LDA for Mouse Tracking

November 4, 2017

Contents

1 Introduction

- In the past ten years, mouse-tracking has become a popular method to target the processes underlying decision making in different domains, ranging from phonetic competition (Spivey, Grosjean, & Knoblich, 2005), and syntactic and pragmatic processing (Farmer, Anderson, & Spivey, 2007; Tomlinson, Bailey, & Bott, 2013) to social cognition (Freeman & Ambady, 2010; Freeman, Dale, & Farmer, 2011).

- In its **basic form**, the standard task used in mouse-tracking experiments is a two-alternative forced choice where participants are presented with a stimulus and then they have to make a choice using the options appearing in the top left or right corner. In each trial, participants have to press a "Start" button, located at the middle bottom of the screen, to see the stimulus. As participants move the mouse cursor to select the response, x, y coordinates are recorded.

- Based on the assumption that motor responses are prepared in parallel to cognitive processing and performed in a cascade manner (i.e., as fast as they can be executed) (Song & Nakayama, 2006; Freeman & Ambady, 2010, Spivey & Dale, 2006), mouse-tracking studies have usually assumed that mouse paths can reveal the dynamics of decision making with fine-grained temporal resolution (Freeman et al., 2011;

Freeman & Ambady, 2010; Hehman, Stolier, & Freeman, 2014). As a result, certain features in mouse trajectories are thought to be indicators of specific decision patterns[1].

- For instance, whenever a decision involves two independent processes –such as a change of decision–, mouse trajectories are expected to be displayed as two movements, whereas a single smooth and graded movement would reflect a commitment with an initial choice (Wojnowicz et al., 2009).

- An example of how a *two-step* processing pattern is concluded from mouse trajectories is given in Dale and Duran (2011).

  - Dale and Duran (2011) compared the processing of negative and affirmative sentences to test the traditional hypothesis that negation is an operator that reverses the truth conditions of the sentence, inducing a shift in the verification strategy: truth conditions for the positive content are first derived and negated only afterwards (Wason & Johnson-Laird, 1972, REF).
  - The authors combined a mouse-tracking parading with a Truth Judgment Task: participants had to decide whether a general statement (e.g. *Cars have (no) wings*) was true or false by clicking on two response buttons located at the top left and right corners.
  - The reasoning behind the paradigm was the following: if negation involves representing the positive argument at an early processing stage, then when participants have to verify a negative sentence such as *Cars have no wings*, they should initially process the positive content and go towards the "false" response, flipping the direction as a second step.
  - Dale and Duran (Experiment 1) found that negation seems to involve a "two-step" verification: mouse trajectories presented more shifts towards the alternative response (i.e. were less straightforward) when evaluating true negative than affirmative sentences.
  - These results suggest that the positive argument may play an important role in negation processing (see Fischler et al. 1983; Hasson and Glucksberg 2006; Kaup et al. 2007a; Ludtke et al. 2008 for similar findings). This, however, is not necessarily the case (Nieuwland and Kuperberg, 2008). Whether or not the positive content of a negative sentence needs to be represented for comprehension seems to depend on the amount of contextual support given for the sentence. Specifically, the "two-step" processing seems to occur for sentences presented out of the blue, whereas no such effect comes about when the right contextual support is provided. How to explain this pattern of results has been at the center of the debate in the negation processing literature (see Y. Tian and R. Breheny XX for review). In this paper, we will not commit to any particular explanation about why negation might or might not induce such processing switch.

1.1 Goals

- As observed, mouse trajectories are assumed to reflect decision making (i.e. infer certain decision from paths). This association, however, has never been explicitly tested. **In this paper, we aim to test this connection between cognition (decision making) and action (mouse trajectories)**. Are decisions reflected on trajectories?

- Instead of taking mouse trajectories as indicators of cognitive processes, we will *manipulate* whether or not our stimuli trigger a flip in decision, and identify the features in mouse trajectories that correlate with these flips. We reasoned that trajectories for trials that involve a change of decision should all share meaningful components.

- A traditional *linear discriminant analysis* (henceforth, LDA) will be fed with these (already labeled) mouse trajectories. The LDA is an optimal solution to classify continuous data into two or more categories. Thus, it will allow us not only to classify mouse trajectories according to whether there

[1] In this sense, mouse trajectories are equivalent to eye movements. However, the main advantage of mouse tracking over eye tracking is the simplicity of the set up, which can be even tested online.

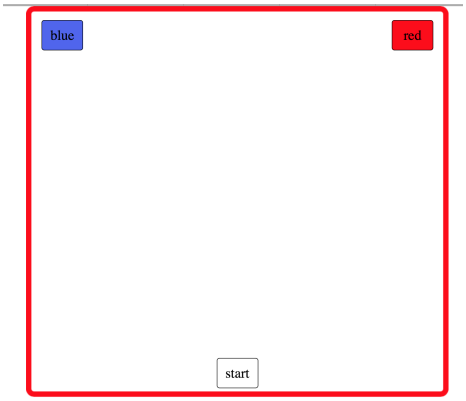


Figure 1: Example of Trial in Calibration Experiment

was a switch of decision, but potentially also to make a second distinction between trajectories that underly a rapid commitment and trajectories that underly uncertainty. We won't explore this issue here.

- In the last section of the paper, the LDA classifier will be further tested with new data, obtained from a replication of Dale and Duran (2011).

