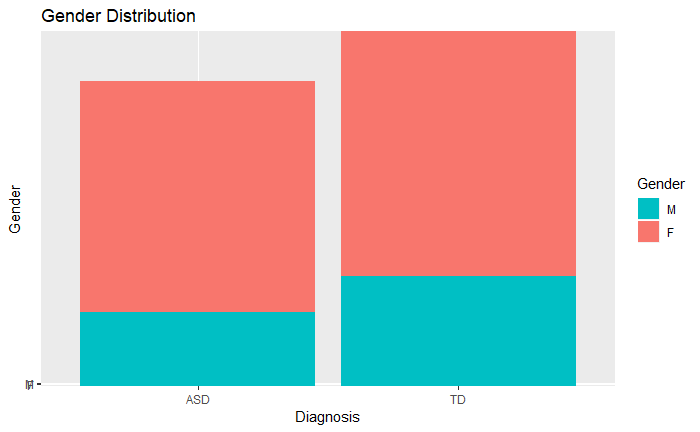
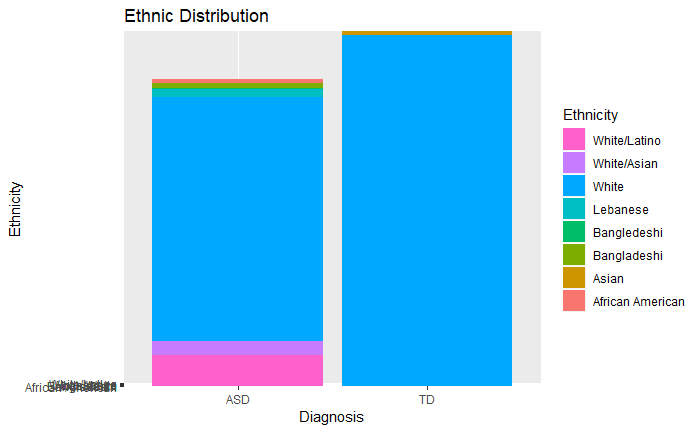
Language Development in ASD

