Experimental Methods 3:

Exam portfolio

by:
Bella Terragni- 201405868,
Peter Andreas Mikkelsen Thramkrongart-
201806892,
Rūta Slivkaitė- 201805872 &
Bianka Szöllősi- 201808610

20.12.2019.

Bella Terragni- 201405868,
Peter Thramkrongart- 201806892,
Rūta Slivkaitė- 201805872 &
Bianka Szöllősi- 201808610

Language Development in ASD

link to github for data cleaning:
https://github.com/szbianka/Experimental-Methods-3-exam
link to github for assignment 2:
https://github.com/bellaterragni/assignment_2

Assignment 2, part 2, Experimental Methods 3

1. Data

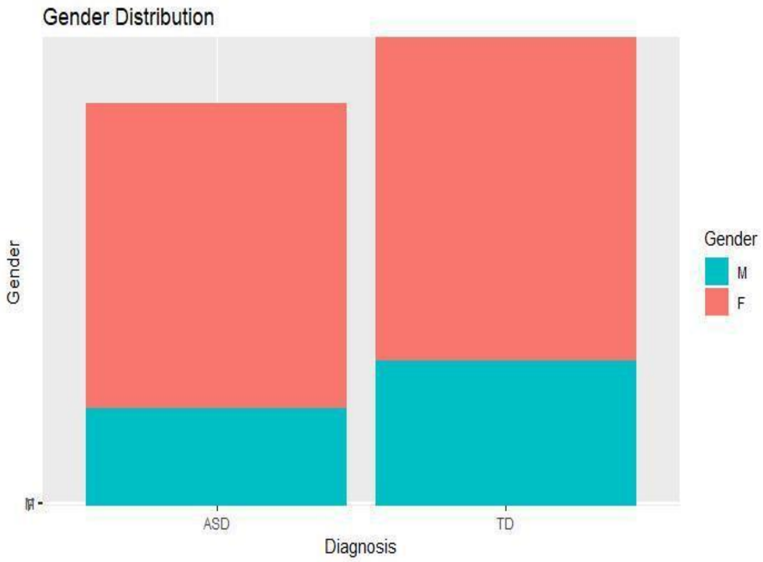


Figure 1

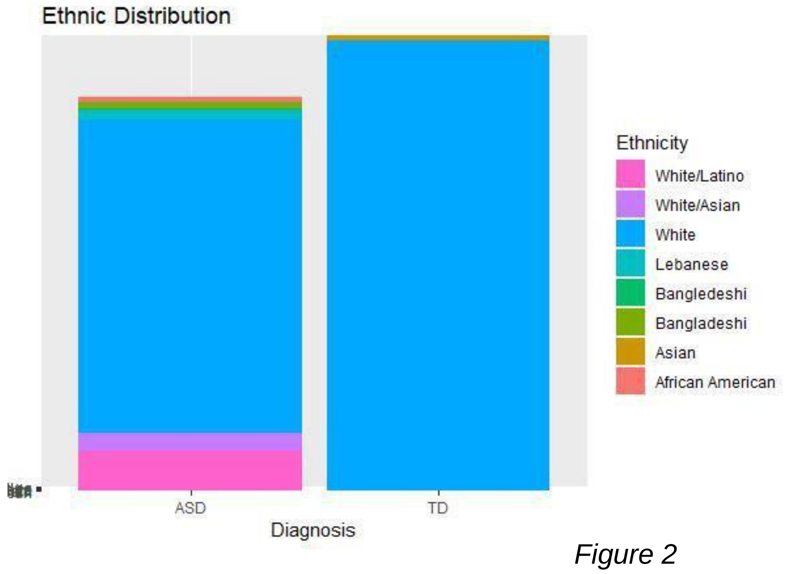


Figure 2

The data used in this paper is collected from videotaped play sessions between a child and their parent. The children were either diagnosed with ASD or typically developing. The parent-child dyads were videotaped over six sessions with four months between each visit. The videos have been transcribed and the dialogue has been analyzed so that mean length of utterances (MLU) is calculated for each session as well as other linguistic specificities. Other than that the children were tested for verbal and non-verbal IQ as well as other factors, which may influence the MLU.

The sample included 61 children, of which 32 were typically developing (TD; Mean age = 20.4 months, F = 26, M = 6) and 29 children had autism disorder (ASD; Mean age = 32.9 months, F = 25, M = 4). The sample sizes and ratio gender are well matched between the two subject groups, but the children's ages are not. This is part of the methodological deliberations, however, since the TD and ASD children are matched according to linguistic performances and not their ages, when the subject groups are formed. A noteworthy concern, though, is the gender ratio. While 61 children participated in the study, only 10 of them were boys. This calls into question any predictions a prospective model might generate about diagnosing male individuals. This is due to tangible differences in how boys' and girls' language develop[1], females having superior verbal abilities at a young age.

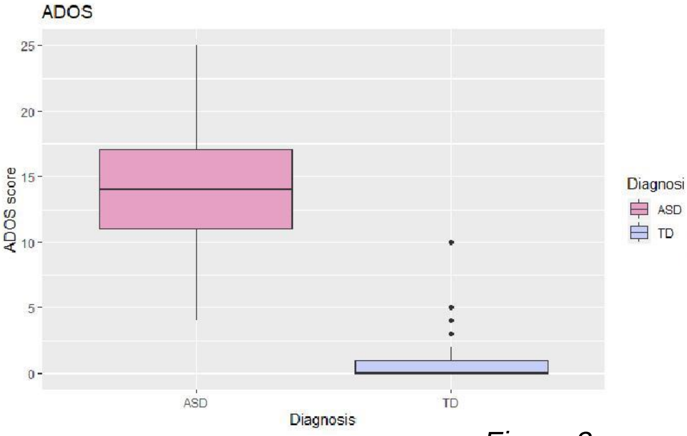


Figure 3

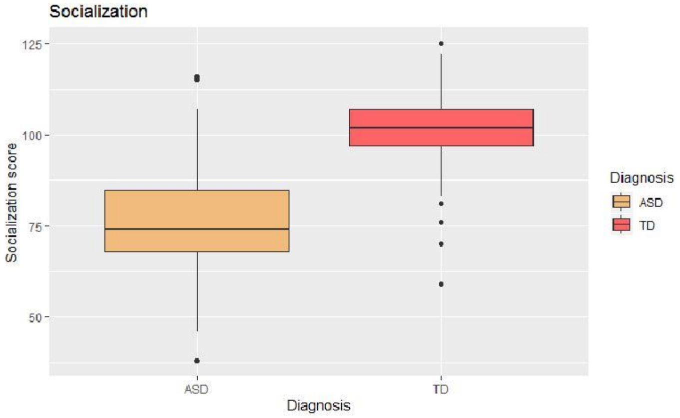
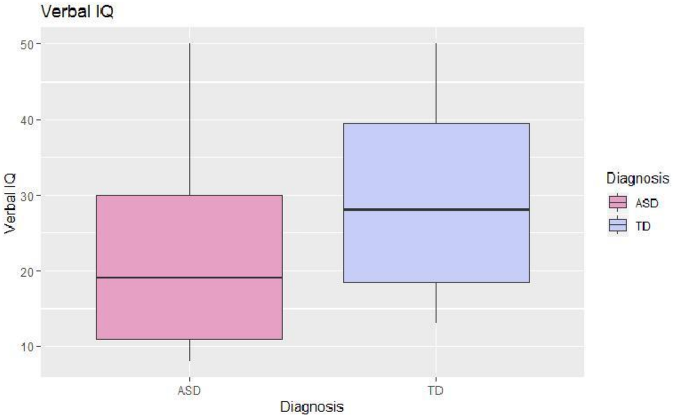


Figure 4



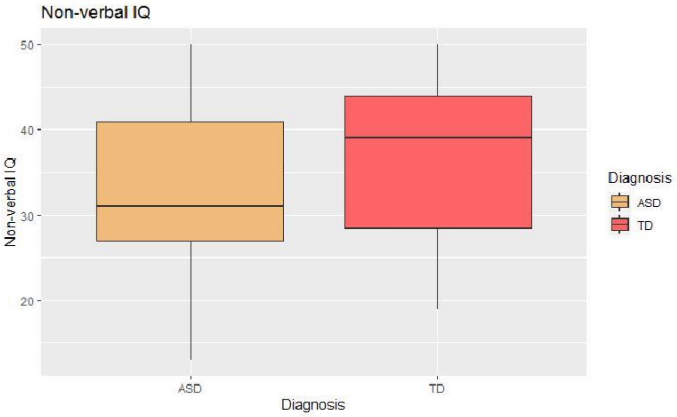


Figure 5 *Figure 6*

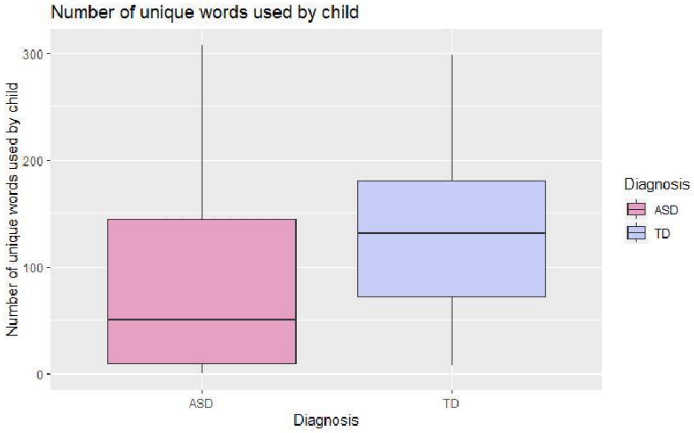


Figure 7

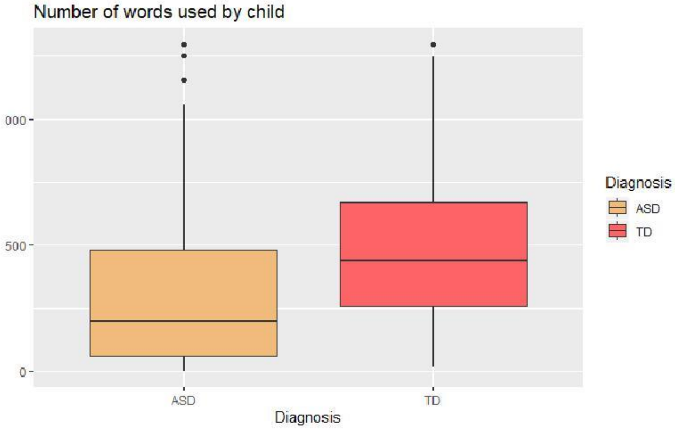


Figure 8

[1] Bouchard, C., Trudeau, N., Sutton, A., Boudreault, M., & Deneault, J. (2009). Gender differences in language development in French Canadian children between 8 and 30 months of age. *Applied Psycholinguistics, 30*(4), 685-707. doi:10.1017/S0142716409990075

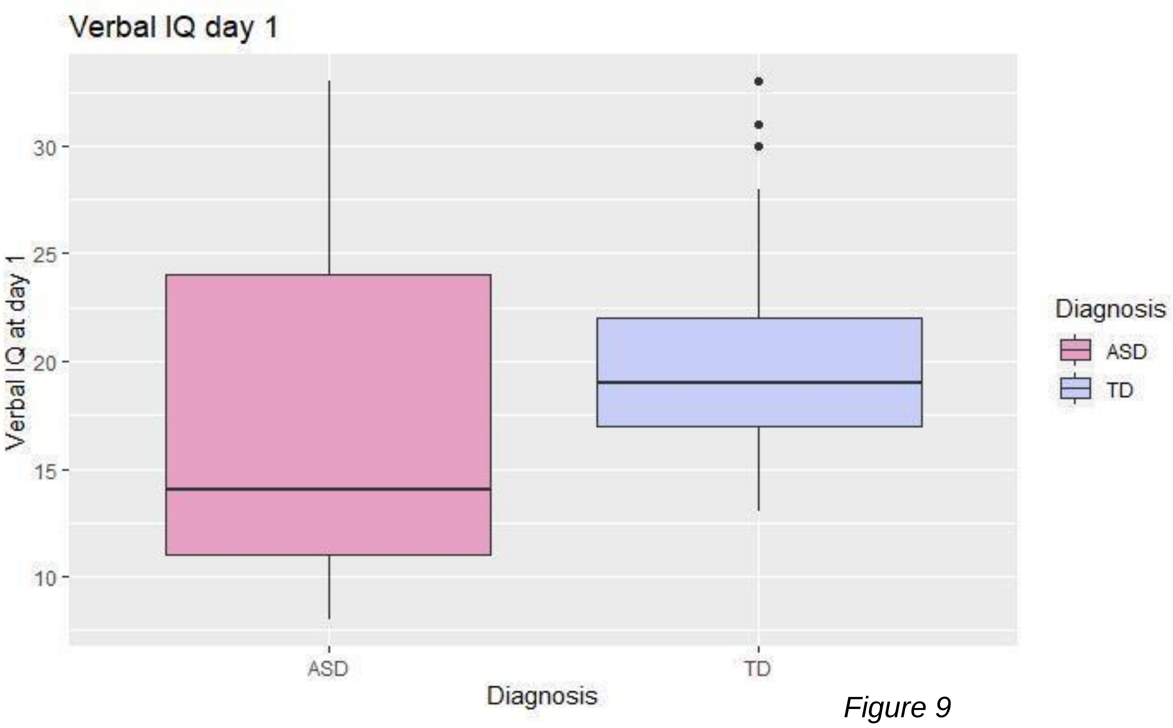


Figure 9

All relevant information about the distribution of characteristics between the two subject groups are visualized in the plots *figure 1-8*. When looking at the barcharts and boxplots, it is evident that there are some differences between the two subject groups. As expected, children with ASD scored lower on nonverbal IQ, verbal IQ, total and unique words used as well as mean length of utterances. On the surface this seems to contradict the fact that the children have been verbally matched, but if we take a closer look at the data from the first visit, we see that the children are, in fact, verbally matched (see *figure 9*). However, due to TD children developing their language proficiencies quicker, ASD children fall behind at later visits. The largest differences between ASD and typically developing children, however, is their ADOS and socialization scores (see boxplots *figure 3 and 4*). The stark difference in ADOS scores is to be expected since it is a tool for diagnostic assessment of autism spectrum disorder and therefore is highly correlated with the diagnosis.

2. Hypotheses

H1A: A child’s MLU changes over time

H1B: A child’s MLU changes according to diagnosis

H2A: A parent’s MLU changes over time

H2B: A parent’s MLU changes according to diagnosis

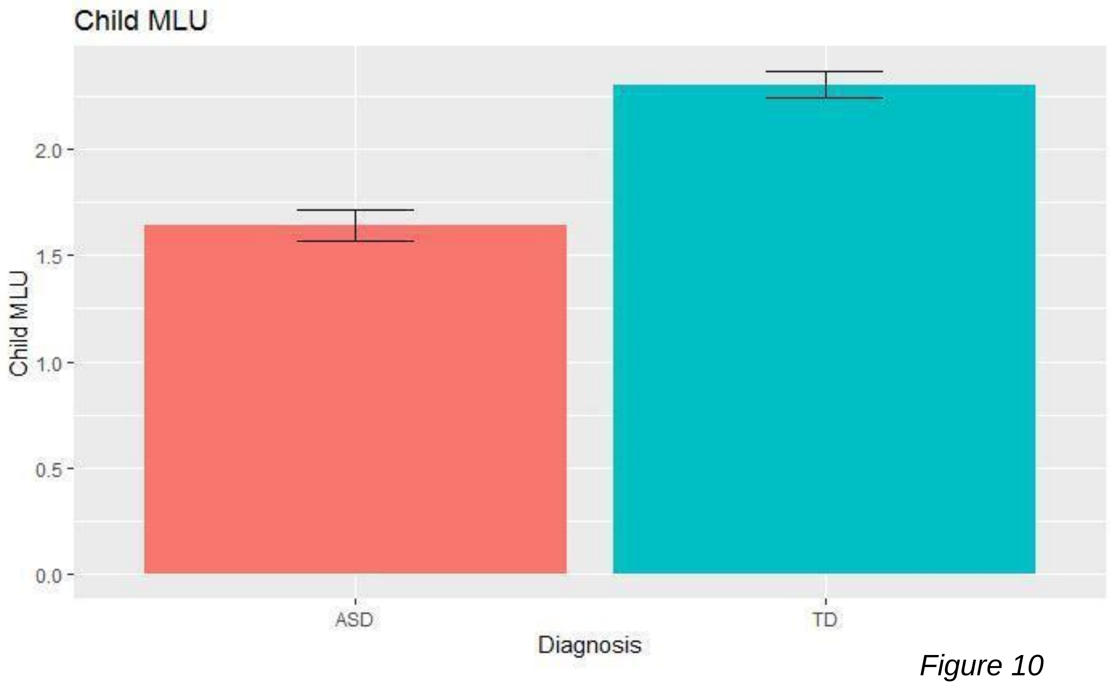


Figure 10

3. Visualization

It is prudent to visualize the data before commencing the analysis. This visualization informs us about the data in a more crude but also easily interpretable way. Figure 11 depicts the MLU of all 61 children using box plots. Looking at the figure it is evident that despite the general tendency of ASD children to have lower MLUs (see also figure 10), individual ASD children score high on the chart. At the bottom of the chart in figure 11 there is a collection of

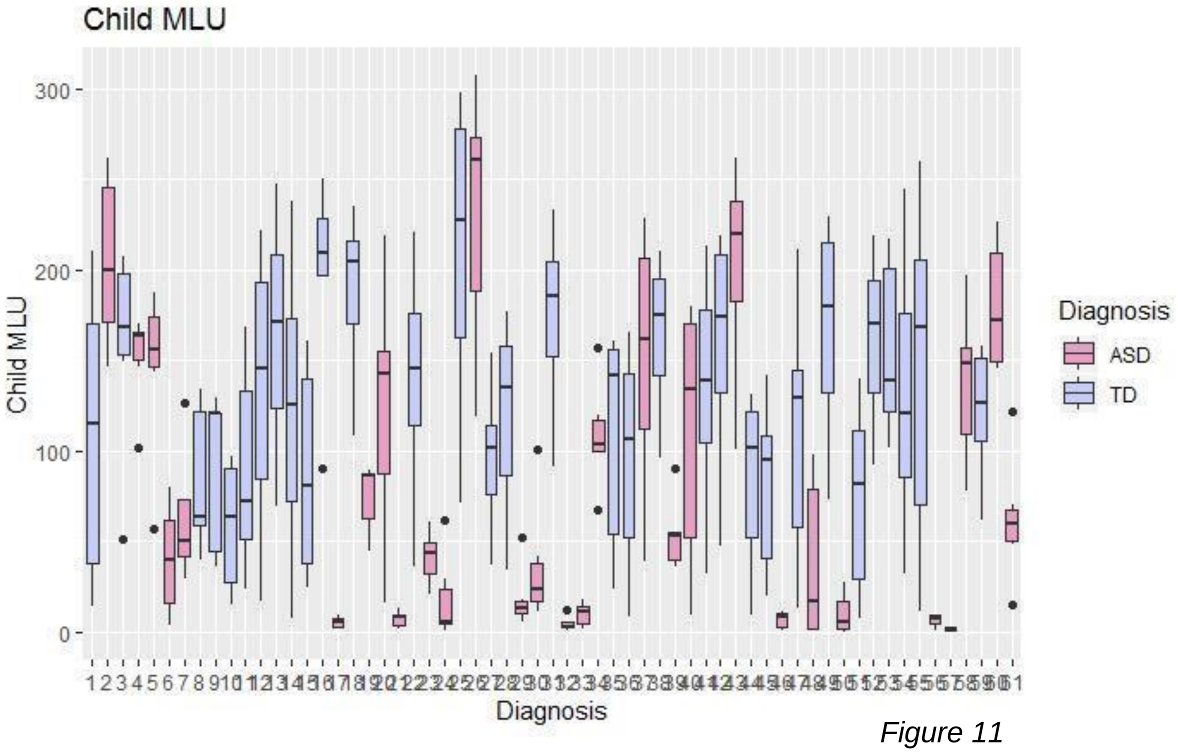


Figure 11

particularly low-scoring ASD children who do not display much spread in their mean length of utterances. These children might be close to non-verbal and their language development might therefore be particularly stunted compared to the other ASD children.

