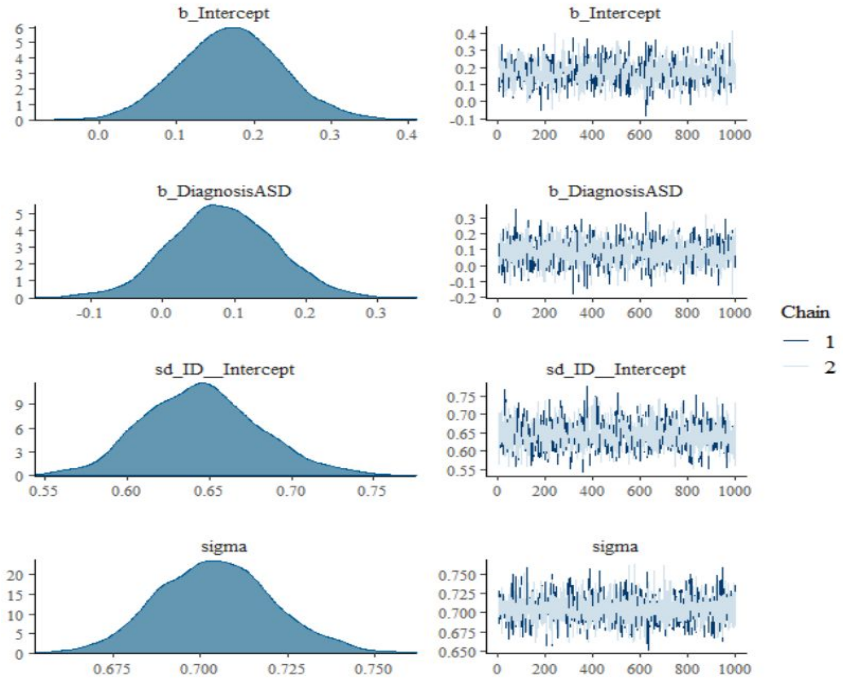
	
Prior predictive plot for model 2	*Posterior predictive check for model 2*

Model 3

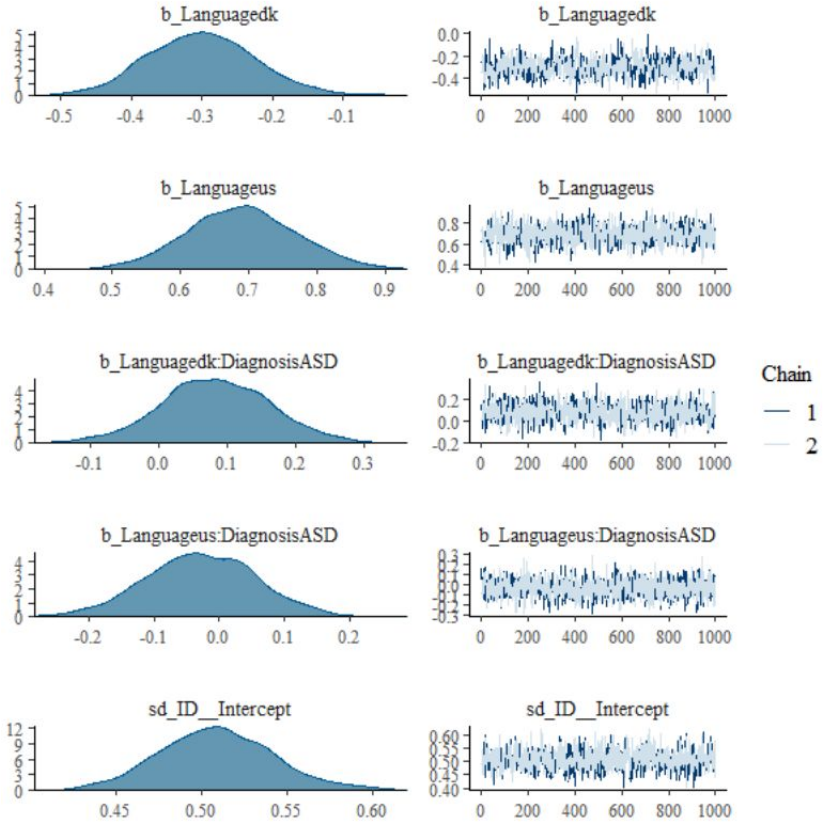
	
Prior predictive plot for model 3	*Posterior predictive plot for model 3*

Appendix 2

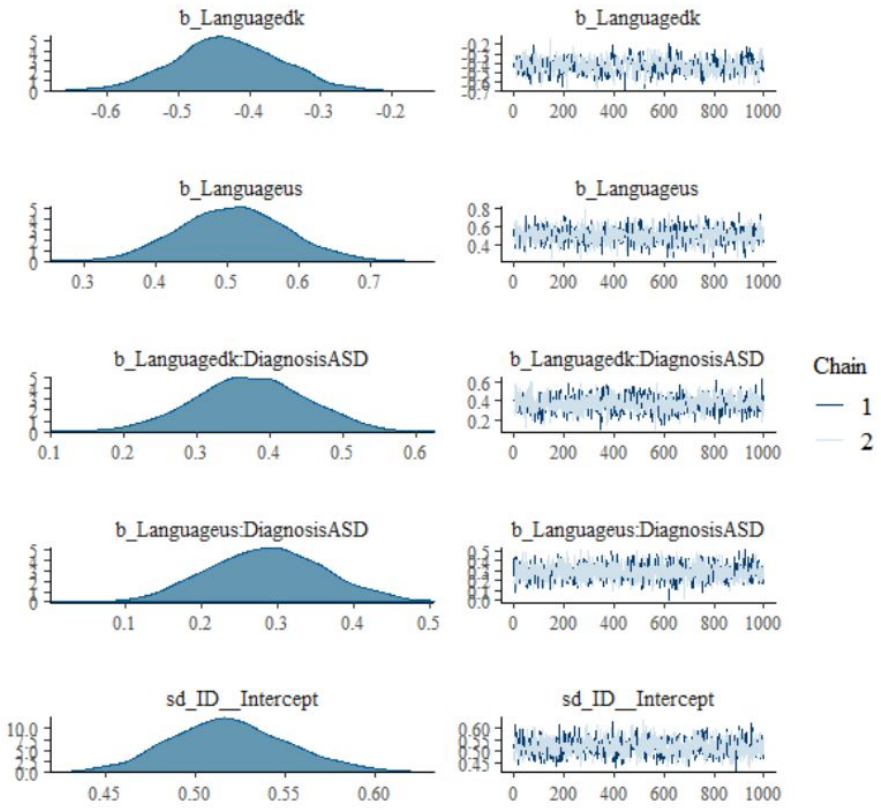
Model 1



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

Computational Modelling, Assignment 5 18/5-2020
Jacobsen, Matilde & Hansen, Eva Sahlholdt
Character count: 18300

Has social conformity increased during COVID-19?

Github: https://github.com/eveandmad/SocKult-Exam.git

Materials and methods

Meta-analysis

Literature search
The data was aggregated manually by 1) searching for relevant keywords and authors informed by advisors on Google Scholar, 2) screening citations in bibliographies of relevant papers, and 3) searching for papers citing relevant papers on Google Scholar (see search terms and platforms in SM table 1). The search was conducted between April 27 2020 and May 11 2020.
Study selection was based on certain inclusion criteria: 1) Use of the specific paradigm for estimating social conformity as described in the following section *Procedure (The conformity task)*; 2) A healthy participant sample.