2 Manipulating decision making: Validation Experiment

We developed an experiment where participants had to perform a *two-alternatives forced task*: at each trial, they were presented with a coloured frame surrounding the screen and they had to determinate whether the frame was blue or red. Responses were made by clicking on the "blue" or "red" buttons, allowing the recording of mouse-movements during each trial. Importantly, responses were considered accurate if they described the color at the moment of the click. In order to mimic decision processing, we manipulated whether the color of the frame remain stable along the trial or changed at some point. An illustration of a trial is provided in Figure 1.

Participants 54 participants were recruited using Amazon Mechanical Turk (F=27). Two subjects were excluded from the analyses because they did not use a mouse to perform the experiment.

Design

- There were two Decision Types: (a) *Straightforward*, where the decision was made at the beginning of the trial; (b) *Switched*, where the decision made at the beginning of the trial changed at some later point.
- The final Frame color (at the response moment) could be red or blue.
- The Decision point depend on the Decision Type. For *switched* trials, there were three possible decision points: early, middle and late. These were determined by the y-position where the color changes.
- The proportion was adjusted so that straightforward trials were the majority (2/3), and people could get used to make their choice right away (without staying in the middle of the screen).

Decision Type	Frame color		Decision point
Straightforward	Blue		–
	Red		–
Switched	Blue	Red	early (y1=.4), middle (y2=.7), late (y3=.9)
	Red	Blue	early (y1=.4), middle (y2=.7), late (y3=.9)

Table 1: Design in Calibration Experiment

Interface The interface was programmed using JavaScript. Mouse movements triggered the extraction of x,y-pixel coordinates (i.e., no constant sample rate). The software was adapted proportionally to the window of the participant's browser, forming a rectangle: the height was covered at 100 percent and the width was 120 percent of the height. Three buttons were displayed during the experiment (*start* and response buttons). Their size was also determined by the browser window (i.e., approximately 20 percent of the total width). The *start* box was placed at the bottom center of the screen. The two response boxes were located at the top left (*blue*) and top right (*red*) corners of window. This location was constant across participants, and handedness was controlled. In each trial, mouse movements were recorded between start-clicks and response-clicks. The x,y-pixel trajectory was saved together with its raw time. Afterwards, the positions were normalized according to participants' window size, to allow comparisons between subjects. The normalization was done by considering the start button at the [0,0] point, the "blue" button corner at [-1,1] and the "red" button at [1,1].

Data treatment Mouse-tracking data are particularly variable trial to trial. On one hand, variations in response times imply different quantity of x,y positions per trial, making difficult the comparisons between items. On the other hand, in our design, positions are extracted based on mouse movements, and devices with more or less sensibility could influence the number of samples taken during the trial. In order to compare mouse trajectories, we normalized the time course into 101 proportional times steps (percentage of trial duration). This normalization, as all the other calculations, was performed in the Spyder environment using Python 2.7.

2.1 Overall performance

- Participants performed the task accurately, making their decision based on the colour of the frame at the time of the response. Inaccurate trials (less than 4%) were removed from the analyses.

- Mean trajectories for each DECISION TYPE and DECISION POINT are illustrated in Figure 2. Temporal information about changes in the x coordinate is provided in the appendix (Figure X). These trajectories suggest that people do react to the change of colour in the expected way; namely, they make a first decision when they see the colour frame, and revise this decision if needed. It's unclear whether the initial decision is made as soon as they press the *start* button and see the frame or if there is some initial uncertainty where participants stay in the middle of the screen.

3 Classifying decision processes with LDA

3.1 Description of LDA classifier

- In order to best recognize the (mouse) patterns characteristic of each class (*switched* vs. *straightforward*), we use a LDA method for classification. This algorithm assumes that different classes have a common covariation matrix, and finds the linear combination of predictors that gives maximum separation between the classes. This linear combination of predictors is obtained as a linear coefficient and it can be used to form a decision rule.

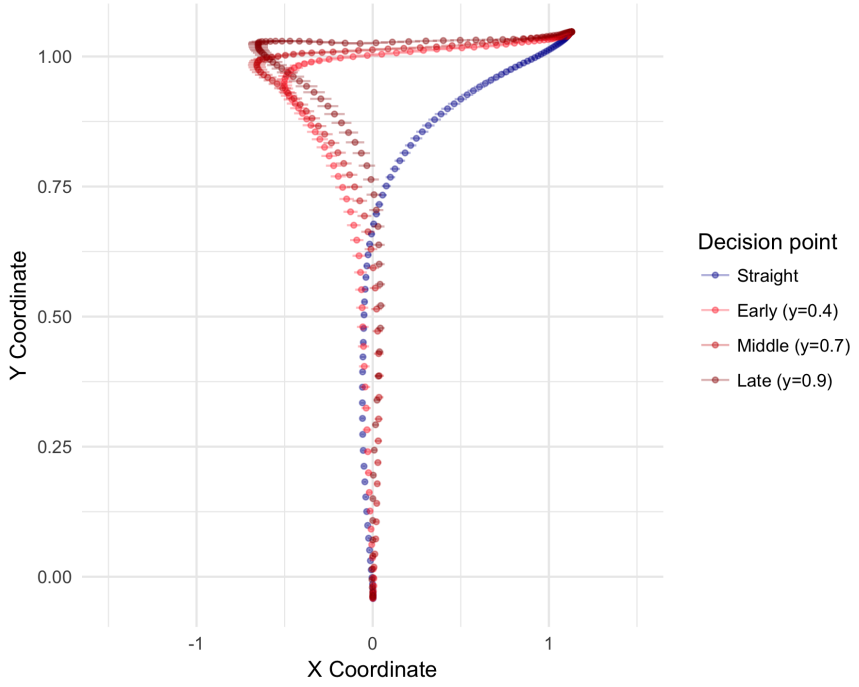


Figure 2: Mean trajectories per class

- The predictors used by the classification algorithm were (a) the *x,y* coordinates, (b) Euclidean-based velocity, and (c) Euclidean-based acceleration. Trajectories were labeled in two classes: *straightforward* and *switched*. The idea being that coordinates give us information about where the mouse is when, while acceleration and speed give us information about how did we arrive there.