Looking at figure 13, which depicts the linear relationship between mean length of utterance and time, the general tendency seems to be that TD children's capacities for creating longer utterances develop quicker than ASD children's. Since ASD children are especially variable (see figure 12) it makes sense to incorporate the individual variance of the children by visualizing the individual linear relationships between MLU and time. Figure 14 depicts each child's individual progression in regards to utterance length as linear functions. When comparing the development of the TD and ASD children in the study (figure 14), the TD children have homogenous trajectories, while the ASD children do not. Overall the linear representations of the ASD children's MLU are set lower and most have only moderate increments, while a few even appear to degenerate.

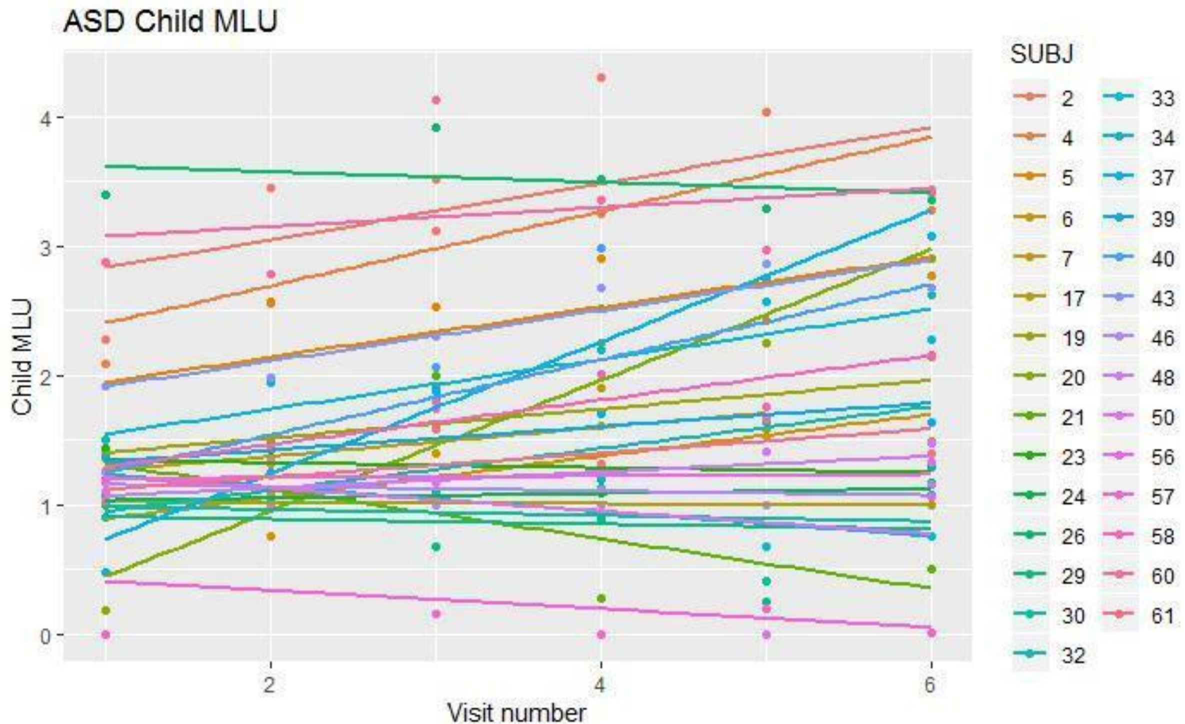


Figure 12

4. Analysis

Hypotheses H1A and H1B

In order to test our hypotheses, different models were constructed. Since we had multiple data points from the same participants, linear mixed effects models were chosen to retain variance within individuals. With model 1a, we tested hypothesis H1A, whether children's mean length of utterance inflates over time, that is over visits. The predictor is thus visit while child MLU is the dependent variable.

Model 1a: child MLU ~ visit + (1|ID)

According to model 1a, visit number has a significant effect on the MLU of a specific child, (= 0.23, t(291.46) = 14.52, p < .0001). The model thus suggests that children's MLU's increase over time, which might be a function of their language development. This is evidence in favor of hypothesis H1A: that a child's MLU changes over time.

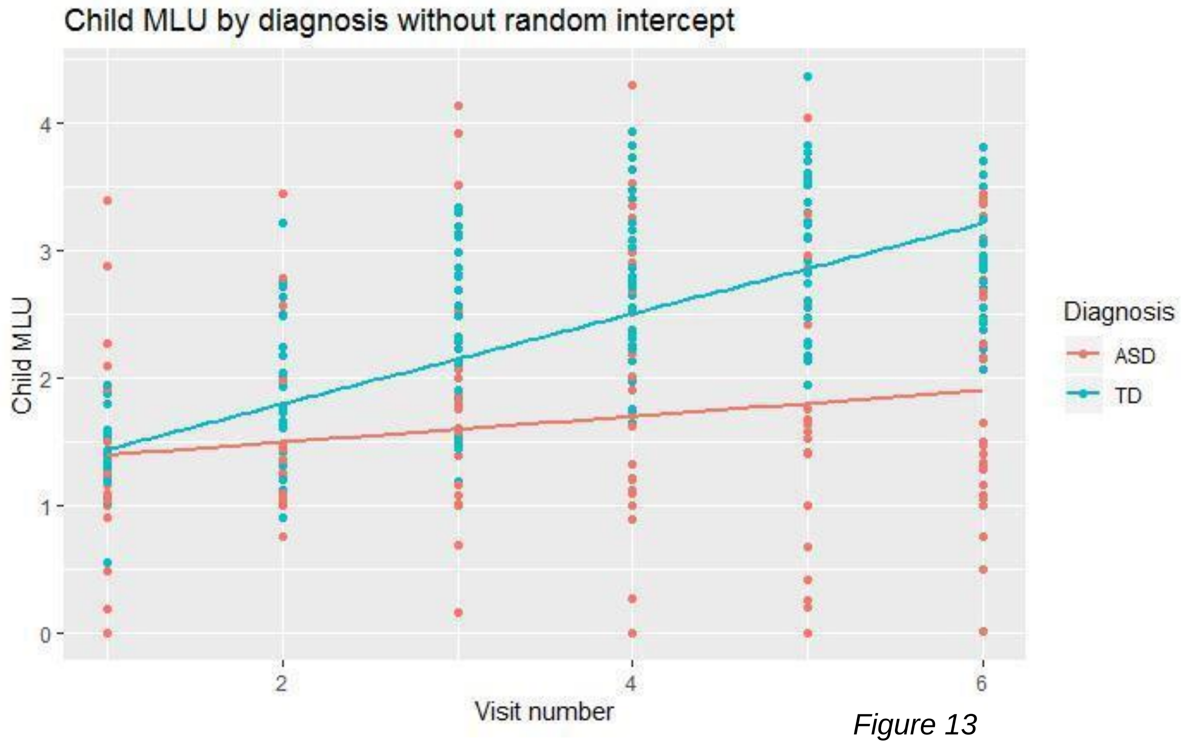


Figure 13

Our second model, model 1b, is a linear mixed effects model, since we are still managing multiple data points from the same participants and want to encompass this variance. In order to construct the model, child MLU is selected as the dependent variable, while diagnosis and visit are added as fixed effects and ID as a random effect. We could also have chosen a model with only diagnosis as predictor, but since visit was a significant predictor in model 1a, we add visit as a fixed effect in model 1b.

Model 1b: child MLU ~ diagnosis + visit + (1|ID)

According to model 1b, child MLU is dependent on diagnosis (= 0.65, t(60.76) = 4.06 p < .0001) and visit number (= 0.23, t(291.36)=14.53, p < .0001). Another variation is model 1c that takes a possible interaction effect between diagnosis and visit into account.

Model 1c: child MLU ~ diagnosis * visit + (1|ID)

The statistical analysis of model 1c shows a significant effect of visit number on child MLU (= 0.09, t(291.23)=4.92, p < .0001). There was likewise a significant result for the interaction effect between diagnosis and visit (= 0.25, t(291.28)=8.96 p < .0001), while there was no significant effect of diagnosis alone on child MLU (= -0.21, t(109.93)=-1.14 p = .26). These results show that the effect of the visit is dependent on diagnosis, that is, the effect of time on MLU is different between ASD and TD children.

A final variation of the model is model 1d, where a random slope has been added dependent on visit number. Now child MLU is the predictor variable, while diagnosis is a fixed effect and visits as well as ID's are random effects.

Model 1d: child MLU ~ diagnosis * visit + (1|ID) + (0 + VISIT|ID)

According to this model, visit is significant (= 0.10, t(80.95) = 3.87 p < .0001) as well as the interaction effect between visit and diagnosis (= 0.25, t(81.68) = 7.01 p < .0001) . In order to test whether model 1c or 1d is superior, we run an anova comparison on the two models.

Model 1d explains a larger percentage of the variance (χ^2(7, 1) = 23.762, *p* < .0001).

These three models, 1b, 1c, 1d, all capture significant parts of the variance, and they have in common that they all support that a child's MLU increases over time (hypothesis H1A) and that diagnosis has a significant effect on the trajectory (hypothesis H1B) whether by itself as model 1b implies or as an interaction effect with visit as model 1c and 1d suggests. The implications of such an interaction effect is that the effect of time on the child's MLU differs depending on diagnosis.

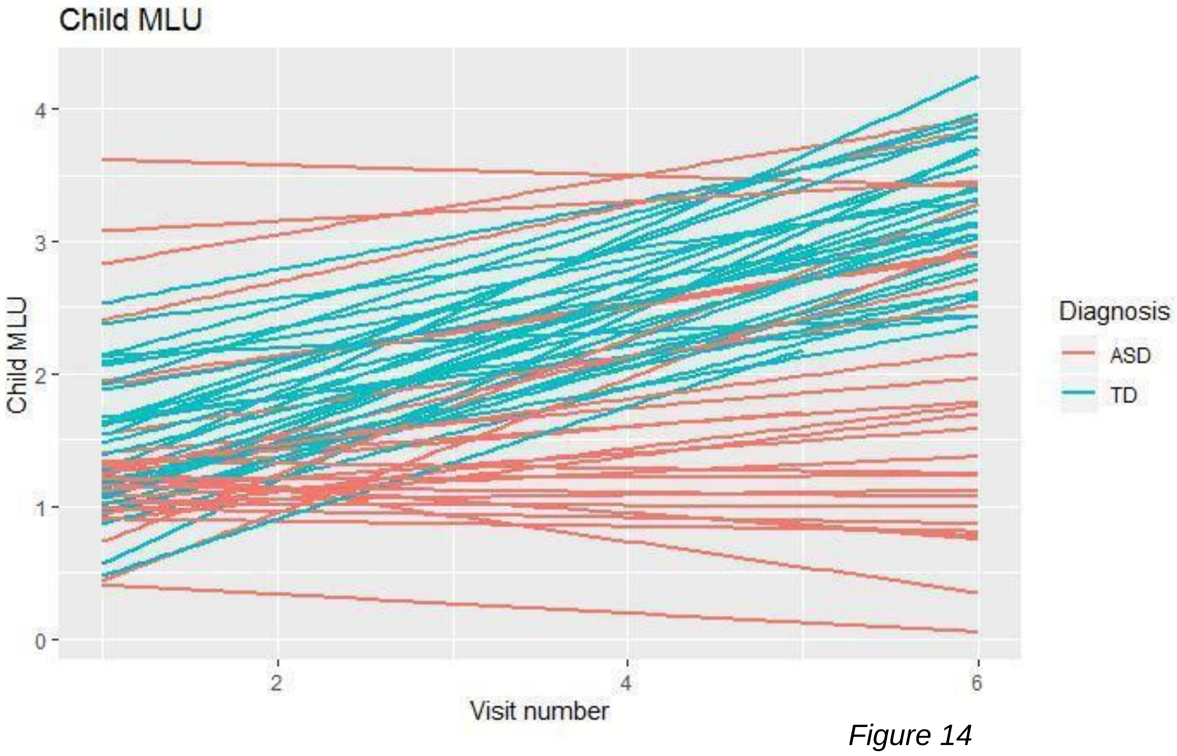


Figure 14

Hypotheses H2A and H2B

When investigating whether the progression of time and child diagnosis affect the parents' mean lengths of utterances, similar mixed effects models are utilized. Firstly, however, it is prudent to take a look at figure 16, which displays the individual trajectories of the parent MLU's. It is quite evident that the ASD parent MLU's are lower on average than those of TD parents. This is also visualized in the boxplot of figure 15.

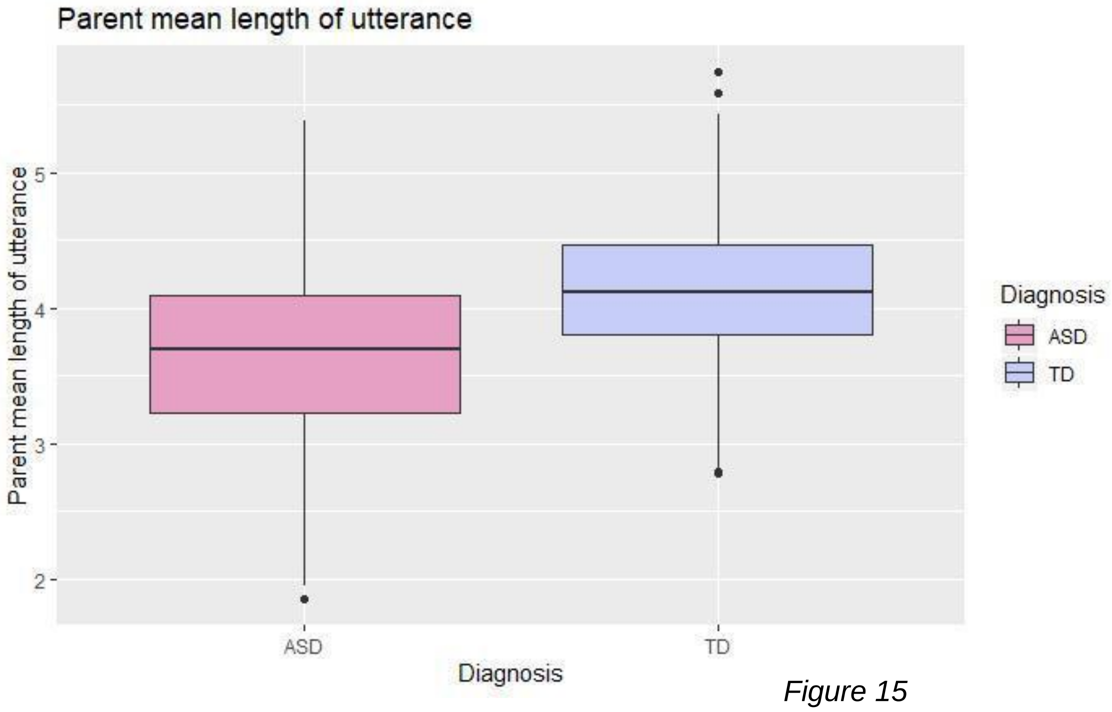


Figure 15

In order to test H2A properly, that is, whether the mother's MLU is dependent on the visit number, a simple model 2a is formulated with participant ID as a random effect and visit number as a fixed effect, while the dependent variable is parent MLU.

Model 2a: parent MLU ~ visit + (1|ID)

According to model 2a, visit number has a significant effect on the parents MLU (= 0.12, t(291.57) = 8.66, SE = 0.014, p < .0001). The model thus suggests that the parent's MLU increase over time, at about half of the pace that their children increase their respective MLU's. These results suggest that hypothesis H2A has validity and that a parent's MLU increases over time.

Model 2b is a mixed effects model with random intercepts. Participant ID is a random effect while diagnosis and time are both independent variables and fixed effects. We have chosen simplicity over complexity in this model, since we would prefer to capture the larger picture and have interpretable results as opposed to possibly overfitting and juggling vague multiway interactions.

