Portfolio 1: data cleaning

https://github.com/StudiegruppeEM3/EM3/blob/master/A1_DataCleaning_template_Astrid.Rmd

Portfolio 2 part 1:

https://github.com/StudiegruppeEM3/methods3_A2_P1/blob/master/Code%20for%20portfolio%202%20part%201.Rmd

Describe the characteristics of the two groups of participants and whether the two groups are well matched.

The sample included 35 (excluding participant 66 because of missing data and only completing visit1) (28 male and 6 female) typical developing (TD) children with a mean age of 43±9 months (mean±sd) and 31 (26 male and 5 female) children suffering from Autistic syndrome Disorder (ASD) with a mean age of 30.6±7.23 months, a t-test showed that the two groups where not matched in age $t(325.71) = 14.41$, $p < 0.01$, (all t-tests reported are reported as ASD vs TD, furthermore we used welch t-test which does not have the assumption of normality). Both TD and children with ASD were mostly white. The TD scored a mean of 0.95±1.8 on the Autism Diagnostic Observation Schedul (ADOS) while the children with ASD scored a mean of 14.11±4.2, a t-test showed a clear difference between the two groups in ADOS score $t(216.81) = 39.99$, $p < 0.01$.

A verbal and non-verbal IQ test from the first visits showed that the TD had a verbal IQ of 20.14±5.1 and a non-verbal IQ of 25.92±3.39 while the children with ASD had a verbal IQ of 17.58±7.4 and a non-verbal IQ of 26.89±5.6. VerbalIQ significantly differed between the two groups $t(303,57)=-3.8$ $p < 0.01$, nonverbalIQ did significantly differ between groups $t(283.42)=1.98$ $p = 0.05$.

Socialization scores from the first visits showed that TD had a socialization score of 100.51±6.75 while the children with ASD had a socialization score of 77.20±9.63, a t-test revealed that socialization differed between the two groups $t(309,83)=-26.68$ $p<0.01$.

Lastly, we checked whether the two groups of children differed in their MLU, word tokens and word types in the first visit. Mean length of utterances for TD was 1.314±0.27 for ASD was 1.31±0.69, a t-test revealed that these values were not significantly different. The number of tokens used by the different groups did differ by a small margin $t(326.28)=-2.1617$ $p< 0.05$, but word type did not significantly differ between the groups.

From this we see that the children weren't matched for a lot of the variables, such as age, socialization score, verbalIQ scores, nonverbalIQ scores, but for the linguistic properties they seem to be quite well matched for instance in their MLU and word types, the significant difference in number of tokens could largely be due to the fact that we made a lot of tests without correcting for multiple comparison, if we did correct for that, using Bonferroni correcting our p-value threshold should have been 0.05*1/number of tests which in this case is 8, giving us a p-value threshold of 0.00625 if we want to keep our type 1 error rate at 5%. Doing this the verbal and non-verbal IQ- scores would also not be deemed significantly different. We therefore conclude that the children were well matched in the current study.

Hypothesis: The child's MLU changes: i) over time, ii) according to diagnosis

Let's start with a simple mixed effects linear model Remember to plot the data first and then to run a statistical test.

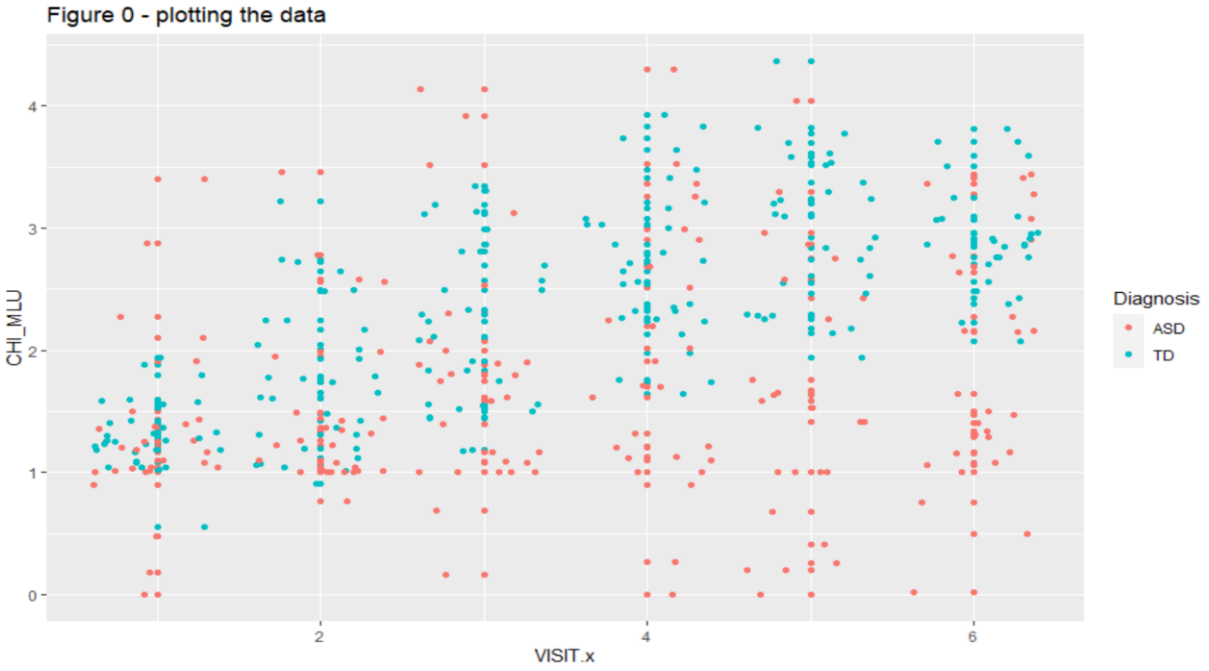


Figure 0; displays mean length of utterances of TD and ASD children compared to the number of visits.

Which variable(s) should be included as fixed factors?

Diagnosis for testing whether it effects children MLU changes. And visit to see if their MLU changes over time, one could include a lot of the other variables as fixed effects however due to our hypothesis we wouldn't include anything else.

Which variable(s) should be included as random factors?

random slope for visit because some children will probably learn faster than others(we did not include this because the model failed to converge) and random intercepts for subject, because every child has a different starting point.

How would you evaluate whether the model is a good model?

First we evaluated the model by checking whether it met the assumptions of a linear mixed effect model.

No patten is to be observed of the residual-plot therefore linearity assumption is met.

The assumption of Homoskedasticity can also be checked from the residual-plot, it seems like there might be a bit of heteroskedasticity, but not too badly.

The output of the model shows the correlation between the two predictors to be 0.369, which is a small to medium correlation, no reason for concern for the assumption of multicollinearity.

normality of residuals assumption is also met by looking at the histogram or the qq-plot.

Then we estimated the R^2 marginal and conditional which were 0.349 and 0.77 respectively, showing that out fixed effects explains 35% of the variance and the whole model explains 77% of the variance. We think that, that is a decent good model.

Growth curve fitting:

We then made 3 new models including quadratic, cubic and quartic aspects of visit, here we used the poly-function which takes care of visit being correlated with itself in the different polynomials, hereby not violated the assumption of multicollinearity.

We then used the anova function to see which model was best, we assed this using the Bayesian information criterion(BIC) we used BIC because it has been shown to be more conservative than AIC, and when doing these kinds of exploratory analyses we think that going with the more conservative is better.

The anova showed that a combination of linear and quadratic was the best model to the data which we then plotted in figure 1.

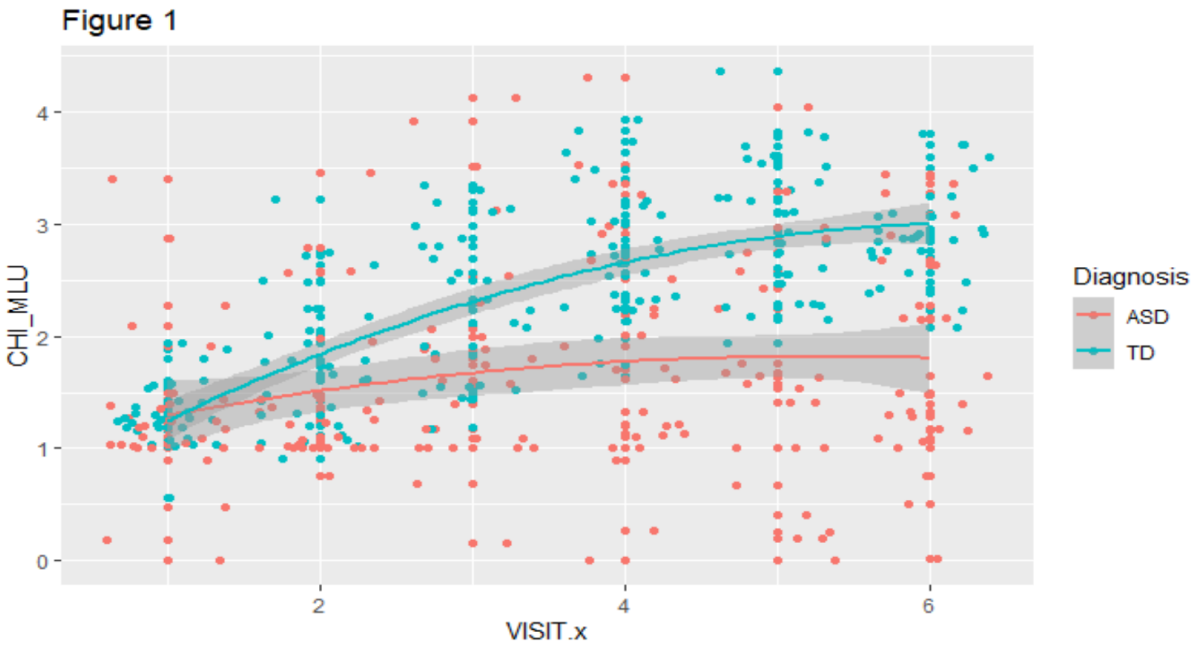


Figure 1; displays the relationship between mean length of utterances in ASD and TD children, compared to visit. Lines fit to the data are growth curves of 2nd order.

Let's check whether the model is doing an alright job at fitting the data. Plot the actual CHI_MLU data against the predictions of the model fitted(model)

See figure 2. This plot shows that the model is pretty good at explaining the data, if all the points were on the straight line the model would be perfect at explaining the data (explaining 100% of the variance), however that doesn't tell us how good the model is at predicting new data.

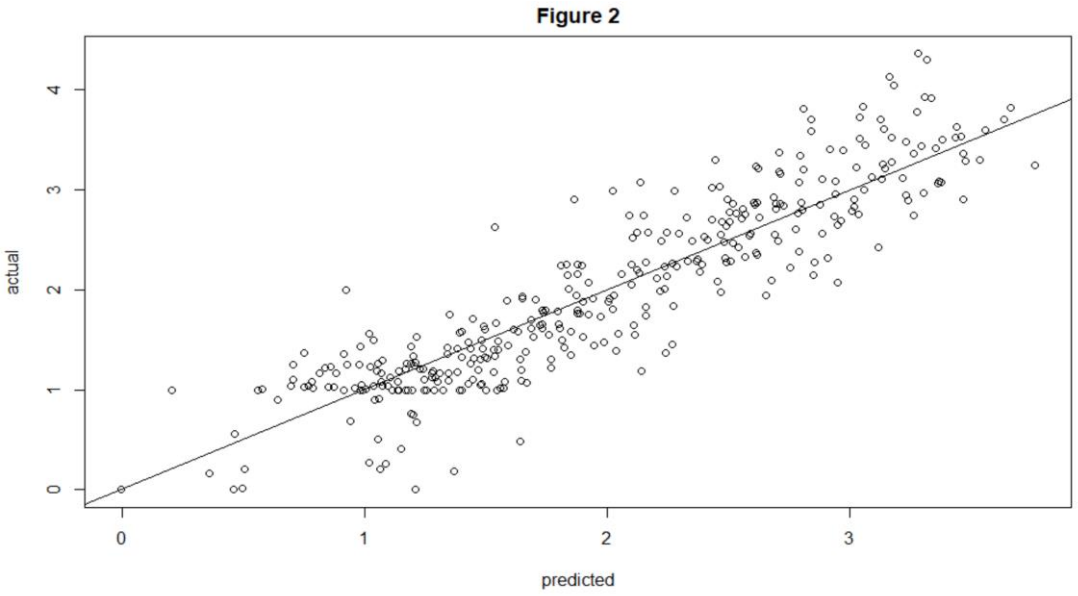


Figure 2; displays the relationship between the predicted value of our model compared to the actual values it tries to predict.

Now it's time to report our results.

There was a significantly difference between ASD and TD in MLU B = 0.6472, se = 0.1636, t(58.8417) = 3.957, p<0.01

when the visit number increases linearly the MLU increases significantly aswell B=3.31, se = 0.6438 t(287.22) = 5.14, p < 0.01

when the visit number increases quadraticly the MLU decreases significantly B=-1.3, se = 0.64 t(287.24) = -2.1, p < 0.05

furthermore there was a interaction effect between diagnosis and visit (linearly) B=8.2 se = 0.89 t(287.27) = 9.2 p < 0.01

- A plain word description of the results

TD has longer length of utterances compared to ASD on average 0.6 longer utterances. The cildren's MLU increases as time passes linearly, however it seems to be the case that this linear increase decreases over time (it becomes less steep). However, these main effects should not be interpreted because of a significant interaction effect. The interaction effect shows that there is a difference in how much the MLU changes over time in the two groups, see figure 3, here its easy to see that the TD group's MLU increases as visits increase whereas the ASD group don't seem to increase their MLUs.

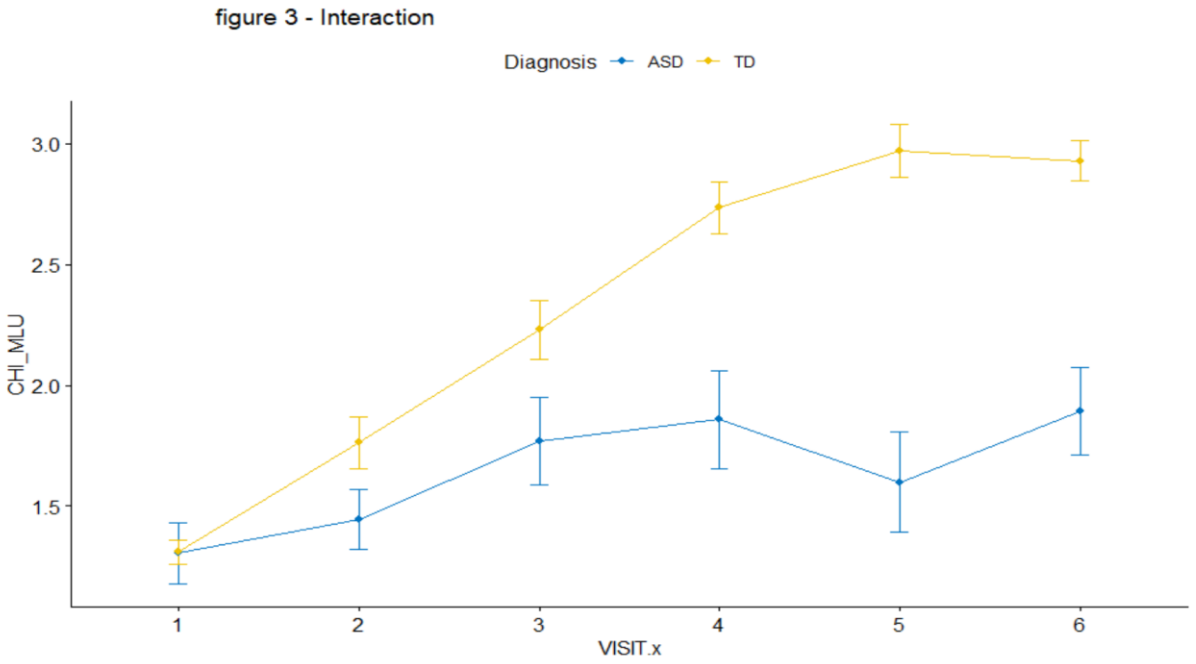


Figure 3; displays the interaction between visit and diagnosis. It can be seen that TD children develop much faster than ASD children.

A plot of your model's predictions (and some comments on whether the predictions are sensible)

a lot of the data falls outside the confidence intervals which arguable makes it less sensible, see figure 4.

Figure 4

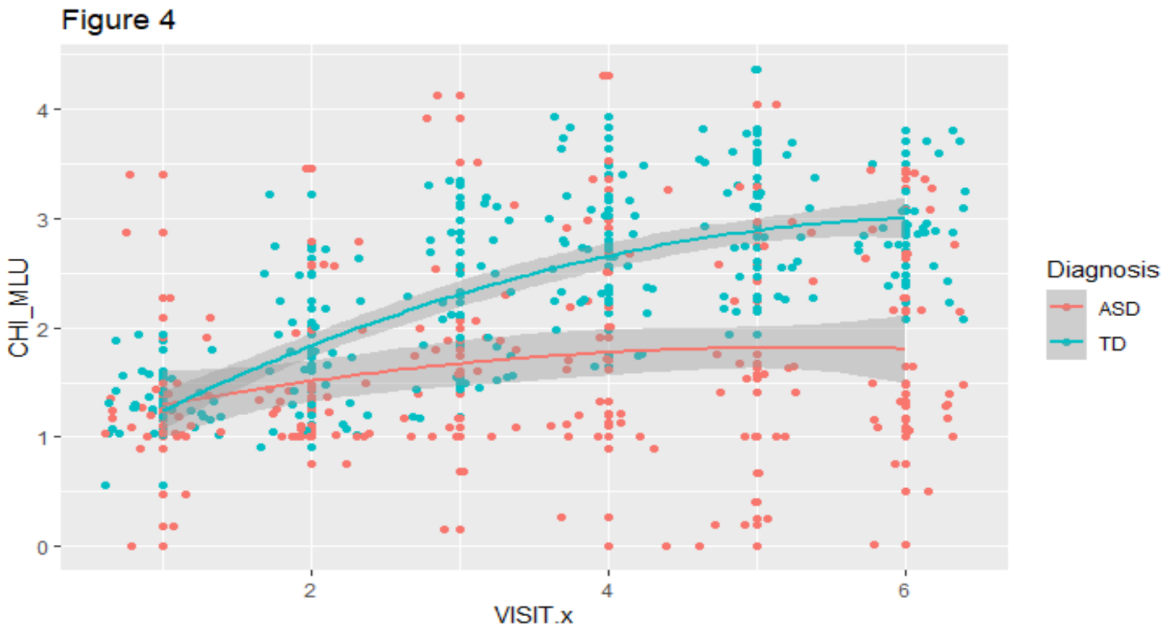


Figure 4; displays the relationship between mean length of utterances in ASD and TD children, compared to visit. Lines fit to the data are growth curves of 2nd order, the shaded area from the curves are 95% confidence intervals.