Data extraction
Data extracted from studies included healthy sample size, study and sample origin, central demographic variables (age, gender, years of education), duration of distractor task, initial and second rating, group rating, and various estimates of change in rating between round 1 and round 2 (rating change) predicted by feedback. Those included hierarchical divisions of rating changes, i.e. it appeared to be common to divide the type of group rating compared to initial rating into three feedback groups (whether the group opinion were lower, i.e. negative feedback: –3 or –2; whether the group opinion were similar, i.e. equal feedback: –1, 0 or 1; whether the group opinion were higher, i.e. positive feedback: 2 or 3). One study reported "raw conformity change", i.e. a mean beta estimate for the difference between rating 1 and 2 as predicted by feedback (see SM table 2 for estimates used in the analysis). Where certain estimates were missing, contact was established to authors in order to access individual-level data.

Study selection and specifications
In total, 11 articles and one unpublished data collection fulfilling the inclusion criteria were assessed. Of those, seven articles provided the estimates necessary for analysis, i.e. estimates of rating change according to feedback. The authors of the remaining four articles and the unpublished data collection were contacted in order to obtain the missing estimates. Of those, the full datasets were provided for one article and the unpublished data collection. Two authors responded that due to the COVID-19 situation, the data could not be accessed. One author didn't respond. Thus, three articles were excluded, resulting in a total of nine papers included in the analysis with an aggregated sample size of $n = 336$.

The participants included had a mean age of 23.6 years (SD = 5.72), an average length of education of 13.43 years (SD = 0.89) and were approximately equally distributed across genders with 54.8 % females. Three studies were sampled in Denmark, two in the United States, one in Russia, one in China, one in the Faroe Islands and one in the Netherlands. All studies were published between 2011 and 2019. Despite all studies applying the same paradigm for testing social conformity, different tasks were used. Thus, four studies rated trustworthiness in faces, four studies rated attractiveness in faces, and one study rated preferences for food. All ratings were conducted on a scale from 1-8 except for one study that provided a scale from 1-7. All studies applied between 150 and 222 stimuli for rating with a median of 153. Individual study demographics are included in SM table 1.

Peri-COVID-19

Data collection
The data was collected between May 7 2020 and May 13 2020. Various platforms were used to obtain participants, including Facebook and word-to-word sharing amongst friends, family and colleagues. In order to obtain sufficient power, we set a lower boundary of 50 participants initiating the experiment as we

expected 1) some having technical issues in data collection, and 2) a certain drop-out rate for completing the second part of the experiment. In total, we aimed to achieve a sample of 30 participants for analysis.

Participants
Before experiment initiation, participants reported age, gender, years of education, country of upbringing, country of residence and currently medically prescribed psychiatric diagnoses. Those variables were collected in order to assess possible bias due to sample variability and increase comparability with the sample included in the meta-analysis. Overall, 66 subjects completed round 1. Of those, 42 subjects completed round 2. Subjects only completing round 1 were excluded from analysis (n = 22). Additionally, two participants were excluded due to technical issues, and two subjects were excluded due to reporting medically prescribed psychiatric diagnoses. In total, 38 participants were included in the analysis. Demography of participants included follows in table 1.

Demographic variables	**Mean (SD)**	**n**	**Percentage**
Age	26.9 (7.46)	-	-
Female : Male	-	24 : 14	63 % : 37 %
Years of education	16.1 (2.2)	-	-
Country of upbringing: Denmark : Not Denmark	-	34 : 4	89 % : 11 %
Country of residence: Denmark : Not Denmark	-	32 : 6	84 % : 16 %

Table 1*: Demographics of participants included in the study as collected from Google Survey*

Procedure
General procedure
Before initiating the experiment, participants were primed to believe the experiment concerned interhuman trust relations during the COVID-19 pandemic to ensure that participants remained naïve of the actual purpose of the experiment. Further, they reported demographic variables as described and signed up to an automatic mail-service to get notification on the second part of the study. In order to match participants correctly, each participant assigned themselves an self-chosen identification term and inserted the URL from the experiment page into the survey. To partake in the second round of the experiment, each participant received an email one hour after initiation with a link to the second round. Here, each participant again reported their identification term and inserted the new URL to enable matching.

The conformity task
In the experiment, participants were instructed to rate 80 facial images individually according to how trustworthy they perceived them to be on a 1-8 scale (1: not trustworthy at all; 8: very trustworthy). After most images, participants were reminded of their own rating and viewed the average rating of previous participants (figure 1). All group ratings were –3, –2, 0, 2 or 3 points from the initial rating of the participant. The group opinion didn't show for some images in order to enable control of regression to the mean. To complete the second round of the experiment, participants received the experiment link after one hour and were instructed to initiate it within 24 hours. The waiting period of at least one hour ensured that subjects were distracted from the task long enough to not explicitly recollect their own ratings, and the upper limit of 24 hours was set to ensure that the period was short enough to expect an effect of social conformity (Huang et al., 2014) and to increase comparability with the meta-analytic estimates. Across participants included in the analysis, there was a mean waiting period of 11.54 hours (SD = 11.42).

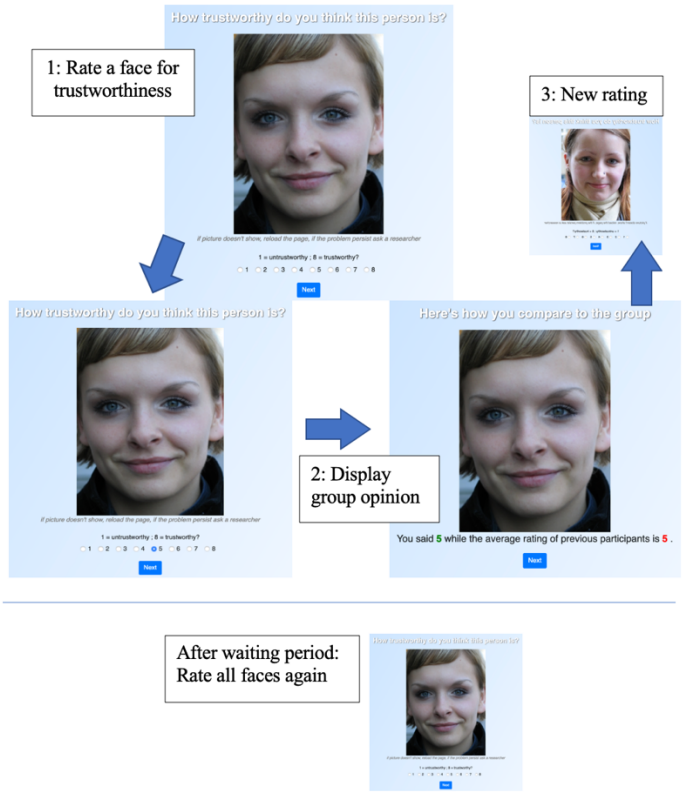


***Figure 1**: Social Conformity task paradigm as adopted from previous studies (Campbell-Meiklejohn et al., 2012; Klucharev et al., 2009; Simonsen et al., 2019). Participants rated 80 female faces for trustworthiness by choosing one number on a scale from 1-8. When proceeding to the next page the participant's rating was highlighted with green color while the group rating was highlighted with red color. Participants unexpectedly had to rate the faces again after at least 1 hour waiting, this time without getting the feedback of group rating.*

Materials and stimuli

The experiment consisted of two OTree scrips (Chen, Schonger & Wickens, 2016) obtained from a previous study (Vermillet, 2020). The scripts were modified to match an online at-home setting. The scripts for both rounds consisted of 80 images of female Caucasian faces from the shoulders up shown in a random order. Only females were used in order to reduce gender-related differences in ratings (Cloutier et al., 2008).
Group feedback was given based on a pseudo-randomization algorithm adopted from previous paradigms. The algorithm ensured that approximately one third of the group ratings matched participant rating (feedback = 0), one third of the ratings were above participant rating (feedback = +2 or +3) and one third of the ratings were below participant rating (feedback = –2 or –3).
Demographics and identification terms were collected using Google Forms. To send the automated emails with links to round 2, the online service Mailchimp was used.

Statistical methods

The entire analysis was conducted in R Studio (R Core Team, 2019).