Model 2b: parent MLU ~ diagnosis + visit + (1|ID)

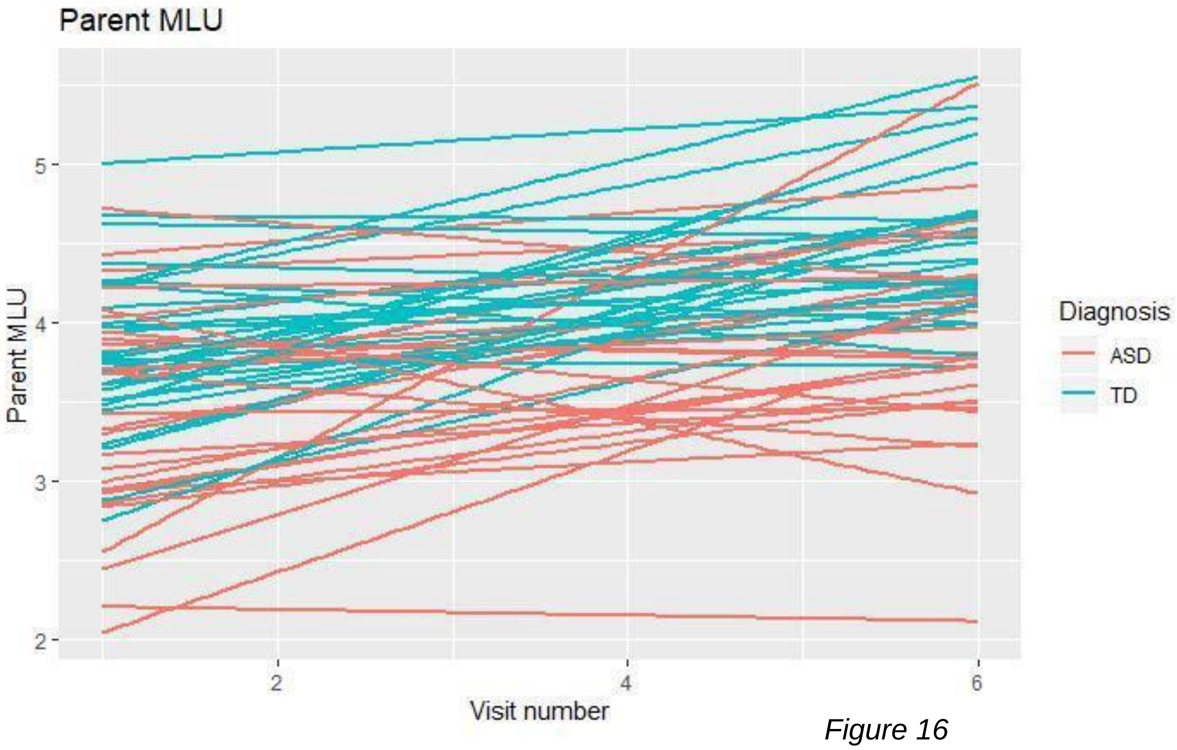


Figure 16

According to the statistical analysis of model 2b, parent MLU is not only dependent on time (= 0.12, t(291.72) = 8.68, SE = 0.014, p < .0001), but also diagnosis (= 0.50, t(60.85) = 4.35, SE = 0.11, p < .0001). Interestingly, the mean length of utterance was shorter for parents with ASD children, which indicates that having ASD children affected how their parents communicated with them. One could conjecture that parents with ASD children simplified their

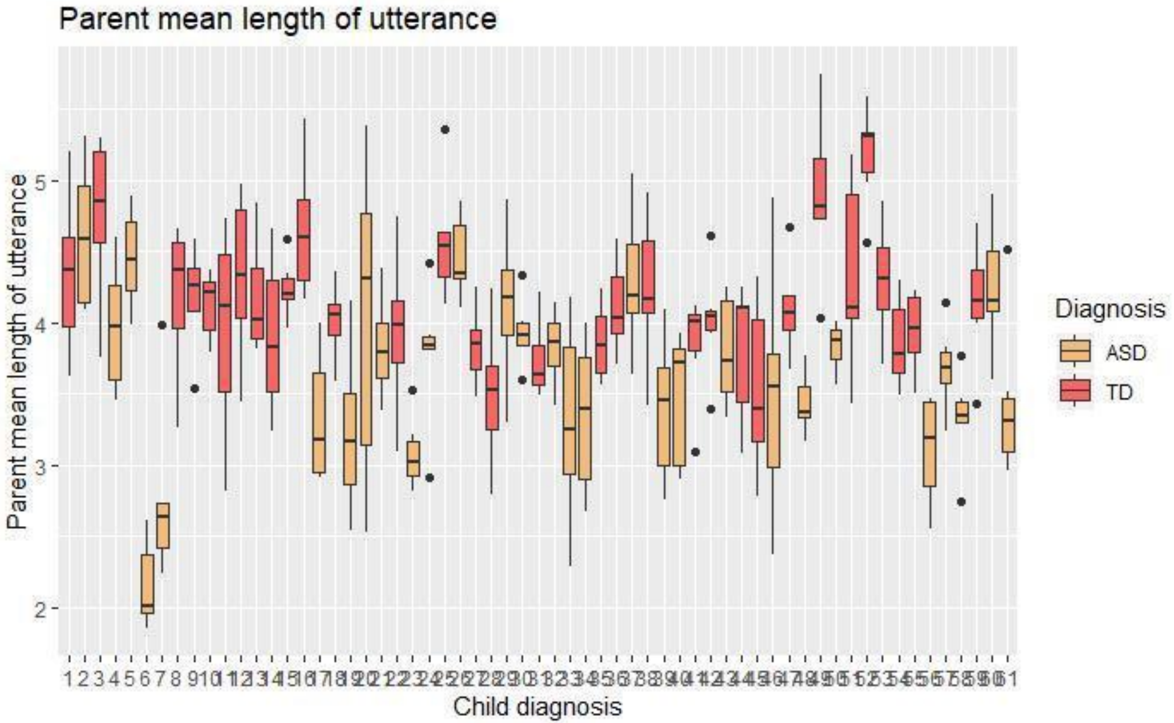


Figure 17

language in order to accommodate their children either by an explicit cognitive decision or due to an implicit reaction to the feedback received from the child. Either way the results strengthen the position of hypotheses H2A and H2B, stating that parent MLU increases over time and that parents with ASD children use shorter sentences. This point is visualized in the boxplot in figure 16, while figure 17 displays the quantile distribution of parents' MLU's between individual children. Both figures can be seen as evidence for differential MLU's dependent on diagnostic groups.

Exploratory Data Analysis

In this next section we will touch upon a more explorative approach to statistics. We have access to several more data points on each participant in the form of verbal and non-verbal IQ scores, number of unique words used, socialization scores, parental MLU, and total number of tokens used by child and parent. If we wanted to create a model which explained the MLU data with high fidelity, we could try to incorporate some of these other data points as additional independent variables in the best fitting model. One approach could be to create several models and compare them to each other before selecting the model which explains the biggest part of the variance. While this method will result in the model which best explains the data, it will presumably not be the model, which best predicts new data. This is due to overfitting, which is a common problem, when hypotheses are altered post hoc.

Despite these precautions, we have attempted to fit the data to different models in order to find one, which strikes the balance between explaining a large portion of the variance and being interpretable. Using AIC or nested F-tests as a criterium, we compared models of increasing complexity and found that including verbal IQ to the previously created model 1d did lead to a significantly improved fit. The improved model 3 has a significantly lower AIC score (AIC = 509.28, $p < 0.001$) than model 1d, and also performs better than our other exploratory models. Model 3 is a mixed effects model with participant ID as a random effect and diagnosis, visit, and verbal IQ as fixed effects. The model includes a random slope and intercept based on ID.

Model 3: child MLU ~ visit * diagnosis * verbal IQ + (1|ID) + (0 + VISIT|ID)

According to the statistical analysis of the model, the child MLU is significantly dependent on the verbal IQ (= 0.063, t(208.13) = 4.82, SE = 0.013, $p < .001$), the interaction between visit and diagnosis (= 0.67, t(291.59) = 7.36 , SE = 0.091, $p < .001$), and the interaction between visit and verbal IQ (= 0.0098, t(291.86) = 3.68 , SE = 0.0027, $p < .001$). The MLU's of the children are also significantly correlated with an three-way interaction effect between all three variables, that is, the visit number, child diagnosis, and verbal IQ score, (= -0.022, t(291.69) = -4.87, SE = 0.0045, $p < .0001$). This model, however much variance it might explain, has low beta values for every predictor except the interaction effect between visit and diagnosis. This makes the model's validity as a predictor questionable.

5. Model assessment

There are several different methods to investigate a model's ability to make accurate predictions. In the case of models that are formed using exploratory analyses of data, it is especially important to check whether the model has predictive validity. If a model is good at making predictions about data it has yet to encounter, we can rest assured that the model explains general tendencies in the data instead of specific noise from the data points on which it was trained. If the model, however, explains a lot of the variance of the training data, but a significantly smaller part of the variance of the test data, there is a good chance that the model has been *overfitted*. If a model is overfitted, it falsely interprets random noise in a sample as important variations. Therefore, overfitting becomes increasingly probable, the more independent variables the model contains.

Taking these deliberations into account, it is sensible to investigate the differences between one of our initial, simple models e.g. model 1c and our best fitted exploratory model, model 3, to see how they compare. Applying the information about overfitting, we can conject that the exploratory model is the one most likely to be overfitted due to its increased complexity.

Model 1c: child MLU ~ diagnosis * visit + (1|ID)

Model 3: child MLU ~ visit * diagnosis * verbal IQ at visit 1 + (1|ID) + (0 + VISIT|ID)

Root Mean Square Error

One way of investigating the predictive power of the two models is via a root mean square comparison between the predicted values from the model and the actual values of the data collected in the experiment. If we utilize the exploratory model 3, the root mean square error, RMSE, between the data points from the training data and the model predictions is 0.372, while the RMSE between the test data and model predictions is 1.129. Since there is a large leap between the RMSE for the training and testing data, it is probable that our model 3 is overfitted to the training data to some degree.

Using the basic model 1c to predict possible MLU values, the RMSE is 0.411 when comparing training and predicted data. When comparing test data with the predicted values, on the other hand, the RMSE is 1.120. These results also indicate a possible overfitting of the model. Interestingly, the RMSE's are quite similar between the basic and exploratory models, but the exploratory model 3 seems to be slightly more overfitted, if we use the absolute difference between training and test RMSE's as a way of assessing this. If we were to choose a model for further analysis in the light of this analysis, we would recommend the basic model 1c. This is partly due to model 1c being slightly better at predicting the test data when looking at RMSE's. Furthermore, increasing the complexity of a model by adding more predictors comes with a cost to the interpretability and applicability of said model. Increased model complexity should only be justifiable by a payoff in terms of better predictions. This is not the case of model 3.

Cross-Validation

Another way of determining the validity of a model's predictions is via cross-validation. The procedure is similar to that of the comparisons made above. When doing a cross-validation, however, the model is trained and tested on the entire data corpus instead of only being trained on one chunk, and only being tested on another, separate chunk.

When doing cross-validation, the dataset is separated into folds, which should contain more or less homogenous groups. We then train the model n times on n-1 of the folds' data, creating predicted values that are afterwards tested on the remaining fold's data points employing root mean square error and AIC scores. Cross-validation ensures that random variances in the testing and training sets do not lead to skewed models. A model should perform equally well on all testing folds, if it is not overfitted to any particular part of the data.

When cross-validating the basic model 1c and best fitted, exploratory model 3, using five folds, model 3 seems to be superior (AIC = 432.4 RMSE = 0.542) to model 1c (AIC = 515.6, RMSE = 0.769) when predicting data. Interestingly, this is the opposite conclusion from the previous comparison of root mean square errors. This highlights the fact that using only one dataset for training the model and one dataset for testing the predictions of the model can be precarious if the datasets are not equally matched. Therefore, cross-validation is a way of making sure that heterogeneity between training and testing datasets does not result in overfitting to the training data. We therefore tentatively conclude that model 3 is superior compared to the basic model 1c, when we do cross validation and only take the AIC scores and RMSE's into account, even if more complex models usually have a greater tendency to overfit. If we compare model 3 to other exploratory models, it still fairs well. However, the complex model CHI_MLU ~ Visit * Diagnosis + Visit * Verbal IQ at visit 1 + Visit * Ados at visit 1 + (1|ID) + (0 + Visit|ID) makes almost as accurate predictions (AIC = 432.4, RMSE = 0.552). Despite this discovery, we will keep on utilizing the basic model 1c and the exploratory model 3 in this paper when analyzing the data.

Predicting individual MLU's

If we look closer at an example from a particular individual, in order to see how the basic model 1c performs in specific real world cases, we could choose a random child from the test set and compare their MLU trajectory with the model's predictions. Such a participant could be the child with ID number 2 from the test set (alias Bernie), who is a TD child.

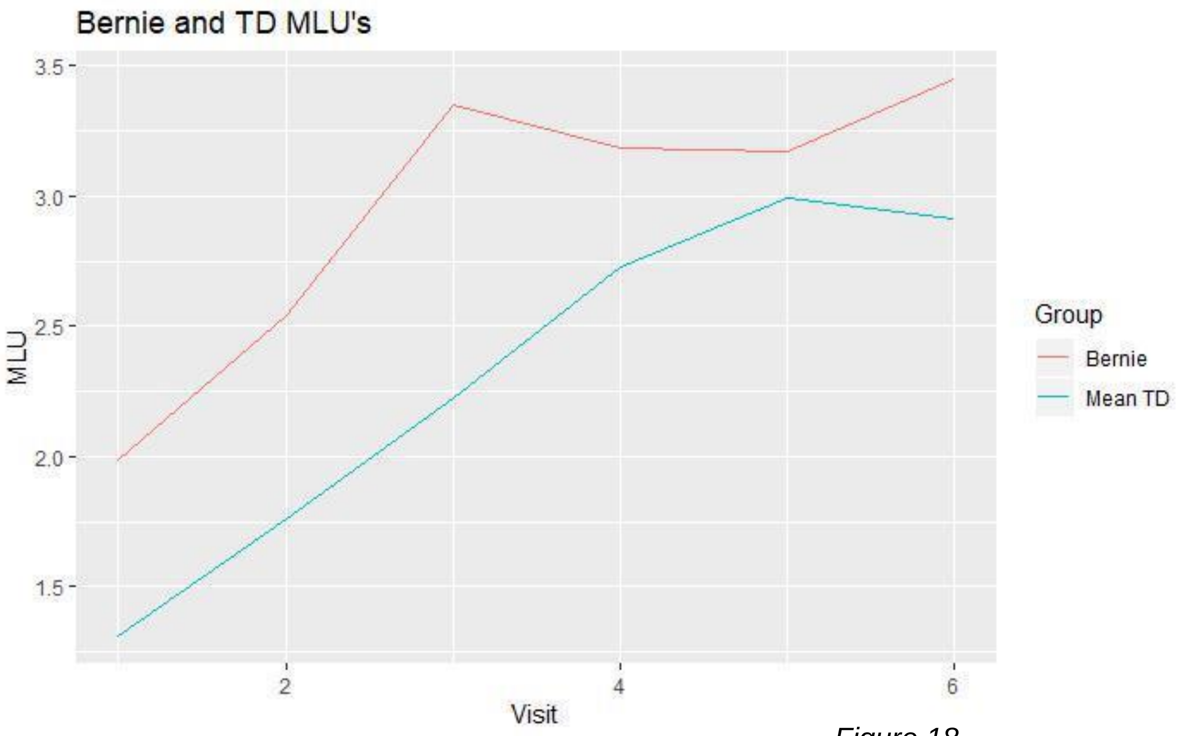


Figure 18

If we set out by examining Bernie's data points from the six different visits, we see that even though the general tendency from the models we have investigated is for TD children to gradually increase their lengths of utterances, neither the TD means nor Bernie's MLU's follow this rule uniformly (figure 18). Moreover, all Bernie's data points lie above the mean MLU's of TD children. This is why including individual intercepts in the mixed effects model is a good idea. If we want to examine more tangible evidence, we could calculate the absolute difference between Bernie's MLU's at the six visits and the mean TD MLU's. The results of these calculations are shown in table 1.

	Visit 1	Visit 2	Visit 3	Visit 4	Visit 5	Visit 6
|Bernie-$\underline{x}$|	0.67	0.78	1.13	0.45	0.18	0.54

Table 1

Another question is how a model specifically performs in predicting one of Bernie's data points at a given visit. We have chosen to investigate visit 6, and in doing so have found that the basic model 1c undershoots its prediction of Bernie's MLU with the absolute value of 0.007. The root mean square error of the predictions from the basic model trained on Bernie's data compared to Bernie's actual data is 0.29. The prediction from the exploratory model 3 also overshoots Bernie's MLU at visit six, but with a value of 0.057. The root mean square error when comparing predicted values of the trained model 3 with Bernie's data for all visits is 0.28. The exploratory model 3 thus seems to be worse at predicting the MLU at visit six, but slightly better in general. We tentatively conclude that the best exploratory model seems to make the most accurate predictions in the specific case of Bernie, while the basic model is best at predicting Bernie's data at visit 6 specifically.

Having tested our best fitted exploratory model and our basic model using several different methods, it becomes evident that the assessments of the models are highly dependent on which methodology we decided use. This is an important point to make, since this flexibility grants researchers a lot of control over the statistical evaluation of their own models.

Assignment 2, part 3, Experimental Methods 3

6. Power analysis

Power Analysis of our Study

To assess power for our effects of interest, which is an interaction between visit and diagnosis, we have fitted our favorite model: child MLU ~ 1 + visit * diagnosis + (1|ID) on both training and testing datasets. Power analysis indicated 100% (500 sim) chance of detecting a medium effect d = .5, which is calculated from the mean SD of TD and ASD (SD_ASD = 0.9, SD_TD = 0.5). Looking at the power curve (figure 19) reveals something suspicious. It seems as though the power function is at 100 % independent of number of participants. If we believe the visualization, the conclusion must then be that a sample size of 10 participants and perhaps fewer is sufficient to detect at effect of 0.5. We admit that these results are not satisfying us, since such a low number of participants should not enable us to make any kind of accurate predictions with the kind of noise we see in the data.

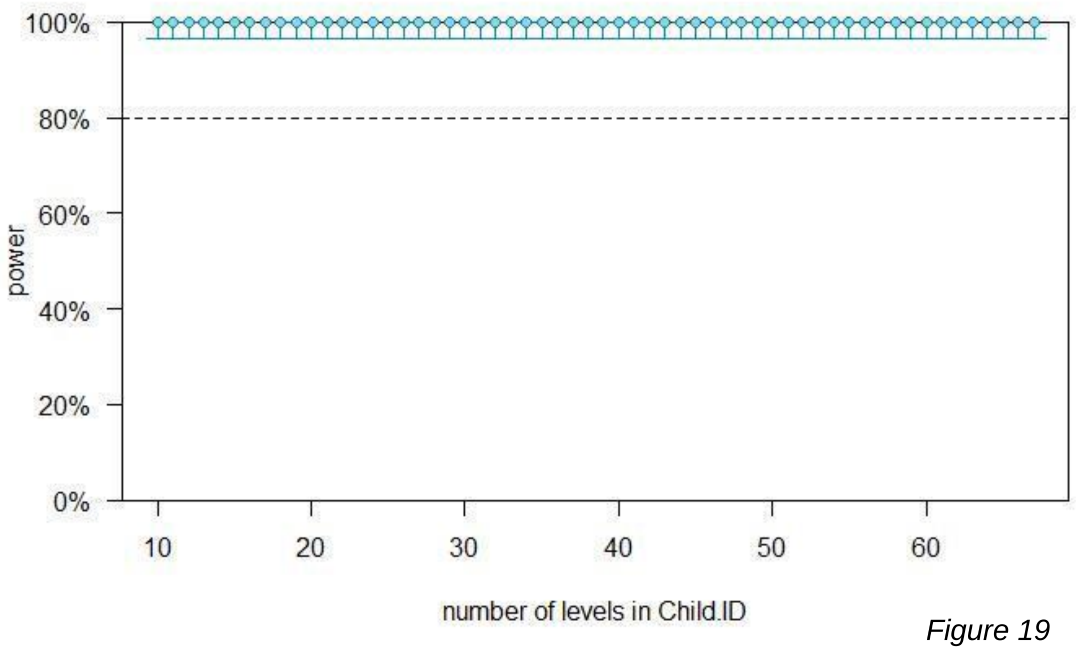


Figure 19

We can use the power analysis estimates for estimating the likelihood of successfully rejecting the null hypothesis (that child's mean length of utterance does not change over time or according to diagnosis) when diagnosing other children. A likelihood that is big enough (>80%) when the minimal interesting effect (aka Cohen´s d) is 0.5 would give us confidence that we will find a moderate sized effect of time(visits) and diagnosis on child's mean length of utterance when it exists. However, if our power analysis does not give us 100% assurance, there is a chance, albeit rather small, of not finding any effect of time or diagnosis on MLU (Type I error) or finding an effect which does not exist (Type II error). Thus, the estimates can be used to reduce the overall rate of data inference errors.