Hypothesis: Parental MLU changes: i) over time, ii) according to diagnosis

We tested this hypothesis with a model including the mothers MLU which was predicted by Visit and Diagnosis. We saw that the mothers MLU changed over time and according to diagnosis, however diagnosis was barely significant, again if we corrected for multiple comparison using bonferroni's correction this would have been non-significant.

Parent MLU is affected by Diagnosis (ASD vs TD) B = 0.364, se = 0.15 t(147) = 2.44 $p < 0.05$ and visist B = 0.098, se = 0.20, t(289.6) = 5, $p < 0.01$. no interaction effect was found.

Your task now is to figure out how to best describe the children linguistic trajectory. The dataset contains a bunch of additional demographic, cognitive and clinical variables (e.g.verbal and non-verbal IQ). Try them out and identify the statistical models that best describes your data (that is, the children's MLU). Describe how you selected the best model and send the code to run the model to Victor and Byurakn.

we have not made models with all combinations, we made 2 overfitted models and compared them. The second model is better both in AIC and the explain variance for the fixed effects. This can be explained by the fact that the second model has more fixed effects, meaning that it explains the data better than the other models but doesn't necessarily mean it predicts new data better.

model 1 has an AIC value og 518 and a R^2 marignal 0.63

model 2 has an AIC value of 211 and a R^2 marignal 0.69

Portfolio 2 part 2:

Language development in ASD - making predictions

Here is the link to our full code

Discuss the differences in performance of your model in training and testing data

Our model:

lmer(CHI_MLU ~ poly(Visit,2) * Diagnosis +verbalIQ1 +(1|Child.ID), data = all)

To evaluate the performance of our model on the training data we calculated the root mean square. The root mean square of the model on the training data is 0.4. Since the children's lengths of utterance are generally quite small, this is RMSE does not indicate a great performance of the model on the training data. However, it is even higher for the test data (1.1).

It makes more sense to evaluate a model using the test data, this makes for a more generalizable model.

Here's a plot showing the predicted values of our model (built on the training data) compared to the observed (actual) value of the test data. This makes sense, since the model has been built on the training data it more closely follows this data.

Figure 1

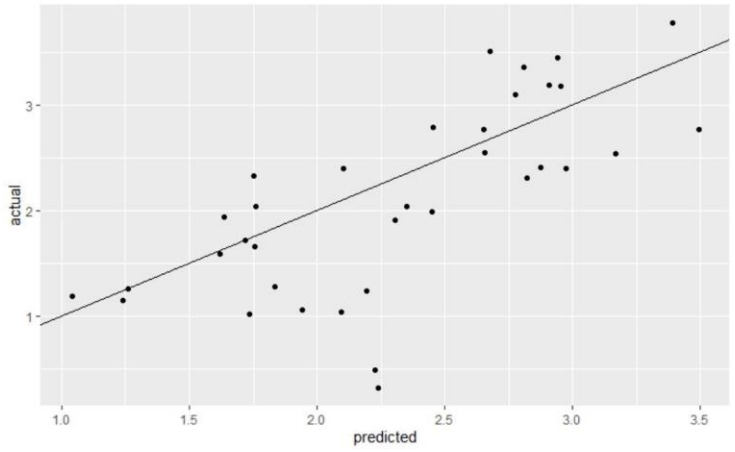


Figure 1; displays the relationship between the predicted values of the model compared to the actual values it tries to predict.

Which individual differences should be included in a model that maximizes your ability to explain/predict new data?

When we cross validate our old model which is child mean length of utterances predicted by visit and diagnosis and their interaction, we get a mean RMSE value of 0.77 from a 10-fold cross-validation. When doing this same procedure (10-fold cross validation) for our model described above, we get a mean RMSE score of 0.52. Therefore, our model has a better fit. This is because the predicted values vary the least from the observed values.

Our model includes diagnosis and verbal IQ. To find the best possible model using any combination of the measures of individual difference, we would repeat the steps done for the base model and our model, looking at all theoretically justified combinations possible of variables and select the model with the lowest RMSE.

Predict a new kid's performance (let's call him Bernie) and discuss it against the expected performance of the two groups

Here's a table showing Bernie's MLU at visits 1 through 6 compared to that of the average child. We also included a column showing the absolute difference between Bernie and the average child.

	Diagnosis	Visit	Mean-MLU for TD	Mean-MLU for Bernie	Absolute difference
1	TD	1	1.3	2.0	0.7
2	TD	2	1.8	2.5	0.8
3	TD	3	2.2	2.4	1.1
4	TD	4	2.7	3.2	0.5
5	TD	5	3.0	3.2	0.2
6	TD	6	2.9	3.5	0.5

Table; displays how Bernie does compared to the average TD child.

Bernie does better than the average child in all visits.

We also used our model to predict how well a child with the predispositions of Bernie is expected to do. Bernie did better than our model predicted We can see this in the plot below. The plot compares the predicted performance of Bernie to his observed (actual) performance.

Figure 2

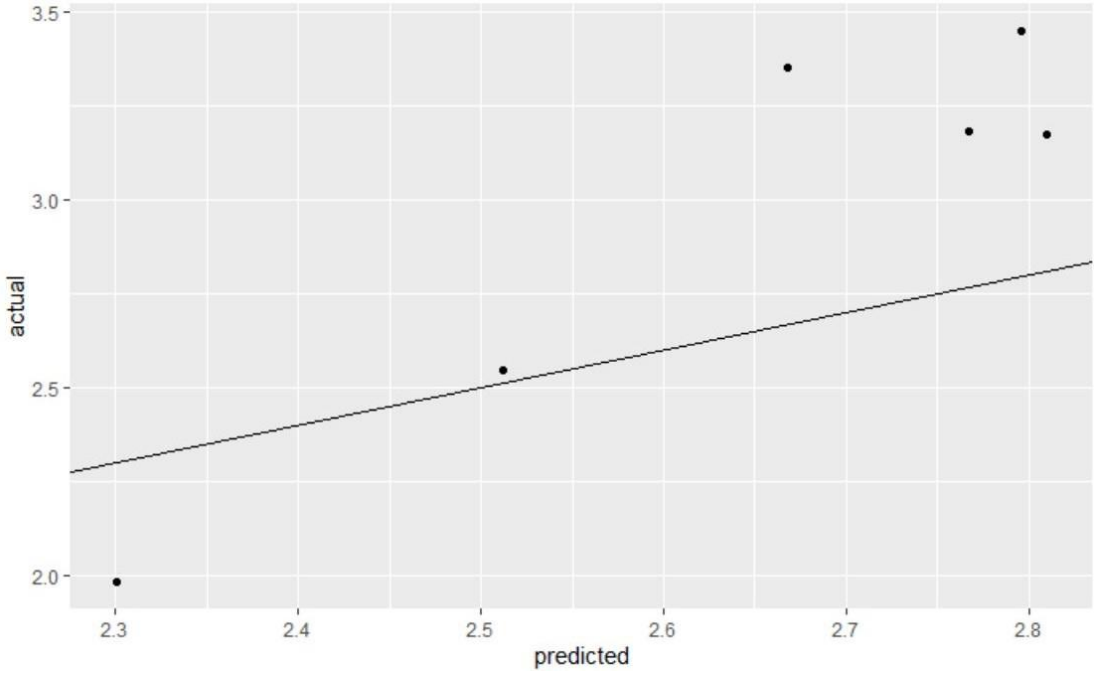


Figure 2; displays the relationship between the models prediction for Bernie compared to his actual values for mean length of utterances.

Who did what?

We all worked on the code on our individual computers in different states of mess. This time we used Jesper's code for the hand-in. We all discussed the answers for the questions in person and on chat. Astrid wrote the sum-up of the discussion in this document and reported the measures from Jesper's code

Portfolio 2 part 3:

Welcome to the third exciting part of the Language Development in ASD exercise

In this part of the assignment, we try to figure out how a new study should be planned (i.e. how many participants?) in order to have enough power to replicate the findings (ensuring our sample size is adequate, our alpha at 0.05 and our beta at 0.8):

Link to code

If we trust the estimates of the current study. Report the power analysis and comment on what you can (or cannot) use its estimates for.

Since our study is concerned with the interaction between the nr. visit and the ADOS score on the child's mean length of utterance (henceforth: MLU), this is what we will interpret, as well as the effect of verbal IQ on the MLU.

The interaction has a power of 100 % (confidence intervals based on 50 simulations) (92.89-100), with an estimate of -0.02.
VerbalIQ has a power of 100% (92.89-100), with an estimate of 0.07.
Visit^2 has a power of 100% (92.89-100), with an estimate of -0.047.
see power curves below:

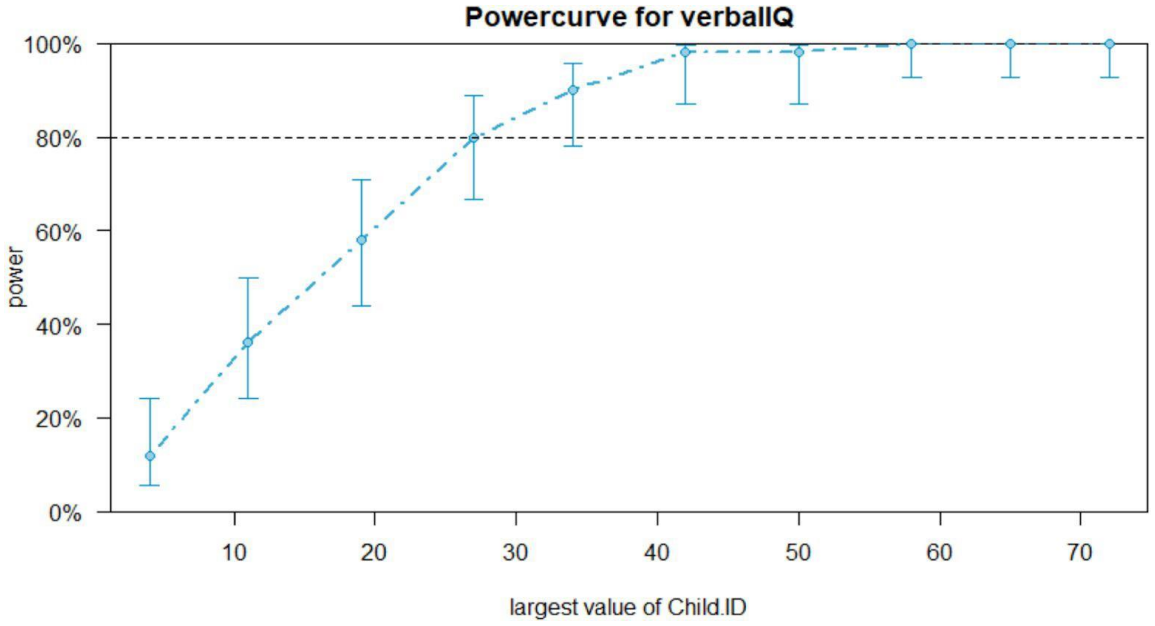


figure 1; displaying power of verbalIQ as a function of number of children.

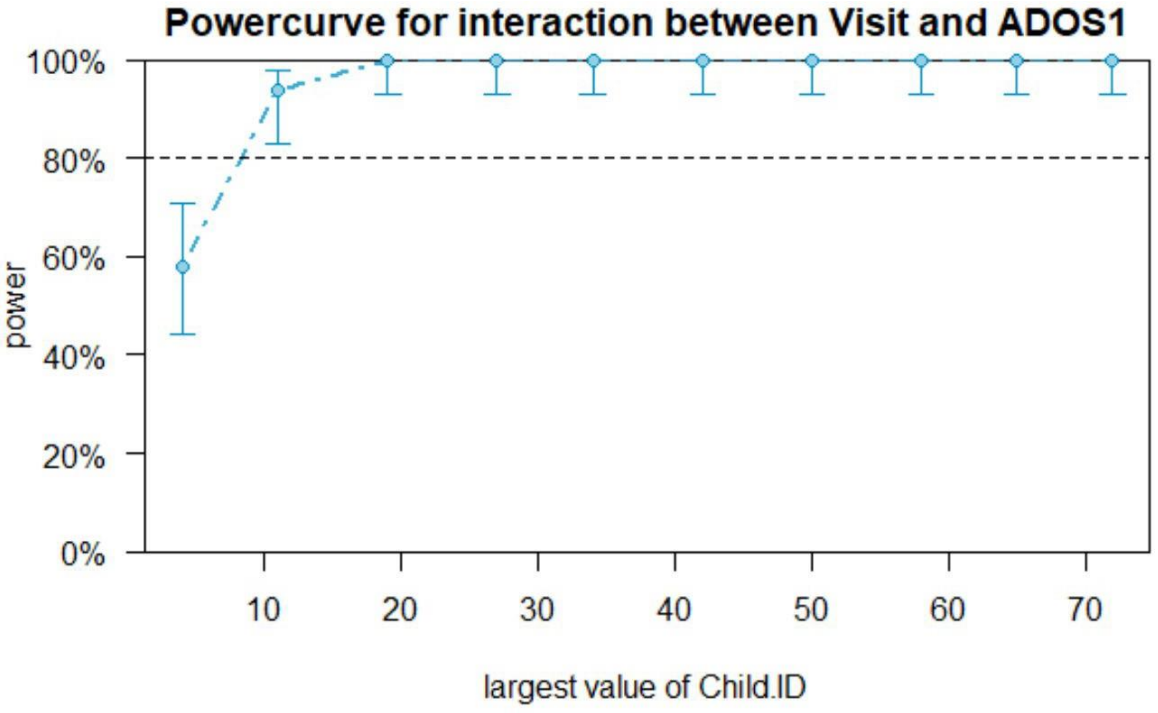


figure 2; displaying power of the interaction between visit and ADOS score as a function of number of children.

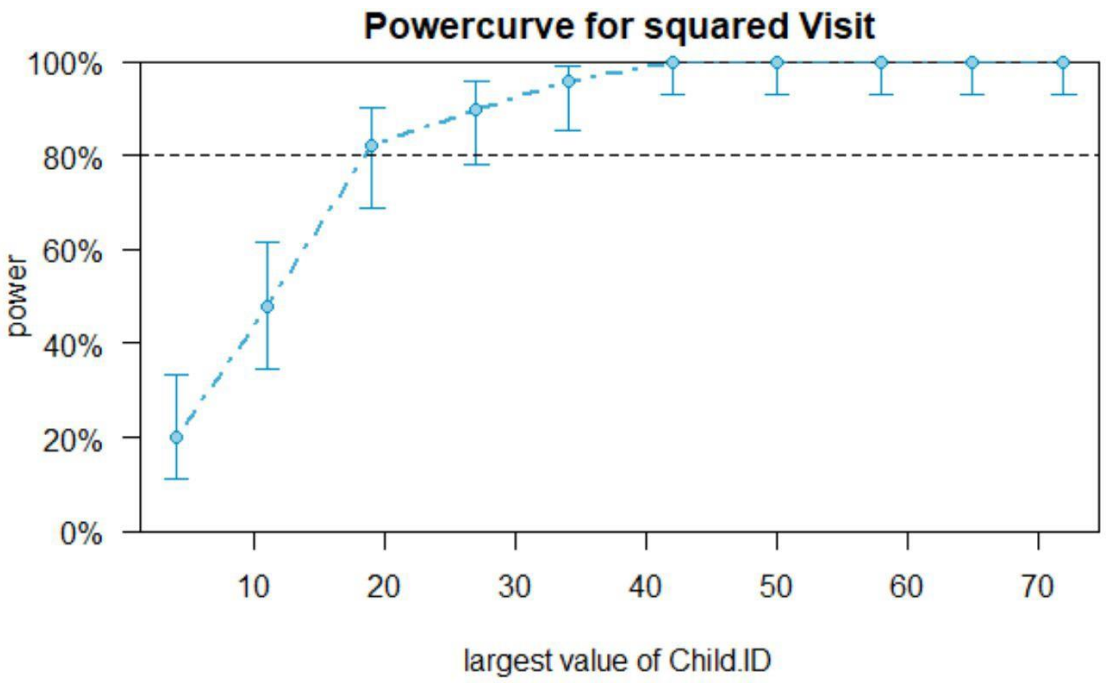


figure 2; displaying power of visit as a second order polynomial, as a function of number of children.

Since the power is high, we are likely to avoid making a type 2 error (false negative). However, conventionally, we should aim for a power of 80% to avoid wasting resources. Therefore, we consider this study overpowered.

If we are skeptical of the current study. Report the power analysis and comment on what you can (or cannot) use its estimates for.

The estimates are not any worse because the study is overpowered. The higher the power, the more we can trust the estimates. However, we are necessarily a lot more sure of these estimates, compared to how many more resources has been used to gain this extra power. It is always a trade-off and depends on the context. For our purposes, it was probably not worth it. However, had this data been based on i.e. web scraping or a meta-study it would not be a problem that it was overpowered.

Identify and justify a minimum effect size for each of your relevant effects. How would you perform a more conservative power analysis?

Here it would make sense to look at prior research, i.e. this present study (read: Riccardo's). Ideally, we should look at other studies as well. We would like to take a conservative approach and set our minimum effect size at half the size of the one used in the present study. Therefore a minimum effect size for verbal IQ is 0.035, the interaction between ADOS score and visit is -0.01, and visit^2 is -0.02.
The risk of setting a too low minimum effect size is that our power gets too high considering the available resources. We would rather risk this than risk having too low power since this is a common problem in psychological research and leads to false negatives.

Assess the power curve by Child.ID, identifying an ideal number of participants to estimate each effect

See the power curves below. These have been completed with 50 data simulations, expect for visit^2, here we ran 100 data simulations because the power was close to 80% (with a confidence interval that was both under and over 80%) for 50 simulations. As to be more confident in our judgment we ran a higher number of simulations.
The following number of participants are needed to find our minimum effect sizes on MLU for each effect with a power of 80%:
Interaction between visit nr. and ADOS: ca. 25 participants
Verbal IQ: ca. 35 participants
Visit^2: More than 70 participants

See power curves below.