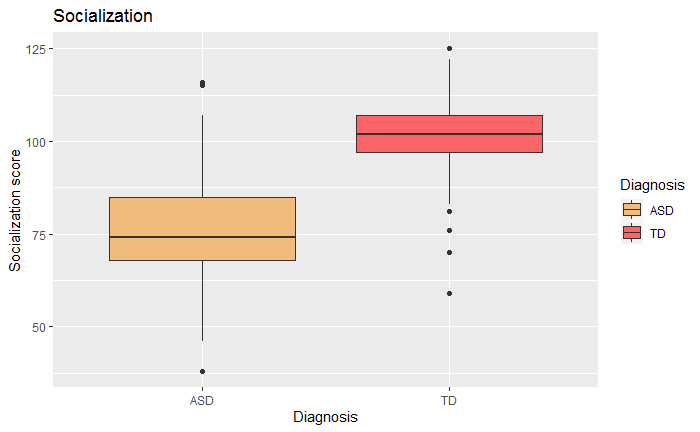
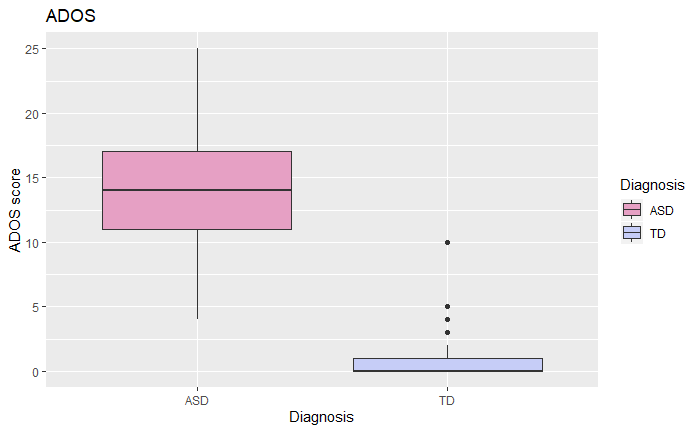
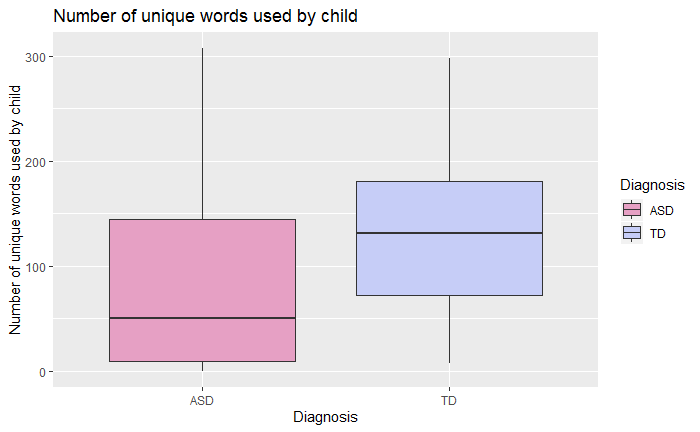
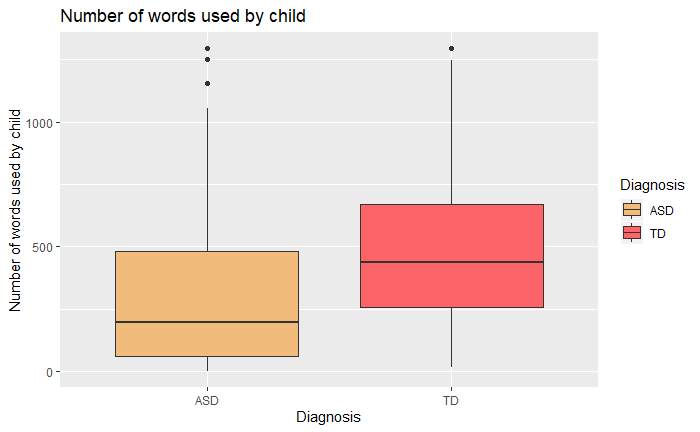
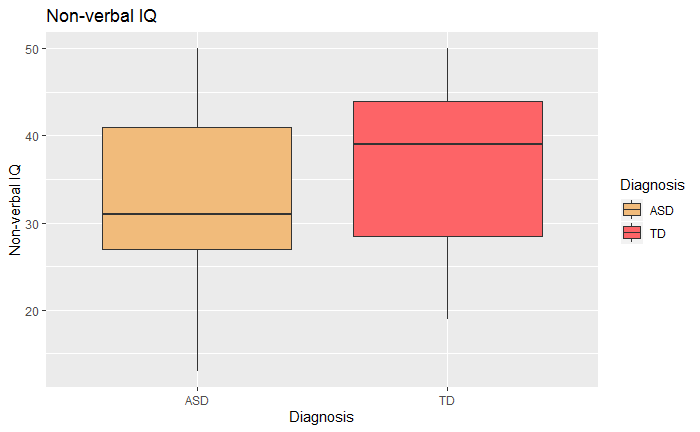
Find code on GitHub: <https://github.com/bellaterragni/assignment_2/blob/master/ASSIGNMENT_2_Bianka_Ruta_Peter_Bella.Rmd>

*Assignment 2, part 1, Experimental Methods 3*

***1. Data***

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.

*Figure 1*

The sample included 61 children, of which 32 were normally 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]](#footnote-1), females having superior verbal abilities at a young age.