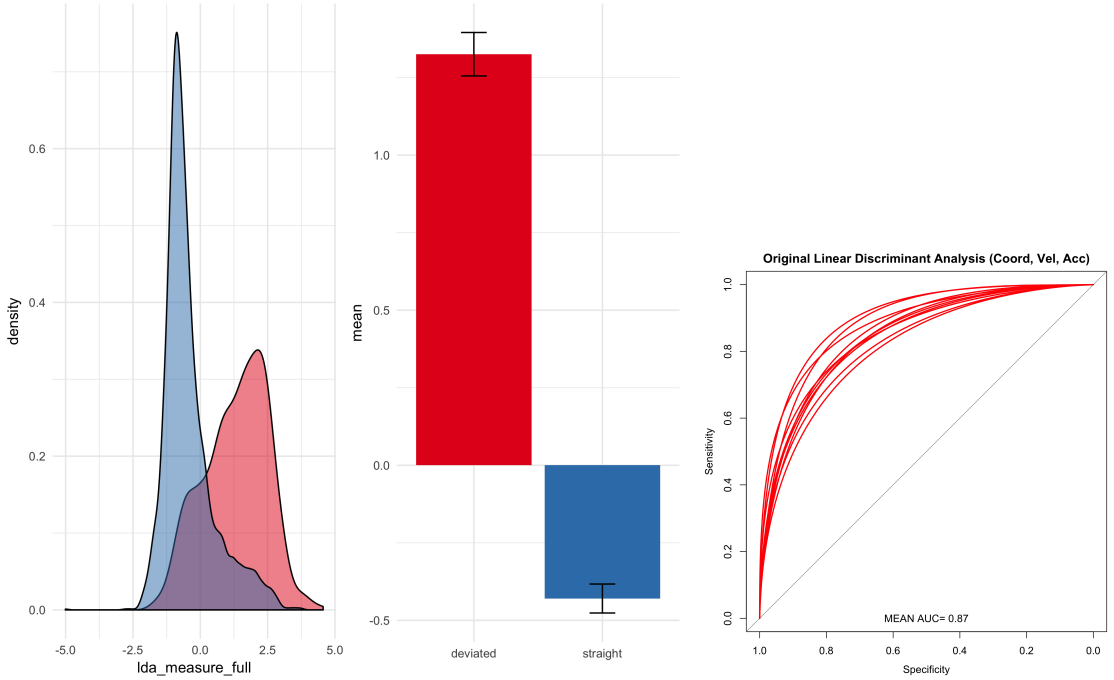
- To avoid collinearity issues, we used a PCA to identified 13 principal components on the features describe above. We perform the LDA on these 13 components. We obtained a *LDA measure* for each trial by calculating the dot product of the linear coefficient of the PCs and the PCs themselves.

3.2 Classifier performance

- The distribution and mean *LDA measures* for the trajectories of each condition are illustrated in Figure 3a.

- The overall performance of the classifier was evaluated by calculating the area under the ROC curve. The ROC curve diagnose the classification ability as a function of the degree of sensitivity (percentage of change-of-mind trajectories correctly identified) and specificity (percentage of rapid-decision trajectories correctly identified) of the classifier. Given that an ideal ROC curve will hug the top left corner, the larger the AUC the better the classifier performance.

- The calibration data was partitionated into 10 bins that kept the proportion of *straightforward* and *switched* trajectories constant (75/25 proportion). For each bin, we took the complementary set of data (the rest 90%) to train the classifier. The data contained in the bin was used as test set to diagnose the classifier performance. In other words, we obtained a ROC curve and its AUC for each test bin (n=10). Figure 3b shows the resulting ROC curves for each of the 10 bins.



(a) Distribution and mean for LDA measure

(b) ROC and AUC for the LDA measure

Figure 3: Performance of the LDA classifier.

- The performance of the LDA classifier was compared to *baseline*, equivalent the worst possible outcome, and a *topline*, which was what we would expect from the classifier under the best possible conditions. For the baseline, we used a random classifier that assigned labels by sampling from a beta distribution centred at the 'straightforward' probability; the topline was computed by testing and training the original LDA classifier with the same set of data.

- To assess whether the performance of the LDA classifier was different from the baseline and from the topline, we calculated the difference in AUC means between each pair of classifiers and tested it against the distribution of the possible mean-differences observable when the outcome is independent from the classifier (null hypothesis). The sampling distribution under the null hypothesis was computed by a resampling the observed data ('permutation distribution').

- Table 2a shows the mean AUCs, the mean differences between the original LDA and each alternative classifier, and the p-value obtained by performing a one-tail test on the data. As expected, our original LDA is significantly better than the random classifier at categorizing trajectories. Conversely, there is no significant difference between the performance of our LDA and the topline, suggesting that our LDA is not different from the best possible classification.