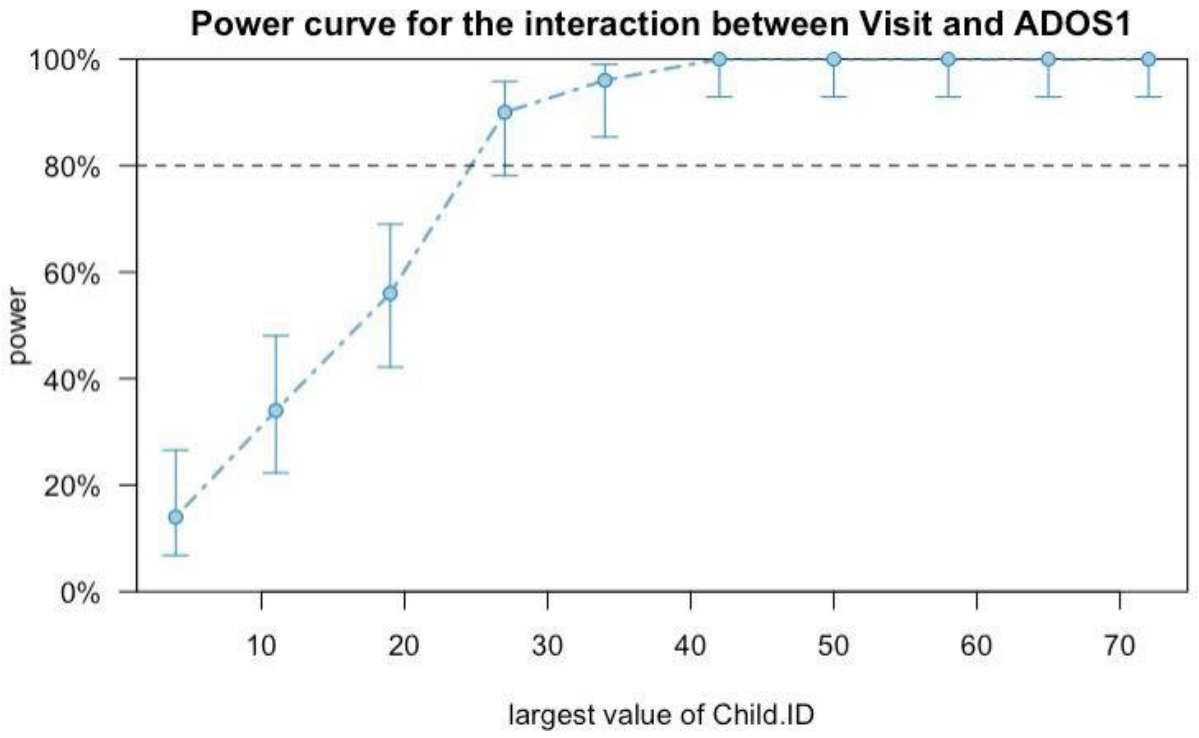


figure 4; displaying power of the interaction between visit and ADOS score as a function of number of children, to reach 80% one would need around 25 participants.

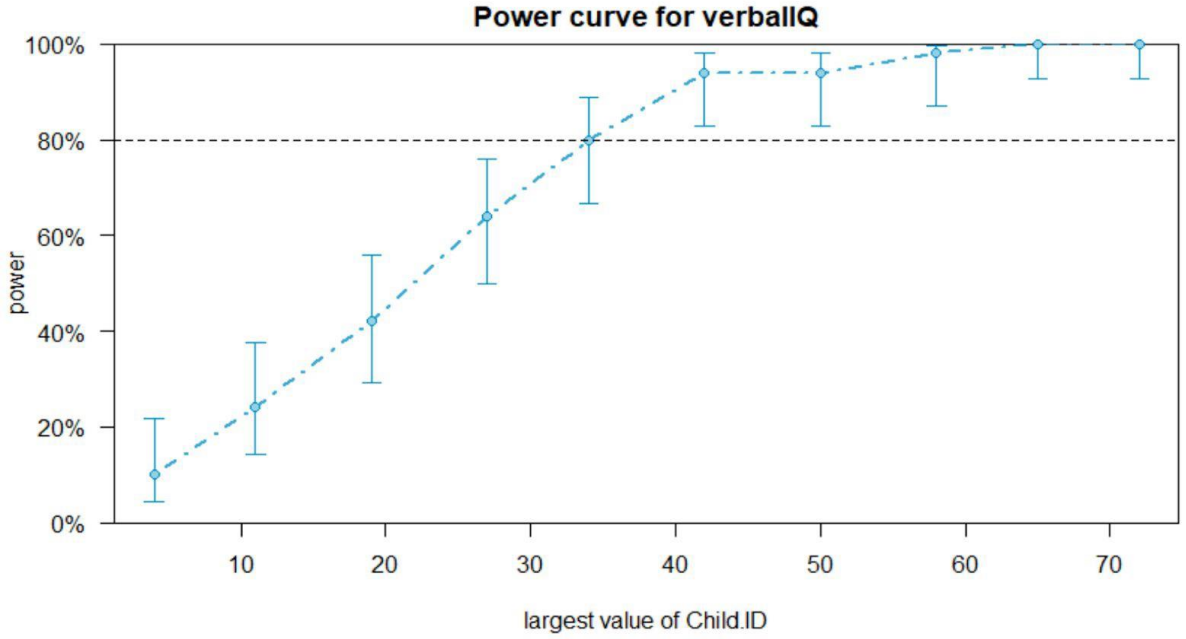


figure 5; displaying power of verbalIQ as a function of number of children, to reach 80% one would need around 35 participants.

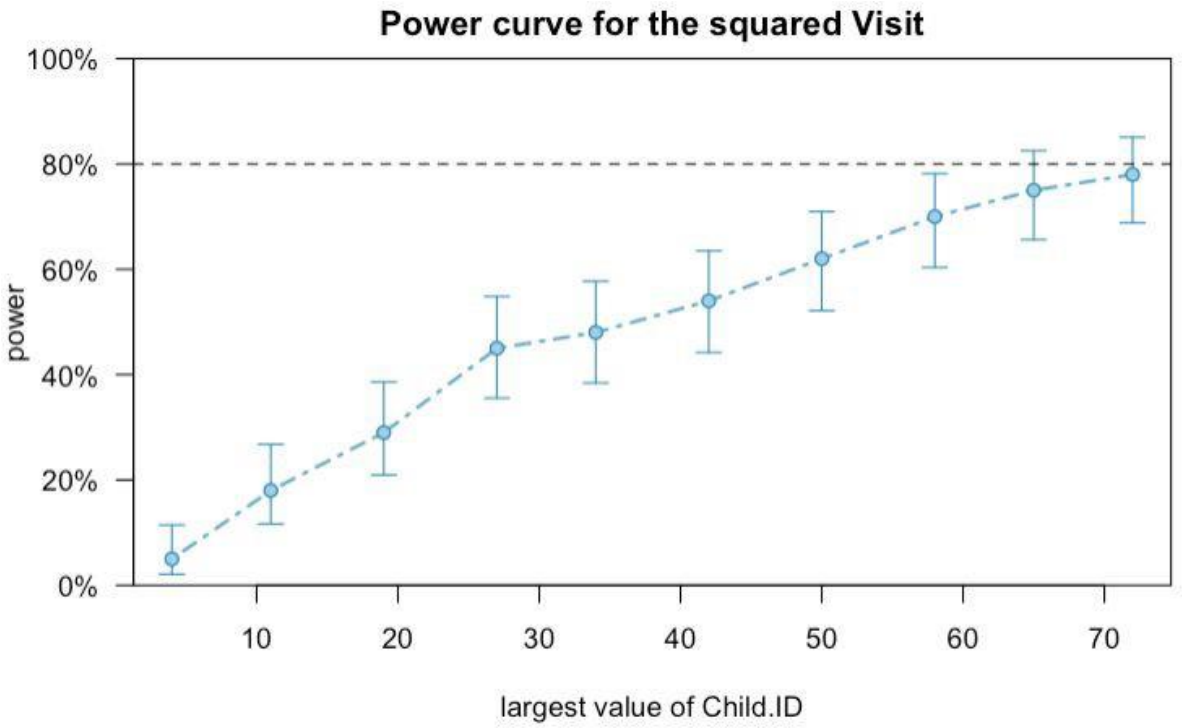


figure 6; displaying power of visit as a second order polynomial as a function of number of children, to reach 80% one would more than 70 participants.

Report the power analysis and comment on what you can (or cannot) use its estimates for.

The power for the interaction and for the verbal IQ is unchanged.
Visit^2 has a power of 82% (ci: 73.05-88.97). This is still an acceptable power, with a more conservative effect size.
We can use our estimates just as well as with the previous study while accepting a lower effect size. This means that if there is a difference, but it is smaller than what has been found in the study (by Riccardo), we are still likely to detect it at this power.

If we only have access to 30 participants. Identify the power for each relevant effect and discuss whether it's worth to run the study and why

We're still using our more conservative minimum effect sizes.

We wanted to avoid selecting a biased group from our data using a subset where we specifically select i.e. the first 30 participants. It could be there is some effect of the order the participants are in.

Therefore, we chose instead to simulate a new group of participants using extend(), based on the data. This group may not consist of exactly 15 ASD and 15 TD kids.

The powers were as follows:

interaction 98% (89.35 - 99.95)

visit^2 28% (16.23 - 42.49)

verbalIQ 86% (73.26 - 94.18)

see power curves below

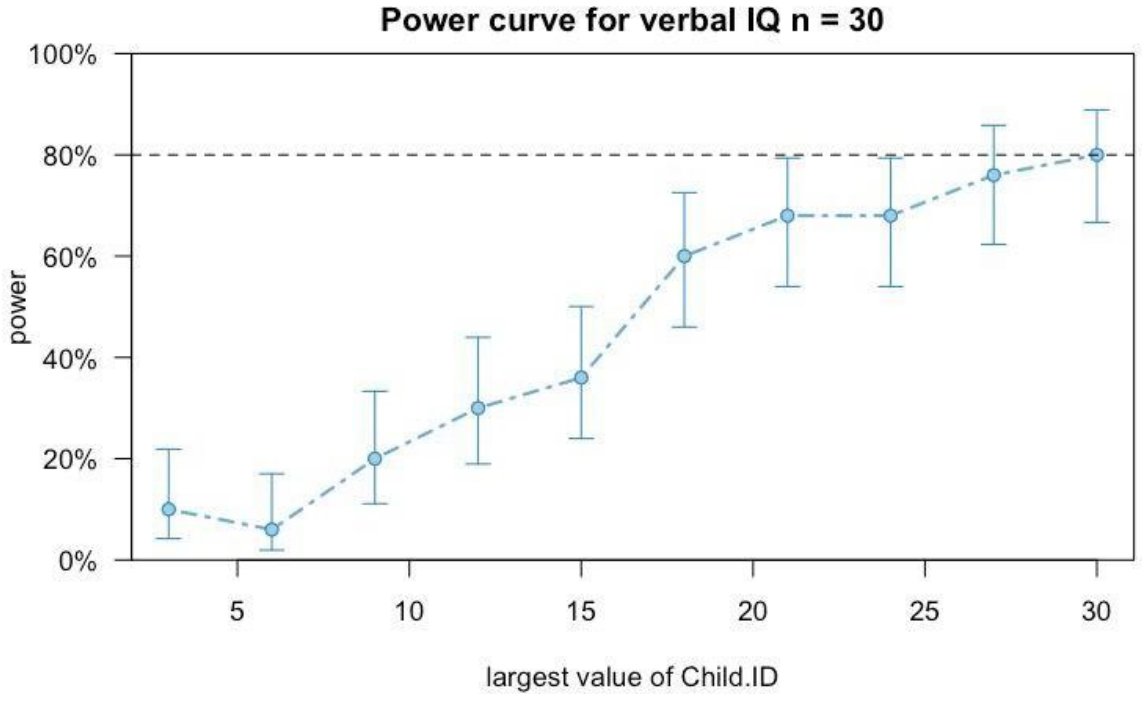


figure 7; displaying power of verbalIQ as a function of number of children, here we let the maximum number of children equal to 30

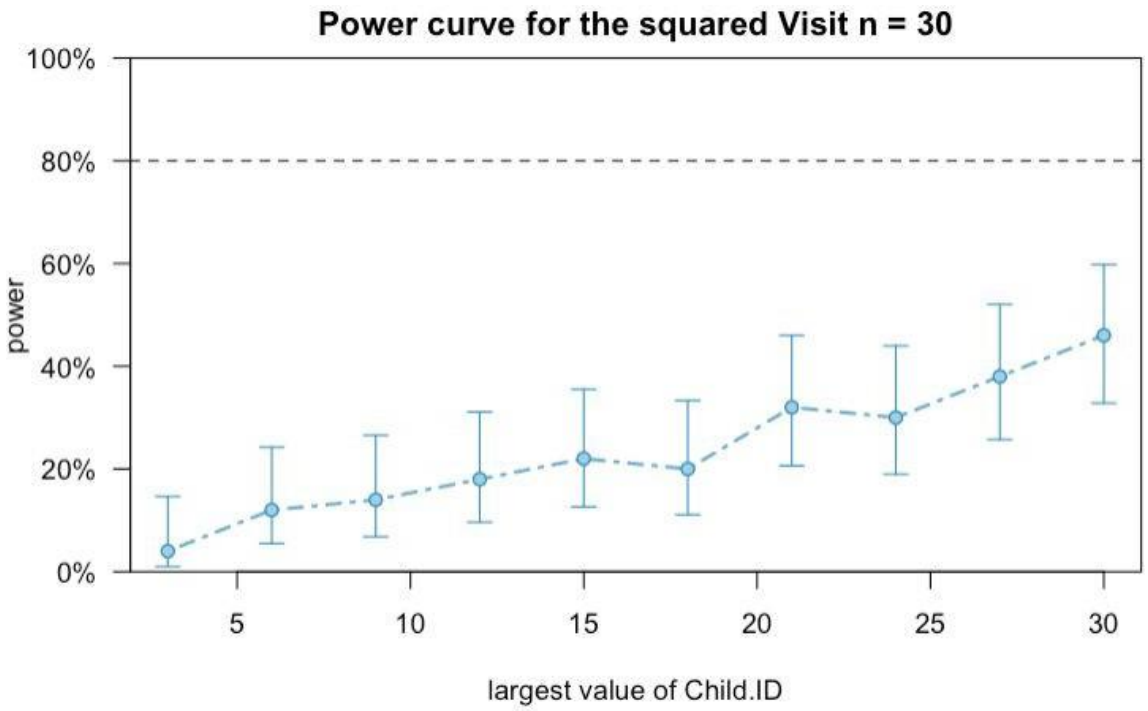


figure 8; displaying power of visit as a second order polynomial as a function of number of children, here we let the maximum number of children equal to 30

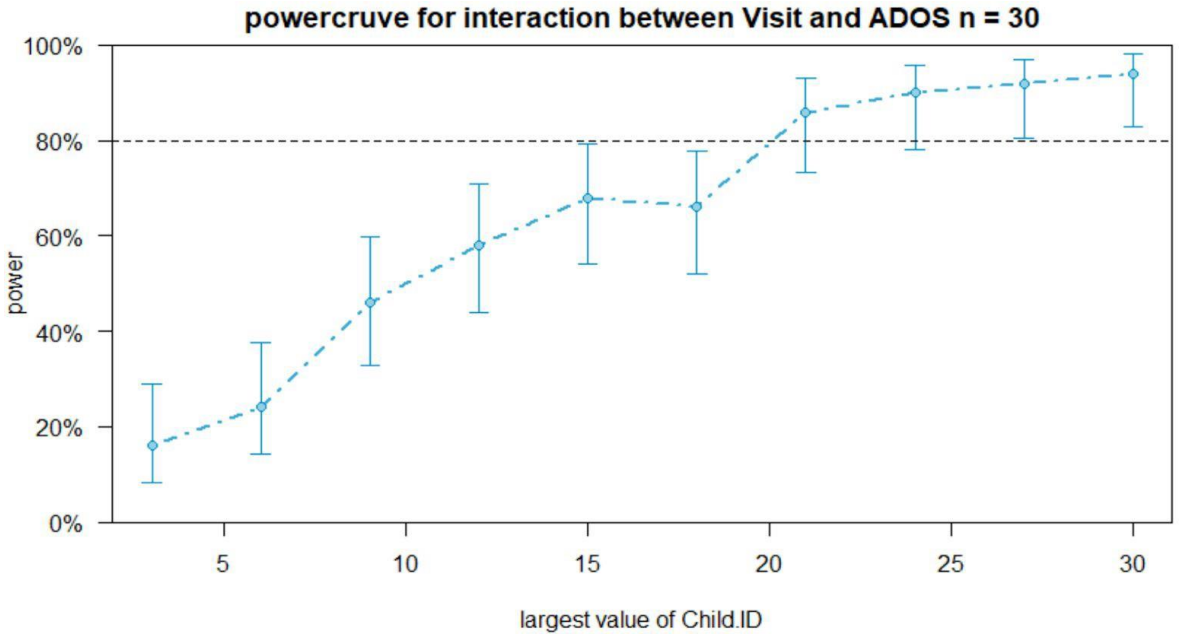


figure 9; displaying power of the interaction between visit and ADOS score as a function of number of children, here we let the maximum number of children equal to 30

Whether it is worth running this study depends on the hypothesis we are interested in. For the interaction and the verbal IQ, the power is high enough to be worth it. But the visit^2 power was too low to be worth it. It is too likely that we would have a false negative.

Who did what?

We completed the portfolio while working all together in the same room. All problems and approaches were discussed amongst the group. Sometimes we split up i.e. googling-tasks for fixing code-specific problems or looking up theoretical issues. Most tasks were complete on all computers. We split up some of the computational heavy runs, to speed up the process. Daniel ran most of the simulations. The code we linked to is mostly based on Jesper's code and commented by Jesper. The report is mostly written by Astrid

Portfolio 3 part 1:

https://github.com/StudiegruppeEM3/methods3_A3/blob/master/A3_P1_SchizophreniaVoice_final_code.Rmd

Methods

Procedure

Every participant watched several videos of triangles moving across a computer screen and were instructed to describe them.

Data

When the participants described the triangles, their voices were recorded and their pitch was extracted every 10th millisecond as a measure of fundamental frequency (hertz).
Each study collected demographic data from the participants, this includes: language, gender, education, age, SANS (total score of negative symptoms), SAPS (total score of positive symptoms, Verbal IQ score , Non verbal IQ score and a total IQ score.
From the recordings the clinical practitioners obtained data of number of syllables, number of pauses automatically inferred from the audio (absence of human voice longer than 200 milliseconds), duration of the full recording, duration of the recording where speech were present, average number of syllables per second, average number of syllables per spoken second and average syllable duration

Sample

Our initial data included 7 studies, 1-4 being danish studies, 5-6 being Chinese studies and the 7th study being japanese participants. Due to lack of data on the japanese participants we excluded them throughout the analysis.
As can be seen in table 1-3, not all of the 6 studies are perfectly balanced in regards to gender or the number of control participants to diagnosed participants. Especially study 5 is unbalanced in both aspects. All the other studies seem to be not perfectly balanced but somewhat balanced. The studies have balanced age between the two conditions very well, the only concern is study 6 where the control group seems to be a bit older. For the studies reporting verbal-nonverbal IQ it can be seen that the control group scores higher than the diagnosed group in both study 1 and 2 however in study 6 it is the opposite way around for verbalIQ. The diagnosed group has high scores of SANS and SAPS which makes perfect sense, since they are diagnosed whereas the control group has either missing data or as in study 4 has very low values. As for education the studies seem to somewhat balance the education level for the two groups, although every control group for every study has higher values for education than their diagnosed counterparts.