*Figure 2*

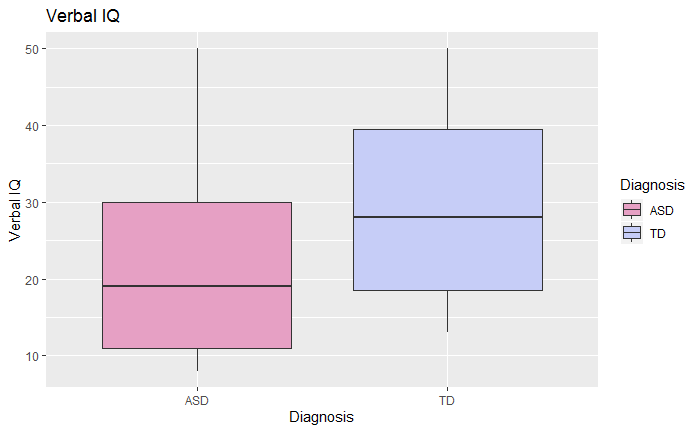
*Figure 4*

*Figure 7*

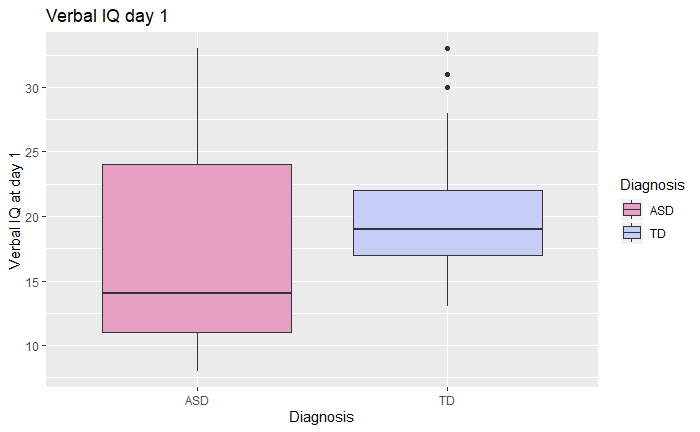
*Figure 6*

*Figure 5*

*Figure 3*



*Figure 8*

******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 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.

*Figure 9*

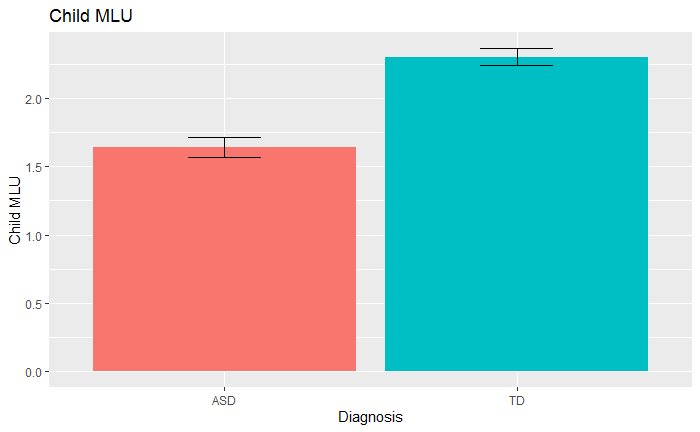
***2. Hypotheses***

H1A: A child’s MLU changes over time

H1B: A child’s MLU changes according to diagnosis

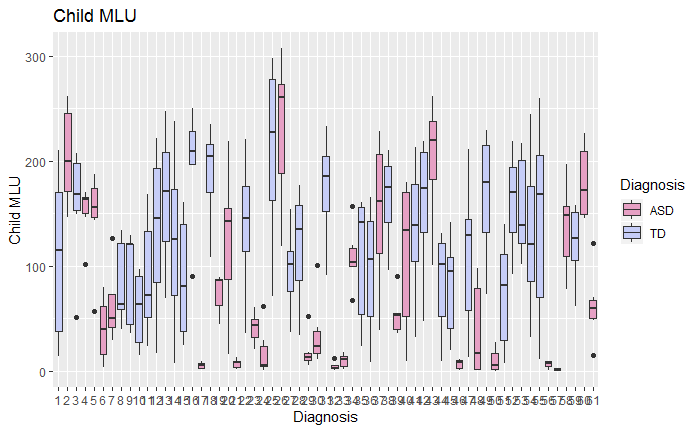
H2A: A parent’s MLU changes over time

H2B: A parent’s MLU changes according to diagnosis



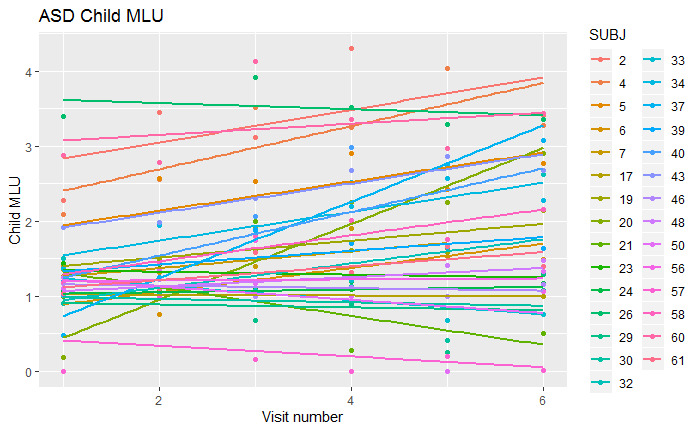
*Figure 10*

***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   
  
  
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.