3.3 Meaningful features and optimal predictors

- Our original LDA classifier takes as predictors both absolute and relative spatio-temporal features (coordinates, speed and acceleration). While x, y coordinates provide information about the absolute mouse position at each specific time step (*where* and *when* questions), speed and acceleration contribute to knowing how one has arrived to a given position (*how* question). Some of these features, however, might not be relevant for the classification. By comparing classifiers trained with different predictors, we

		(a)		(b)				
	Original LDA (coords, speed, acc)	Baseline	Topline	LDA with different predictors				
				coords, vel	vel, acc	coords	vel	acc
AUC (mean)	.87	.52	.87	.87	.83	.87	.82	.67
Mean Difference	-	.35	-.002	-.0004	.04	-.006	.04	.2
p **value**	-	<.001	0.58	.5	<.001	.68	<.001	<.001

Table 2: Cross-validation results for the LDA classifier. The performance of the LDA was compared to the one of (a) Random and Topline classifiers and (b) LDA classifiers with different predictors.

can gather information about which features of mouse trajectories are more relevant for distinguishing between decision processes.

- The performance of these additional classifiers was diagnosed in the same way as before, by computing the AUC for each of the 10 test bins. Pair-wise comparisons with the original LDA were done by testing whether the observed mean differences could have been drawn from the distribution under the null hypothesis. (by permutation test).

- We trained five additional classifiers obtained by subsetting the three original LDA predictors. If both absolute and relative features are required to predict the decision type, we would expect our original LDA classifier to be better than any other classifier that takes only a subset of these original predictors.

- Table 2(b) summarises the comparisons between each of these classifiers and our original LDA. The original LDA does not significantly differ from classifiers that contain the coordinates among their predictors, suggesting that the distinction between *straightforward* and *switched* trajectories can be solely explained by the information contained in the x, y coordinates. In contrast, the original LDA is significantly better than classifiers which use only speed and acceleration as predictors.Should we say something about the fact that the difference with the vel,acc and vel classifiers is significant but not particularly big?. These comparisons therefore reveal that, for classifying our data, absolute spatio-temporal features (x, y coordinates) are generally better predictors than relative features (speed and acceleration). That is to say, it seems to be more relevant to know where are you when than how did you get there.

- Note that this does not mean that changes in decision across the board do not have an impact on speed and acceleration. Indeed, it has been suggested that the speed and acceleration components can capture the level of commitment towards specific responses, such that a change of decision (*swiched* trajectories) might have associated certain speed/acceleration dynamics (Hehman et al 2014). However, our data are not well classified by the these components. This could be due either to the specific type of decision processes we are capturing or by the fact that our data are too noisy to be able to see differences in speed and acceleration (relation with the fact that we have online data).

4 LDA *versus* traditional mouse-tracking analyses

- The LDA classifier is *a priori* the optimal solution to the type of discrimination problem examined here. However, when addressing the question of whether a change of decision has occurred, previous studies have used alternative techniques to analyse mouse trajectories. In what follows, we will compare the performance of our LDA-classifier to the one of other five measures commonly used in mouse tracking studies.

- Following the review in Hehlman et al. (2014), we can distinguish two main types of measures, those based on spatial analyses and those based on temporal analyses.

  - **Spatial analyses** rely on the x, y coordinates themselves and their distance to each of the possible responses (target response and alternative response). They evaluate the degree of unpredictability and complexity of the path.

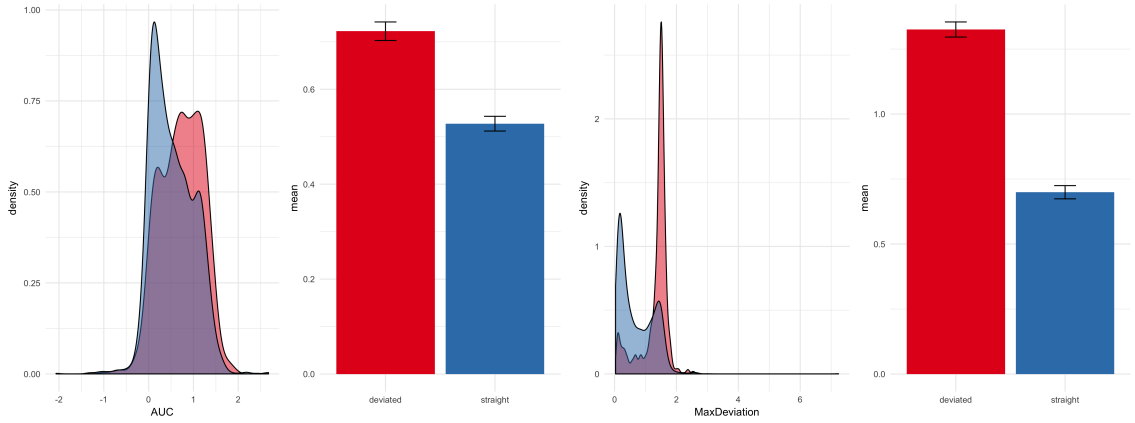
    The two most popular spatial measures are the *Area Under the Curve* and the *Maximal deviation point*. The AUC is the geometric area between the observed mouse-trajectory and an idealised straight-line trajectory drawn from the start to the end point. Higher values are associated with higher deviation peaks towards the alternative; values closer to zero (or below) suggest trajectory close to ideal. The Maximal Deviation (MD) is the point that maximizes the distance between the path and the ideal trajectory. The number of times the trajectories goes back and forth along the x-axis has been also been used as an indicator of the complexity of the trajectory (*number of x-flips*, Dale and Duran 2011).