Meta-analysis

A Bayesian Meta-Analysis was conducted using the package "brms" (Bürkner, 2017) with procedures described in the literature (Harrer et al., 2019). To estimate differences in rating change, mean estimates of change according to feedback were extracted. As rating scales were equal across studies and group feedback were, too, assigned in a similar manner, no standardization procedure was applied. For studies reporting estimates of change in relation to type of feedback (negative, equal or positive), estimates were assigned as rating change according to feedback type -2.5, 0 or 2.5. For the study reporting raw conformity score, an estimate of rating change for a feedback of 1 was modelled.
To assess the pooled effect size we used a random effects model relying on a Gaussian likelihood. The outcome variable consisted of estimated rating change and the standard error of these estimates as predicted

by feedback. A random intercept for feedback and a random slope for each article was included as we expected the effect to vary between studies (see formula 1).

Change | se(Standard Deviation of Change) ~ 1 + Feedback + (1 + Feedback | Article)

Formula 1: *Model formula for meta-analysis*

We used weakly informative priors for our model estimates. We expected rating change to lie between -1 and 1 and thus defined a normally distributed prior with a mean of 0 and variance of 0.5 ($\mu \sim N(0,0.5)$). For between study variance, we expected a positive deviation with high density close to zero while not constraining the probability of larger values. Thus, we defined it as a Half Cauchy distributed prior with a mean of 0 and variance of 0.3 ($\tau \sim HC(0,0.3)$) (Williams et al., 2018). We expected the slope for change according to feedback to lie between -0.5 and 0.5 (as observed in the literature included in the Meta-Analysis) and thus defined a normally distributed prior for the beta with a mean of 0 and variance of 0.25 ($\beta \sim N(0,0.25)$). Finally, we defined a prior for the correlation within varying effects to be an LKJ distribution with $\eta = 5$ in order to constrain the plausibility of a correlation of 1 ($r = LKJ(5)$).

To assess model quality we performed prior and posterior predictive checks (SM figure 1). Additionally, we ensured model convergence by looking at trace plots and ranked Markov chains (SM figure 2 and 3), and assessing that Rhat estimates were smaller than 1.05 and that effective sample sizes for both bulk and tail were larger than 200 (McElreath, 2020). Additionally, we performed a hypothesis test of the results as we expected a conformity effect larger than zero across studies. This test was based on the evidence ratio (the amount of evidence supporting the hypothesis compared to evidence against the hypothesis). An evidence ratio above 3 was interpreted as substantial evidence for the alternative hypothesis, whereas a ratio below ⅓ was interpreted as substantial evidence for the null-hypothesis.

Regression to the mean

To estimate the hidden impact of regression to the mean, the method of control in the respective analyses of the studies included in the meta-analysis was assessed. Most studies did not report proper control for this confound, suggesting inflation of the meta-analytic pooled effect size. To estimate the underlying true effect, further analysis was conducted. First, an estimate of approximate regression to the mean was computed with linear regression predicting second rating from first rating (formula 2) from one meta-analytic study where we had access to the required variables (Unpublished in-class experiment 2020). Where the model failed to converge, the varying structure was simplified.

SecondRating ~ 1 + FirstRating + (1 + FirstRating | ID)

Formula 2*: Model used to estimate regression to the mean*

The approximate estimate for regression to the mean was inserted as a fixed effect in a simulation based on the social conformity paradigm, built to reveal the true conformity effect when different levels of regression to the mean was inferred (Fusaroli, 2020). A simulated sample size of 300 participants was used to run 1000 simulations.

Comparative analysis: Peri-COVID-19

To test the hypothesis that private social conformity increased during the COVID-19 pandemic a Bayesian multilevel interaction model relying on a Gaussian likelihood was built using the package "brms" (Bürkner, 2017). To control for regression to the mean, a subset of data from the meta analysis with sufficient information provided was used for analysis (including Simonsen et al. 2019 and Unpublished in-class experiment 2020). In total, 121 subjects were included in this analysis.

Rating change as predicted by first rating, feedback and condition (whether the data was obtained pre-pandemic outbreak or peri-pandemic outbreak) was modelled, including first rating and feedback varying by subject and facial stimuli (formula 4).

Change ~ 0 + FirstRating:Condition + Feedback:Condition + (1 + FirstRating + Feedback | ID) + (1 + FirstRating + Feedback | FaceID))

Formula 3*: Model formula for comparative analysis interacting by condition*

We used weakly informative priors for our model estimates. We expected the effect of rating change according to feedback for both conditions to lie between -0.5 and 0.5 and thus defined a normally distributed prior for the beta of feedback with a mean of 0 and variance of 0.3 ($\beta \sim N(0,0.3)$). For the slope of change according to first rating we expected a slightly bigger variation, as regression to the mean likely has influence in this effect and thus defined a normally distributed prior with a mean of 0 and variance of 0.5. For the varying effects of the model we defined a similarly broad prior with a normally distributed mean of 0 and variance of 0.3 as we did not have much prior knowledge about the estimate ($\sigma \sim N(0,0.3)$). Finally, we defined a prior for the correlation within varying effects to be an LKJ distribution with $\eta = 5$ in order to constrain the plausibility of a correlation of 1 ($r = LKJ(5)$).

To assess model quality, prior and posterior predictive checks were performed (SM figure 5). Model convergence was ensured following the same procedures as described above (SM figure 6 and 7). A hypothesis test relying on evidence ratio was conducted to test whether the effect of feedback on social conformity was bigger in the peri-COVID-19 than pre-COVID-19 condition.

Additionally, the new data was included in the meta-analysis to assess heterogeneity from meta-analytic studies. Importantly, neither the meta-analytic nor the new data was corrected for regression to the mean. Estimates of mean rating change according to type of feedback of the new data collection is included in SM table 2 and a new forest plot (SM figure 8) was performed to assess the difference.

Results

Meta-analysis

The Bayesian meta-analysis model revealed a significant effect of feedback on rating change, (0.17, 95 % CIs: 0.11, 0.21). Additionally, evidence testing of the hypothesis that rating change as predicted by feedback is above zero proved highly evident, (ER = 5999, 95 % CIs: 0.13, 0.2), see figure 2. A forest plot (SM figure 4) was generated to assess the heterogeneity between studies. In general, low heterogeneity was found (τ = 0.03, 95% CIs: 0.00, 0.09), likely as a result of highly similar methods and population samples.