Conservative Power Analysis

Considering that the range of recorded Child MLU is from 0 to 4.3 we can conclude that having the minimum effect of d = 0.5 is probably too big. Furthermore, having the effect size of d = 0.05 is probably too small. Considering the minimum effect size for our relevant effects (interaction between visit and diagnosis) within the range from 0.05 to 0.5, we chose to run the conservative power analysis with the effect size of d = 0.1. We argue that a child who improves by 0.1 MLU per visit would be a small but still significant effect in the context of our study. Power analysis indicated 96% chance of detecting a small effect of 0.1.

After the assessment of the power curve by Child.ID (figure 20) with an effect size of 0.1 and 100 numbers of simulations, we identify an ideal number of 38 participants to estimate our interaction effect.

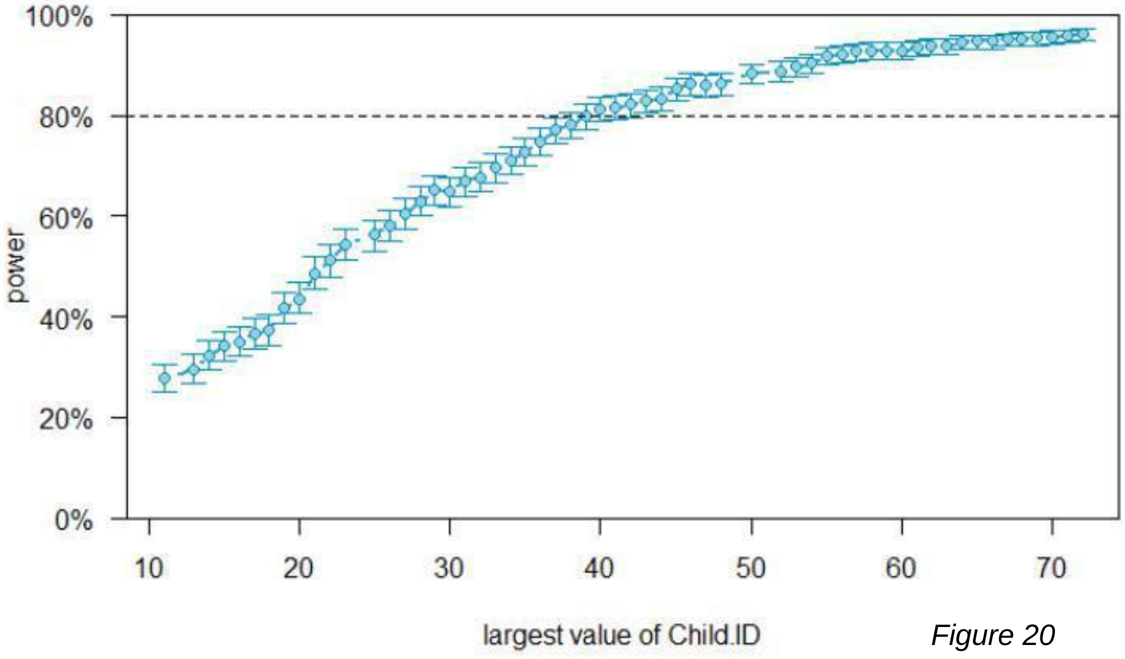


Figure 20

Power Analysis on a Subset of Participants

Assuming that we only have access to 30 kids (15 with ASD and 15 TDs), we evaluate that it would still be worthwhile to run the study, since we found the model to have 96 % power in our power simulation (nsim = 1000), when estimating the interaction effect between diagnosis and visit. This point is illustrated by the power curve (nsim = 1000) in figure 21.

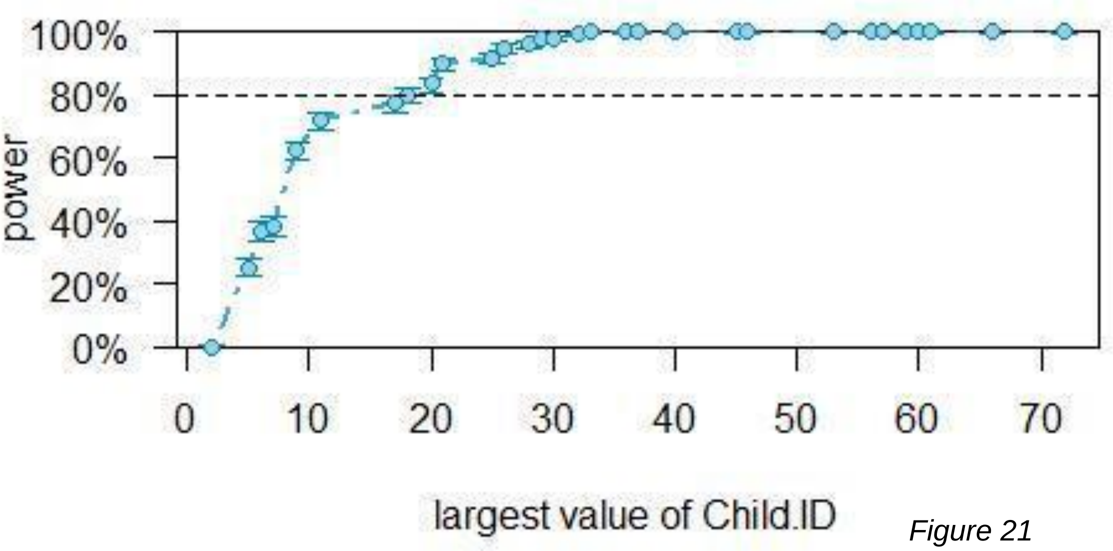


Figure 21

Bella Terragni- 201405868,
Peter Thramkrongart- 201806892,
Rūta Slivkaitė- 201805872 &
Bianka Szöllősi- 201808610

Diagnosing schizophrenia from voice data

link to github: https://github.com/szbianka/assignment-3

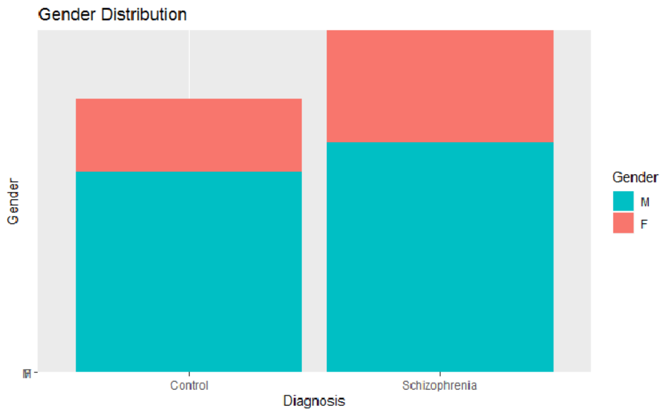
Our intention with this analysis is to determine whether voice data can be utilized to predict a psychiatric diagnosis like schizophrenia. We use an exploratory approach, where we test the performance of different vocal features as predictors of schizophrenia. Lastly, we intend to test whether our voice model is a good classifier of schizophrenia by examining the predictive performance of the model on our data.

Methods and data

The data used in this paper was acquired from voice recordings from people diagnosed with Schizophrenia and their matched controls (on gender, age and education). The participants were asked to watch and describe several videos of triangles moving across the screen. The pitch of their voice was extracted every 10 milliseconds along with several duration related features (e.g. number of pauses, number of syllables, etc.).

The original sample included seven studies, of which 1-4 was conducted in Danish, 5-6 was conducted in Mandarin Chinese and 7 was conducted in Japanese. After clearing the data and omitting all the missing values, we were left with only three studies: Study 1 and Study 2 in Danish, and Study 6 in Chinese. In this cleared data there were a total of 1992 participants, 1676 Danish, of whom 813 had Schizophrenia (mean age= 22,98 years, F= 356, M= 457), and 863 controls (mean age= 22,62 years, F= 366, M= 497), 316 Chinese, of whom 316 had Schizophrenia (mean age= 28,71, F= 204, M= 112), unfortunately with missing controls.

Therefore, we decided to only split the data according to diagnosis, with the assumption that there would not be too big differences between the two languages in relation to acoustic voice features. This left us with 1129 Schizophrenia patients (mean age= 24,58 years, F= 560, M= 569) and 863 controls (mean age= 22,62 years, F= 366, M= 497). As a result, although we have less controls, the sample sizes, age and gender are rather well matched between the controls and schizophrenic subjects.



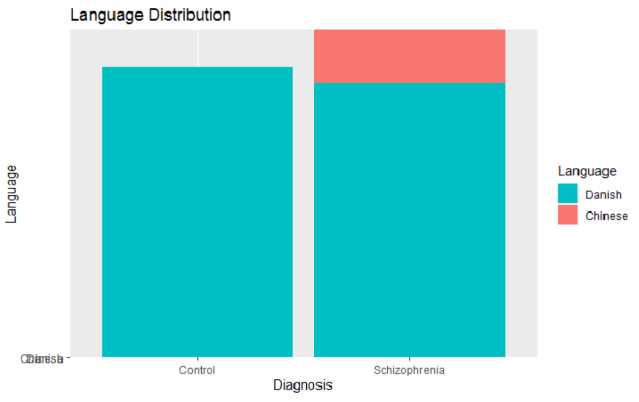


Figure 1: Gender distribution of the subjects   *Figure 2: Language distribution of the subjects*

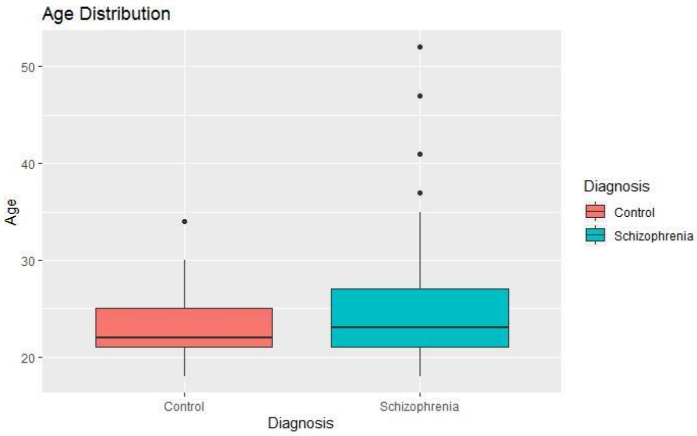


Figure 3: Age Distribution of the subjects

Analysis and results

Determining the acoustic profile of a schizophrenic voice

After cleaning up the data and omitting all the missing data points we were left with two languages: Danish and Chinese and three studies: Study 1, Study 2 and Study 6.

In order to find out the acoustic profile of a schizophrenic voice we set up four generalized linear mixed effects models.

We have decided to analyze both languages together, with the assumption that they should not reveal big differences regarding the acoustic voice features. Some data points are dependent – come from the same participant, therefore, each participant is likely to have his/her own base level in relation to values on the dependent variable. We will be modeling random intercepts per subject in all our models. In addition, we have included a random effect by study in order to replicate the meta-analysis findings.

To compare our results to the meta-analytic findings we used two measures: p values and standardized mean difference. P-values were obtained by likelihood ratio tests of the full model with the effect in question against the model without the effect in question. To estimate the differences in vocal patterns between individuals with schizophrenia and HC (healthy controls) we extracted the standardized mean difference (SMD; also known as Hedges' g.). The outcomes of the models were scaled before running the analysis.

Model 1: Pitch variability model

sd ~ Diagnosis + (1|Subject) + (1|Study)

This model measures difference in pitch variability between two diagnosis levels: schizophrenia and healthy controls. We want to see to what degree and in what direction pitch variability (sd) is different in schizophrenia patients compared to controls. We model pitch variability as a dependent variable, while diagnosis as a predictor.

Our results revealed that diagnosis (Schizophrenia) affected pitch variability ($\chi 2$ (1)= 8.72, p< 0.05) decreasing it by about -0.11 (SE = 0.04). SMD revealed significant effects of diagnosis (in terms of Hedges' g) on pitch variability (lower, -0.11, 95% CIs: -0.19, -0.02)

Model 2: Proportion in spoken time model

phonationtime..s. ~ Diagnosis + (1|Subject) + (1|Study)

This model measures difference in proportion of spoken time between two diagnosis levels: schizophrenia and controls. We model phonation time in seconds as a dependent variable, while diagnosis as a predictor.

Our results revealed that diagnosis (Schizophrenia) affected duration of spoken time ($\chi 2$ (1)= 3.36, p=0.07) increasing it by about 7.679e-02s (SE= 4.312e-02). SMD revealed insignificant effects of diagnosis (in terms of Hedges' g) on proportion of spoken time (higher, Hedges' g: 0.16, 95% CIs: 0.07, 0.24).

Model 3: Speech rate model

speechrate..nsyll.dur. ~ Diagnosis + (1|Subject) + (1|Study)

This model measures difference in speech rate between two diagnosis levels: schizophrenia and controls. We model speech rate in average number of syllables per second as a dependent variable, while diagnosis as a predictor.

Our results revealed that diagnosis (Schizophrenia) affected speech rate ($\chi 2$ (1)= 7.11, p<0.05), increasing it by about 0.12 s SE = 0.05). SMD revealed significant effects of diagnosis (in terms of Hedges' g) on speech rate (faster, 0.18, 95% Cis: 0.09, 0.27).

Model 4: Pause duration model

PauseDuration ~ Diagnosis + (1|Subject) + (1|Study)

This model measures difference in pause duration between two diagnosis levels: schizophrenia and controls. We model pause duration in second as a dependent variable, while diagnosis as a predictor.

Our results revealed that diagnosis (Schizophrenia) affected pause duration ($\chi 2$ (1)= 2.67, p=0.10), increasing it by about 6.231e-02 s (SE = 3.813e-02). SMD revealed insignificant effects of diagnosis (in terms of Hedges' g) on duration of pauses (longer, 0.09, 95% Cis: 0.01, 0.18).

After the analysis of our created models we are able to describe the acoustic profile of a schizophrenic voice: to evaluate which features are different compared to the control group.

People with schizophrenia have lower pitch variability, prolonged proportion of spoken time, increased speech rate and increased duration of pauses compared to controls. The effect sizes for all of the features were quite small and the only significant effects of diagnosis were on pitch variability (lower, Hedges' g: -0.11, 95% CIs: -0.19, -0.02) and speech rate (faster, Hedges' g: 0.18, 95% Cis: 0.09, 0.27).

Looking at the meta analysis findings we can see that diagnosis had a significant effect on all of these features. Moreover, the effect sizes of tested features are bigger and some of the features reveal different directions of the effects. The main differences in the directions of the effects were in proportion of spoken time, which is lower in meta-analysis and higher in our analysis and speech rate, which is slower in meta-analysis and faster in our analysis.

Choosing the best acoustic feature

After the comparison to the meta-analysis findings (see *Table 1*) we choose pitch variability as the best acoustic feature. Our analysis has revealed the same direction of the effect of diagnosis on pitch variability as in meta-analysis and it is also significant. While speech rate has an effect size which is a little bit bigger compared to pitch variability, it has a different direction of the effect of diagnosis compared to meta-analysis. Thus, striving to replicate the meta-analysis findings, we have decided to not choose speech rate as our best feature.

Acoustic feature of the voice	**Our analysis**	**Meta-analysis**
Pitch variability	lower, -0.11, 95% CIs: -0.19, -0.02	lower, Hedges' g: -0.55, 95% CIs: -1.06, 0.09
Proportion of spoken time	higher, Hedges' g: 0.16, 95% CIs: 0.07, 0.24	lower, Hedges' g: -1.26, 95% CIs: -2.26, 0.25

Speech rate	faster, 0.18, 95% Cis: 0.09, 0.27	slower, Hedges' g: -0.75, 95% CIs: -1.51, 0.04
Pause duration	longer, 0.09, 95% Cis: 0.01, 0.18	longer, Hedges' g: 1.89, 95% CIs: 0.72, 3.21

Table 1 Comparing our effect sizes and directions to meta-analysis findings

Investigating the predictive power of our best feature

Inclusion and exclusion criteria for data

Our team has access to voice data collected from seven different studies in three different languages. The question lies in determining which datasets to include in our analysis and which, possibly, to exclude. First off, having more data could improve our trained model's power and make smaller effects noticeable and significant. However, attaining an inflated sense of confidence in our model is not our end all target; we are far more interested in creating a model, which can be applied as a classifier. This entails that the model must make good predictions. A model might show effects with very significant p-values, but if these effects are miniscule, they will not contribute much to a binary classification. Furthermore, not all of the studies have recorded the same vocal features. We could discard all features which are not universal, but our team assessed that sampling down, only choosing to fit our model to the extensive data from the two Danish studies, would be the best course of action. This way we also avoid any problems, which might arise from comparing vocal features like pitch between as widely different languages as non-tonal Danish and tonal Chinese. It naturally follows from this decision that, if successful, we will end up with a Danish classifier of schizophrenia, which, while it may not be universal, hopefully, is more specific to the patients in the Danish health sector.

Pitch variability as a binary classifier

According to our analysis, pitch variability is the best feature to model diagnosis. However, as alluded to earlier, this information is not sufficient to inform us about the utility of the feature as a single distinguisher between schizophrenic and non-schizophrenic people. We only know that according to our analysis, pitch variability is significantly different between people who have schizophrenia diagnosis and those who do not. We have yet to find out whether pitch variability is predictive of the diagnosis.