    (1) a. $\text{AUC} = \sum(x_t - x_{t-1})(y_t + y_{t-1})/2 - \text{AUC}_{\text{ideal trajectory}}$
    b. $\text{x-flips} = \sum H[(x_t - x_{t-1})(x_{t-1} - x_{t-2})]$
    c. $\text{Acceleration flips} = (\sum H[(a_t - a_{t-1})(a_{t-1} - a_{t-2})]) - 1$

  - While these measures can asses the degree of complexity of the path, they might not be able to distinguish between 'two-step' and 'uncertain' decision processes –i.e., trajectories with a true deviation to the alternative or centred on the middle of the screen.
    Moreover, these measures do not have very good temporal resolution. For instance, a late medium-size deviation towards the alternative could underly a two-step decision whereas an early but big-size deviation towards the alternative might very well be considered just noise. However, measures such as the maximal deviation point might not be able to make a significant distinction between them.
  - A slightly better way of assessing whether trajectories have a meaningful deviation towards the alternative response is to measure the *ratio of the target distance to the alternative distance* for each x, y position. While ratio values closer to 1 suggest a position near the middle, higher values indicate a deviation towards the alternative response. Here, we extract the maximal log-ratio value per trial.
  - **Temporal measurements** such as speed and acceleration have served to address the degree of competition at different time steps. Stronger competition between alternative responses is typically translated into steeper acceleration peaks (Hehman et al., 2014). In addition, the decision point has also been associated with temporal measures such as acceleration peak and time of highest spatial deviation, either maximum ratio or maximum deviation. Importantly, in the former case, the decision is considered to appear just after the maximal deviation/maximal ratio point, whereas last acceleration peaks are typically after. The *acceleration component* is generally analized (Dale and Duran 2011, and other studies on error correction) as a function of the number of changes in acceleration (NB: This is not the same as computing the number of times the acceleration changes direction, going from positive to negative acceleration, as D&D claim).
    Following the procedure in Dale and Duran (2011), trajectory velocity and acceleration were computed using the distance (in pixels) covered per second over a moving window of six x, y pixel coordinates. The number of acceleration flips were calculated on this smoothed acceleration.

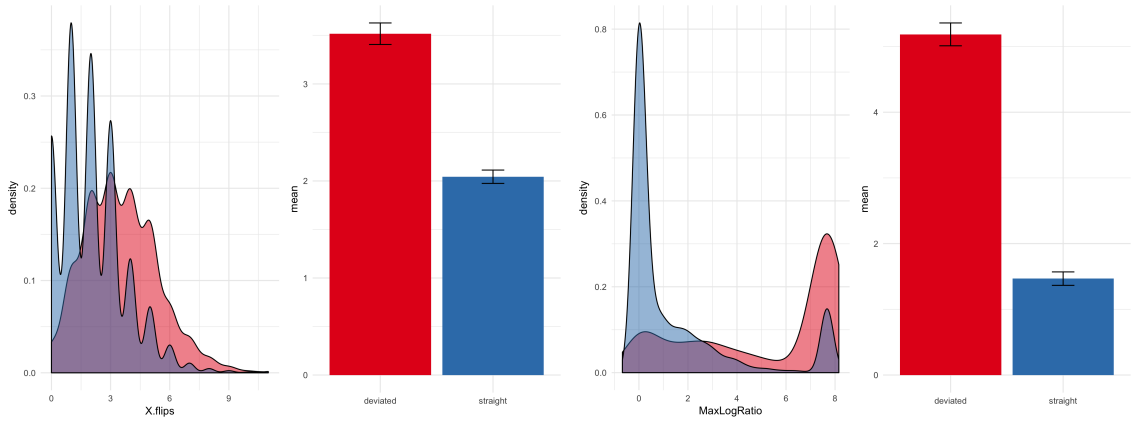
- Validation data was analysed for each of these measures. Figure 4 illustrates the distribution and mean values for each of our classes.

- Following the cross-validation procedure described in the previous section, we diagnosed the performance of each of these measures as classifiers for our 10 bins of data. Table 3 summarizes the results of these comparisons.



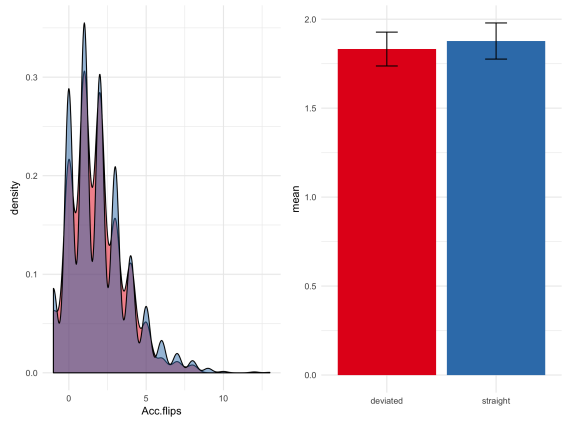
(a) Area Under the Curve

(b) Maximal Deviation



(c) X-coordinates Flips

(d) Maximal LogRatio



(e) Acceleration Flips

Figure 4: Different measures applied to our calibration data.

