1 Method

1.1 Participants

The present study involved 86 subjects with age ranging from 86 months ($\sim$ 7 years) to 205 months ($\sim$ 17 years). At the time of the recruitment the mean age of the subjects was 144.08 months ($\sim$ 12 years), with a standard deviation of 30.90 ($\sim$ 2.60). Reclutamento: essere sotto le 2 deviazioni standard in almeno un valore della dde e/o rientrare nella diagnosi della dislessia con la lettura di brano.
Osservazione: quali sono i criteri di inclusione? Se il criterio di inclusione e' essere carente in uno degli score della dde (quindi punteggio <=-2.0), allora i soggetti 509 e 734 perché sono stati reclutati?

```
    id class comorb age wspeed wacc nwspeed nwacc errtc errsub QI VCI PRI WMI
58 509     0      0 190  -1.15    1   -1.72   1.0     5      0 96 102  92 115
77 734     0      0 146  -0.60    1   -0.90   0.6     0      0 95 100  99 115
   PSI dc so mc cf vc co rs errtot perc.errtc perc.errsub bakker
58  97  8  9 12 10  8 13  9      5          1           0      P
77  85 10 11 15  5 10 12 10      0          0           0      M
```

1.2 Procedure

Since we based our classification on several criteria, such as DSM-V, clinical evaluation and DDE, firstly we proceeded with a qualitative data visualization, trying to assess the relationship between the classes and the DDE scores. We plotted the accuracy (both for words and non-words) against the speed (both for words and non-words) and we observed the general trend divided by class. The general trend for all the three classes is decreasing when plotting non-word accuracy against non-word speed. Subject "673" was clearly outside the distribution and we decided to discard it, treating it as an outlier. Hence we reduced our database from 96 subjects to 95 subjects. Moreover it can be observed that the subjects of third class, that of the subjects classified as "severe", assume spreader values than the subjects classified as "low" and "medium". It is straightforward that the classes with a lower impairment tend to display subjects closer to each other and with a less steep decreasing trend. When inspecting the classes distribution in the case of the plot of word accuracy against word speed we observed that the trend of the "severe" class is increasing, whereas the trend of both the "low" and "medium" classes is increasing. Hence, higher the accuracy, higher the speed for the third class, while higher the accuracy, lower the speed for the first and second classes. It is straightforward that the values of the third class tend to be lower than the other and in most cases lower than the cutoff of -2.0. In this case too it was possible to observe that the two higher classes tend to be closer to each other, while the third class spreads its values assuming a wider range of values.
From these graphs we could observe that accuracy and speed modulate the three classes and show that classes 0 and 1 behave in the same way and assume close values. It can be hypothesized that class 1 can be seen as a class 0, dividing the dataset into a two-classes classification. In order to state that the Kruskal-Wallis test was made. The Kruskal-Wallis test is a non-parametric method for testing whether samples originate from the same distribution. We tested the DDE scores for classes 0 and 1 and for classes 1 and 2 respectively. We expected to have the same distribution (accept the null hypothesis) in the case of classes 0 and 1 and to reject the null hypothesis, hence statistically different distributions between classes 1 and 2. The results obtained are shown in Tables 1,2. As

we can observe the results are in line with what we expected. In fact we can reject the alternative hypothesis given the p-values in Table 1, hence the DDE scores correspondent to classes 0 and 1 have the same distribution. Moreover, from the p-values in Table 2, we accept the alternative hypothesis stating that the DDE scores correspondent to classes 1 and 2 have different distribution.

dde.score	chi.sq	p.value
word speed	3.91	0.05
non-word speed	3.12	0.08
word accuracy	3.89	0.05
non-word accuracy	0.59	0.44

Table 1: Classes 0 and 1.

dde.score	chi.sq	p.value
word speed	7.880	0.005
non-word speed	9.920	0.002
word accuracy	5.920	0.015
non-word accuracy	4.860	0.028

Table 2: Classes 1 and 2.

We merged the "low" and "medium" classes and we test the new classification, in order to prove that the distribution still differ with the new classification, hence we run the Kruskal-Wallis test again. Results are shown in Table 3.

dde.score	chi.sq	p.value
word speed	37.73	$<$0.001
non-word speed	31.70	$<$0.001
word accuracy	22.63	$<$0.001
non-word accuracy	15.70	$<$0.001

Table 3: Merged 0 and 1 classes and class 2.