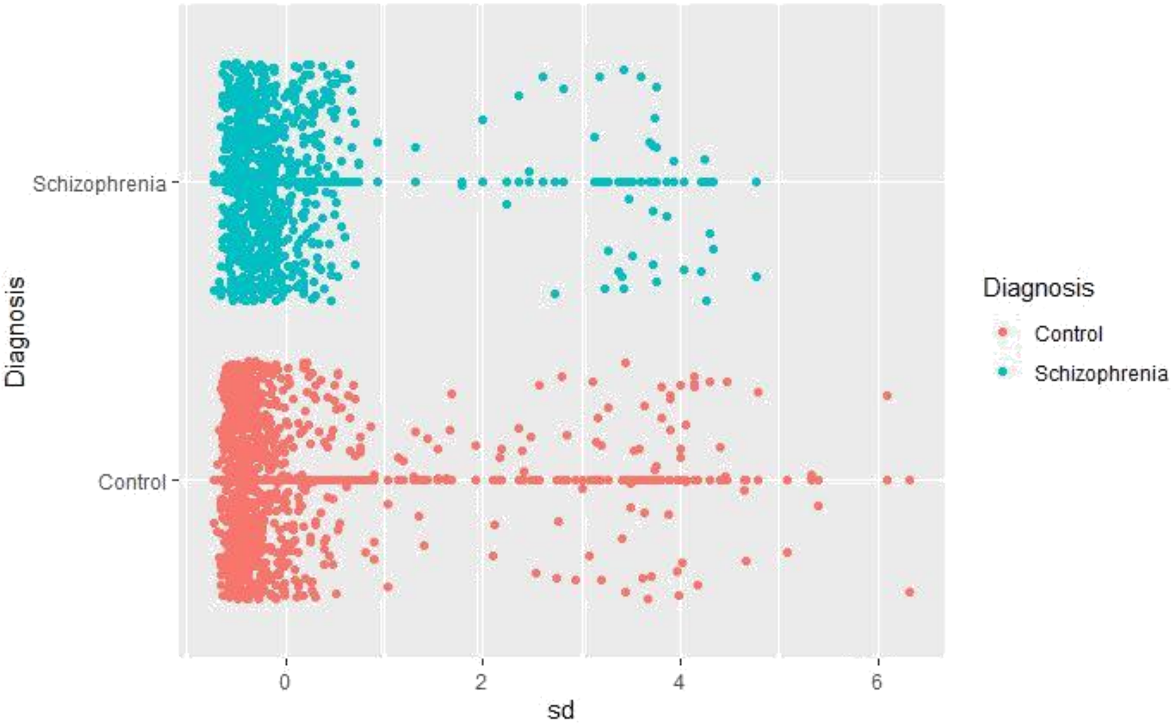


Figure 4: the distribution of pitch variability (sd) between diagnostic groups

In order to test how well pitch variability can be used to distinguish between individuals with and without the diagnosis, we fit model 5 to our data using logistic regression. Diagnosis is the dependent variable, while pitch variability is the predictor. We would have preferred to use random intercepts based on individual participants, but were unable to get the model to converge with mixed effects. This means we have a confound, since our data is not completely independent:

Model 5: Diagnosis ~ Pitch variability

The outcome confusion matrix is depicted in *Table 2*.

	Reference		
Prediction		Control	Schizophrenia
	Control	663	679
	Schizophrenia	191	131

Table 2

Looking at the predicted results compared with the actual, the conclusion must be that while pitch variability is significant in our general linear mixed effects model analysis, the feature fares badly when classifying diagnosis. Accuracy is 0.477 and when examining the sensitivity and specificity of the model, sensitivity is 0.162, while specificity is 0.776. The model therefore makes more type 2 than type 1 errors and is thus more inclined to underpredict the schizophrenic diagnosis. Our model is therefore fairly good at classifying controls, but quite awful at classifying schizophrenic patients. If we wished to apply this classifier in a clinical context, it would be very problematic, since the model will generate a lot of false negatives. We argue that failing to detect sick individuals in a clinical setting is worse than having more false positives. Therefore we suggest that the threshold for distinguishing between controls and schizophrenic patients might be shifted, so that the model becomes more sensitive to schizophrenic patients.

When we do cross-validation on the logistic regression model using 10 folds over 10 iterations, we make sure that diagnosis is evenly distributed between folds and that individuals don't exist in more than one fold each iteration. From this analysis we get an accuracy of 0.442, a specificity of 0.597, and a sensitivity of 0.288. Again, these scores are low, with specificity being close to chance and sensitivity being quite dreadful. This trend is also evident in *Figure 5* that depicts a ROC curve from the logistic regression. The curve depicts the relation between type 2 and type 1 errors at different set thresholds. We can see that our model does not make good predictions, since the area under the curve is close to 0.5 or chance. If we examine the gain curve from the logistic model (*Figure 6*), it is likewise evident that our model's predictions are not much better than chance.

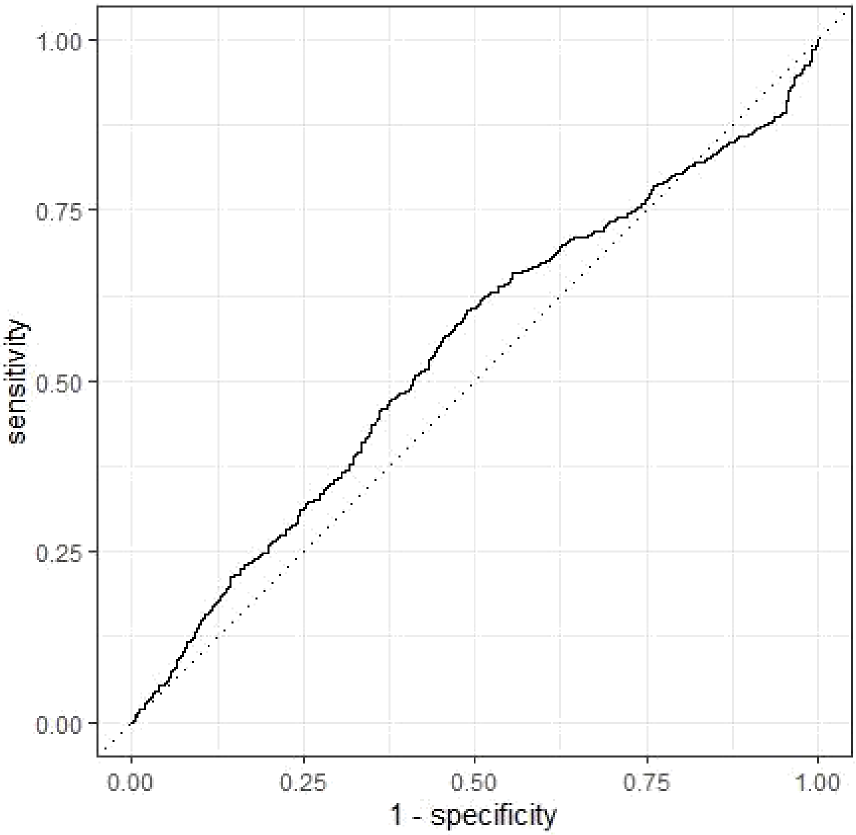


Figure 5: GLM logistic regression ROC curve

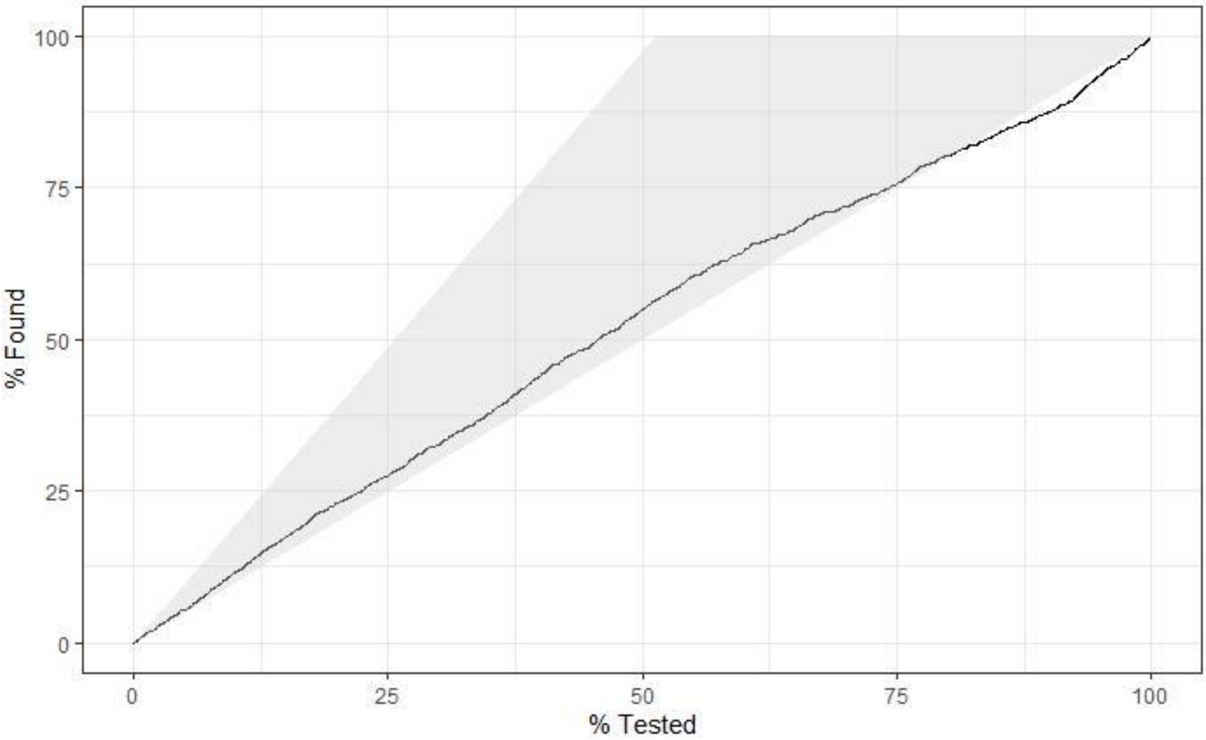


Figure 6: GLM logistic regression gain curve

Generating best performing model to classify schizophrenia

Since it seems like a single feature is not enough to classify schizophrenic patients, we will investigate whether a combination of features might improve the performance of our classifier model. In order to do this, we applied the Tidymodels framework to two different kinds of classification methods. Namely logistic regression and support vector machine. Tidymodels creates the best performing model by assigning each predictor different weights. It uses all features as predictors if not otherwise stated. This is useful as we don't know much about what feature combination is the best. It is crucial to consider which features the model is allowed access to, as it will exploit any predictor it can, for the best fit. Therefore, we removed everything from our data, but Diagnosis, Gender, Age, nsyll, duration of clip, phonation time, speech rate, articulation rate, average speaking duration, mean pitch, pitch sd, pitch IQR, min pitch, max pitch, npauses and range.

When investigating the output of our analysis, we find the accuracy of the predictions were 0.56 (kappa = 0.11) for the logarithmic model and 0.64 (kappa = 0.27) for the svm model. These results demonstrate that while our logistic regression tidymodel is an improvement to our single feature logistic regression model, the support vector machine model is superior in classifying schizophrenic patients. The kappa value for the svm model is tolerable, but not good enough for clinical use.

We suggest trying alternative machine learning models in order to find the best classifier of diagnosis. As our mixed effects models failed to converge, a logical next step would be to successfully train mixed effects models. *Figure 7: tidymodels logarithmic ROC curve*

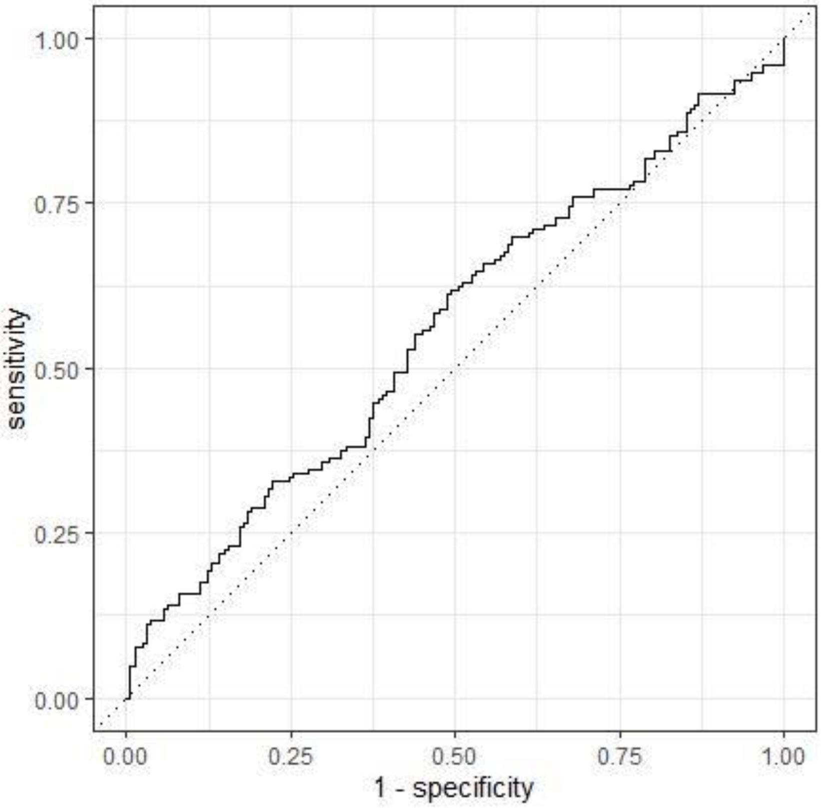


Figure 7: tidymodels logarithmic ROC curve

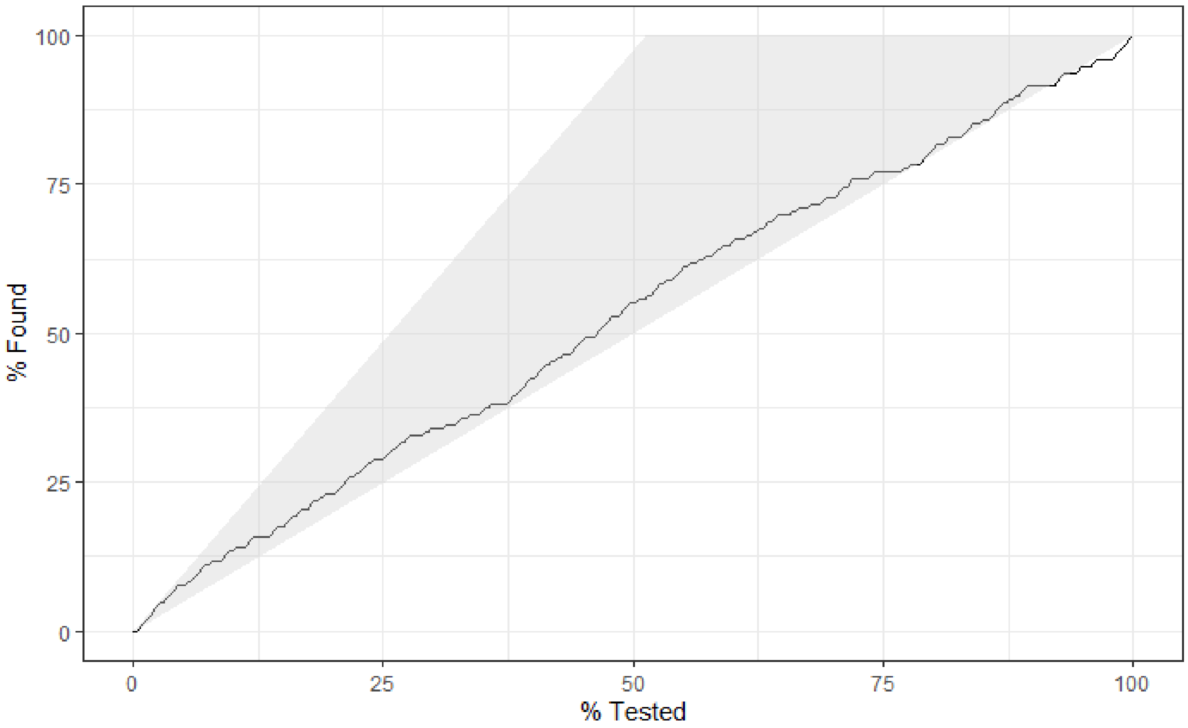


Figure 8: tidymodels logarithmic gain curve

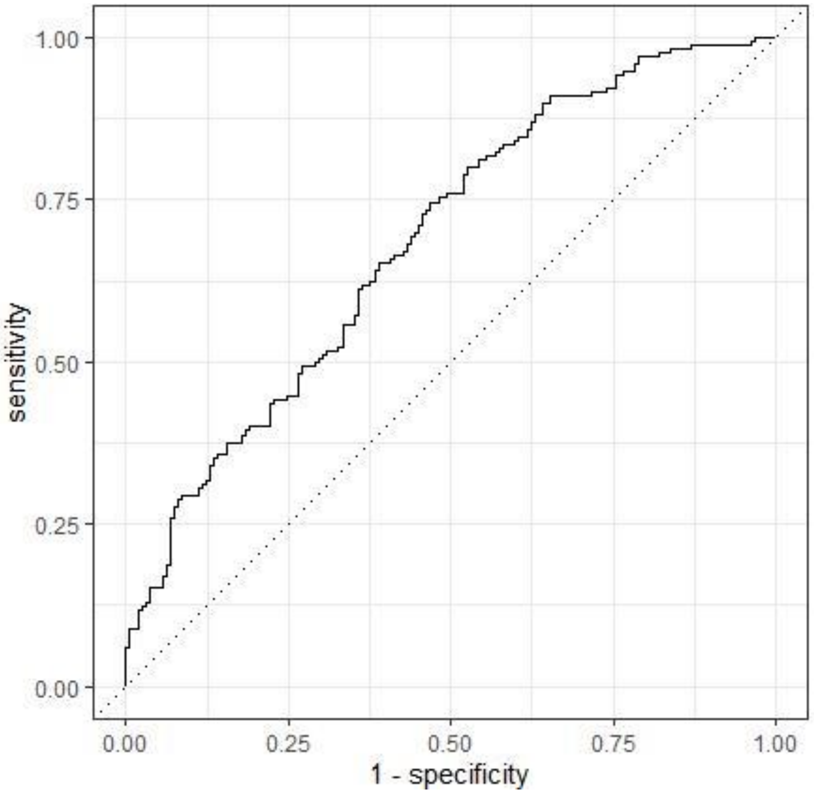


Figure 9: tidymodels support vector machine ROC curve

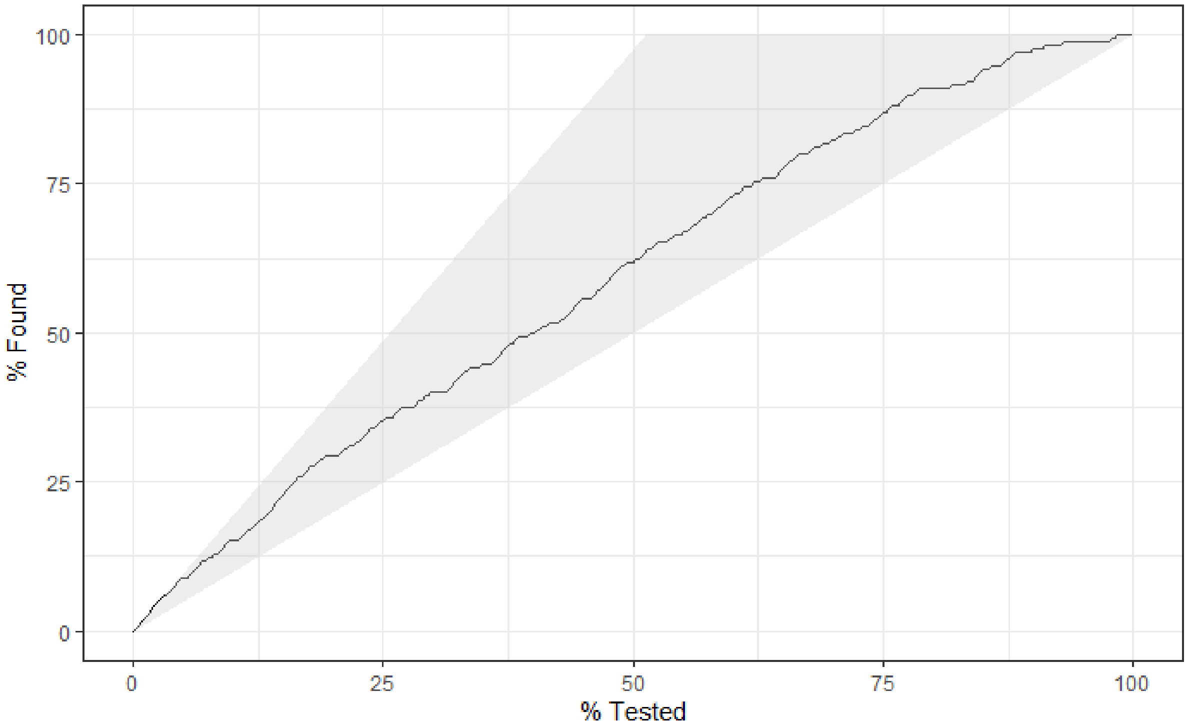


Figure 10: tidymodels support vector machine gain curve

Bella Terragni- 201405868,
Peter Thramkrongart- 201806892,
Rūta Slivkaitė- 201805872 &
Bianka Szöllősi- 201808610

Assignment 4: Physiological data

link to github: https://github.com/szbianka/assignment-4

1) How do you preprocess heart rate and respiration data? Describe the process. If any data needs to be excluded, list the excluded data and motivate the exclusion.