	Original LDA	AUC	Maximal Deviation	Maximal Log Ratio	X-Coord. Flips	Acc. Flips
AUC (mean)	.87	.62	.81	.81	.73	.53
Mean Difference	-	.24	.06	.06	.14	.34
p **value**	-	<.001***	<.001***	<.001***	<.001***	<.001***

Table 3: Cross-validation results for the LDA classifier. The performance of the LDA was compared to the one of five commonly used measures in mouse-tracking studies.

- The results of these comparisons reveal that the LDA is significantly better at classifying our data than other commonly used measures. While the difference is in all the cases significant, the comparison between mean AUCs values suggests that measures such as Maximal Deviation and Maximal LogRatio are better at distinguishing decision processes than AUC, X and Acceleration flips. In order to understand such pattern, it should be noticed that these two measures are the only ones that are calculated based on coordinates themselves and therefore give more importance to spatio-temporal information. In other words, both the deviation from the ideal trajectory and the log-ratio give different weight to positions depending the moment when they occurred, and therefore are more sensitive to the moment at which deviation occurred.

5 Extension to linguistic data

- Can our LDA also classify more complex decision processes, such as the ones involved in sentence verification tasks? Dale and Duran found differences in the processing of true positive and negative sentences when people performed a simple truth-value judgement task. In this section, we aim to (1) replicate Dale & Duran results when performing the same analyses as them, (2) to use our classifier (trained with validation data) to test these processing differences, and (3) compare the performance of our classifier with the one of other measures.

- Our LDA classifier is expected to dissociate between mouse trajectories that underlie a change of decision and trajectories that do not (*switched* vs. *straightforward*). If, as proposed by Dale and Duran, processing negation involves an abrupt shift in cognitive dynamics, then mouse trajectories corresponding to negative trials should pattern with trajectories involving a decision change, like the ones we included in the validation experiment. This should not be the case for trajectories corresponding to positive trials. If our classifier can make a distinction between negative and positive trials, we will be providing additional support to the hypothesis that, at least in this context, processing negation involves a two-step derivation (i.e. an unconscious change of decision).

- It should be, nevertheless, noticed that the validation data used to train the LDA is only an approximation of what should happen during an unconscious change of decision, such as the one expected for negation processing. Therefore, we expect some additional variability on the negation results (since there are some aspects of the decision process that the LDA won't be able to capture).

5.1 Experiment

- Our experiment presented some minor methodological changes with respect to Dale and Duran's (Experiment 1), but it tested the same contrast; namely, the interaction between truth value (true or false statements) and sentence polarity (affirmative and negative sentences).

- Participants were asked to perform a truth-judgment task, where they had decide whether a statement is true or false according to common knowledge.

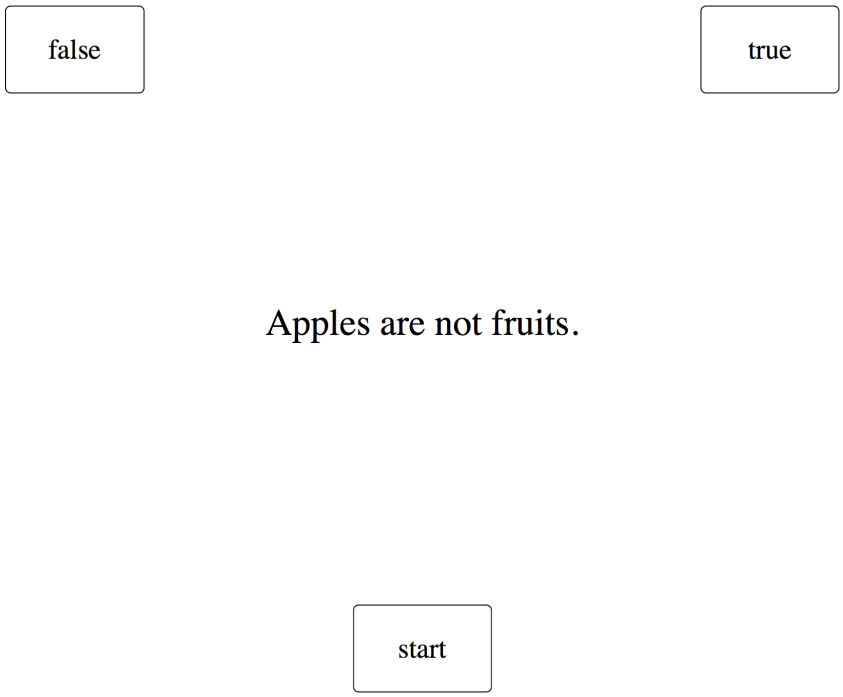


Figure 5: Trial Example Dale & Duran Replication

- Unlike Dale and Duran's experiment, the complete statement was presented in the middle of the screen after participants pressed "Start" (i.e. no self-paced reading). The "True" and "False" boxes appear at the top-left or top-right corners of the screen, in the same way as in our validation experiment.
- Each of the sentences could be either a true or a false statement in its negated or non-negated form. An illustration of the sentences used as examples is provided in Table 4.

Participants 53 English native speakers were tested using Amazon Mechanical Turk. They were rewarded for their participation.

Design The experimental design consisted in two fully crossed factors: TRUTH (true, false) and POLARITY (negative, positive). We had a total of 4 conditions, and each participant saw 4 instances of each condition (16 sentences).