*Figure 11*

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 visualize 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*

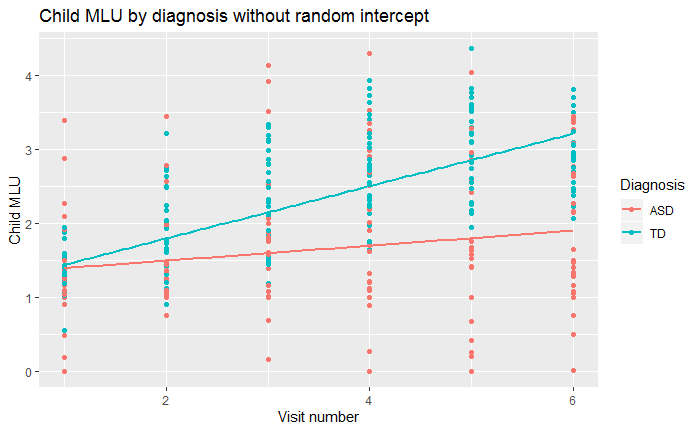
***4. Analysis***

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, (𝜷 = 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 (𝜷 = 0.65, t(60.76) = 4.06 p < .0001) and visit number 𝜷 = 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)

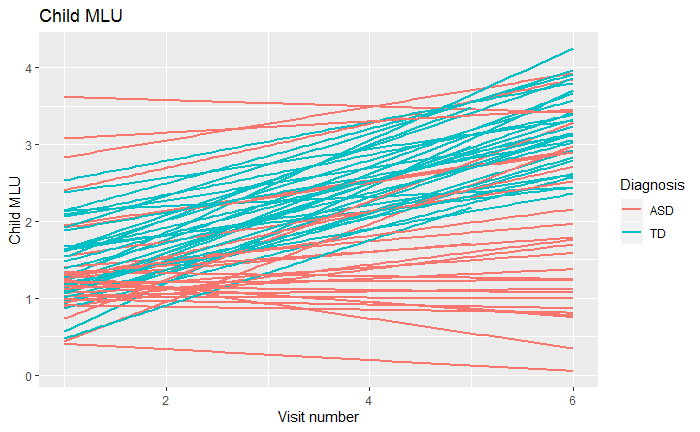
The statistical analysis of model 1c shows a significant effect of visit number on child MLU (𝜷 = 0.09, t(291.23)=4.92, p < .0001). There was likewise a significant result for the interaction effect between diagnosis and visit (𝜷 = 0.25, t(291.28)=8.96 p < .0001), while there was no significant effect of diagnosis alone on child MLU (𝜷 = -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 (𝜷 = 0.10, t(80.95) = 3.87 p < .0001) as well as the interaction effect between visit and diagnosis (𝜷 = 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 (*χ2*(7, 1) = 23.762, *p* < .0001).

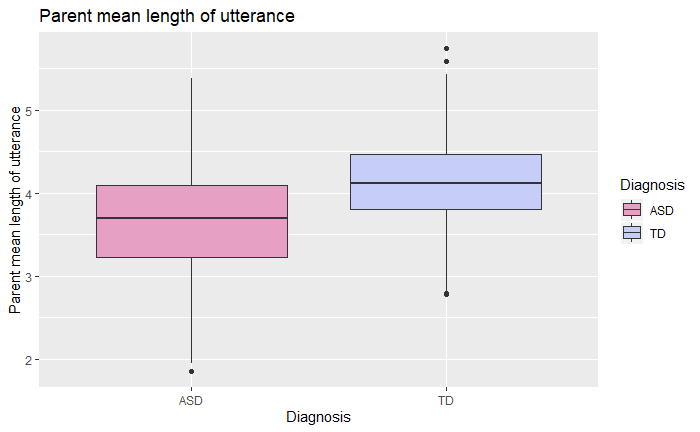
These three models, 1b, 1c, 1d, all capture significant parts of the variance, and they have in common that they all support that a child’s MLU increases over time (hypothesis H1A) and that diagnosis has a significant effect on the trajectory (hypothesis H1B) whether by itself as model 1b implies or 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’s diagnosis affects 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*