The process of cleaning the collected physiological data and preprocessing it before analysis involves, first off, the removal of outliers. Outliers can be tricky, since there is no universal rule about how much variability we should allow for in our data. In the specific case of this analysis, we set the cut-off threshold at 2.5 standard deviations. This threshold is chosen because it balances the need for removing artifacts and for retaining most of our true variance. Compare left side of *Figure 1,* where the data is depicted before removal of outliers, with the right side of the figure which illustrates the data without outliers. Instead of omitting the data points deemed outliers completely, we replace them with 2.5*SD. This method is chosen so that we don't coerce the data into appearing more mean-centered.

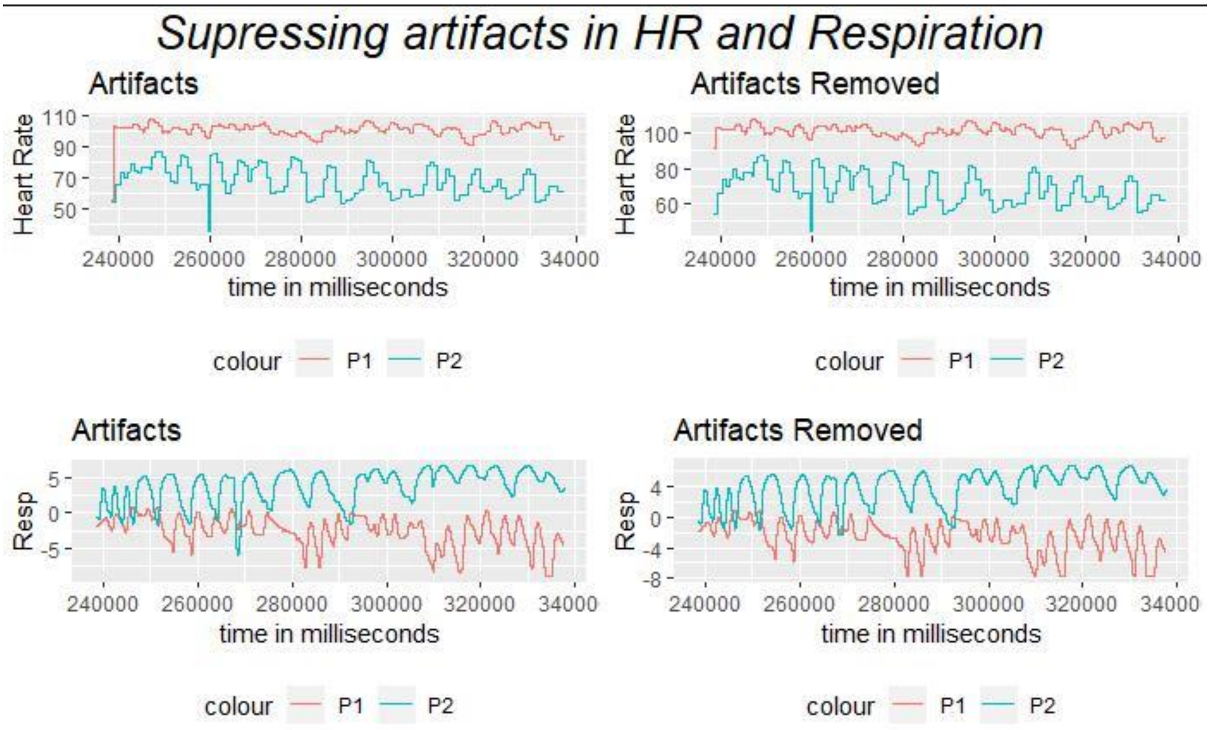


Figure 1 - Removing artifacts from one participant pair

The next major step during preprocessing is scaling. It is crucial to scale physiological data when we want to compare results between individuals, since participants' physiological responses are quite variable and dependent on the participants' age, gender, physical health and the measuring tool. When we scale the heart rate and respiration data, the amplitude, baseline, and range is altered to make the data more comparable. This is illustrated in *Figure 2*, which depicts the scaling for a single participant. *Figure 3*, on the other hand depicts scaling for multiple participants. It is evident that the scaling has influenced the ordinate axis.

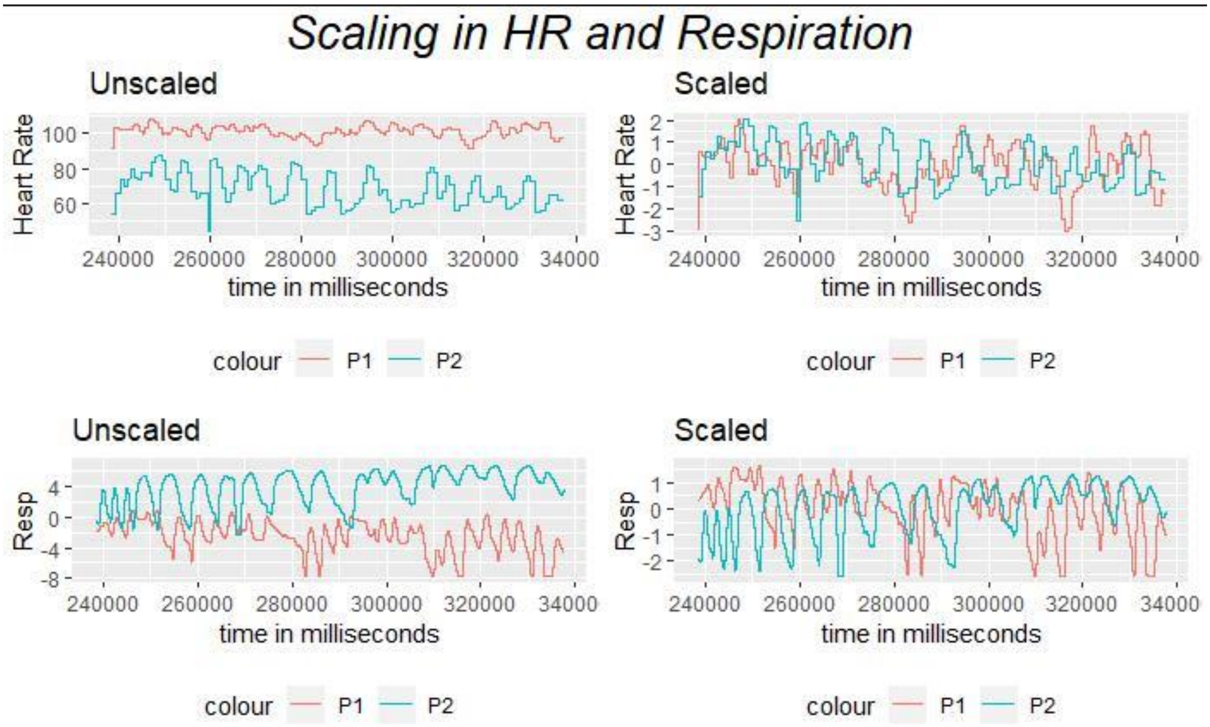


Figure 2 - Scaling heart rate and respiration makes data comparable

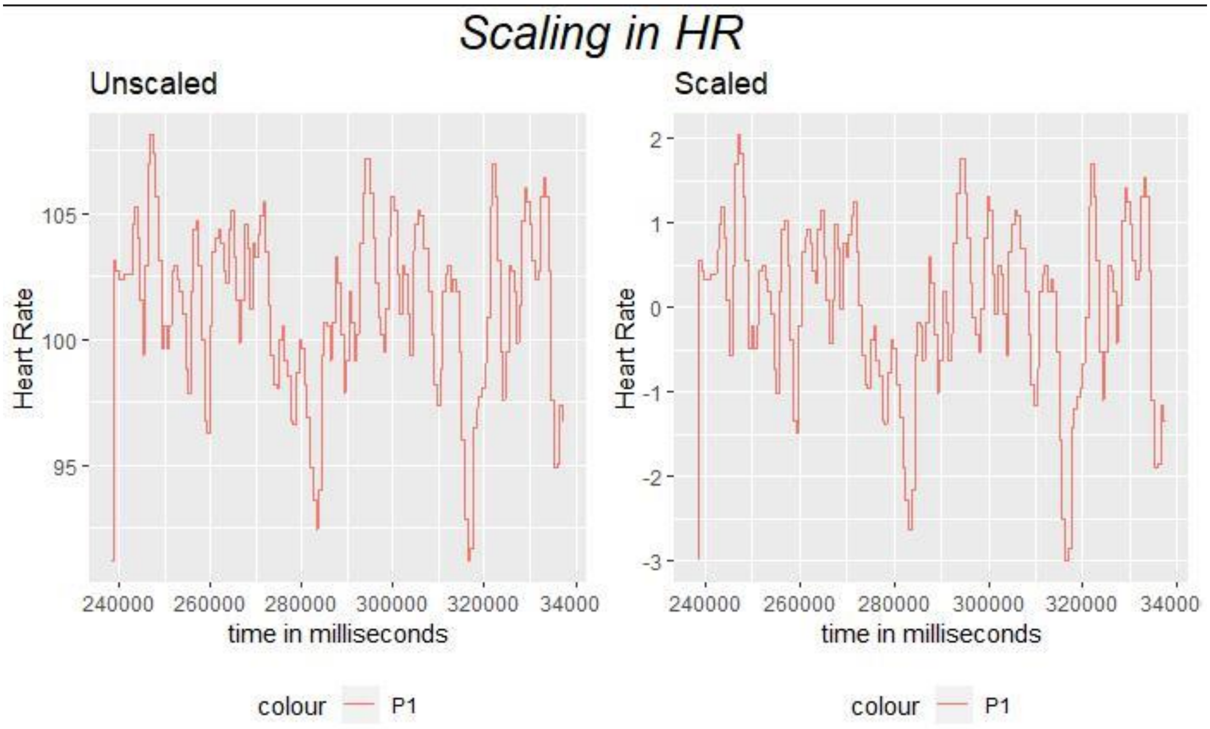


Figure 3 - Scaling heart rate in all participants changes the axis

The second to last step in preprocessing the data is downsampling. We tried to do it by groups of 100 and 1000. We are interested in downsampling, because the sheer number of data points collected per second is so large that many of the data points are essentially redundant. When downsampling with 1000, we group 1000 individual data points into one, finding the mean respiration and heart rate. This is helpful in big data sets where the values don't vary particularly fast. See *Figure 4.*

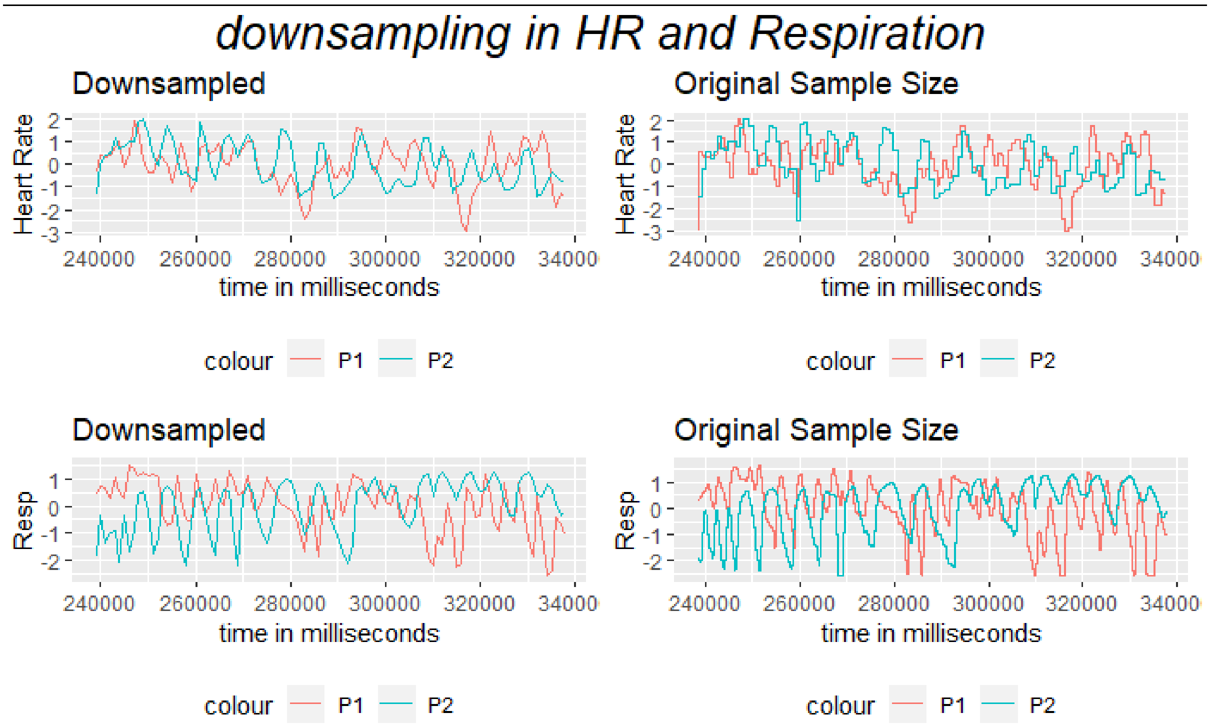


Figure 4 - Downsampling data makes peaks and troughs easier to see

The last crucial step of the preprocessing is detecting any artifacts, which have not been rectified by any of the previous steps. When detecting bad data, a simple, but effective course of action is to visualize the data. Using this method, we only found one participant, who we deemed as having a stretch of contaminated data, which was hence removed. This is illustrated in *Figure 5* and *Figure 6*.