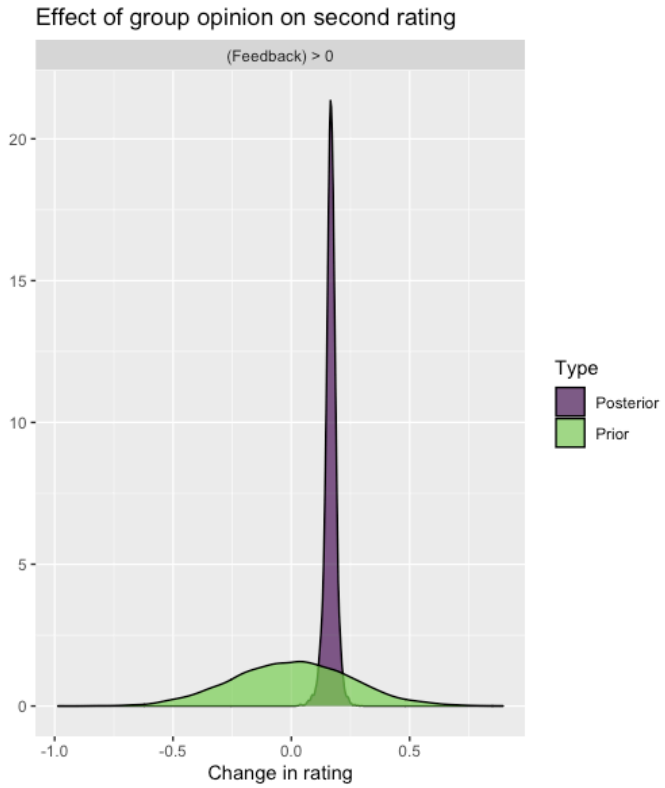


***Figure 2**: Evidence ratio indicating a great deal of certainty that there is a rating change > 0*

Regression to the mean
The regression model (formula 2) revealed a significant effect of first rating as a predictor of second rating ($\beta = 0.666$, SE = 0.033, $t = 19.95$, $p < 0.001$). Thus, a fixed conformity estimate of 0.66 was included in the simulation. With this estimate, a fixed true conformity effect of -0.17 was necessary to obtain an estimated conformity effect of 0.17 as found in the meta-analysis (see SM table 4 for an overview of simulated values). Inversely, when controlling for regression to the mean (following SM formula 1) on the subset of data used for the regression analysis, a true conformity effect of 0.02 was obtained (ER > 3, 95% CI's: 0,0.04), identical with controlled results from other studies (Simonsen et al., 2019). Based on the simulation, this effect should show an estimated conformity of 0.52. As this is not the case in the meta-analysis as visualized in the forest plot (SM figure 4, Unpublished in-class experiment 2020), a decreased regression to the mean estimate, other confounds or masking effects might also play a role in the pooled meta-analytical effect size.

Comparative analysis: Peri-COVID-19

The Bayesian multilevel model revealed similarly small main effect of feedback for both conditions (pre-COVID-19: $\beta = 0.03$, 95% CI's = 0.01,0.05, peri-COVID-19: $\beta = 0.03$, 95% CI's = 0.00, 0.06) as well as similarly negative effects of initial ratings for both conditions (peri-COVID-19: $\beta = -0.19$, 95% CI's = -0.26, -0.13, pre-COVID-19: $\beta = -0.20$, 95% CI's = -0.26, -0.14) on change of trustworthiness ratings. When testing the hypothesis that a conformity effect above zero existed in both conditions, an evidence ratio above 3 provided substantial evidence of the alternative hypothesis. For the hypothesis that the effects of feedback has increased peri-COVID-19 compared to the pre-COVID-19 effect, no substantial evidence was found ($\beta = 0$, 95% CI's = -0.02, 0.03, ER = 1.63). See full details in table 6.

When not correcting for regression to the mean, means and variance of raw change according to feedback of the peri-COVID-data was highly compatible with those obtained for the pre-COVID-data included in the comparative model (figure 3).

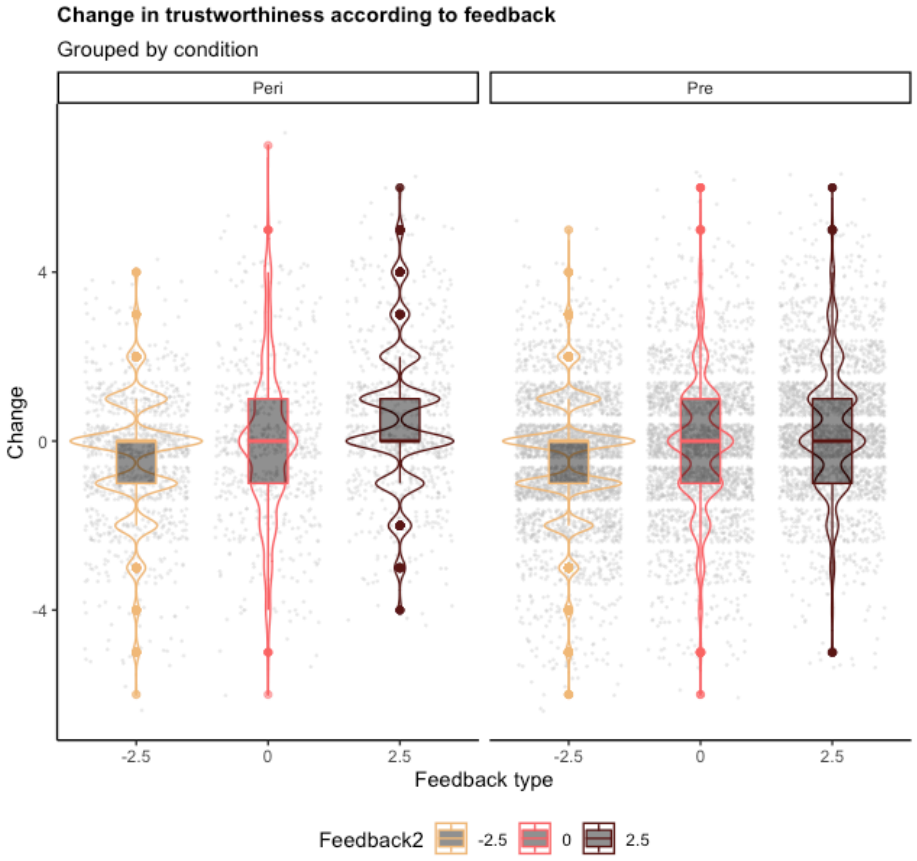


***Figure 3**: A violin distribution of the effects of change divided into different groups of feedback.*

References

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Campbell-Meiklejohn, D. K., Simonsen, A., Jensen, M., Wohlert, V., Gjerløff, T., Scheel-Kruger, J., Møller, A., Frith, C. D., & Roepstorff, A. (2012). Modulation of Social Influence by Methylphenidate. *Neuropsychopharmacology*, *37*(6), 1517–1525. https://doi.org/10.1038/npp.2011.337

Cloutier, J., Heatherton, T. F., Whalen, P. J., & Kelley, W. M. (2008). Are Attractive People Rewarding? Sex Differences in the Neural Substrates of Facial Attractiveness. *Journal of Cognitive Neuroscience*, *20*(6), 941–951. https://doi.org/10.1162/jocn.2008.20062

Harrer, M., Cuijpers, P., Furukawa, T. A., & Ebert, D. D. (2019). *Doing Meta-Analysis in R: A Hands-on Guide*. PROTECT Lab. https://bookdown.org/MathiasHarrer/Doing_Meta_Analysis_in_R/

Huang, Y., Kendrick, K. M., & Yu, R. (2014). Conformity to the Opinions of Other People Lasts for No More Than 3 Days. *Psychological Science*, *25*(7), 1388–1393. https://doi.org/10.1177/0956797614532104

Klucharev, V., Hytönen, K., Rijpkema, M., Smidts, A., & Fernández, G. (2009). Reinforcement Learning Signal Predicts Social Conformity. *Neuron*, *61*(1), 140–151. https://doi.org/10.1016/j.neuron.2008.11.027

McElreath, R. (2020). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan* (2nd ed.). CRC Press.

Simonsen, A., Fusaroli, R., Skewes, J. C., Roepstorff, A., Mors, O., Bliksted, V., & Campbell-Meiklejohn, D. (2019). Socially Learned Attitude Change is not reduced in Medicated Patients with Schizophrenia. *Scientific Reports*, *9*(1), 992. https://doi.org/10.1038/s41598-018-37250-x

Williams, D. R., Rast, P., & Bürkner, P.-C. (2018). *Bayesian Meta-Analysis with Weakly Informative Prior Distributions* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/7tbrm

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

Chen, D.L., Schonger, M., Wickens, C., 2016. oTree - An open-source platform for laboratory, online and field experiments. Journal of Behavioral and Experimental Finance, vol 9: 88-97

Vermillet, 2020: Did not respond in reference to assess citation. Contact: arnault@cc.au.dk

Fusaroli, 2020. Code for simulation adapted from script.