Tabel 1

Study	N	Schizophrenia / Control	Male/ female	Age(mean ± sd(ci)) Schizophrenia	Age (control)
1	70	34/36	37/33	22.8±3.1 (23-22.7)	22.7±3.2(22.8-22.6)
2	46	23/23	33/13	23.3±3.8 (23.5-23.1)	23.7±3.5(23.8-23.5)
3	47	19/28	23/24	41.1±12.2 (41.8-40.4)	37.5±12.9(38.2-36.9)
4	58	29/29	33/25	24.7±3.7(24.9-24.5)	24.3±4.5(24.5-24.1)
5	26	20/6	19/7	25.5±4.8(25.6-25.4)	27±1.3(27-26.9)
6	51	23/28	25/26	27.3±7.2(27.4-27.2)	35.7±9.3(35.9-35.6)
Total	298	148/150	170/128		

Tabel 2

Study	Nonverbal-IQ Schizophrenia	Nonverbal-IQ Control	VerbalIQ Schizophrenia	VerbalIQ Control
1	87.8±16.8 (88.3-87.2)	99.3±12.9 (99.7-98.8)	86.9±16.4 (87.4-86.3)	96.4±14.8 (96.9-95.9)
2	90.6±20.7 (91.7-89.5)	107.6±11.7 (108.2-107)	94.2±21.7 (95.3-93.1)	113±12.4 (113.7-112.4)
3	NA	NA	NA	NA
4	NA	NA	NA	NA
5	NA	NA	NA	NA
6	92.8 ± 18.9 (93.1- 92.5)	108±22.1 (108.5-107.6)	96.8±13 (97-96.6)	93.8±13.5 (94-93.5)

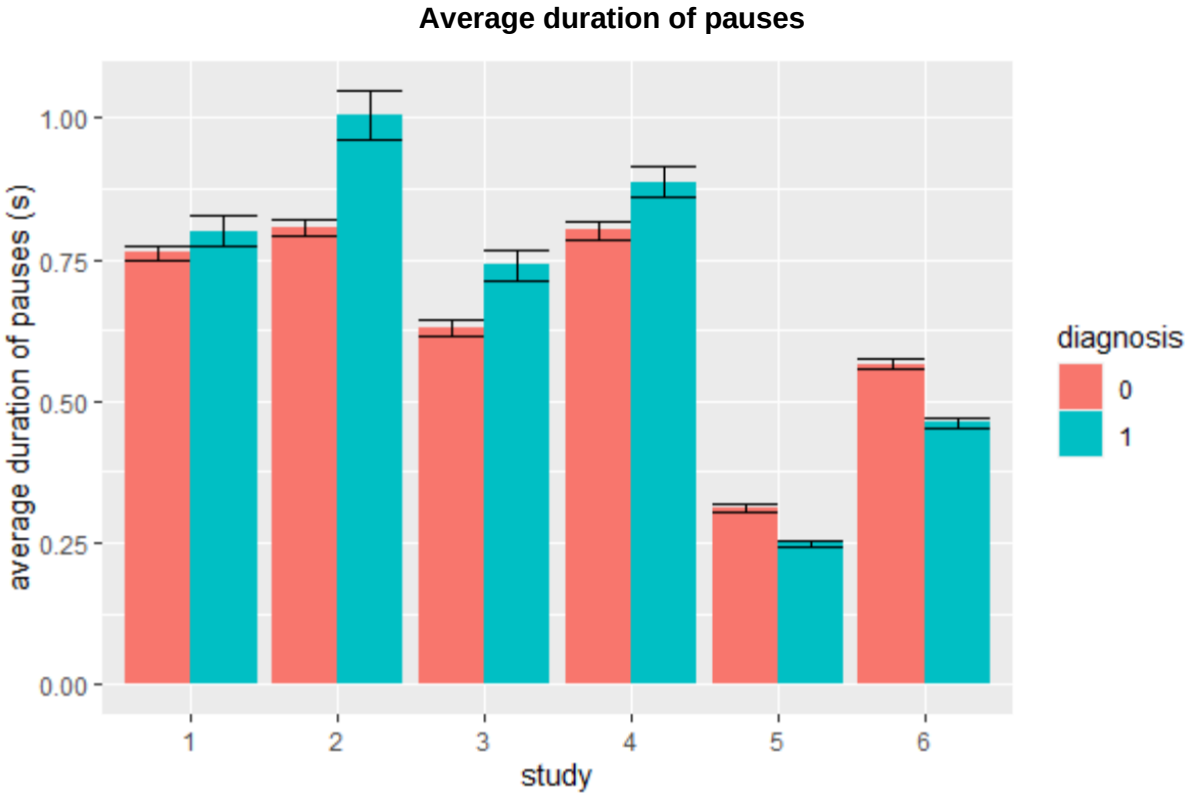
Tabel 3

Study	SANS Schizophrenia	SANS Control	SAPS Schizophrenia	SAPS Control	Education schizophrenia	Education Control
1	10.2±4.3 (10.4-10.1)	0	10.5±4.0 (10.6-10.4)	0	12.1±2.3 (12.2-12.0)	13.4±2.2(13.5-13.3)
2	9.9±5.2 (10.1-9.6)	0	14.5±4.3 (14.7-14.3)	0	12.2±2.5 (12.3-12)	15.2±2.6(15.3-15.1)
3	NA	NA	NA	NA	12.7±2.7 (12.9-12.6)	15.9±2.7(16-15.8)
4	8.6±3.6 (8.7-8.4)	1.3±1.9 (1.4-1.2)	7.0±3.9 (7.1-6.8)	0.3±0.8 (0.3-0.2)	14.7±2.6 (14.8-14.6)	15.8±2.2(15.9-15.7)
5	7.8±2.5 (7.8-7.77)	NA	11.6±4.0 (11.7-11.6)	NA	13.1±3 (13.2-13.1)	16.9±0.3(16.9-16.9)
6	6.6 ± 3.5 (6.6- 6.5)	NA	9.8±3.67 (9.8-9.7)	NA	12.5±2.3 (12.5-12.4)	12.7±2.5(12.7-12.6)

Acoustic profile

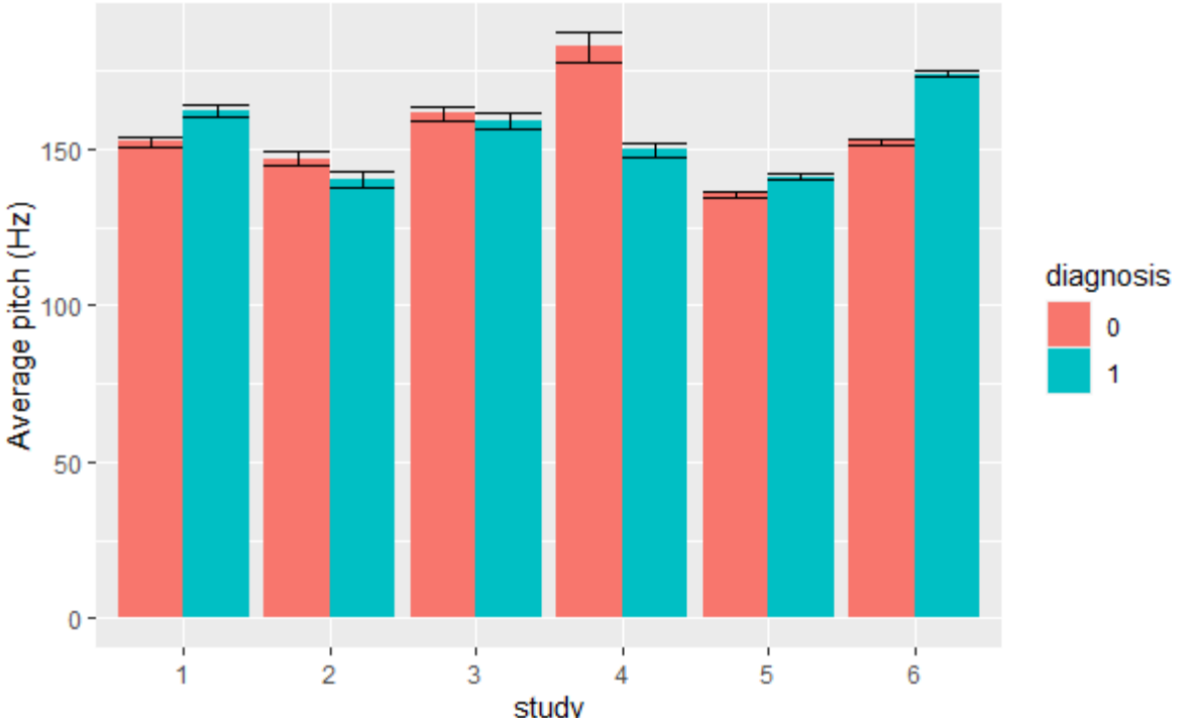
All the following plots error bars are displayed as 95% confidence intervals.

Plots.



The plot shows a difference between diagnosed and control groups of their average duration of pauses. The hypothesis from the meta-analysis stated that the duration of pauses would be longer for patients diagnosed with schizophrenia than those not diagnosed. We see a trend of this in the first four studies (Danish studies), and an opposite effect at study 5 and 6 (Chinese studies).

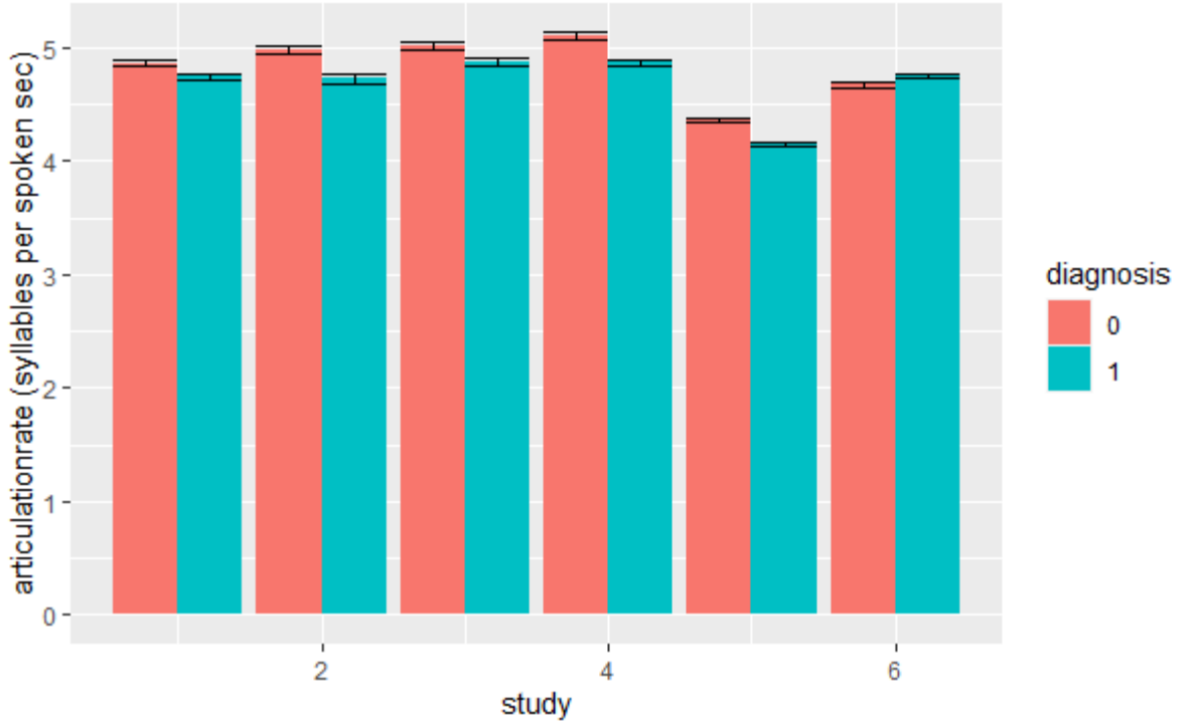
Average pitch



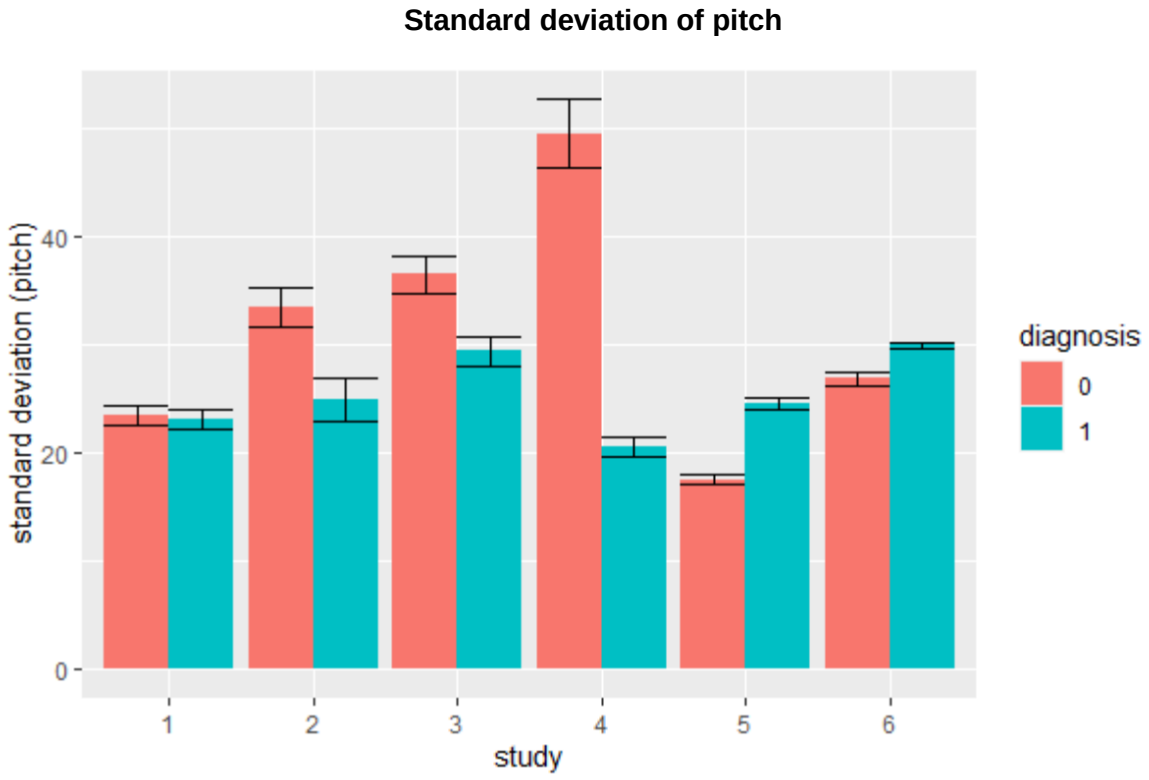
The plot shows a difference between the diagnosed and control group of their average pitch, measured in hertz.

In three of the six studies the diagnosed group has a higher pitch (1, 5 and 6), but in three of the four danish studies the opposite effect is found - the control group has a higher pitch than the diagnosed group.

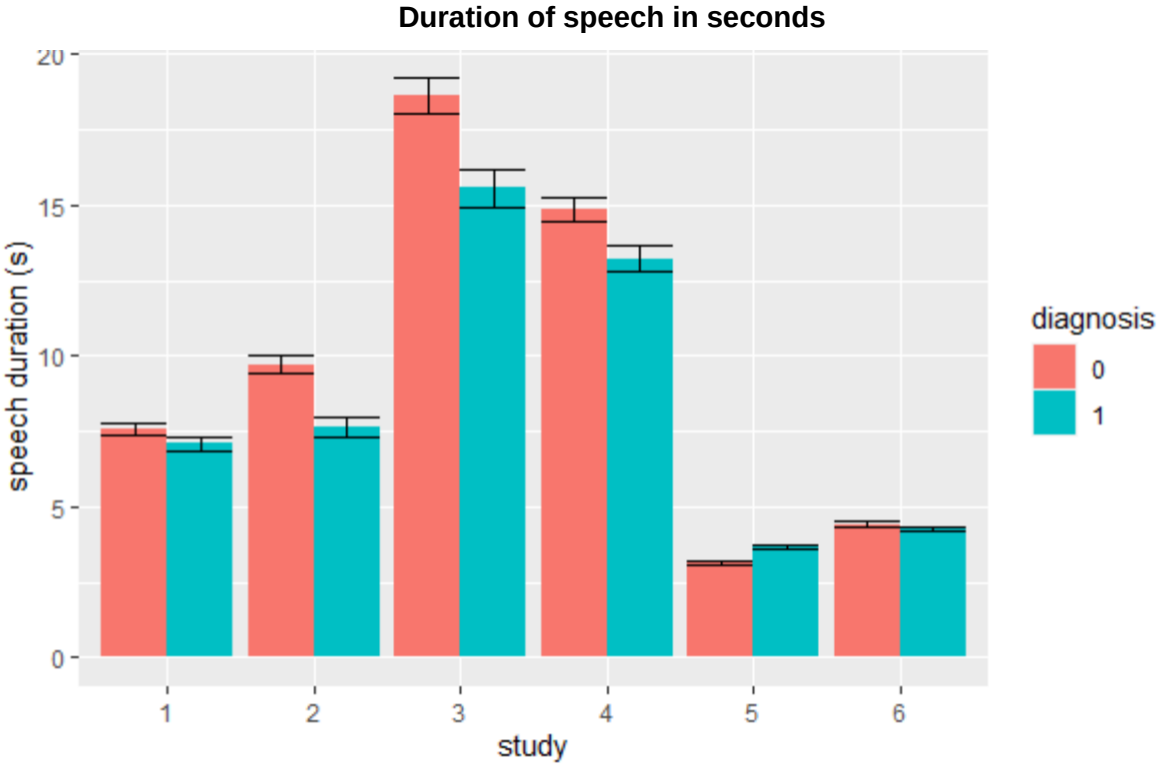
Syllables per spoken second



The plot shows a difference between diagnosed and control group of their average number of syllables per spoken second. The hypothesis from the meta-analysis stated that the speech rate for the diagnosed would be slower than the control group. The higher the number of syllables per spoken seconds is the measure of how fast or slow one speaks.
We see that in all the four danish studies and one of the chinese studies (five out of six studies) the rate at how quick one speech is different between the two groups, showing that people with schizophrenia are slower than the ones without.



The plot shows a difference between the diagnosed and control group of the standard deviation of the measured pitch. The hypothesis from the meta-analysis stated that the diagnosed participants with schizophrenia vary less in pitch than the control group.
In all the four danish studies the participants that are not diagnosed varies more in pitch than the ones who are. The two chinese studies do not indicate this.