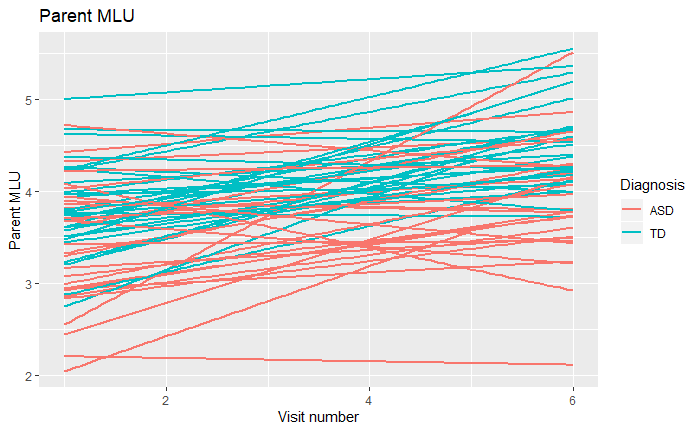
In order to test H2A properly, that is, whether the mother’s MLU is dependent on the visit number, a simple model 2a is formulated with independent values participant ID as a random effect and visit number as a fixed effect, while the dependent variable is parent MLU.

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

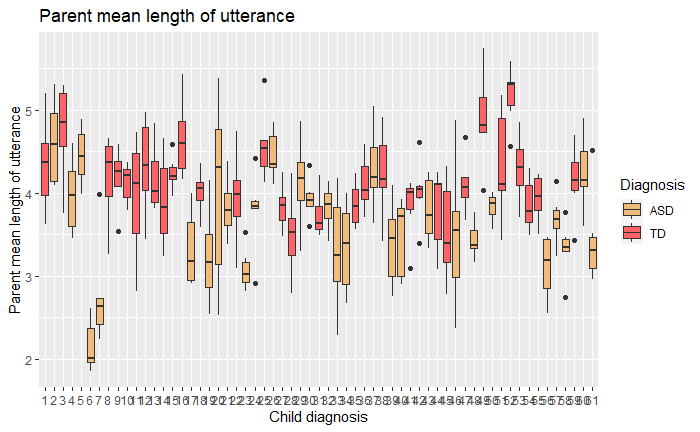
According to model 2a, visit number has a significant effect on the parents MLU (𝜷 = 0.12, t(291.57) = 8.66, SE = 0.014, p < .0001). The model thus suggests that the parent’s MLU increase over time, at about half of the pace that their children increase their respective MLU’s. These results suggest that 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 (𝜷 = 0.12, t(291.72) = 8.68, SE = 0.014, p < .0001), but also diagnosis (𝜷 = 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   
  
  
  
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.

*Figure 17*

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 interpretability. 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 (𝜷 = 0.063, t(208.13) = 4.82, SE = 0.013, p < .001), the interaction between visit and diagnosis (𝜷 = 0.67, t(291.59) = 7.36 , SE = 0.091, p < .001), and the interaction between visit and verbal IQ (𝜷 = 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, (𝜷 = -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