This merge could be done, and we obtained a dataset with 63 subjects of class 0 and 32 subjects of class 1.

We were then able to perform *binomial logistic regression* in order to predict the binary classification of our data based on the regression variables: DDE speed scores, WISC subscales. We divided the subscales in performance subscales (cf, rs, dc), verbal subscales (so, vc, co) and digit span (mc).
Firstly we split the data into two chunks:training and testing set. The training set is used to fit our model, which is tested over the testing set. The data partition was made with a training set that accounted for the 60% of our data. The remaining 40% became the test set. Thirty-eight subjects from class 0 and twenty from class 1 were selected in the training set, while twenty-five from class 0 and twelve from class 1 formed the test set. We fit the model with DDE speed scores and performance subscales, with binomial family, and we obtained the following result:

```
Call:
glm(formula = form.vis, family = binomial, data = db.train)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
```

```
-1.48409  -0.19005  -0.00893   0.09326   2.38543

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -23.82548   13.67072  -1.743   0.0814 .
nwspeed      -4.40989    2.55988  -1.723   0.0849 .
wspeed       -1.47476    1.98300  -0.744   0.4571
cf           -4.24600    2.15112  -1.974   0.0484 *
dc            1.43139    1.02859   1.392   0.1640
rs            2.54168    1.28095   1.984   0.0472 *
nwspeed:cf   -0.01648    0.19989  -0.082   0.9343
nwspeed:dc    0.04685    0.21106   0.222   0.8244
nwspeed:rs    0.32815    0.20358   1.612   0.1070
wspeed:cf    -0.93838    0.44929  -2.089   0.0367 *
wspeed:dc     0.36764    0.26414   1.392   0.1640
wspeed:rs     0.19889    0.22830   0.871   0.3837
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 74.726  on 57  degrees of freedom
Residual deviance: 23.436  on 46  degrees of freedom
AIC: 47.436

Number of Fisher Scoring iterations: 9
```

As we can see, the *cf* subscale, the *rs* subscale and the interaction *cf, wspeed* are statistically significant ($p < 0.05$). The negative coefficient for the cf predictor suggests that, all other variables being equals, subjects with higher values of the cf predictor, are less likely to be classified as “severe”. Moreover, the positive coefficient for the rs predictor suggests that, all other variables being equals, subjects with higher values of the rs predictor, are more likely to be classified as “severe”. Eventually, the negative coefficient for the interaction term, we recall that the DDE scores values are negative standard deviations, suggests that the variation of the two predictors reduces the log odds by 0.94 of being classified as “severe”.
We decided to consider these predictors because of what we stated in our hypothesis. Moreover we tested, through the analysis of deviance, the model against the reduced model with as predictors only the speed scores of DDE and we preferred the model with WISC performance subscales ($p < 0.05$). The model selected is preferable even to the null model ($p < 0.001$).
Now we would like to see how the model is doing when predicting the class on a new set of data. In order to evaluate the performance of the classification we computed the Matthews Correlation Coefficient (MCC), this because we have an unbalanced classification. We obtained an MCC of 0.56.
Iterating the procedure with the verbal models we obtained that this model is preferable than the reduced ($p < 0.001$) or null models ($p < 0.001$), but nothing significant was found. On the other hand the model with the digit span as predictor is not preferable than the reduced model ($p = 0.53$), hence we did not take it into account.
Since the MCC result is somehow dependent on the split of the data, we performed a 10-fold cross validation (CV) and we obtained an MCC (mean of the MCCs) of 0.53, hence similar to the one obtained before. We recall that the MCC index assumes val-

ues between -1 and 1, with 1 that is the perfect prediction, 0 that is no better than random prediction and -1 that represents total disagreement. In order to check that our classification results actually depend on the predictors, in particular the ones that resulted significant, and it does not depend on fate, we performed a random label classification. We randomized the class labels for our data, a hundred times, and we refitted the logistic regression on training data (accounting for the 60% of the dataset) and we tested it on the test set. We computed the MCC for each loop and then we computed the mean, obtaining a value of -0.023. Hence we can conclude that our classification is quite good.