The plot shows a difference between the diagnosed and control group of their duration of speech. How much they speak during a session. The hypothesis from the meta-analysis stated that the speech rate would be lower for diagnosed participants, indicating that schizophrenia participants speak less than the control group.
In five out of the six studies this effect is found, stating that the control group speaks more during a session than participants diagnosed with schizophrenia.

We have separated the studies according to language because as the plots show, study 5-6 (the chinese studies) seem to display different relationships between acoustic profile and diagnosis this could also be argued that it is due to the fact that the chinese studies are taken from a completely different population than the danish studies. We therefore only look at the danish studies.

Statistical analysis

We made a mixed effect logistic regression to try and replicate the meta-analysis findings.

First we look how correlated our predictor variables are in a correlation matrix see table 2.

Table 2:

	speechdur	articulationrate	fsd	pausedur
speechdur	1.00	0.06	0.05	-0.14
articulationrate	0.06	1.00	-0.08	0.02
fsd	0.05	-0.08	1.00	0.00
pausedur	-0.14	0.02	0.00	1.00

As can be seen in the correlation matrix, none of our predictor variables are especially correlated, the highest correlation coefficient is -0.14 which raises no concerns. We therefore include all 4 fixed effects in our model. We scale all our fixed effects so that we can compare our estimates to the meta-analysis findings, when scaling the fixed effects the estimates becomes a standardized effect size called standardized mean difference or hedges' g.
In the model we include random intercept for each participant, this ensures that we account for the fact that the studies are done using a repeated measure design. When having participants as a random intercept we also account for gender, age, education and study because every participant gets their own intercept.

Therefore our model was:

diagnosis ~ scale(speechdur)+ scale(fsd)+scale(articulation rate)+scale(pausedur)+(1|name)

Results

Fixed effect (scaled)	Estimate	Std.error	t-value	p-value
Intercept	-16.6	1.0	-16.79	>0.001
Proportion of spoken time	-0.05375	0.61365	-0.088	0.930
Pitch variability	-0.19020	0.59796	-0.318	0.750
Speechrate	-0.16750	0.56040	-0.299	0.765
pauseduration	0.17489	0.74894	0.234	0.815

The meta-analysis findings

Fixed effect	**Estimate**	**CI lower**	**Ci upper**
Pitch variability	-0.55	-1.06	0.09
Proportion of spoken time	-1.26	-2.26	0.25
Speech rate	-0.75	-1.51	0.04
Pause duration	1.89	0.72	3.21

Our estimates for the model are in the same direction as the meta-analysis findings but not to the same magnitude.

Assignment 3 - Part 2

Diagnosing schizophrenia from voice

Link to code:
https://github.com/StudiegruppeEM3/methods3_A3/blob/master/A3_P2_DiagnosingSchizophrenia_code.Rmd

Methods - models (Astrid)

We analyzed the data from 4 studies of speech patterns in Danish schizophrenic patients and controls. We chose to only use Danish speakers to keep the analysis simple. Schizophrenia may affect speech patterns differently in different languages. Furthermore when trying to predict schizophrenia in repeated measure models we decided to include study as a random intercept and not participant, this was due to the fact that including participant as a random intercept, explained an enormous amount of variance which basically made our acoustic features neglectable.
We made 10 models using mixed effect logistic regression. Each model predicted diagnosis from a single acoustic feature, with either name or study as a random intercept. We also included a full mode, with all 4 acoustic features. Then we compared these models using ANOVA where we looked for the model with the lowest AIC and BIC value. Furthermore, we made an importance plot of our 4 acoustic features (see figure 1). Both analyses indicated that pitch variability is our best acoustic feature.

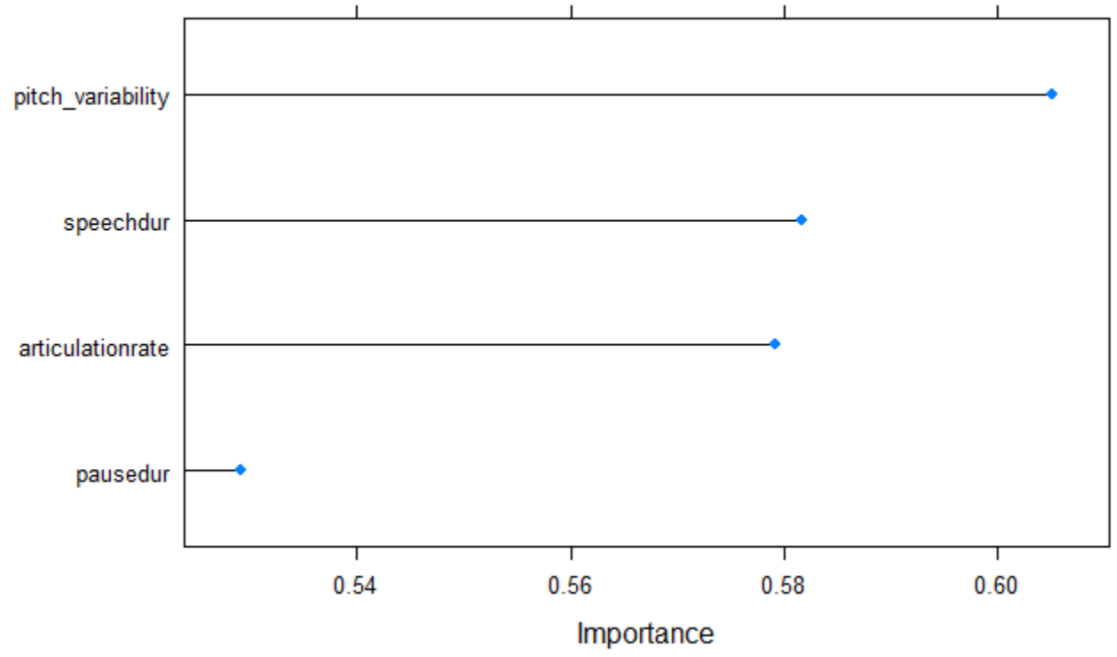


figure 1, shows the importance of our 4 acoustic features in predicting schizophrenia, results are 10-fold cross-validated, while the predictors have been scaled and

centered. The metric used for importance was the area under the receiver operating characteristic curve (henceforth: ROC).

Methods - confusion matrix and ROC (Daniel)

We examined this model's (diagnosis ~ pitch-variability) confusion matrix and calculated the accuracy, sensitivity, specificity, positive and negative predictive value (henceforth: PPV and NPV). We also plotted the ROC (see figure 2). This was firstly done when training and testing on the same data set and afterwards done with a 10-fold cross validation.

Then, we calculated the accuracy, sensitivity, specificity, positive and negative predictive value, and plotted the ROC curve for the cross-validated model where every acoustic feature was included in the model. We did this for several different models to find the best possible model for predicting schizophrenia with our 4 features, we tried general logistic regression(GLM), elastic net logistic regression, repeated random forest, k-nearest neighbors(KNN), Support Vector Machine with a Radial Basis Kernel Function(SVM), Linear Discriminant Analysis(LDA), Classification and Regression Trees (CART) and one ensemble model. Our ensemble model first consisted of GLM, KNN, SVM, CART and LDA however due to highly correlated models we excluded LDA and GLM because of correlation above 0.8, because highly correlated predictors can have detrimental effects on predictive power, see table 2. When stacking the ensemble model we used the method svmradial, because that was the best single best method (in respect to accuracy and area under the ROC) of the remaining 3 methods.

Correlation matrix for different classification methods

Model	IDA	CART	GLM	KNN	SVM
IDA	1	0.3	0.98	0.67	0.89
CART	0.3	1	0.35	0.003	0.46
GLM	0.98	0.35	1	0.69	0.92
KNN	0.67	0.003	0.7	1	0.5
SVM	0.89	0.46	0.92	0.51	1

tabel 2: Shows the correlation coefficients between the different models in the ensemble model.

Results (Jesper)

The best acoustic feature to predict schizophrenia diagnosis was pitch variability.

mean values of the model on pitch variability (Pernille)

Accuracy	0.58
Sensitivity	0.45
Specificity	0.69
PPV	0.57
NPV	0.58
Area under ROC	0.61
Kappa	0.15

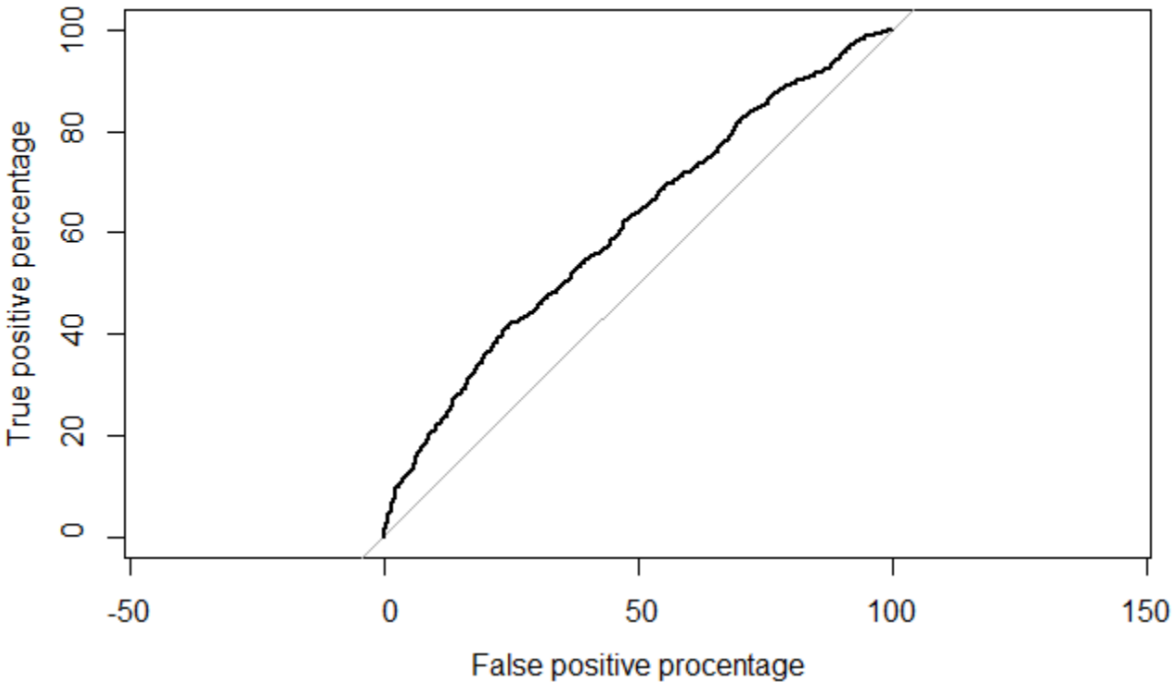


figure 2; ROC showing the relationship between 1-specificity in percent and sensitivity in percent for the model diagnosis ~ pitch variability

Mean values of the Cross-validated pitch-variability model. (Jesper)

Accuracy	0.54 (0.60-0.48)
Sensitivity	0.52 (0.61-0.44)
Specificity	0.56 (0.65-0.46)
PPV	0.52(0.59-0.48)
NPV	0.56 (0.63-0.49)
Area under ROC	0.60 (0.65-0.54)

Kappa 0.08 (0.21-(-0.05))

Mean values of several cross-validated models. (Jesper)

metric/ model	Accuracy	Sensitivity	Specificity	Kappa	Area under ROC
Glmer	0.58 (0.63-0.53)	0.49 (0.57-0.42)	0.67 (0.74-0.6)	0.16 (0.26-0.063)	0.61 (0.57-0.66)
elastic-net logistic regression	0.61 (0.65-0.56)	0.48 (0.54-0.42)	0.73 (0.79-0.67)	0.21 (0.31- 0.12)	0.61 (0.65-0.56)
Glm	0.60 (0.65-0.56)	0.49 (0.55-0.42)	0.72 (0.78-0.66)	0.21 (0.31-0.11)	0.60 (0.65-0.56)
CART	0.60 (0.62-0.58)	0.64 (0.71-0.58)	0.55 (0.59-0.5)	0.20 (0.24-0.15)	0.62 (0.65-0.60)
KNN	0.58 (0.61-0.56)	0.65 (0.66-0.63)	0.48 (0.52-0.45)	0.16 (0.21-0.11)	0.61 (0.63-0.58)
SVM	0.62(0.65-0.60)	0.68 (0.71-0.65)	0.54 (0.58-0.50)	0.24 (0.29-0.18)	0.65 (0.68-0.63)
LDA	0.61 (0.63-0.58)	0.71 (0.74-0.68)	0.50 (0.47-0.53)	0.21 (0.26-0.16)	0.63 (0.65-0.61)
Ensemble of KNN, SVM and CART	0.62 (0.64-0.60)	0.66 (0.69-0.63)	0.57 (0.62-0.5)	0.23 (0.27-0.19)	0.64 (0.66-0.61)

table 3. Showing different performance measures, as mean (upper-95%-confidence interval- lower-95%-confidence interval), from different models. The Glmer is with study as a random intercept

measure/ model	df	t-value	p-value
Glmer	9	4.67	0.0012
elastic-net logistic regression	9	4.60	0.0013
Glm	9	4.28	0.002
CART	9	10.64	0.000002
KNN	9	8.01	0.000022
SVM	9	13.8	0.00000023
LDA	9	14	0.0000002
Ensemble of KNN, SVM and CART	9	12	0.00000066

table 4 shows measures from a one-sided t-tests performed on areas under the ROCs for the given model on the values compared to a mean of 0.5. (see discussion to why we compare the values to 0.5) all p-values reported are uncorrected for multiple comparison.

Discussion

Discuss the "classification" process: which methods are you using? (Astrid)
For the first model we analyzed, we tested and trained the model on the same data. This is a problem because it leads to estimates that are unreliable if we use that model for predicting new data, that the model has not seen yet. We therefore continued to use a 10-fold cross-validation on every model made afterwards. As expected our estimates when cross-validating the pitch variability model, became significantly worse, but probably more accurate, if trying to predict new data.

Which confounds should you be aware of? (Daniel)

Some of the variables in our full model, including all acoustic features may be correlated. Normally, when we're trying to explain a phenomenon this would be a problem, since many models assume absence of multicollinearity. Since we're only trying to predict a diagnosis, this is only a problem if the prediction is worsened.
We checked to see if the model's predictors are correlated see table 5. They are not strongly correlated. If they had been, we would have checked how this affected the prediction of the models.

We also decided to scale and center all features used in every model, this can help with model convergence as well as processing time, one thing to be aware of when scaling and centering is that one has to scale the test data based on the scaling of the training data, else information about the test data will be leaked into the training data, ultimately skewing our estimates to be better than they actually are.

acoustic feature	pitch variability	articulation rate	pausedur	speech duration
pitch variability	1	-0.13	-0.05	0.06
articulation rate	-0.13	1	0.02	0.06
pausedur	-0.05	0.02	1	-0.14
speech duration	0.06	0.06	-0.14	1

table 5, showing the correlation matrix between the acoustic features.

What are the strengths and limitations of the analysis? (Jesper)

Our analysis shows that different models produce different results and can be used to solve problems, based on whether one thinks that sensitivity or specificity is more important. In our case we can see that the glm model produces models that have a high average specificity whereas the LDA- model produces models that have high average sensitivity. Our analysis only included Danish subjects which means that we do not know whether these estimates can be used to diagnose Chinese subjects, however from our representation of the data in assignment 3 part 1 we saw that the acoustic features did not have the same trend for Chinese individuals as compared to Danish individuals. Therefore, one can assume that even our best model predicting

schizophrenia in the Danish subjects will do poorly in predicting schizophrenia in the Chinese subjects.

The biggest limitation of the analysis at hand is that the acoustic features are bad at predicting schizophrenia, values for the area under the ROC was at the highest 0.65 which could be categorized as being at least poor if not bad. This is because an area under the ROC of 0.5 corresponds to the model having no discriminatory ability at all. Therefore, when we compared our models to this value of 0.5, we evaluated if our models were better than randomly guessing.

Portfolio 4

Daniel, Jesper, Pernille KJ and Astrid

Link to code:

https://github.com/StudiegruppeEM3/Portfolio_4/blob/master/A4_P1_PhysiologicalCoordination_instructions.Rmd

Preprocessing (P)

It is important to clean heart rate (HR)-data, because there can be quite a lot of artifacts, that effect our data. Also, it is useful to downsample the dataset, to make it easier to work with.

Downsampling (A)

Our heat-rate measures are on a millisecond scale, but the human heart doesn't beat that fast, so many of the data points are redundant. Therefore, we down-sample, which makes the datasets more manageable. We do this by grouping the datapoints into groups of 100 and calculate the means of these. These means are our new data-points. This makes the dataset 100 times smaller.

Outliers (A)

Next, we consider outliers. Outliers are data-points are outside the normal range of our data. The most obvious case of outliers in HR-time-series, is if your HR changes more than is physically possible. However, some artifacts fall within the physically possible range.

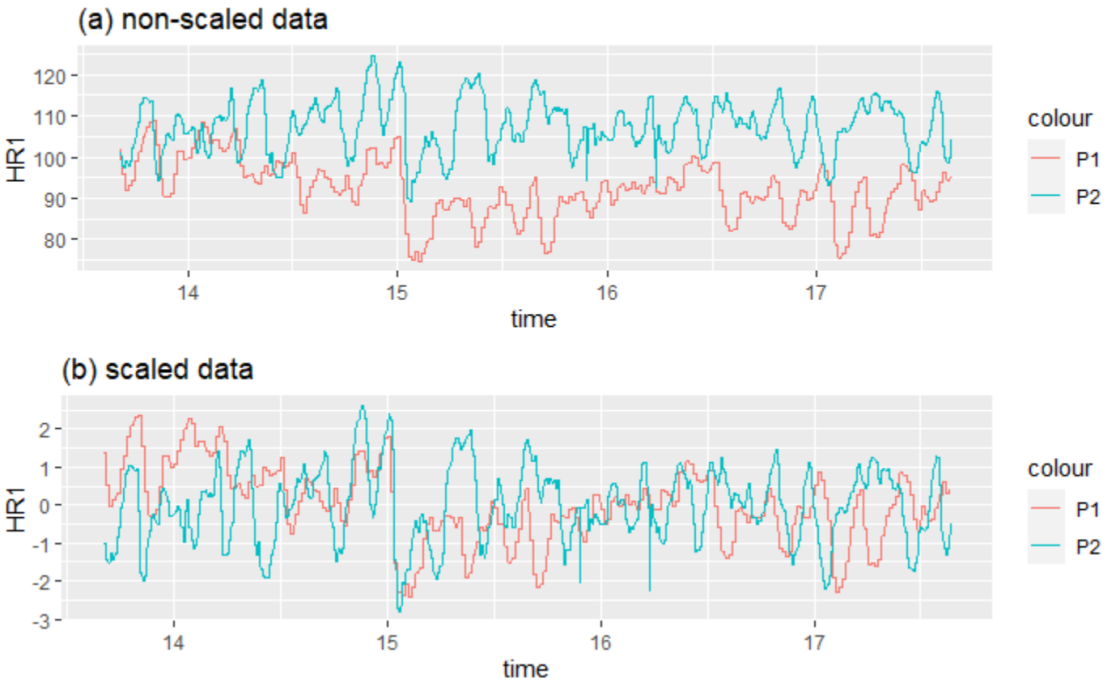
To deal with these outliers, we replaceable datapoints that are more than 2.5 standard deviations (SD) from the mean with the mean-value. We also considered the approach where one makes the values over 2.5 SD from the mean to the exact value of 2.5 SD over or under the mean, one might argue that this approach could keep the integrity of the data, but we think that the outliers are due to artifacts and not reflect actual changes in heart rate.