Supplementary Materials

Meta-analysis

Paper	Simonsen et al. (2014)	Campbell-Meiklejohn et al. (2012)	Zhao et al. (2016)	Klucharev et al. (2011)	Simonsen et al. (2019)	Unpublished in-class experiment (2020)	Shestakova et al (2012)	Nook and Zaki (2015)	Zaki et al. (2011)
Title	Serotoninergic effects on judgments and social learning of trustworthiness	Modulation of Social Influence by Methylphenidate	Investigating the Genetic Basis of Social Conformity: The Role of the Dopamine Receptor 3 (DRD3) Gene	Downregulation of the Posterior Medial Frontal Cortex Prevents Social Conformity	Socially Learned Attitude Change is not reduced in Medicated patients with schizophrenia	-	Electrophysiological precursors of social conformity	Social Norms Shift Behavioral and Neural Responses to Foods	Social Influence Modulates the Neural Computation of Value
Authors	A. Simonsen & J. Scheel-Krüger & M. Jensen & A. Roepstorff & A. Møller & C. D. Frith & D. Campbell-Meiklejohn	D. Campbell-Meiklejohn & A. Simonsen & M. Jensen & V. Wohlert & T. Gjerløff & J. Scheel-Kruger & A. Møller & C. Frith & A. Roepstorff	C. Zhao, J. Liuf, P. Gongh J. Hua, X. Zhoua	V. Klucharev, M. A. M. Munneke, A. Smidts, G. Fernández	A. Simonsen, R. Fusaroli, J.C. Skewes, A. Roepstorff, O. Mors, V. Bliksted, D. Campbell-Meiklejohn	-	A. Shestakova, J. Rieskamp, S. Tugin, A. Ossadtchi, J. Krutitskaya, V. Klucharev	E.C. Nook & J. Zaki	J. Zaki, J. Schirmer, and J.P. Mitchell
Search specification	Google Scholar	Followed citation from Simonsen et al 2019	Google Scholar	Google Scholar	Professor advice	Professor advice	Google Scholar	Followed citation from Wu et al. (2016)	Followed citation from Wu et al. (2016)
Search term	"social conformity task face trustworthiness"	-	"facial trustworthiness"	"group opinion social conformity"	-	-	"Klucharev"	-	-
Area	Denmark	Denmark	China	The Netherlands	Faroe Islands	Denmark	Russia	USA	USA
Sample size (healthy)	20	19	149	15	39	44	15	21	14
Female	20	19	60	15	12	25	15	18	0
Age: Mean (sd)	23.5 (2.5)	23 (2.7)	22.5 -	21.1 -	39.2 (10.6)	22 -	19.9 -	20.1 -	21.8 -
Years of education: Mean (sd)	13.8 (1.9)	13.8 (1.8)	-	-	14.2 (3.1)	-	-	-	-
Specification of education	-	-	University students	-	-	Cognitive Science students	Students	Undergraduates from Stanford	-

Task	Trustworthi-ness	Trustworthi-ness	Attractive-ness	Attractive-ness	Trustwor-thiness	Trustwor-thiness	Attractive-ness	Food rating	Attractive -ness
Rating scale	1-8	1-8	1-8	1-8	1-8	1-8	1-8	1-8	1-7
Number of stimuli	153	153	120	222	153	153	222	150	180
Distraction time (min)	30	30	30	30	60	15	30	5	30

Table 1*: Individual demographic variables of studies included in Meta-Analysis*

Paper	**Change Low: Mean (SD)**	**Change Same: Mean (SD)**	**Change High: Mean (SD)**	**Raw Change: Mean (SD)**	**First Rating: Mean (SD)**	**Second Rating: Mean (SD)**
Simonsen et al. (2014)	-0.39 (0.05)	-	0.46 (0.05)	-	4.87 (0.10)	4.94 (0.14)
Campbell-Meiklejohn et al. (2012)	-0.31 (0.11)	-	0.50 (0.10)	-	4.88 (0.11)	4.93 (0.15)
Klucharev et al. (2011)	-0.40 (0.25)	-0.05 (0.28)	0.28 (0.24)	-	-	-
Simonsen et al. (2019)	-0.41 (1.53)	-0.08 (1.59)	0.46 (1.59)	-	5.0 (1.6)	5.0 (1.6)
Unpublished in-class experiment (2020)	-0.54 (1.46)	-0.22 (1.49)	0.14 (1.41)	-	4.63 (1.76)	4.40 (1.81)
Shestakova et al (2012)	-0.74 (0.42)	-0.18 (0.27)	0.43 (0.32)	-	-	-
Nook and Zaki (2015)	-0.11 -	0.03 -	0.07 -	-	-	-
Zaki et al. (2011)	-0.33 -	-	0.10 -	-	-	-
Zhao et al. (2016)	-	-	-	0.26 (0.095)	-	-

Table 2*: Individual estimates from studies included in Meta-Analysis*

Estimate type	**Estimate**	**Est.Error**	**Q2.5**	**Q97.5**
b_Intercept	0.031	0.043	-0.062	0.115
b_Feedback	0.166	0.024	0.113	0.210
sd_Article__Intercept	0.046	0.043	0.002	0.157
sd_Article__Feedback	0.027	0.025	0.001	0.094

