# Reporting results

## Model performance compared to baseline and topline

The topline LDA model that have been trained and tested on the full dataset from the replication experiment had a mean AUC-ROC score of 0.604 with 95% CI [0.598, 0.611], while the baseline model had a mean AUC-ROC score of 0.500 with 95% CI [0.500, 0.500]. The full LDA model that have been trained on the validation data from Maldonado et al. (2019) and tested on the full dataset from the replication experiment had a mean AUC-ROC score of 0.575 with 95% CI [0.564, 0.585].

The full LDA-model does perform close to the topline performance, but not that much higher than chance, which is just assigning random labels to the trajectories. The full LDA-model only has a mean AUC-ROC score 0.075 higher than chance. This paper will argue that it is not that much better than chance.

A distribution of the assigned LDA-value for each trial condition can be seen in figure X.

|  |  |  |
| --- | --- | --- |
|  | Mean | 95% CI |
| Topline LDA | 0.604 | [0.598, 0.611] |
| Full LDA | 0.575 | [0.564, 0.585] |
| AUC | 0.575 | [0.564, 0.585] |
| MAD | 0.569 | [0.557, 0.580] |
| X-reversals | 0.558 | [0.547, 0.569] |
| Coords LDA | 0.541 | [0.535, 0.547] |
| X-flips | 0.541 | [0.535, 0.547] |
| Baseline | 0.500 | [0.500, 0.500] |

Chart

Description automatically generated

## Model performance compared to other traditional mouse-tracking measures

Chart, line chart

Description automatically generatedAs can be seen in figure X and table X the full LDA-model performs quite similar to almost all of the traditional mouse-tracking measures. AUC, MAD, and X-reversals all have overlapping 95% CI with the full LDA-model. X-flips performs the worst of them all with a mean AUC-ROC score of 0.541 and 95% CI [0.535, 0.547]. An LDA-classifier only trained on the x- and y-coordinates from the validation data has a performance indistinguishable from x-flips, with a mean AUC-ROC score of 0.541 and 95% CI [0.535, 0.547].

## Optimisation of number of PCA-features

Notes:

* Compare to validation experiment (perhaps also negation?)
* Mention that it is performance on bootstrapped samples with increasing number of participants?