Plot 1; shows different ways of controlling for outliers, (a) shows no outlier correction. (b) shows how outliers, defined as 2.5 standard deviations from the mean, gets moved to be the mean value instead. (c) shows how the data would look if one would make outliers, again 2.5 standard deviations from the mean, to the exact value of 2.5 times the standard deviation over or below the mean.

Scaling (A)

Next, we scale the data. This is to ensures that the two participants are on the same scale, see plot2.

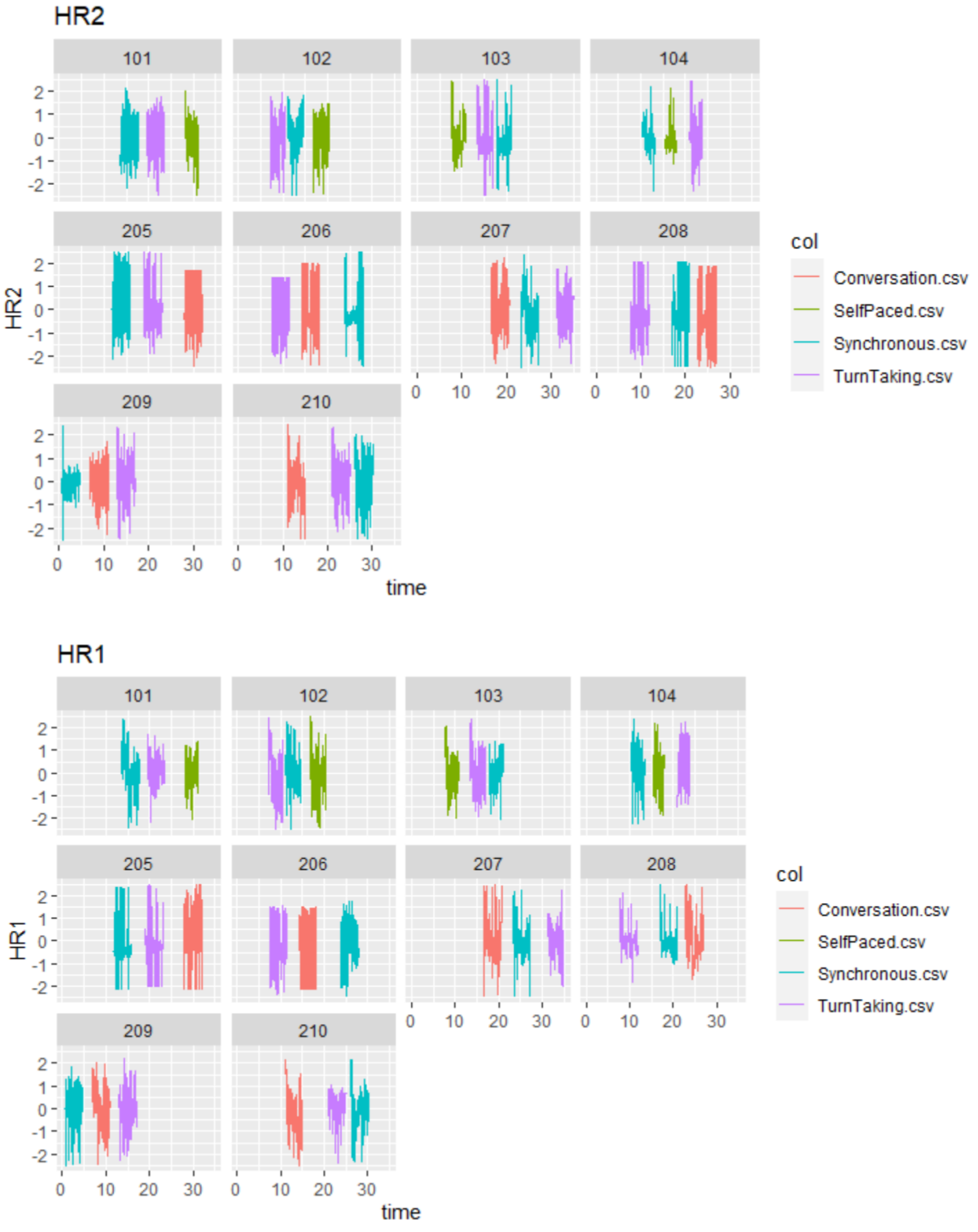
For example, it could be that P1 is in better shape, and therefore generally has a lower heartrate.



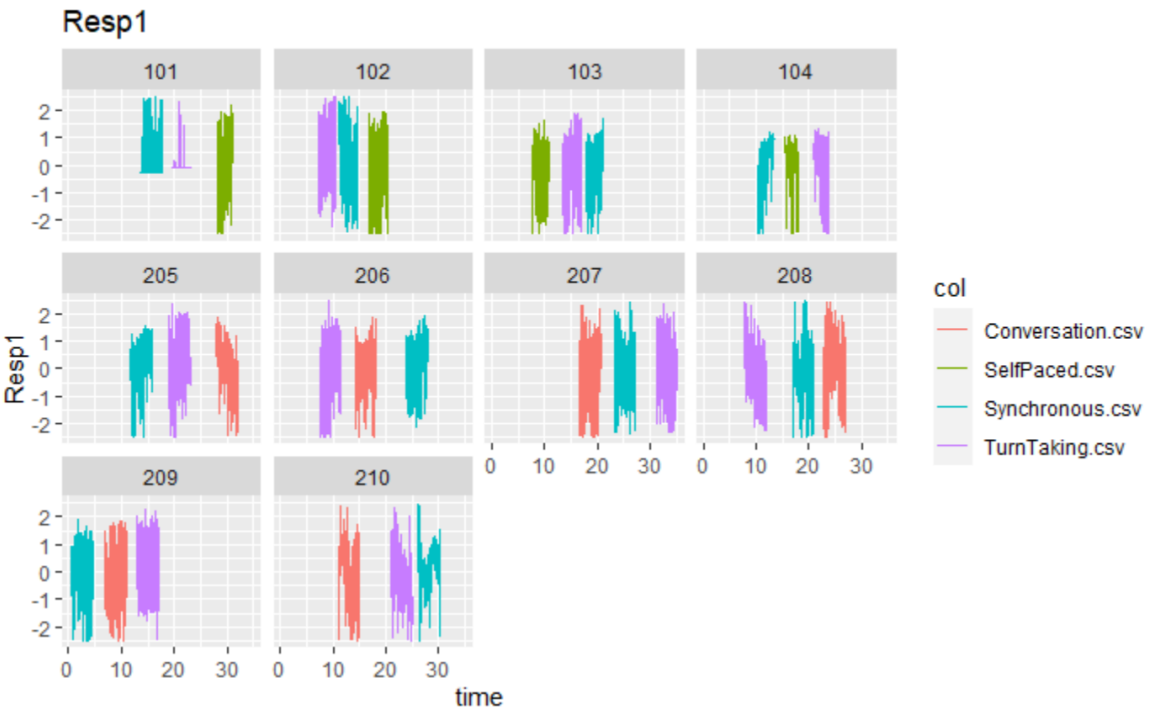
Plot2; shows how scaling the data, puts the two participants heart rate data on the same scale. (a) shows non-scaled data, whereas (b) shows scaled data.

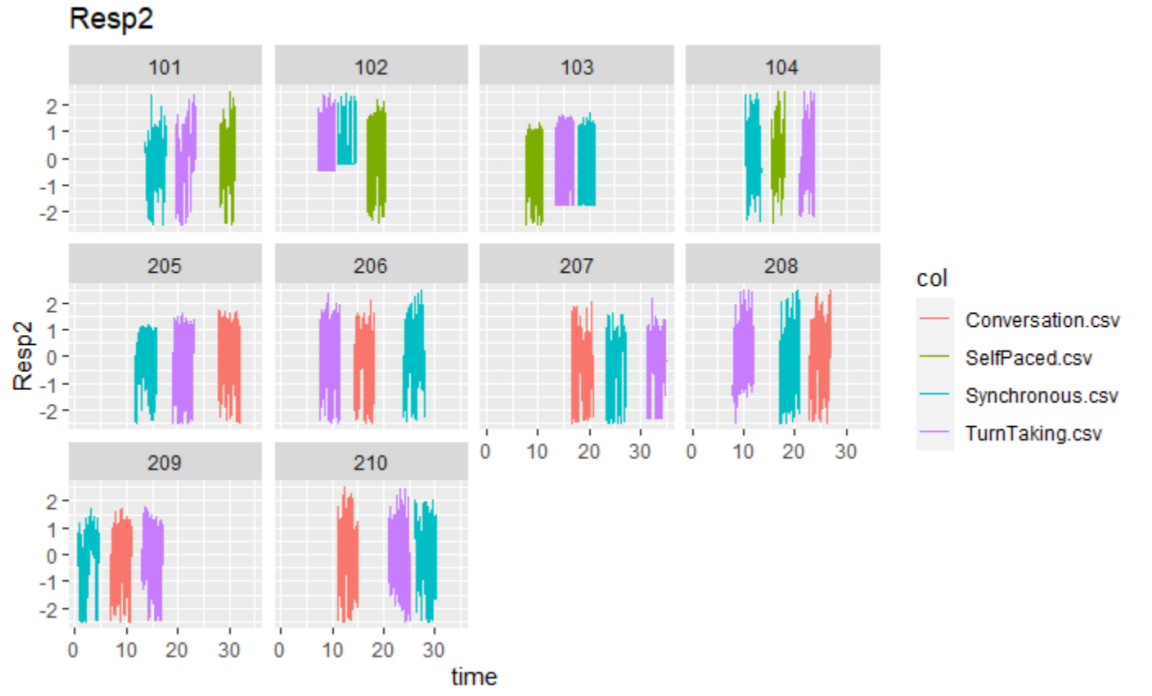
Whole dataset (P)

After we have pre-processed the whole dataset, we plot the data to visually inspect it for errors. In plot 3 we display only the data from study 1 and 2, to save space.



Plot 3a; displays scaled heart rate for the individual participants (HR1 / HR2) in studies 1 and 2 in the different conditions.





Plot 3b; displays scaled respiration data for the individual participants (Resp1 / Resp2) in studies 1 and 2 in the different conditions.

The Resp1 data looks a bit odd for participant 101 (study 1, group 1) during the synchronous and turn taking conditions (see Plot 3b). We remove the respiration data for this participant in these two trials.

Analysis (D)

When analyzing the data, we use two differential equations from Ferrer & Helm, 2013.

$$\frac{dHR_{self}}{dt} = a_1 * (HR_{ideal} - HR_{self_{lag}}) + a_2(HR_{other_{lag}} - HR_{self_{lag}})$$

$$\frac{dHR_{other}}{dt} = b_1 * (HR_{ideal} - HR_{other_{lag}}) + b_2(HR_{self_{lag}} - HR_{other_{lag}})$$

If a_1 and b_1 are positive the change in heartrate (either in oneself or the other person) will move towards the ideal HR for that participant (HR_ideal). Therefore, the values a_1 will tell us whether our participants self-regulate and b_1 will tell us whether our participants co-regulate.

For the purpose of our analysis, we will combine these two equations into one. This means we only get two estimates instead of four. We assume that our first estimate is an average of a_1 and b_1 and our second estimate will be an average of a_2 and b_2. Furthermore, in our models we assume that the HR_{ideal} term is the mean of the heart rate, and since we have scaled our heart rate data this term is 0.

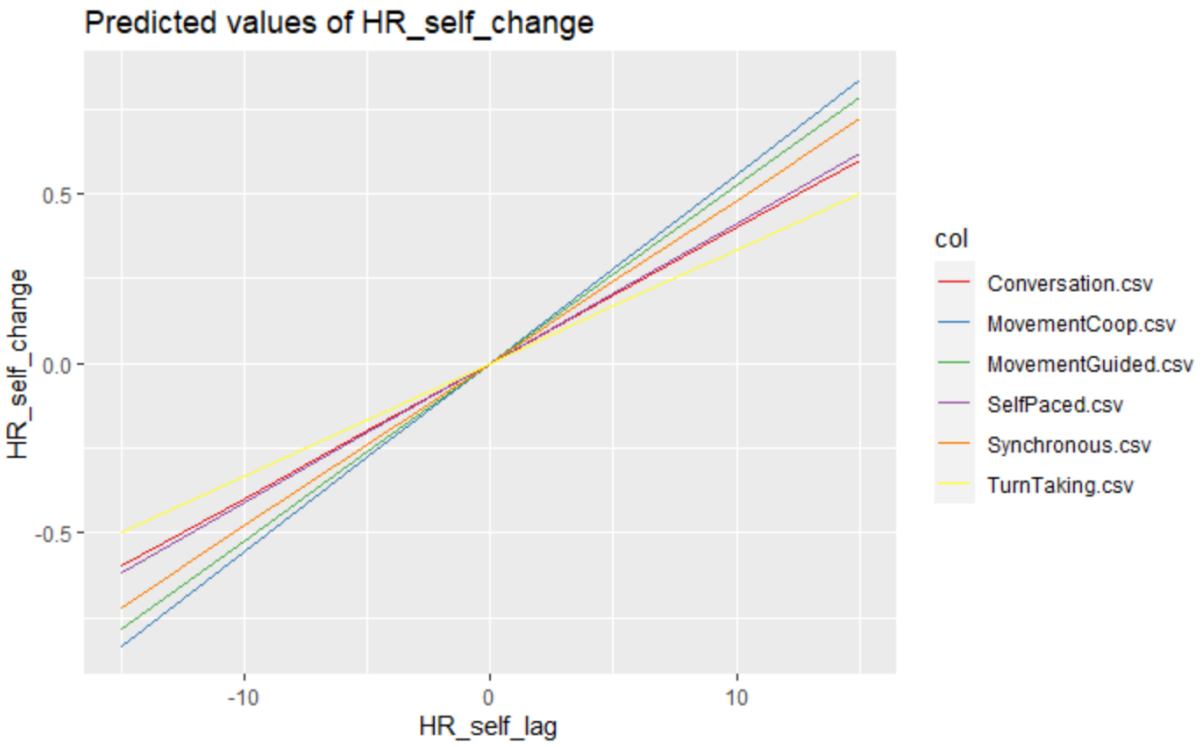
First we made a model that predicted change in heart rate from $-HR_{self}\_lag$ and from the difference between the others heart rate and HR_{self} (dif). We also included participant as a random intercept.

	Estimate	Std. error	Df	t-value	p-value
Intercept	0.00014	0.00046	356100	0.3	0.76
self-regulation	0.0414	0.00067	356100	62.12	0
Co-regulation	0.0013	0.00046	356100	2.866	0.004

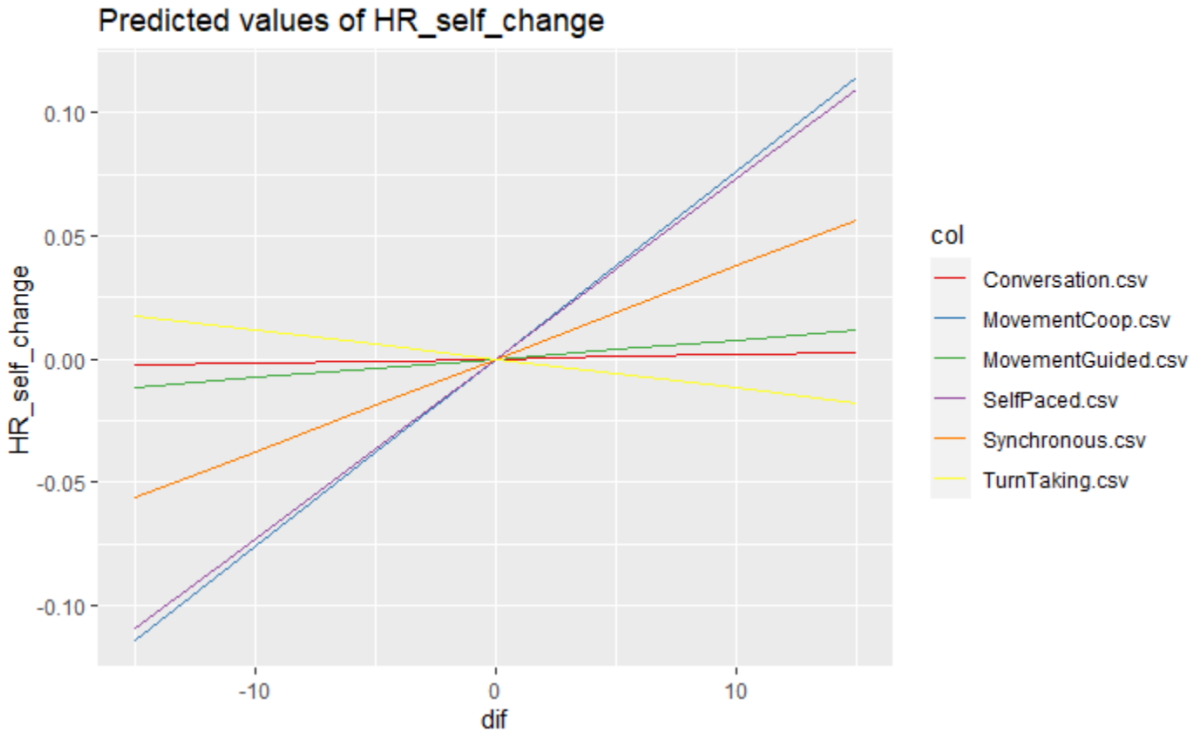
Table 1

Table one shows the output from this model. Since both estimates are positive, it seems like the participants self-regulate and co-regulate.

Next, we included condition as an interaction effect as well as random slope per participant. In this model it is important to note that all the interaction estimates are compared to 0 (and not a baseline condition).