Table 3*: Summary of estimates from meta-analysis*

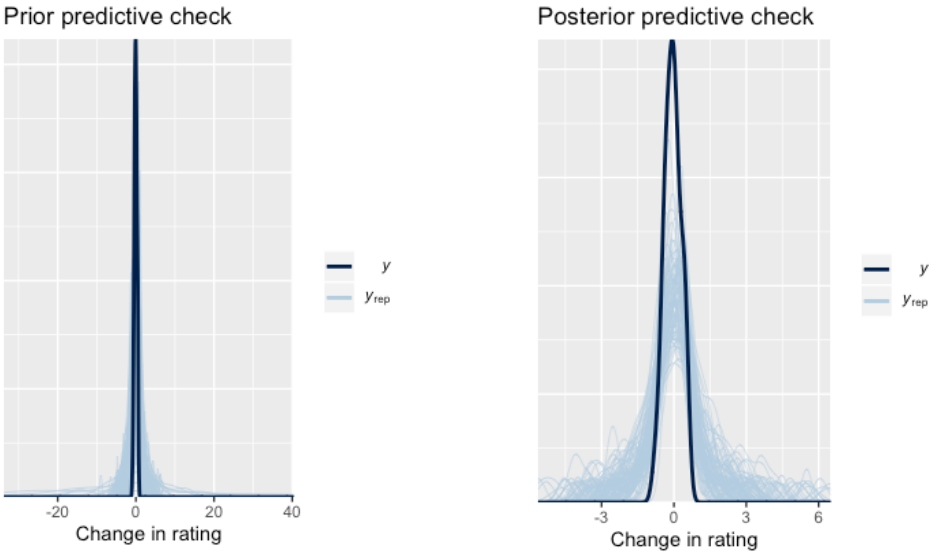


Figure 1*: Prior and posterior predictive checks of meta-analytical model*

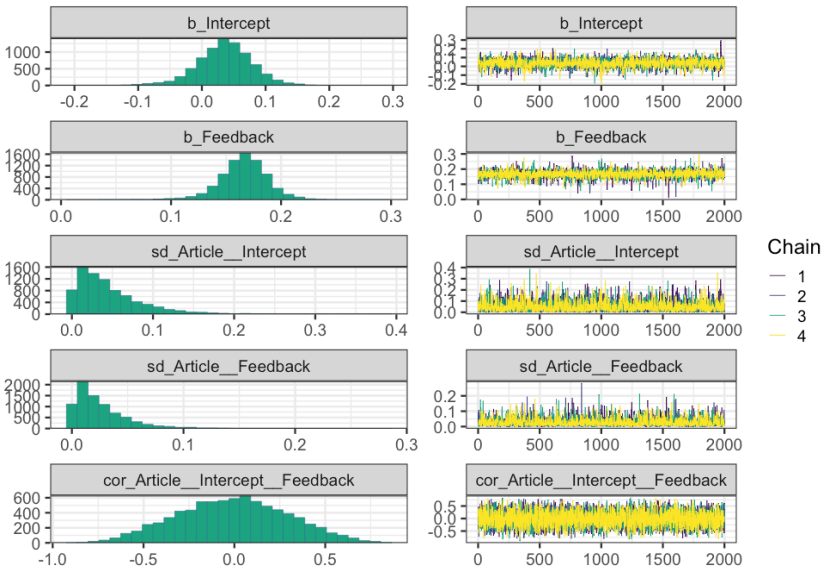


***Figure 2**: Model assessment I: Histogram and trace plots of meta-analytical model estimates*

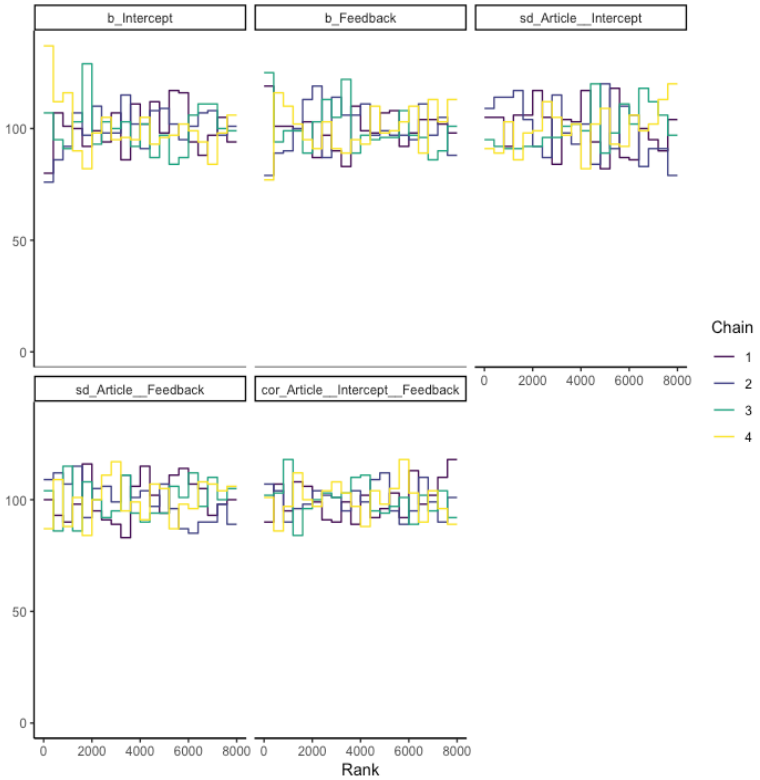


***Figure 3**: Model assessment II: Ranked Markov chains of meta-analytical model convergence*

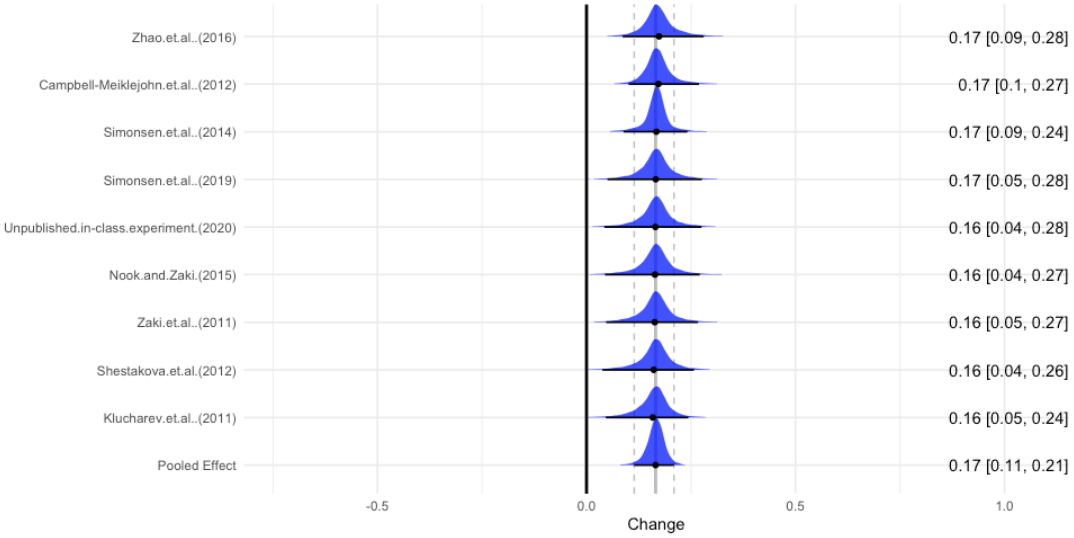


***Figure 4**: Forest plot of meta-analytical effect sizes*

Regression to the mean

True regression to the mean	True conformity	Estimated conformity
0.66	-0.20	0.09
0.66	-0.17	0.17
0.66	-0.10	0.28
0.66	-0.05	0.47
0.66	0.02	0.52
0.52	-0.20	0.26
0.52	-0.17	0.35
0.52	-0.10	0.40
0.52	-0.05	0.44
0.52	0.02	0.55

***Table 4**: Estimates as obtained from a simulation assessing causal mechanisms of the social conformity paradigm. "True regression to the mean" estimates were obtained through a regression analysis of real data, whereas the "True conformity" scores were changed manually in the simulation to get the estimated conformity effect as explained by the simulation (see provided code for more details on the simulation).*

Comparative analysis: Peri-COVID-19

Across participants (n = 38)	Mean (SD)	Number of ratings
First rating	4.90 (1.83)	2971
Second rating	4.89 (1.79)	2971
Change in rating with low group rating (feedback = -2, -3)	-0.412 (1.49)	1090
Change in rating with same group rating (feedback = 0)	-0.076 (1.78)	668
Change in rating with high group rating (feedback = +2, +3)	0.52 (1.74)	840

***Table 5**: Means of ratings and means of change according to feedback with count estimates*

Estimate type	Estimate	Est.Error	Q2.5	Q97.5
b_FirstRating:ConditionPeri	-0.194	0.033	-0.262	-0.130
b_FirstRating:ConditionPre	-0.199	0.030	-0.260	-0.141
b_ConditionPeri:Feedback	0.031	0.015	0.003	0.061
b_ConditionPre:Feedback	0.027	0.010	0.009	0.048
sd_FaceID Intercept	0.540	0.059	0.431	0.661
sd_FaceID FirstRating	0.072	0.012	0.049	0.096
sd_FaceID Feedback	0.016	0.010	0.001	0.037
sd_ID Intercept	2.162	0.109	1.953	2.377
sd_ID FirstRating	0.320	0.027	0.268	0.374
sd_ID Feedback	0.013	0.009	0.001	0.034