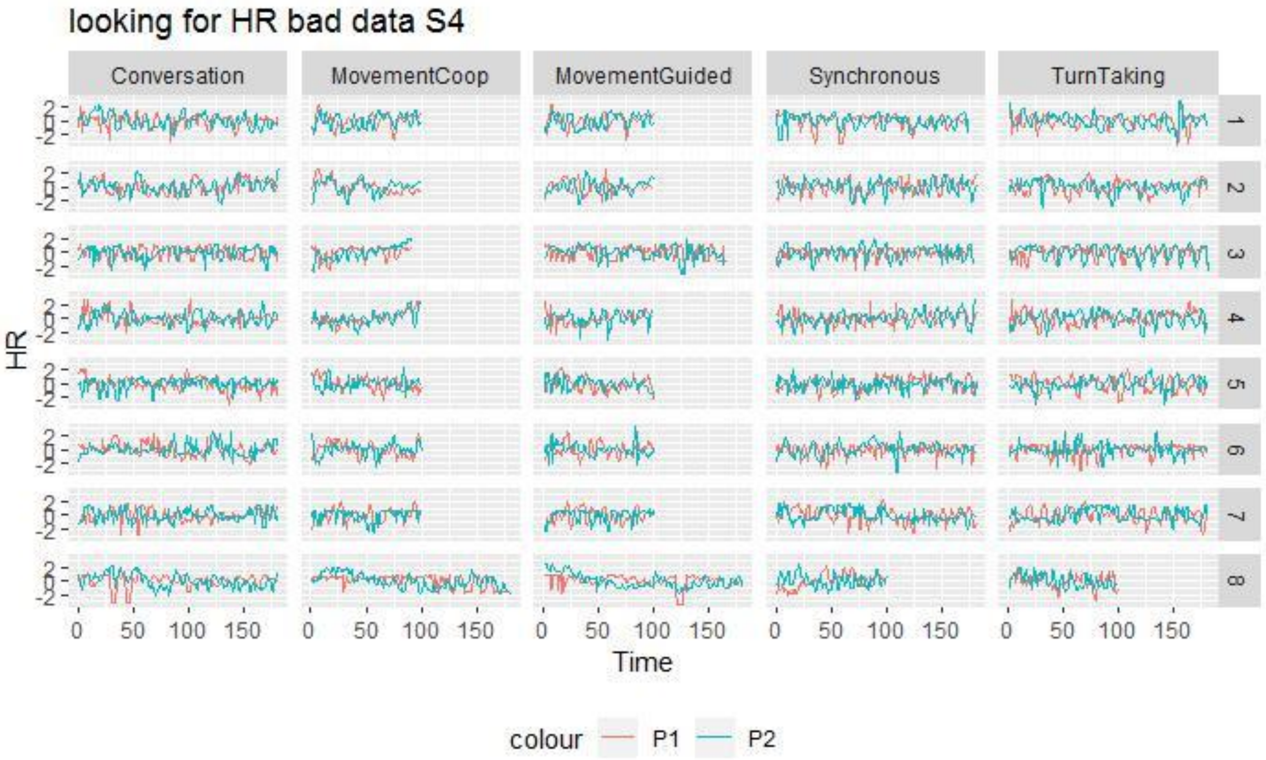


Figure 5 - Visualization of heart rate data before removal of artifacts

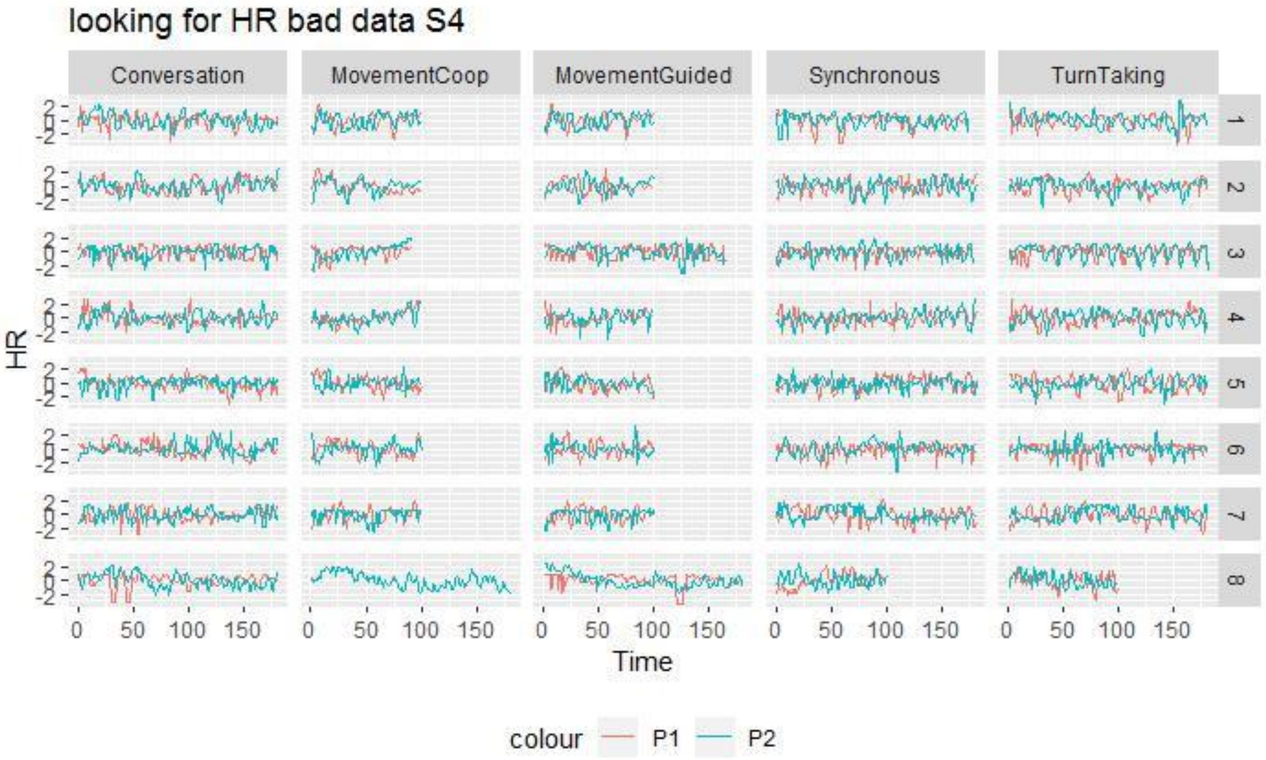


Figure 6 - Visualization of HR data after removal of P1 in Group 8, condition MovementCoop

2) Do you observe interpersonal coordination in heart rate and respiration? Describe your control baseline, the method used to quantify coordination, and the statistical models used to infer whether coordination was higher than in the baseline. Report the results of the models.

What would be a suitable baseline for answering whether our different conditions elicit synchronization of heart rate and respiration, that is interpersonal behavioral coordination? Since we have not collected empirical control data in the traditional sense during the experiment, we, first off, need to simulate a control dataset. This is necessary to evaluate whether the physiological responses are truly coordinated or if they are merely similar due to these responses oscillating within relatively stable standard ranges.

Two types of controls are created. The first type is shuffled controls, where the data points which were prior in a fixed time series are shuffled. The temporal structure will thus be disrupted inside each trial. Significant p-values using this control signifies that the dyad is in fact synchronous, but not because of interpersonal coordination. Rather, it appears the pairs are breathing and their hearts beating synchronously because of the two signals being within the same normal range and rhythm, and not due to any direct causality between the interlocutors of the dyad. If the synchronization of the respiration and heart rate is a real driver of interpersonal behavioral coordination, then we would observe higher coordination between dyads in their real time series compared to the shuffled controls.

The other type of controls we employ are surrogate pairs. In creating this type of control data, we split all our real pairs and join them in different combinations of surrogate pairs. When we compare our original pairs to these surrogate pairs, we expect that the real pairs will be more synchronous and display a larger degree of interpersonal coordination. If this is not the case, we must consider that the reason why our dyads might appear to be synchronized is *not* because there is an existing physiological interpersonal mechanism between the conversationalists, but due to people performing the same actions during conditions. We therefore selected participants from different groups, who share the same task or condition, but are not actually co-present or interacting. If behavioral coordination is driving shared HR dynamics and co-presence is necessary for facilitating shared HR dynamics, we will observe higher coordination in real pairs than in artificially constructed pairs.

In order to test for interpersonal coordination in heart rate we created a linear *Model A***,** where the change in the heart rate was modeled as a function of heart rate of the self and heart rate of the other participant, with an interaction with the condition and with the type (real pairs vs shuffled controls or surrogate controls). *Model A* contained the estimates between conditions. We have also accounted for different intercepts across the different conditions. If interpersonal coordination is happening, there should be a coupling between the change in heart rate or respiration of the self and the previous heart and respiration rate of the partner.

Model A:
HR_change_self ~ (0 + Condition + (HR_self + HR_other) : Condition) : Type

Condition	Beta	SE	P value
Conversation:HR_other:TypeOriginal	0.016	0.015	0.300
Synchronous:HR_other:TypeOriginal	0.001	0.016	0.964
TurnTaking:HR_other:TypeOriginal	-0.019	0.015	0.194
Conversation:HR_other:TypeSurrogate	-0.011	0.005	0.045 *
Synchronous:HR_other:TypeSurrogate	-0.018	0.006	0.002 **
TurnTaking:HR_other:TypeSurrogate	-0.024	0.005	3.6e-05 *

Table 1 - output from analysis of model A

As we see from *Table 1,* we get more negative beta coefficient values of coupling parameter in surrogate pairs compared to the original pairs. We also get more significant p values in surrogate pairs, which we find a bit puzzling. Nevertheless, the values we get are rather low and don't really show any clear evidence for interpersonal coordination. We cannot conclude that interpersonal coordination is a driving factor for shared HR dynamics and that co-presence is necessary for facilitating shared HR dynamics.

To test the interpersonal coordination in the heart rate even further, we employed two different mixed effects linear models, controlling for participant and group (in our case not for study variability as we only used Study 4) because we assume that the difference in interpersonal coordination can also be affected by individual differences in participants and different pair dynamics.

To try out a more advanced model we have created a linear mixed effect *Model B*, where we added an interaction with a condition and type, as well as random intercepts for ID and Group.

Model B:
HR_change_self ~ (HR_self + HR_other) : Condition : Type + (1 | ShuffleID) + (1 | Group)

Condition	Beta	SE	p-value
HR_other:ConditionConversation:TypeSurrogate	-0.01142	0.005508	0.038 *
HR_other:ConditionSynchronous:TypeSurrogate	-0.01853	0.005928	0.002**
HR_other:ConditionTurnTaking:TypeSurrogate	-0.02479	0.005792	1.88e-05***
HR_other:ConditionConversation:TypeOriginal	0.01568	0.01568	0.283
HR_other:ConditionSynchronous:TypeOriginal	0.0007226	0.01525	0.962
HR_other:ConditionTurnTaking:TypeOriginal	-0.01988	0.01479	0.178

Table 2 - output from analysis of model B

As we see from *Table 2,* we get similar results to *Model A.* We still get more negative beta coefficient values of coupling parameter in surrogate pairs compared to the original pairs, as well as more significant p values in surrogate pairs. We still find this a bit puzzling, since our model suggests that surrogate pairs are more correlated than the original pairs.

For *Model C* we included random intercepts for condition and random slopes for participant, and Condition and Group.

Model C:
(HR_change_self ~ (0 + Condition + (HR_self + HR_other) : Condition) : Type +
(0 + Condition + (HR_self + HR_other) : Condition|Group) : Type

This model was very heavy on our computers and it never finished running. We suspect that we would also run into convergence issues, if we were to keep running the model.

Model D:
HR_change_self ~ (0 + Condition + (HR_self + HR_other) : Condition) : Type +
(0 + Condition + (HR_self + HR_other) : Condition|ShuffleID) :Type +
(0 + Condition + (HR_self + HR_other) : Condition|Group) : Type

Model D is even more complex than *Model C*, and, not surprisingly, we had the same problems fitting our enormous dataset to this model. The benefit of *Model D*, however, is that it takes into account that the particular participant or group can influence heart rate change differently according to condition.

3) Do you observe differences in coordination between conditions? Report the models and results.

We have used the lmer model comparing different conditions to the baseline condition (conversation) as a method to infer whether coordination in other conditions was significantly different from the baseline condition. The change in the HR of self was modeled as a function of the heart rate of self and heart rate of the other and the interaction with the condition. We included random intercepts for condition and random slopes for participant.

Model E:
HR_change_self ~ 0 + Condition + (HR_self + HR_other) : Condition

Condition	Beta	SE	p-value
ConditionMovementCoop:HR_other	0.017	0.021	0.438
ConditionMovementGuided:HR_other	0.001	0.021	0.991

ConditionSynchronous:HR_other	-0.014	0.019	0.444
ConditionTurnTaking:HR_other	-0.035	0.019	0.065

Table 3 - output from analysis of model E where the conversation condition is a baseline

4) Is respiration coordination a likely driver of heart rate coordination? Describe how you would test for it.

First, we would calculate heart rate coordination by adding together the absolute value of the heart rate change of the self and that of the other. We would take the absolute value because two people can be either negatively or positively coordinated. Then, applying the same rationale, we would calculate respiration coordination and use it as a predictor of the heart rate coordination. We might include random intercepts to account for the variability of participants.

HR_coordination= |HR_change_self | + | HR_change_other|
HR_coordination ~ |Resp_change_self | + |Resp_change_other|

Bella Terragni- 201405868,
Peter Thramkrongart- 201806892,
Rūta Slivkaitė- 201805872 &
Bianka Szöllősi- 201808610

Meta analysis

Pitch Variability in Patients with Schizophrenia

link to github: https://github.com/PeterThramkrongart/Assigment-5

Introduction

When one examines the literature on distinctive vocal patterns in Schizophrenia, the current evidence implies that patients diagnosed with Schizophrenia seem to display atypical voice patterns, such as increased pauses, distinctive tone, intensity of voice and poor speech (Parola, Simonsen, Bliksted & Fusaroli, 2019).

Therefore, the objective of this report is to analyze the existing evidence for pitch variability in patients with a diagnosis within the Schizophrenic spectrum. We approach this question using meta-analyses that either compare mean pitch measures from studies or measures of standard deviations. In order to illustrate how the different studies each impact the metaanalysis and final results, we use forest plots. Lastly, our intent is to assess the validity of the theory in general by using funnel plots to investigate publication bias within the field.

Meta-analysis of pitch variability in Schizophrenic patients

Taking a look at the Prisma 2009 Flow Diagram, we find that after the removal of duplicate records, there are a total of 4341 studies collected, out of which 46 studies are included in the quantitative synthesis. This is roughly 1,06% of the total studies. Out of these 46 studies, we found 6 studies that matched our requirements for our mean model, and we found 15 studies that matched our requirements for our SD model.

For this meta-analysis we chose to analyze only pitch variability in the f0 fundamental, because this is the dominant pitch frequency. From the studies, we extracted certain features and measures depending on the model we want to train on the data. The features are mentioned below in *Table 1* and include different measures of pitch frequencies, type of task performed during recording, patients' diagnoses, and sample sizes.

Mean Model	SD Model
The type of task	The type of task
Diagnosis	Diagnosis
Sample size of diagnosed participants	Sample size of diagnosed participants
Sample size of control participants	Sample size of control participants
Mean of pitch means, Schizophrenia	Mean of pitch standard deviations, Schizophrenia
Standard deviation of pitch means, Schizophrenia	Standard deviation of pitch standard deviations, Schizophrenia
Mean of pitch means, healthy controls	Mean of pitch standard deviations, healthy controls
Standard deviation of pitch means, healthy controls	Standard deviation of pitch standard deviations, healthy controls

Table 1 - Elements of the two models

The two models we create in our paper in order to do our meta-analysis, are designated Mean Model and SD Model, respectively. The Mean Model uses the mean measures of pitch data from the selected studies, while the SD Model utilizes standard deviations of pitch in its analyses.

We extracted the standardized mean difference in pitch variability (SMD) between schizophrenia patients and healthy controls. As a measure of SMD, we used Cohen's D. We then conducted a meta-analysis consisting of mixed effects regression models. The models used were based on the following parameters:

Model: $\sim 1 + (1|$; $h = 1/(^2)$

Figures 1 and *2* display two forest plots for the mean model and SD model, respectively. Looking at the plots, it is evident how the model output is influenced by the different studies from which it is built. Investigating the forest plot for the model built on mean measures of pitch, we see that the model estimates a summary effect size of 0.16 (SE = 0.16, p = .2). On the other hand, the forest plot created from the model of standard deviations of pitches estimates a summary effect size of -0.25 (SE = 0.30, p = .4).

The results from the SD Model implies a reduction in pitch variability in Schizophrenic patients, but since the standard error is large and the interval crosses 0, the results are not dependable. On the other hand, the results from the Mean Model implies a higher variability in Schizophrenic patients' pitches. This is contradictory to the results from the SD Model, but these results are quite uncertain. The conclusion must be that the difference in pitch variability is small if not insignificant and that studies even disagree whether Schizophrenic patients have higher or lower pitch variability.

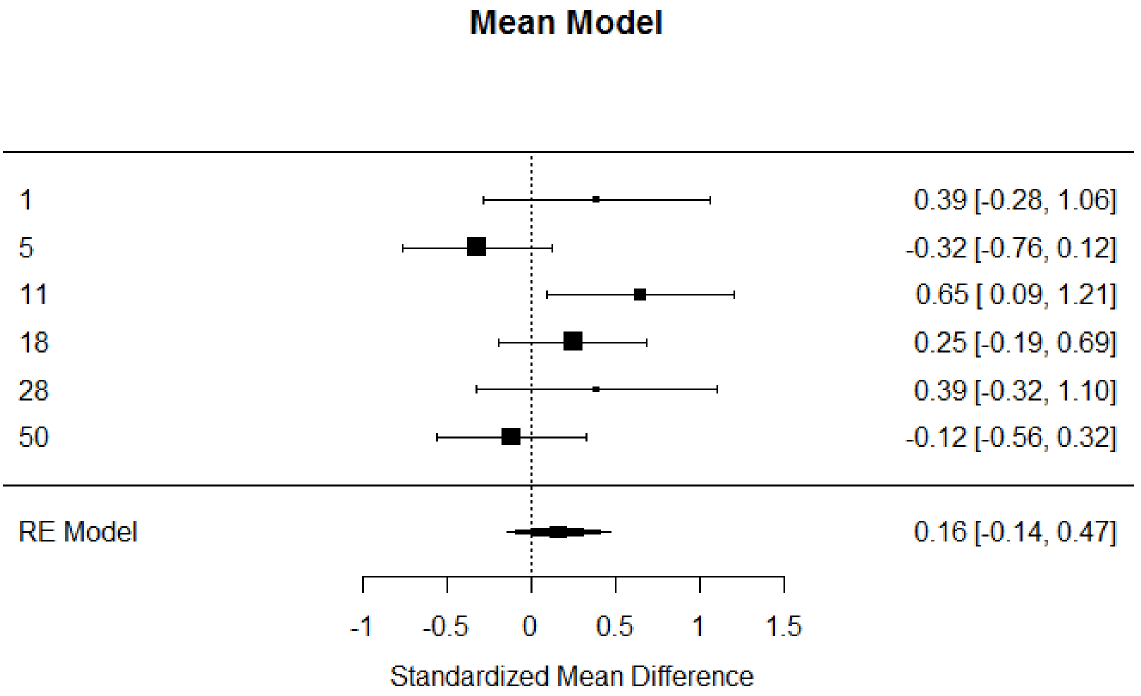


Figure 1 - Forest plot of the Mean Model

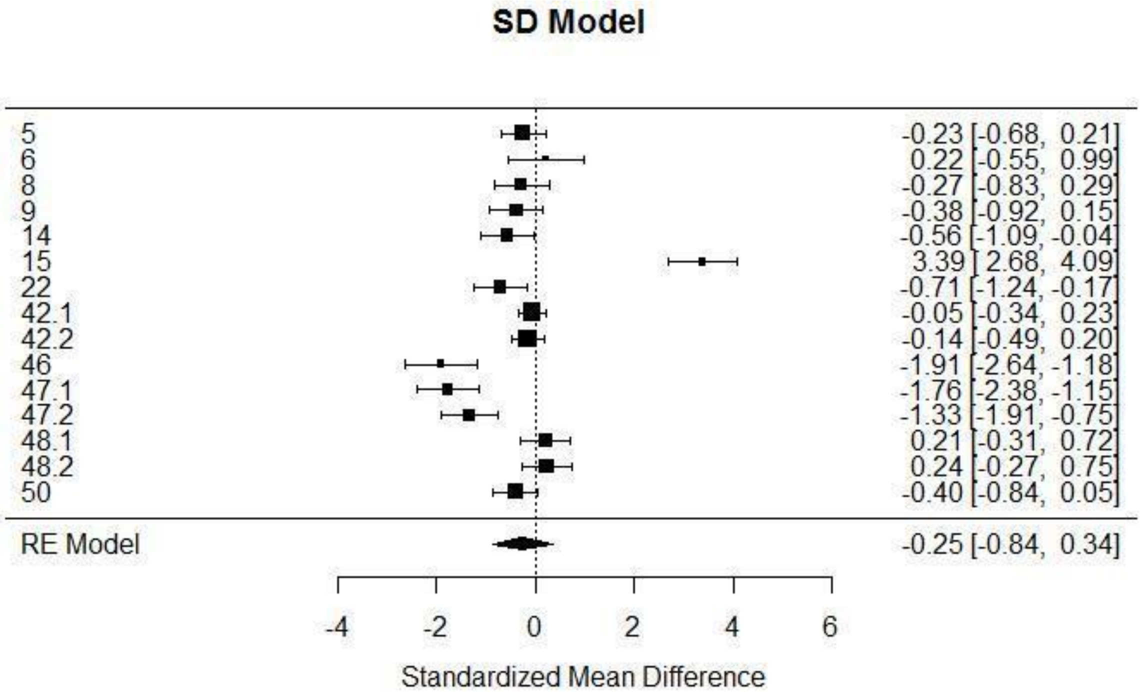


Figure 2 - Forest plot of the Standard Deviation Model

Incorporating own study to meta-analysis

In order to investigate how our study from assignment 3 would influence the meta-analysis, were it included in the calculations, we need to extract features from this study equivalent to those mentioned in *Table 1*.