Plot 4a; visually displays the slope coefficients of self-regulation from table2, where positive slopes indicate self-regulation



Plot 4b; visually displays the slope coefficients of co-regulation from table2, where positive slopes indicate co-regulation

	Estimate	Std. Error	Df	t-value	p-value
Intercept	0.00014	0.00046	356100	0.3	0.76
Self-regulation, conversation	0.0399	0.0013	356100	30.8	>0.001
Self-regulation, movement coop	0.0556	0.0033	356100	16.8	>0.001
Self-regulation, movement guided	0.0523	0.0029	356100	18.0	>0.001
Self-regulation, self-paced	0.0413	0.0036	356100	11.4	>0.001
Self-regulation, synchronous	0.0479	0.0012	356100	38.6	>0.001
Self-regulation, turn taking	0.0332	0.0011	356100	28.3	>0.001
Co-regulation, conversation	0.00018	0.0009	356100	0.2	0.84
Co-regulation, movement coop	0.0076	0.0021	356100	3.562	>0.001
Co-regulation, movement guided	0.00077	0.0020	356100	0.38	0.7
Co-regulation, self paced	0.0073	0.0023	356100	3.14	>0.01
Co-regulation, synchronous	0.0038	0.0008	356100	4.44	>0.001
Co-regulation, turn taking	-0.0012	0.0008	356100	-1.4	>0.16

Table 2

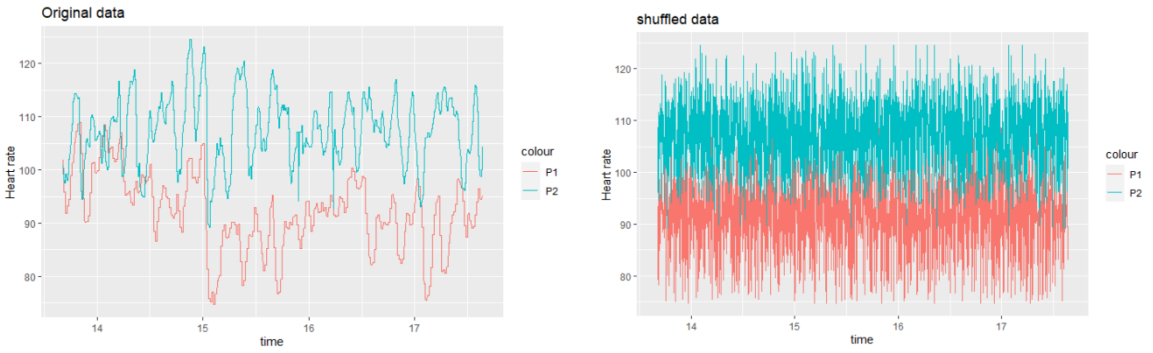
As can be seen from table 2 and the plots 4a and 4b, the participants self-regulate in all conditions (all estimates for self-regulation are positive)

Next, we see that in some of the conditions the two persons co-regulate (synchronous, self-paced and movement-Coop conditions). We also see that the two persons (anti) co-regulate in the condition Turn-Taking, this can be seen from the negative estimate, however this effect is non-significant.

Shuffled dataset (J)

It could be that these effects are due to the sequence of the data. To break this temporal dependency, we create a new random order of the heart rate datapoints (shuffle). We analysis this new dataset in the same way, that we did for the original dataset. Finally, we make a model that compares that uses shuffle/original as an interaction effect, with shuffle as the baseline.

We expect the self-regulation to be much higher in the shuffled dataset, since the effect on oneself is more drastic, when the order is random. See plot 5. Here its easily visible that in the shuffled data-set it looks like the participants self-regulate more dramatically which we would see with increased values of the slopes for self-regulation. Due to this fact we can't use the shuffled control data to compare self-regulation, but only co-regulation.



Plot 5; shows the difference in the same participants in the shuffled data-set compared to the original data-set

Plot 6b: visually displays the slope coefficients of co-regulation from table2, where positive slopes indicate co-regulation in both shuffled controls and the original data.

	Estimate	Std. Error	Df	t-value	p-value
Intercept	0.0258	0.0042	28.68	6.1	>0.001
Co-regulation, conversation	0.0005	0.0033	712200	0.15	0.88
Co-regulation, movement coop	0.0016	0.0075	712200	0.21	0.83
Co-regulation, movement guided	0.0137	0.0078	712200	1.7	0.08
Co-regulation, self-paced	0.0219	0.0090	712200	2.4	0.02
Co-regulation, synchronous	0.0023	0.0032	712200	0.7	0.47
Co-regulation, turn taking	-0.0060	0.0032	712200	-1.9	0.06

Table 3: regression estimates for co-regulation with shuffled controls as baseline.

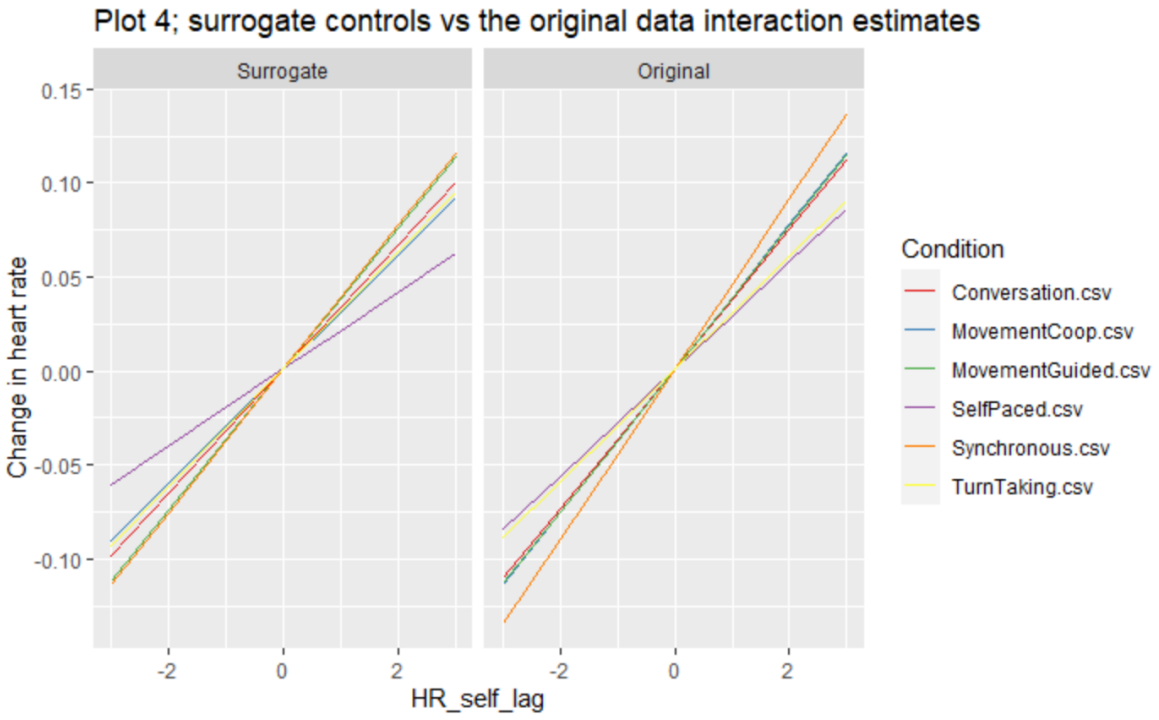
Co-regulation is now only significant in the self-paced condition (see table 3). This co-regulation is positive, meaning that the person does co-regulate more with their partner compared to a shuffled control. One can also see that movement-guided also is close to significant p = 0.08 also in a positive direction. One can also see that the turn-taking condition is close to being significant this however in a negative direction, again meaning that the two person anti-co-regulate.

Surrogate pairs (J,D,P,A)

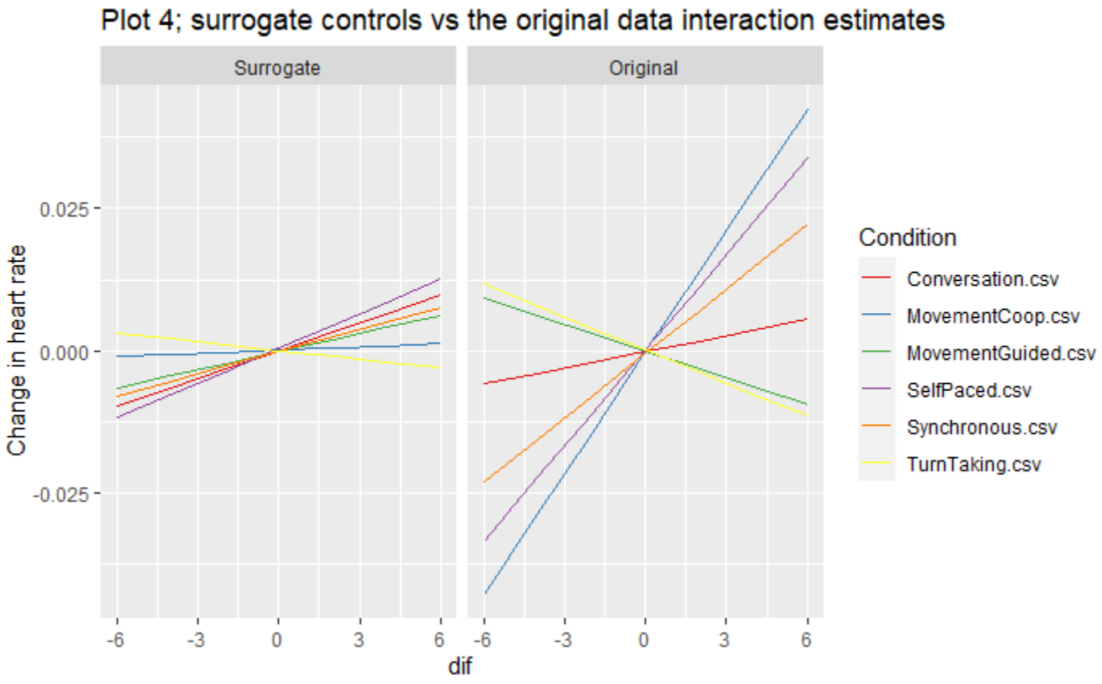
Finally, to check that the participants are effect by each other and not just the task that they are performing, we created surrogate pairs. Surrogate pairs are combinations of datasets from two participants who completed the same task but with different partners.

Next, we created a model, that compares the surrogate pairs to the original dataset, just like we compared the original dataset to the shuffled dataset.

We expected our self-regulation estimates to be identical to the original data, since participants were still being compared to themselves. However due to some data loss in the preprocessing stages the estimates vary a little.



Plot 7a: visually displays the slope coefficients of self-regulation from table2, where positive slopes indicate self-regulation in both surrogate controls and the original data.



Plot 7b: visually displays the slope coefficients of co-regulation from table2, where positive slopes indicate co-regulation in both surrogate controls and the original data.

	Estimate	Std. Error	Df	t-value	p-value
Intercept	0.0012	0.0001	2490000	8	>0.001
Co-regulation, conversation	-0.0007	0.0009	2490000	-0.7	0.46
Co-regulation, movement coop	0.0068	0.0021	2490000	3.3	>0.001
Co-regulation, movement guided	-0.0026	0.0021	2490000	-1.2	0.22
Co-regulation, self paced	0.0035	0.0027	2490000	1.3	0.19
Co-regulation, synchronous	0.0025	0.0009	2490000	2.8	0.005
Co-regulation, turn taking	-0.0014	0.0009	2490000	-1.6	0.10

Table 4: regression estimates for co-regulation with surrogate pairs as baseline

From table 4, we can see that participants seem to co-regulate in the conditions: movement coop, synchronous.

All in all our analyses suggest that the effect of co-regulation is very small and not very consistent. In the shuffled control analysis we found significant effects for self-paced, which was non-significant in the surrogate pairs control condition. In the surrogate pair analysis we found significant co-regulation in the synchronous and movement coop condition, which was not found in the shuffled control analysis.

Another interesting finding was that in the turn-taking condition all the estimate were negative, both in the shuffled control analysis and the surrogate pairs' analysis. This means that the participants anti-coregulate.

We aren't quite sure why we don't find better results, it may be due to the fact that there is no effect present or maybe our use of inadequate analysis that could provide us with to low statistical power, one would preferably use the package "CRQA" in R.

Limitations of surrogate pairs as control baseline:

There may be some issues with time when we combine two participants who did not do the task at the same time. For example, for the turn-taking condition pairs might have switched at slightly different times.

Effects of respiration coordination on heart rate coordination (J,D,P,A)

To test the effects of respiration coordination on heart rate coordination we would need a longer data format. So, like we merge the equations (1) and (2), we would need to include the respiration data as well. The data-structure would look like;

X
HR_self
HR_other
Resp_self
Resp_other

Then we would make columns for change and lag and make a model that would look something like this:

Change of X ~ (HR_other-HR_self) * (Resp_other-Resp_self)*condition

We would use this model to test whether respiration coordination effects heartrate coordination.

If the interaction effect is significant, respiration coordination effects heartrate coordination. Adding the condition interaction allows us to check if this effect is different for different conditions.

We are unsure how we would do this in R and how the other columns (fixed effects) would look like.

Bibliography.

Ferrer, E., & Helm, J. L. (2013). Dynamical systems modeling of physiological coregulation in dyadic interactions. *International Journal of Psychophysiology*, *88*(3), 296–308. https://doi.org/10.1016/j.ijpsycho.2012.10.013

Portfolio 5: Building on the shoulders of giants: meta-analysis

by Astrid, Daniel, Jesper and Pernille KJ

https://github.com/StudiegruppeEM3/Portfolio_5/blob/master/A5_MetaAnalysis.rmd

What is the current evidence for distinctive vocal patterns in schizophrenia? Report how many papers report quantitative estimates, comment on what percentage of the overall studies reviewed they represent (see PRISMA chart) ***your method to analyze them****, the estimated effect size of the difference (mean effect size and standard error) and forest plots representing it. N.B. Only measures of pitch mean and pitch sd are required for the assignment (all the other measures have been removed from the dataset for the sake of simplicity).*

(Astrid, Pernille KJ)

From the PRISMA 2009 flow-diagram, we know that the meta-analysis included 46 papers out of 4341 studies identified which were not replicates, this means 1.06% of the total amount of studies were included.
46/4341 * 100% = 1.06%

We decided to investigate the mean value of pitch and the pitch variability. Out of the 46 studies, we found 15 studies eligible for our pitch variability analysis and 6 studies that matched for mean pitch analysis. For both analyses, we included the article ID, study ID, sample size, and type of task.

To analyze these studies, we made two different models: one for mean pitch and one for pitch variability. Firstly we calculate the standardized mean difference and the sampling variance for all the studies in both analyses. We then plot the effect sizes to investigate our data as boxplots:

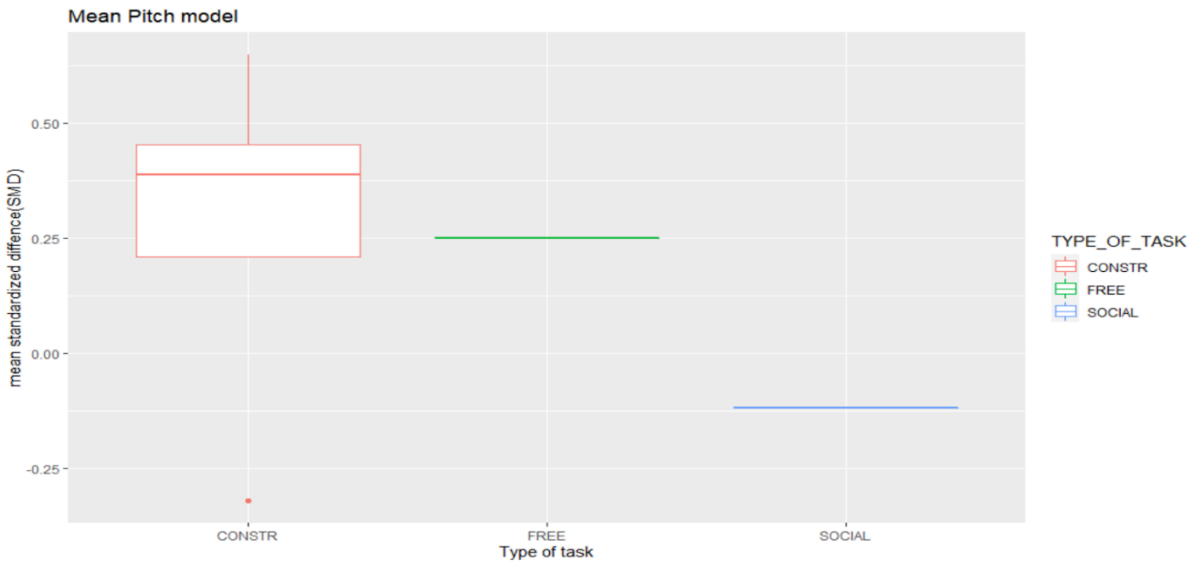


figure 1; showing a boxplot of the effect sizes for the different types of tasks for the mean pitch model

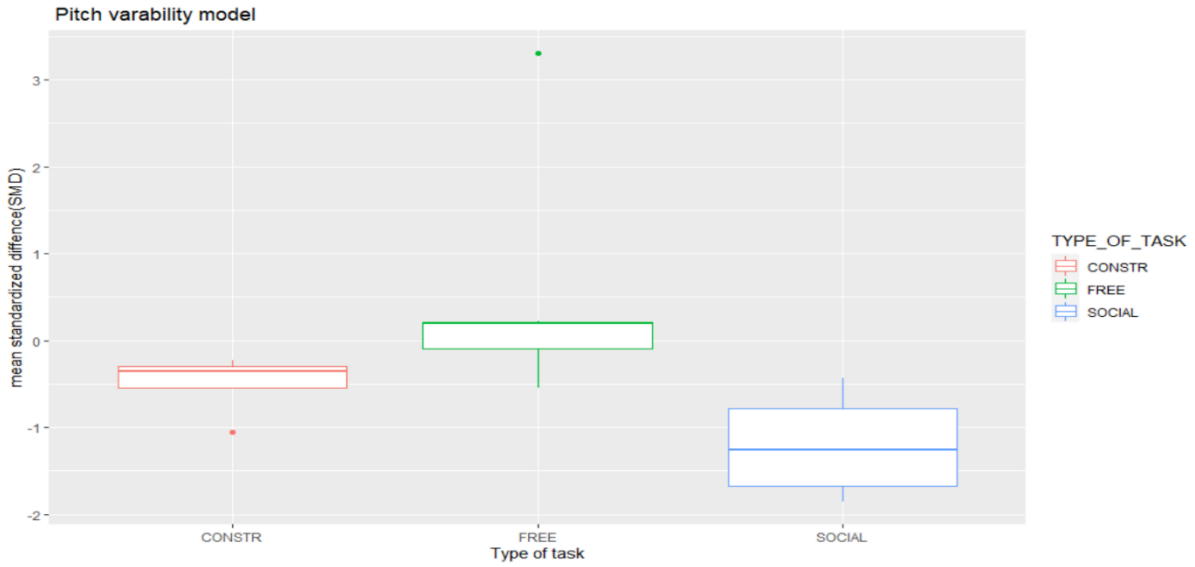


figure 2; showing a boxplot of the effect sizes for the different types of tasks for the pitch variability model

As can be seen from the boxplots above, it looks like there is an outlier in the pitch variability model in the FREE type task. We will keep this in mind when interpreting our forest plot.