Truth value	Polarity	Example
True	Positive	Cars have wheels.
	Negatives	Cars have no wings.
False	Positive	Cars have wings.
	Negatives	Cars have no wheels.

Table 4: Design

Interface and data treatment The interface and data treatment were the same as the ones used for the calibration experiment. Mouse trajectories' time course was normalised into 101 time steps.

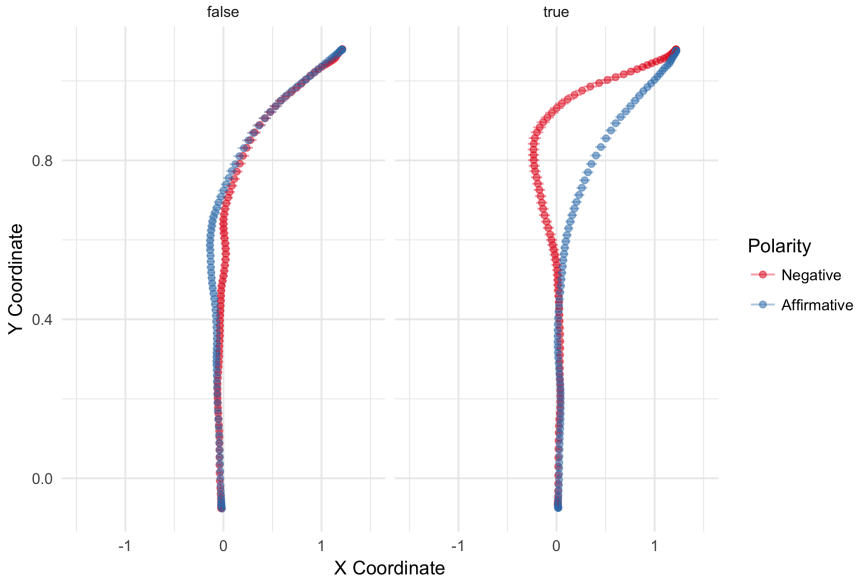


Figure 6: Mean trajectories for accurate trials

5.2 Results

Replicating Dale and Duran (2011)

- All participants responded correctly more than 75% of the time. No participant was discarded based on accuracy. Only accurate trials were taken into account for the analysis. Mean trajectories for the four possible trial conditions are shown in Figure 6.

- To assess whether we replicate Dale and Duran's results, we calculated the x-flips (see above) and analysed them with a linear mixed-effects model (Baayen, Davidson, and Bates, 2008), taking Truth, Polarity and their interaction as predictors. We included random intercepts per subject and a random slope with the interaction of both factors. P-values were obtained by comparing the omnibus model to a reduced version of itself, where the relevant factor was removed. This pipeline mirrors the model performed by D&D.

- Unlike Dale and Duran, we did not perform analyses based on the acceleration component (acceleration flips). Our validation experiment suggest that this measure is unable to dissociate trajectories involving a change of decision from straight trajectories. This seems to be related with the fact that our data is very noisy. We reasoned that if the different decision processes involved in a rather simple task were not captured by the acceleration component, this measure would also be unable to classify more complex processes, such as the ones at play in a sentence verification task.

- The model revealed a main effect of Polarity, such that negation significantly increases x-flips by 0.76 ($\chi^2 = 10.11; p = .0014$), and a significant interaction Truth $\times$ Polarity ($\chi^2 = 22.7; p < .001$), such that the difference between negative and positive sentence is bigger for the true than for the false statements. There was no significant effect of Truth ($\chi^2 < 1; p = .5$). Table 6 summarises the pattern of means and estimates for both ours and Dale and Duran's results.

- These results seem to replicate Dale and Duran's findings: verifying true negated sentences produces less straightforward trajectories than true positive sentences. It should, however, be noticed that the values obtained in the two experiments are slightly different; namely, our results present higher range

Table 5: default

Condition	x-flips	x-flips in D&D
T/no negation	2.22	1.13
T/negation	3.67	1.71
F/no negation	2.82	1.24
F/negation	2.9	1.34
Estimate Polarity	.76	0.35
Estimate Truth	.07	0.13
Estimate Truth×Polarity	1.35	0.47

Table 6: Mean and effect estimates

of values. As for the acceleration component, this might be due to the fact that our data is generally noisier than theirs, probably because of x, y coordinates were sampled from mouse-movements and not at a fixed sampling rate.

- More generally, our findings pattern with a broader set of psycholinguistic studies, which have used different techniques and shown that verifying negative sentences involve computing the positive content at an early processing stage (CITE).

Classifier performance

- Two different LDA classifiers, trained with validation data, were applied to the new experimental data. The first classifier was our original LDA, which had as predictors x, y coordinates as well as velocity and acceleration. The second LDA had only x, y coordinates as predictors. Validation results (see above) suggest that the simpler model, which only relies on coordinates information, might be sufficient to classify the data. In other words, the simple model might fit the data as well as a more complex model, and be interpreted more straightforwardly.

- Since the relevant different between the positive and negative sentences is expected to arise specifically for *true* statements, we will analyse specifically these cases (NB: This will facilitate the analyses).