***Table 6**: Summary of estimates for interaction model*

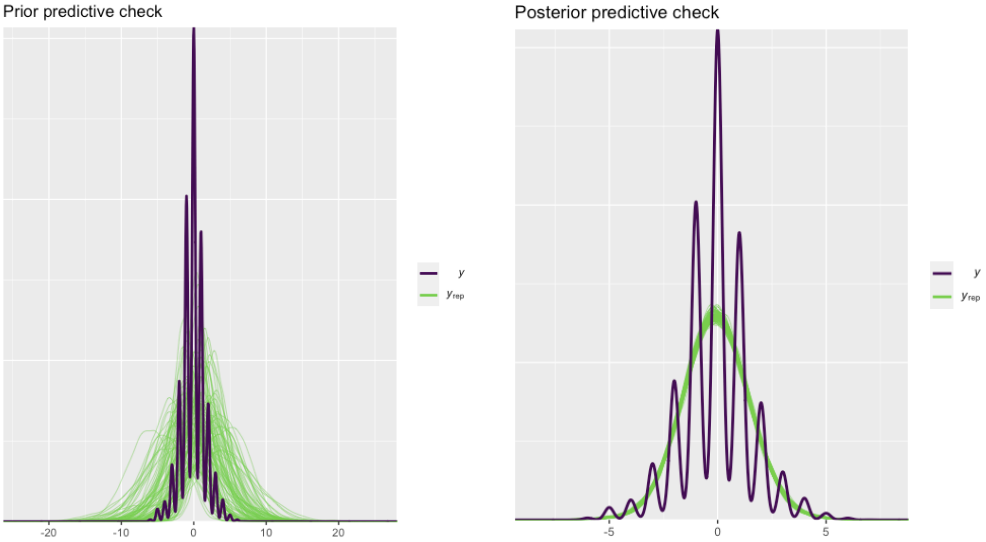


***Figure 5**: Prior and posterior predictive check of model interacting with condition*

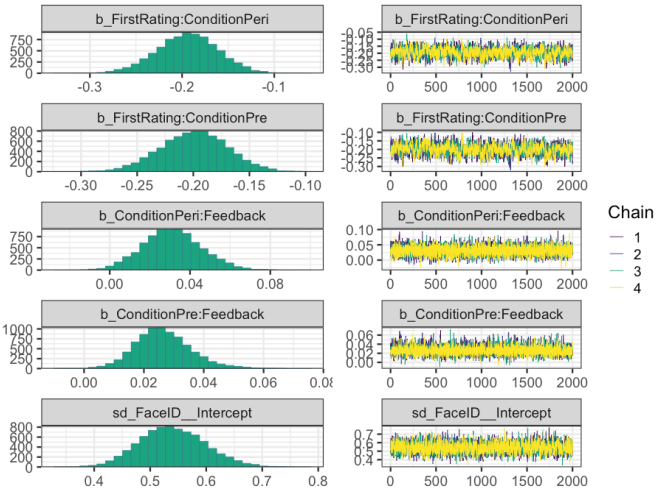


***Figure 6**: Model assessment 1: Histogram and trace plots for estimates of interaction model*

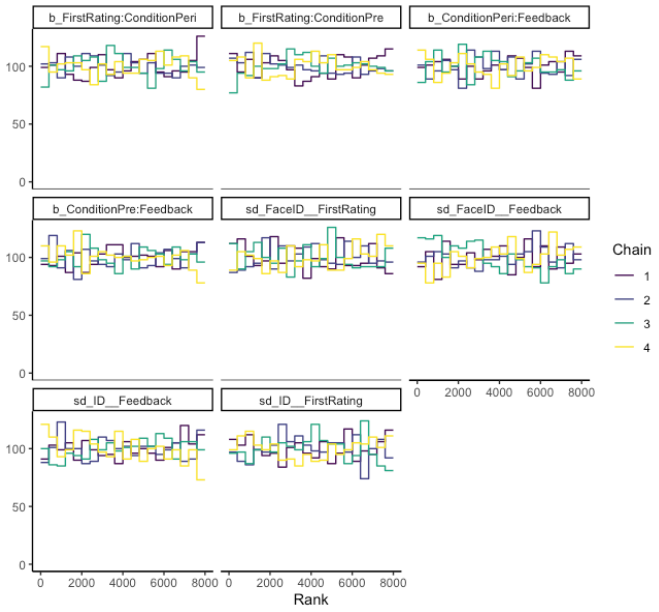


***Figure 7**: Model assessment II: Ranked Markov chains of model convergence of interaction model*

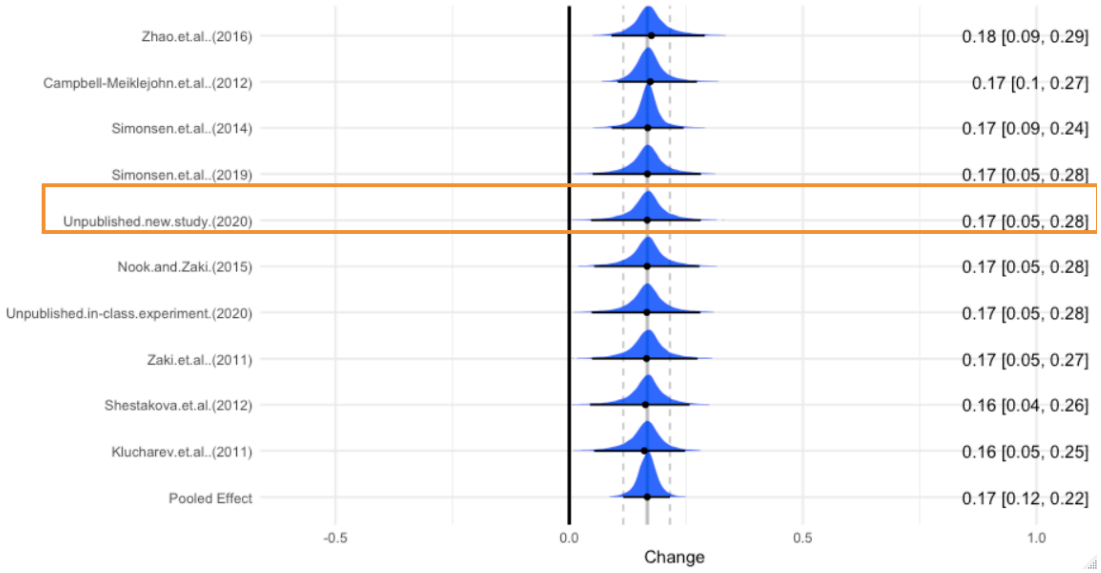


***Figure 8**: Forest plot including own study*

Individual analysis: Peri-COVID-19

A Bayesian multilevel model was built using the package "brms" (Bürkner, 2017) based on procedures described in the literature (McElreath, 2020). To estimate differences in change from rating 1 to 2, a subset of the data-set collected in the new study was assessed according to estimates of change where participants received a low feedback (-2 or -3) or high feedback (2 or 3) reducing the amount of rated images to 1930. This was done to avoid that the effect potentially would be drawn closer to zero due to an invisible conformity effect in the feedback type where participants received the same group rating as their own (e.g. if first rating and belonging group rating was low for a particular case, a similar second rating that would have been high if not for conformity was to be observed as regression to the mean rather than conformity and thus shrinking the effect rather than increasing it).

The following random effects model with a random slopes for each participant and stimuli was used:

Change ~ 1 + FirstRating + Feedback + (1 + FirstRating + Feedback | ID) + (1 + FirstRating + Feedback | FaceID))

Formula 1*: model formula for own study of peri-COVID-19 effect of private social conformity*

We used weakly informative priors for our model estimates. We expected rating change to lie between -2 and 2 and thus defined a normally distributed prior with a mean of 0 and variance of 1 ($\mu \sim N(0,1)$). We expected the slope for change according to feedback to lie between -0.5 and 0.5 (as observed in the literature included in the Meta-Analysis) and thus defined a normally distributed prior for the betas with a mean of 0 and variance of 0.25 ($\beta \sim N(0,0.25)$). For the varying effects of the model we defined a similarly broad prior with a normally distributed mean of 0 and variance of 0.3 as we did not have much prior knowledge about the estimate ($\sigma \sim N(0,0.3)$). Finally, we defined a prior for the correlation within varying effects to be an LKJ distribution with $\eta = 5$ in order to constrain the plausibility of a correlation of 1 ($r = LKJ(5)$).