Models

We made linear mixed effect models of the standardized mean difference, weighted with their sampling variance with study as a random intercept. We also made models including the type of task as a fixed effect. The results can be seen below.

linear mixed effect models with no fixed effects:

In the mean pitch model the effect size was 0.21 with a standard error of 0.15 which was calculated from 5 degrees of freedom resulting in a non-significant effect because of the p-value of 0.22

In the pitch variability model the effect size was -0.2 with a standard error of 0.36 which was calculated from 10.5 degrees of freedom resulting in a non-significant effect because of the p-value of 0.59.

linear mixed effect models with type of task as fixed effect we chose to have the constrained condition as baseline.

The mean pitch model:

	Estimate	standard error	df	t-value	p-value
Intercept	0.28	0.21	3	1.3	0.28
Task = free	-0.03	0.47	3	-0.06	0.96
Task = social	-0.40	0.47	3	-0.85	0.46

tabel 1; displays the summary output of the linear mixed effects model for mean pitch with type of task as fixed effect and study as a random intercept.

The pitch variability model:

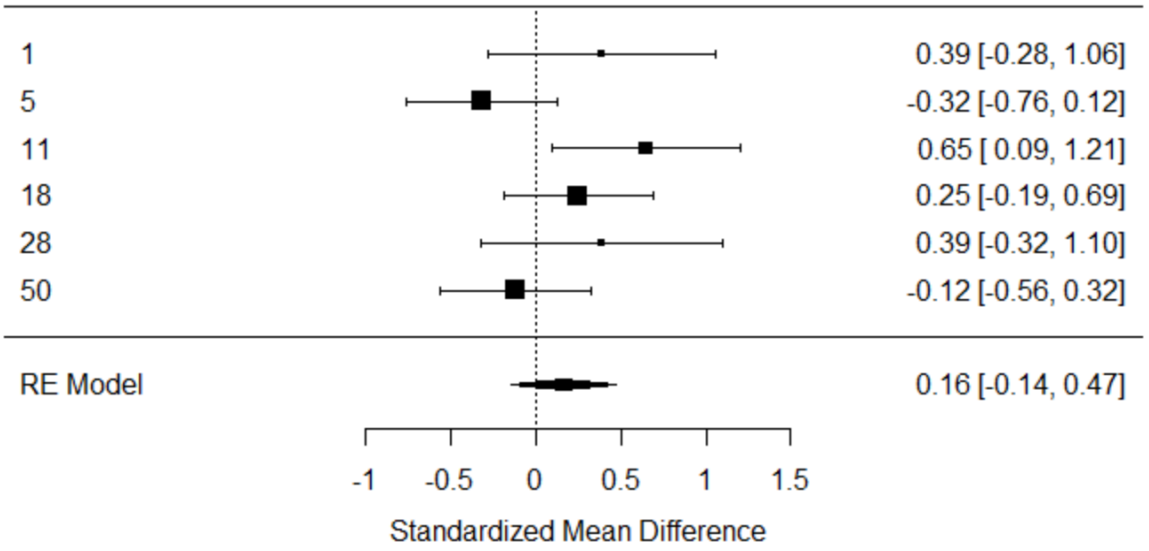
	Estimate	standard error	df	t-value	p-value
Intercept	-0.5	0.4	10.4	-1.2	0.24
Task = free	1.1	0.61	10.1	1.8	0.10
Task = social	-0.6	0.12	2.1	-4.9	0.03

tabel 2; displays the summary output of the linear mixed effects model for pitch variability with type of task as fixed effect and study as a random intercept.

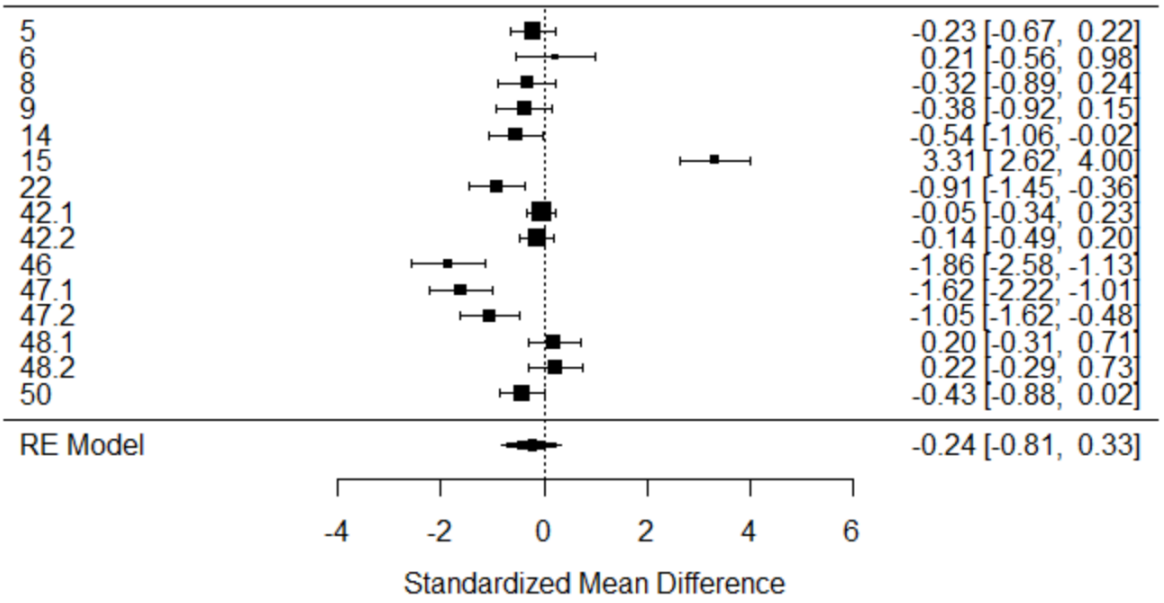
We then made use of the metaphor package to use the function rma. This function completes a meta-analysis via linear (mixed effects) models and compares the outputs of these results to the normal linear mixed effect models here we excluded type of task as a fixed effect for simplicity in our forest plots.

Using the rma function the mean pitch model's effect size became 0.16 with a standard error of 0.16
The pitch variability model's effect size became -0.24 with a standard error of 0.29

We then made forest plots for these two models.



Plot 1: forest plot of the standardized mean differences and their confidence intervals for each study for the mean pitch model.



Plot 2: forest plot of the standardized mean differences and their confidence intervals for each study for the pitch variability model with type of task as a fixed effect.

Interpretation:

Firstly when interpreting our results we can see an obvious outlier which one might want to look further into, in the pitch variability model. Here it would be a good idea to check the specific study to see whether a typo or some other error has occurred. While including this outlier the results from our analyses show that schizophrenic patients seem to have a higher mean pitch, but this effect is very small SMD = 0.16, with a lot of uncertainty, indicated by the confidence interval of (-0.14 , 0.47).

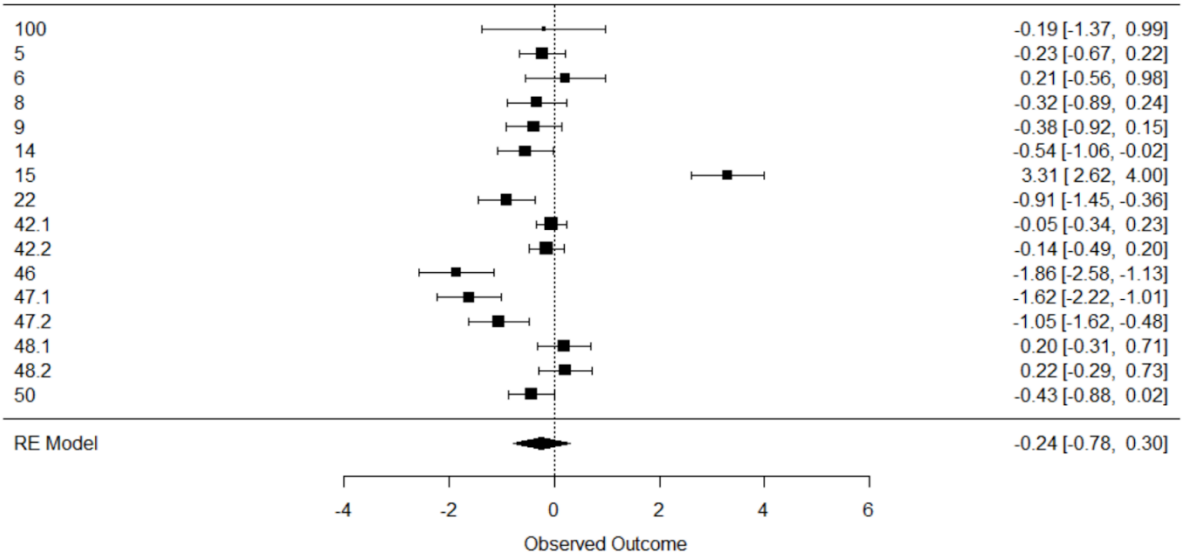
The results also show that schizophrenic patients seem to have a lower pitch variability in contrast to what is normally hypothesised, but this effect is again very small SMD = -0.24, with a lot of uncertainty, indicated by the confidence interval of (-0.81 , 0.33).

From this initial analysis it should be concluded that there is no difference in mean pitch and pitch variability in schizophrenic patients compared to health controls. However an argument could be made for collecting more data because in the pitch variability model it seems to be the case that there could be big differences between types of tasks.

Do the results match your own analysis from Assignment 3? If you add your results to the meta-analysis, do the estimated effect sizes change? Report the new estimates and the new forest plots.

(Jesper)

Our results from assignment 3 showed that the effect size of pitch variability in schizophrenic patients compared to healthy controls was -0.19 with a standard error of 0.6, this analysis showed that this effect size was -0.24, which is very comparable, this can also be seen in the forest plot below where our results from assignment 3 has been included.



plot 3; Showing the forest plot of the meta-analysis with our assignment 3 results displayed as study 100.

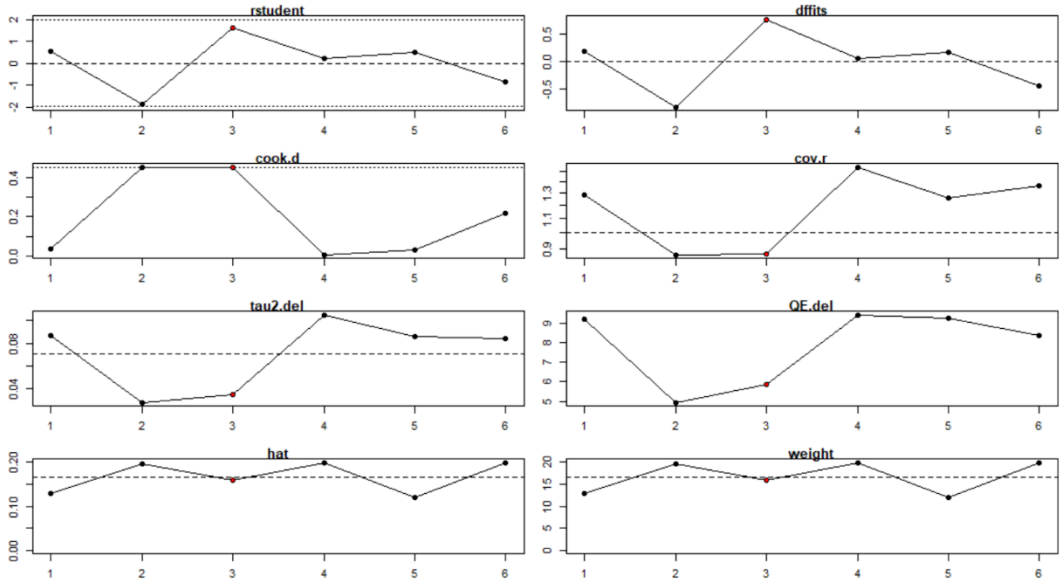
when including this new data the effect size of the meta-analysis did not change SMD = -0.24, however the confidence intervals became a little tighter, (-0.78 , 0.30).

Quality assessment of the literature

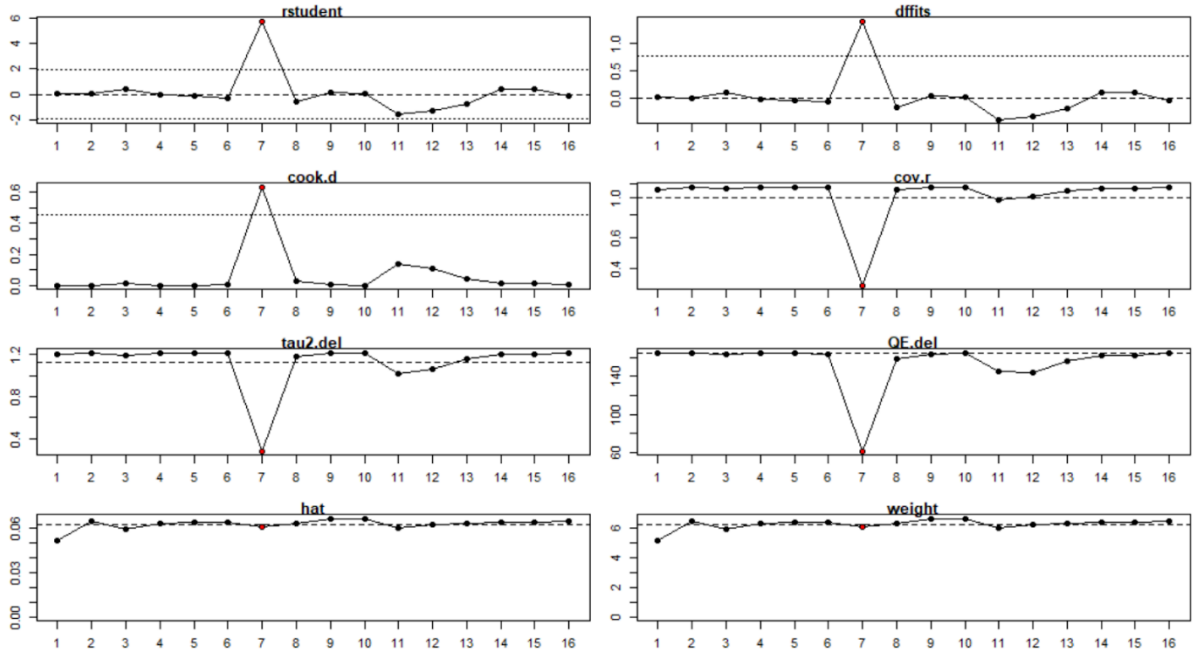
Assess the quality of the literature: report and comment on heterogeneity of the studies (tau, I2), on publication bias (funnel plot), and on influential studies.

(Daniel)

To get the quality of the literature and comment on the heterogeneity of the studies we looked at the output of our meta-analysis model. For our mean pitch model the values of tau and I^2 came out to be, tau = 0.267 and I^2 = 50% while for the pitch variability model the outcome was tau = 1.1 and I^2 = 94.4%. What these values tell us is that there is a moderate amount of heterogeneity in the mean pitch model and very high heterogeneity in the pitch variability model. This very high variance in the studies in the pitch variability model is partly due to the outlier we detected earlier. We can also see this outlier in plot XXX for the pitch variability model.

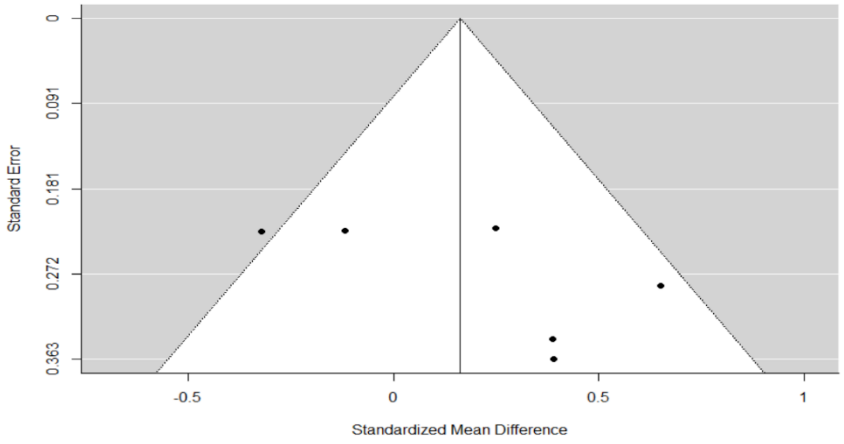


plot 4; influence plot for the studies on the mean pitch model. One can see that there is no clear outliers

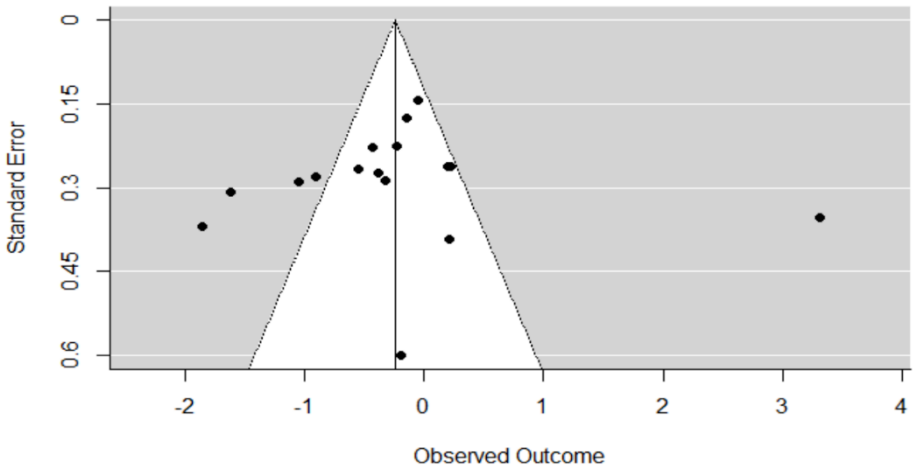


plot 5; influence plot for the studies on the pitch variability model. One can see that there is a clear outlier present (study no. 7)

To investigate whether the literature contains publication bias we created funnel plots for both our models. Looking at the funnel plots below we see that it looks like there seem to be no publication bias present, this can be seen because the funnel plots seem to be more or less symmetrical. however the outlier mentioned earlier is also present here, in the pitch variability model.



plot 6: Funnel plot of the mean pitch model



plot 7: Funnel plot of the pitch variability model