After extracting these features, we want to calculate the effect size of the study in order to make a comparative analysis. These calculations, in concert with the recalculated forest plots, reveals that our study is not particularly influential, using the SD model, and slightly influential when employing the Mean model.

In the model relying on pitch means, the effect size of our study is -0.05, while the summary effect size is 0.11. This is a point 0.05 difference from prior estimations. However, in the standard deviation model, the summary effect size is -0.25, while the individual effect size for our study is -0.22. Here there is no difference from prior estimates.

Furthermore, if we interpret the forest plots depicted in *Figure 4*, we become aware of how our study in the SD model is close to the effect size of the entire model. Our study therefore supports and reinforces the general tendencies of the SD model meta-analysis thus far, even though the measurements are not particularly influential in changing the mean effect size. On the other hand, *Figure 3* depicts the mean model estimates. Here our study interval clearly pulls the general interval to the left, since our study effect size is more positive than the summary effect size.

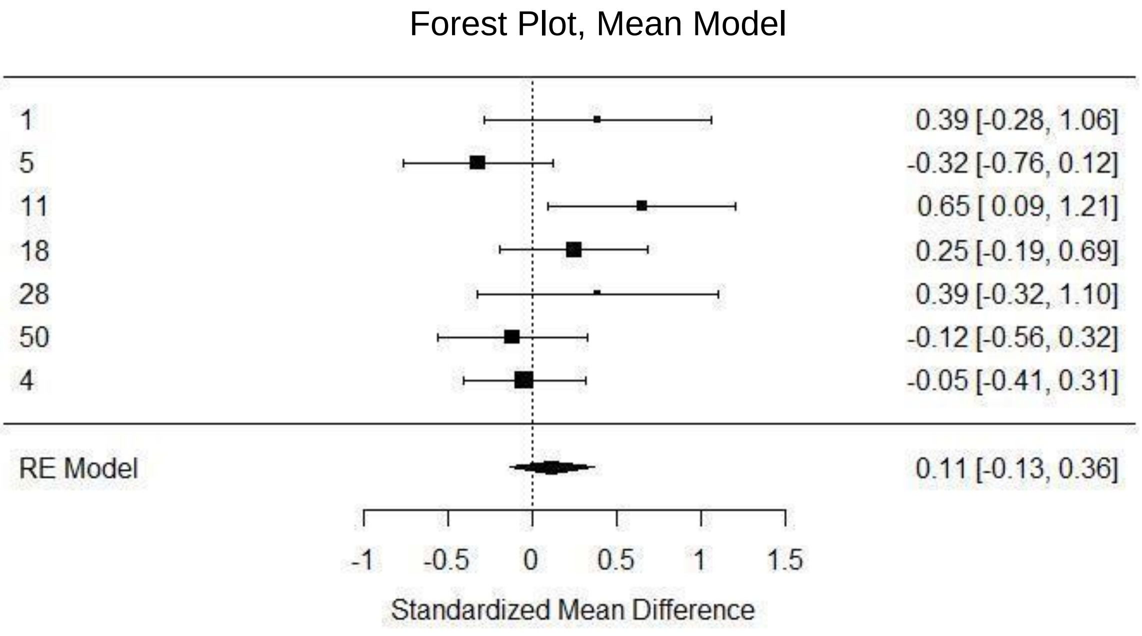


Figure 3 - forest plot of the Mean Model including our study 4

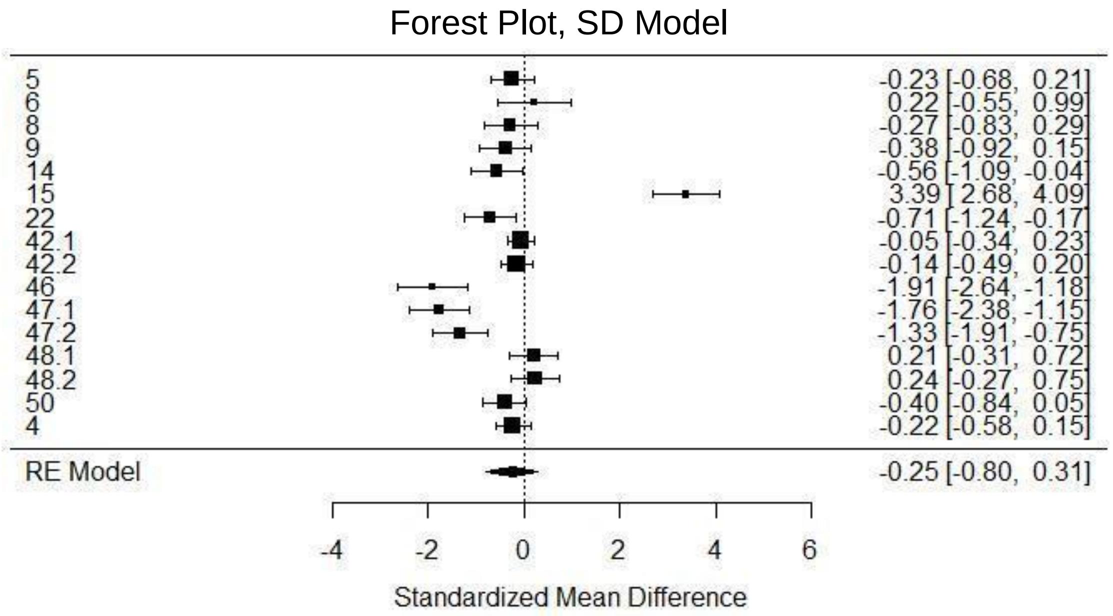


Figure 4 - forest plot of the SD Model including our study 4

Quality assessment of the literature

In order to check the heterogeneity of the chosen experimental literature, we plotted its data (which included the value of tau for each study among other measures) to see whether any particular study had an unwanted influence on the others. From looking at *Figure 6*, we can clearly see that in our SD model, study number 6 was highly abnormal and therefore quite influential on the metaanalysis. Therefore, if we removed this study from our analysis, we would get significantly different values. Values, which probably would improve the variability of the litterature.

From the diagrams depicted in *Figure 5,* made from the data from our Mean Model, we can see that these studies vary in a more random manner, and that no single study is responsible for creating large discrepancies in the data. However, the studies 1 and 5 are weighted in a manner, so that they have the capacity to influence the results more than the other studies.

Further, we extracted the actual values of tau and I^2 from our models. For our mean model tau = 0.071, I^2 = 50.29, while for our SD model tau = 1.30, I^2 = 95.37. These numbers signify that we see a moderate amount variance between studies when comparing mean measures of pitch, and a very high amount variance between studies when comparing SD measures of pitch. The very high amount variance in the SD model was probably due to the outlier that is study number 6 in *Figure 6*.

We have only a few studies to compare, and since we have not read the individual articles making up the meta-analysis, we argue that we are not able to make an informed decision about whether it would be justified to remove the mentioned outlier or not.

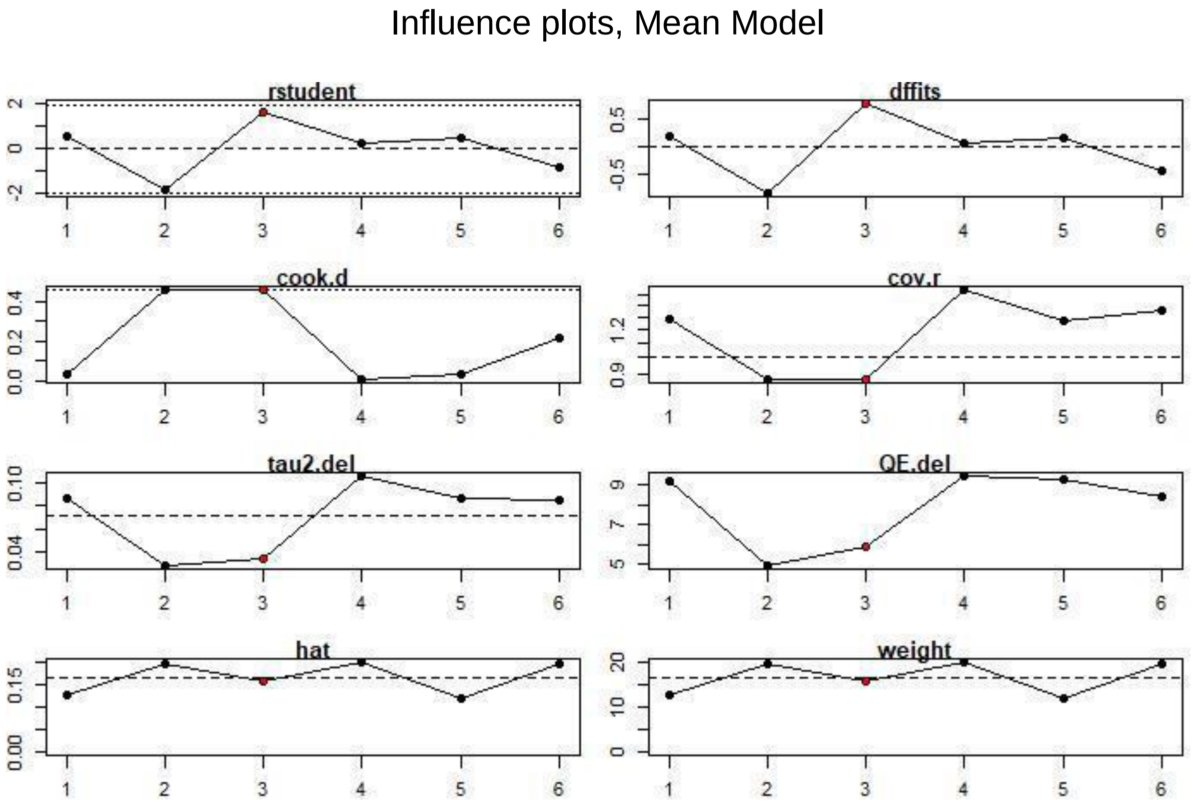


Figure 5 - Influence diagrams from the mean model

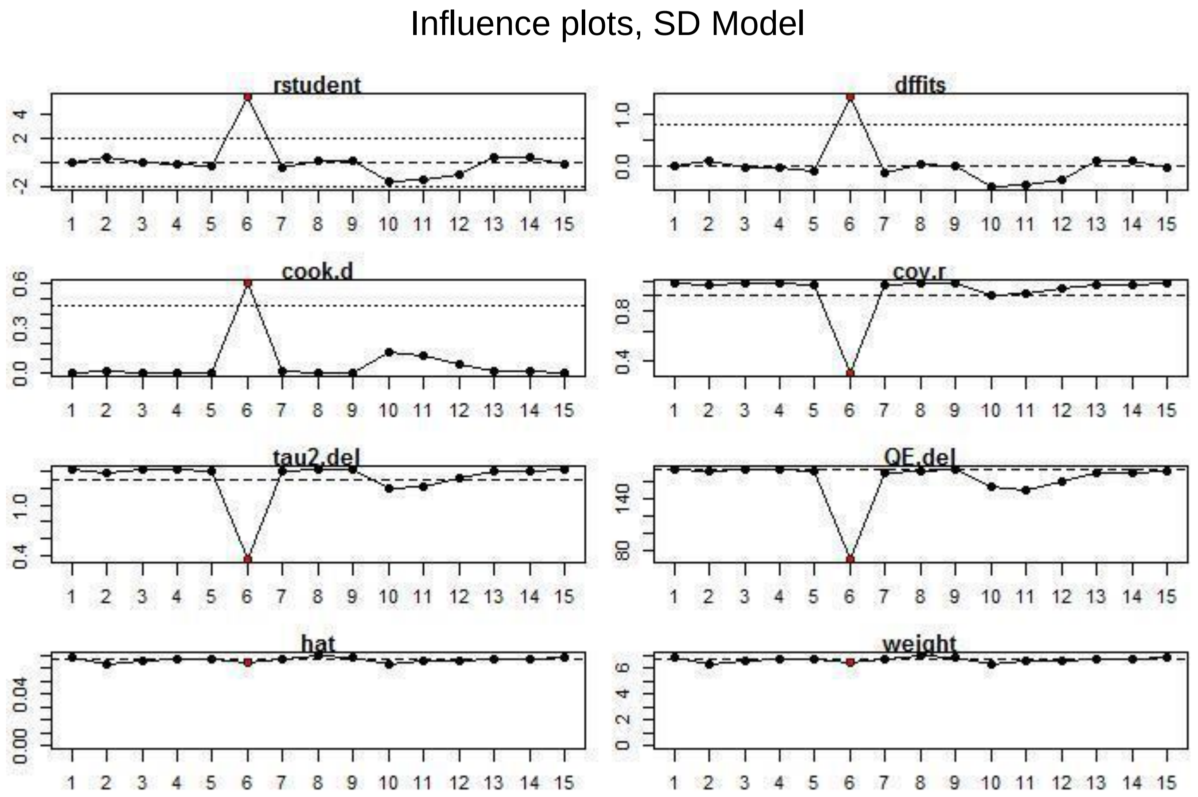


Figure 6 - Influence diagrams from the SD model

In order to test for publication bias, we create funnel plots to illustrate whether there is a tendency to only publish studies with significant results. Looking at the funnel plots in *Figure 7 and 8,* we can conclude that there is not a particularly big publication bias on distinctive vocal patterns in schizophrenia, as the funnel plots seem to include more or less the same amount of studies on each side of the central line of the plot. The figures also illustrate that studies with both higher and lower power were included in the metaanalysis, since higher-powered studies are placed towards the top of the funnel. The figures also indicate that studies with small standard errors were published alongside studies with larger standard errors. We can also observe that one study in the SD model's funnel plot seems to be an outlier, pulling the model towards higher values of standardized mean difference. This outlier might very well be the same as we identified in the influence plots in *Figure 6*.

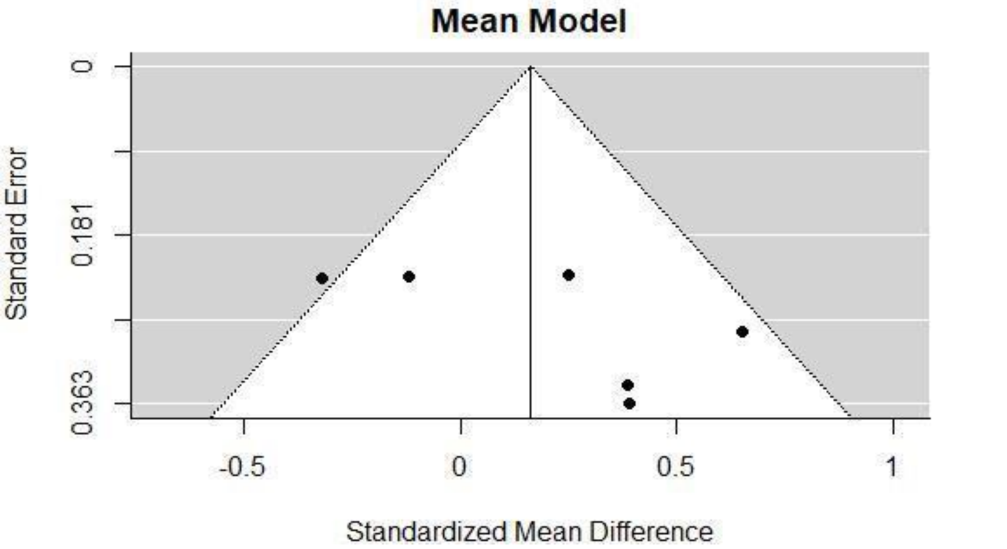


Figure 7 - Funnel plot of the Mean Model studies

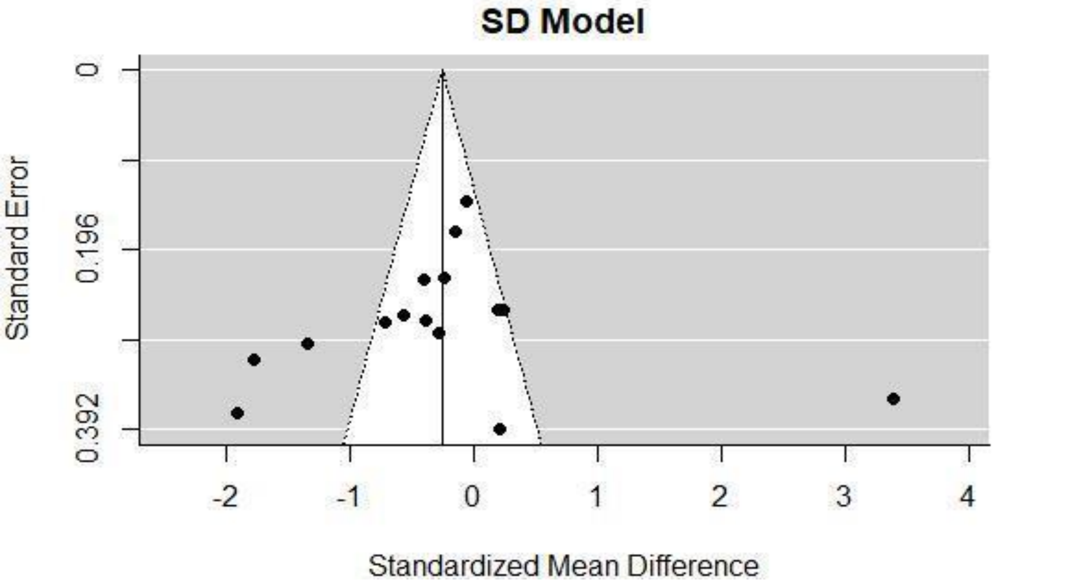


Figure 8 - Funnel plot of the SD Model studies