*Assignment 2, part 2, Experimental Methods 3*

***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

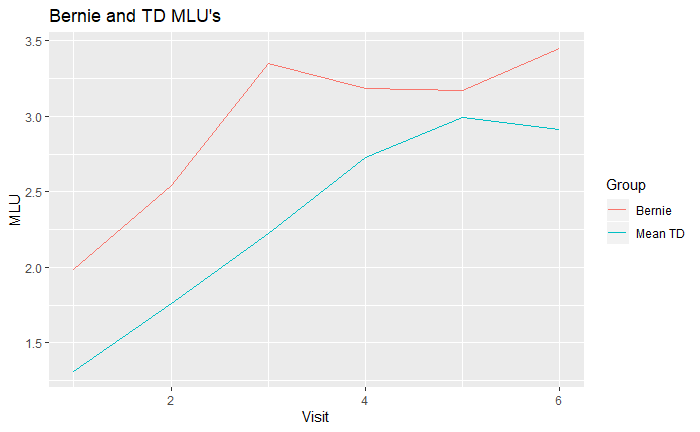
Another way of determining the validity of a model’s predictions is via cross-validation. The procedure is similar to that of the comparisons made above. When doing a cross-validation, however, the model is trained and tested on the entire data corpus instead of only being trained on one chunk, and only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

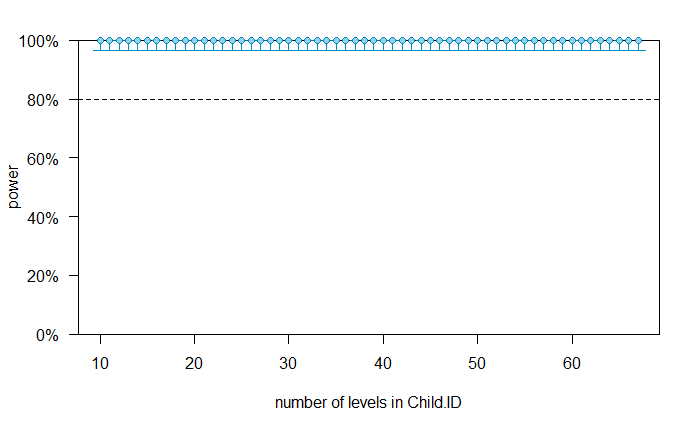
Having tested our best fitted exploratory model and our basic model using several different methods, it becomes evident that the assessments of the models are highly dependent on which methodology we decided use. This is an important point to make, since this flexibility grants researchers a lot of control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

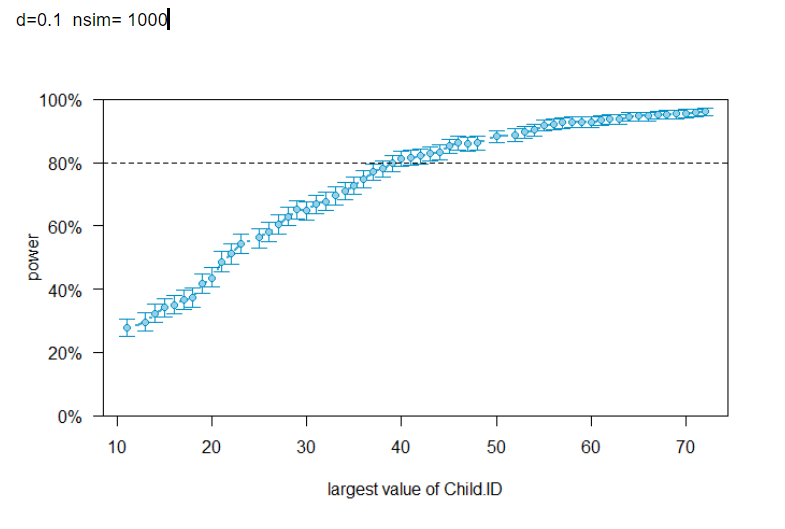
We can use the power analysis estimates for estimating the likelihood of successfully rejecting the null hypothesis (that child's mean length of utterance does not change over time or according to diagnosis) when diagnosing other children. The likelihood that is big enough (>80%) when the minimal interesting effect (aka Cohen´s d) is 0.5 would give us confidence that we will find the moderate size 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.

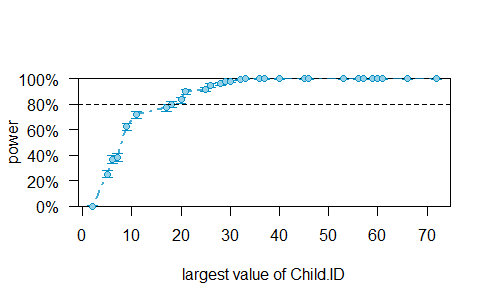
*Figure 20*



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*



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 [↑](#footnote-ref-1)