To assess model quality we performed prior and posterior predictive checks (see SM figure 9). Additionally, we ensured model convergence by looking at trace plots, unranked and ranked (SM figure 10 and 11) and assessing that Rhat estimates were smaller than 1.01 and effective sample sizes for both bulk and tail were smaller than 200 (McElreath, 2020). Individual level estimates of participants were explored to ensure homogeneity within the data (SM figure 12). Additionally, we tested whether the effect was larger than chance (Feedback > 0) based on the evidence ratio.
To assess that the experiment worked as expected means and standard deviations of ratings were summarized across participants revealing a highly similar mean from first to second rating. Means and standard deviations of the raw change in rating were divided into three categories according to type of feedback (low, same, high) and summarized, revealing a general tendency to go down in rating when feedback was low and go up in rating when feedback was high. This tendency is most likely to be an effect of regression to the mean, though. The numbers of stimuli with certain feedback type as generated by the algorithm revealed a slight overweight in negative feedback, but still balanced enough to deem the automated feedback generation successful (see table 4).
As expected a small effect of private social conformity was observed based on the main effect of feedback after controlling for regression to the mean ($\beta = 0.02$, 95% CI's = 0 - 0.05, ER = 15.23). See full details of the model in table 7.

Estimate type	Estimate	Est.Error	Q2.5	Q97.5
b_Intercept	2.985	0.197	2.605	3.378
b_FirstRating	-0.615	0.032	-0.678	0.552
b_Feedback	0.019	0.014	-0.008	0.046
sd_FaceID__Intercept	0.440	0.059	0.325	0.561
sd_FaceID__FirstRating	0.027	0.018	0.001	0.066
sd_FaceID__Feedback	0.021	0.015	0.001	0.055
sd_ID__Intercept	0.955	0.105	0.764	1.174
sd_ID__FirstRating	0.150	0.023	0.109	0.198
sd ID Feedback	0.021	0.015	0.001	0.055

Table 7*: Summary of estimates from individual analysis of new study*

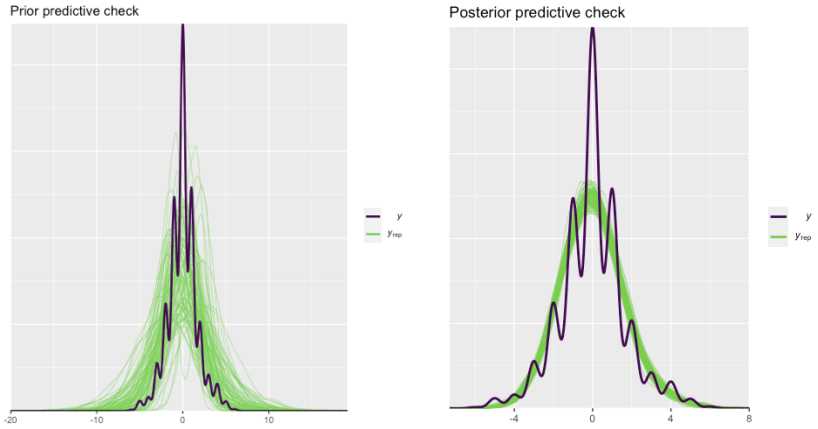


***Figure 9**: Prior and posterior predictive check of peri-COVID19 study on high and low group feedback*

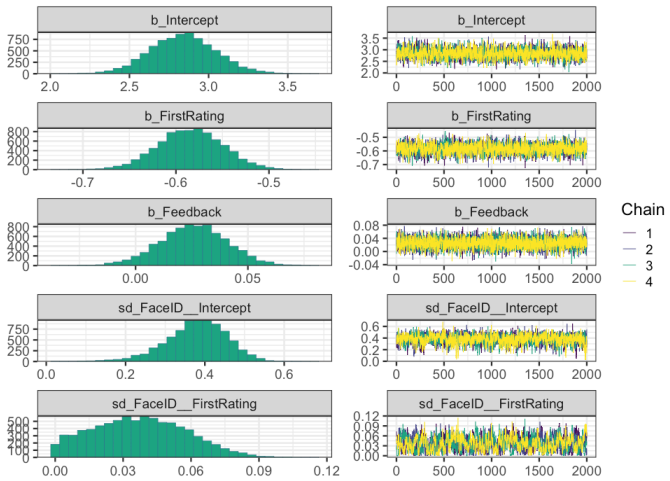


***Figure 10**: Model assessment 1: Histogram and trace plots for estimates of peri-COVID19 study on high and low group feedback*

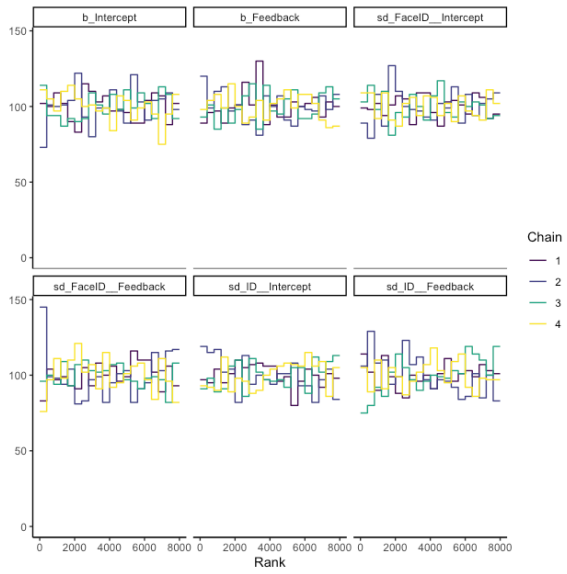


***Figure 11**: Model assessment II: Ranked plots of model convergence of peri-COVID19 study on high and low group feedback*

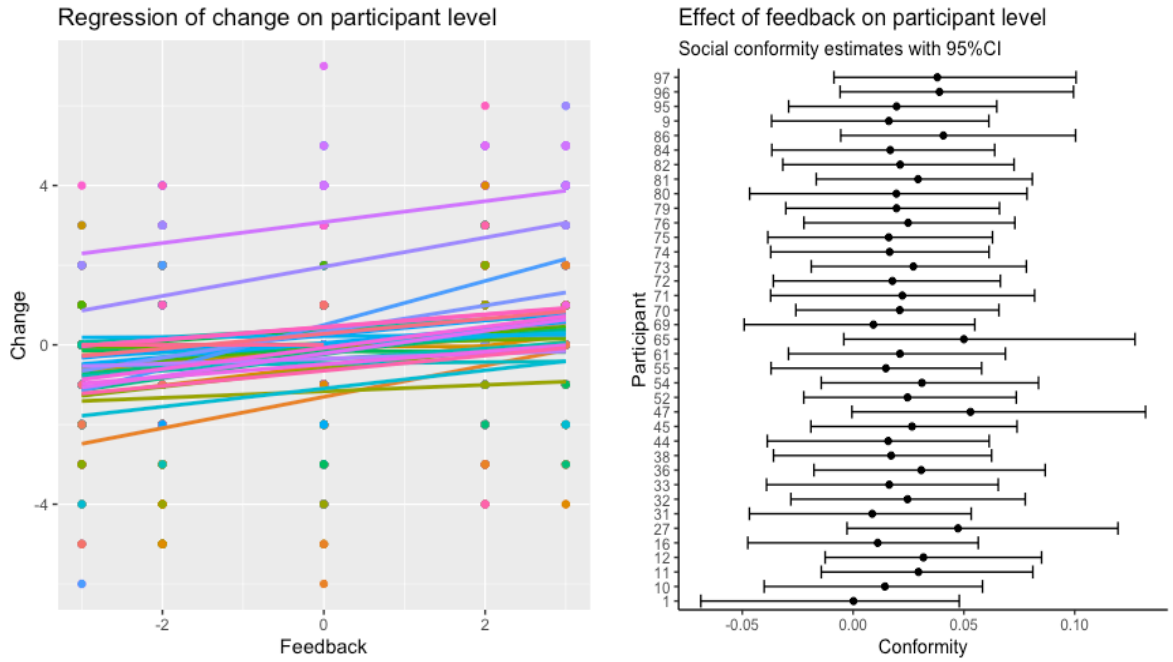


***Figure 12**: Assessment of individual level estimates from peri-COVID-19 study*