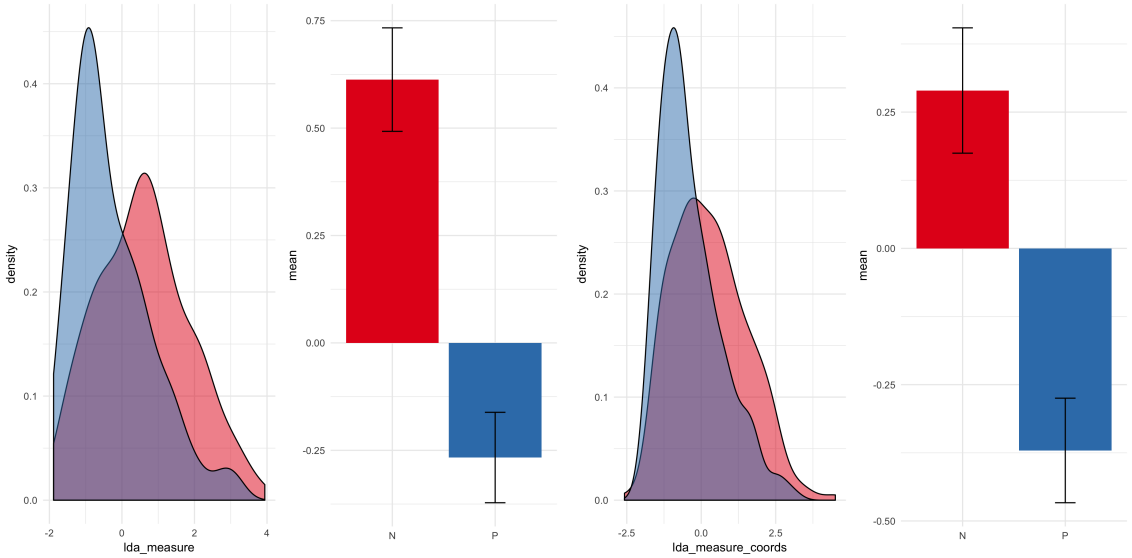
- Distribution and means of the resulting *LDA measure* for *true* trials and for both classifiers are illustrated in Figure 7. As before, the overall performance of the classifier was evaluated by calculating the area under the ROC curve (Original LDA=.7; Coords LDA=.66). Add LMM for these measures?

Measure comparison

- What is the difference in performance between the different measures? (Do the different measures differ on their ability to finding the effect when the effect is there?)

- Our results so far suggest that the LDA trained with validation data can distinguish negative and positive trials. The question that arises is whether this LDA classifier trained with controlled data can make a better distinction than other commonly used measures, such as the ones used by Dale and Duran. Indeed, it could be the case that the LDA classifier is only a better strategy than other measures when is trained and tested with an homogenous set of data (i.e. when training and testing sets correspond to the same type of decision processes). In other words, our classifier might not be able to extrapolate to different types of decision processes.

- Figure 8 shows the results for values obtained for different measures for the contrast between positive and negative true statements.



(a) Original LDA (coords, speed and acceleration) (b) LDA only with coordinates

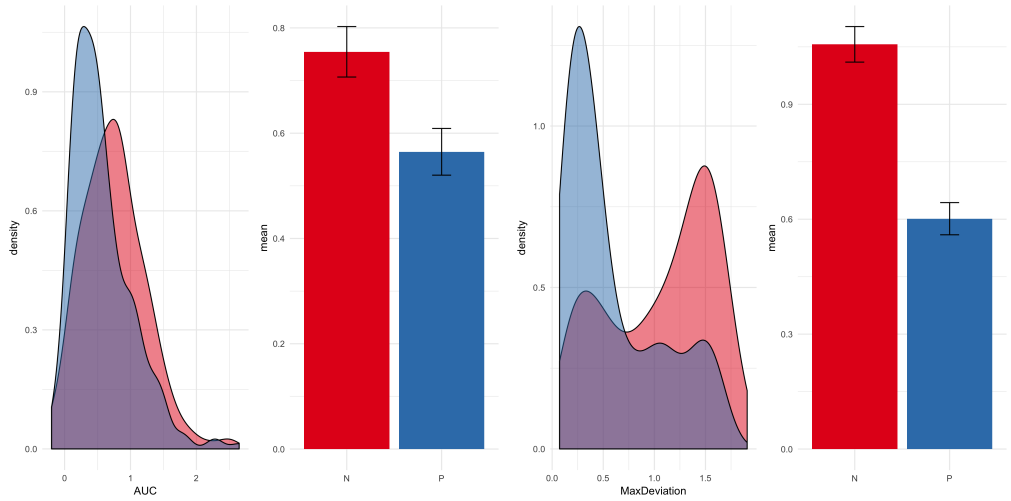
Figure 7: Two LDA classifiers applied to *true* trials.

- To estimate the power/robustness of these measures, we asked whether they could accurately detect the effect when we reduce the sample size. We calculate the area under the ROC curve for different sample sizes by bootstrapping from the original data set. Figure 9a shows a power curve: for each sample size the percentage of iterations where the AUC is above .5 (above random performance). Figure 9b illustrates the mean AUC values obtained for each sample size and each measure.

- The results in Figure 9 suggests that: (a) none of the measures performs a classification as good as the one observed with validation data; (b) the ranking based on power does not correspond exactly to what we see in the AUC statistics (ie. how good they are at detecting the effect is not necessarily equivalent to having the highest AUC values): (i) the Original LDA is as powerful as the one of the Maximal Deviation and Maximal Log Ratio; (ii) the AUC statistics suggest the existence of a ranking between these three measures such that Maximal Deviation gives rise to overall higher AUC values. (c) Finally, the performance of the Coords LDA, x-flips, the area under the curve seems to be weaker across the board.

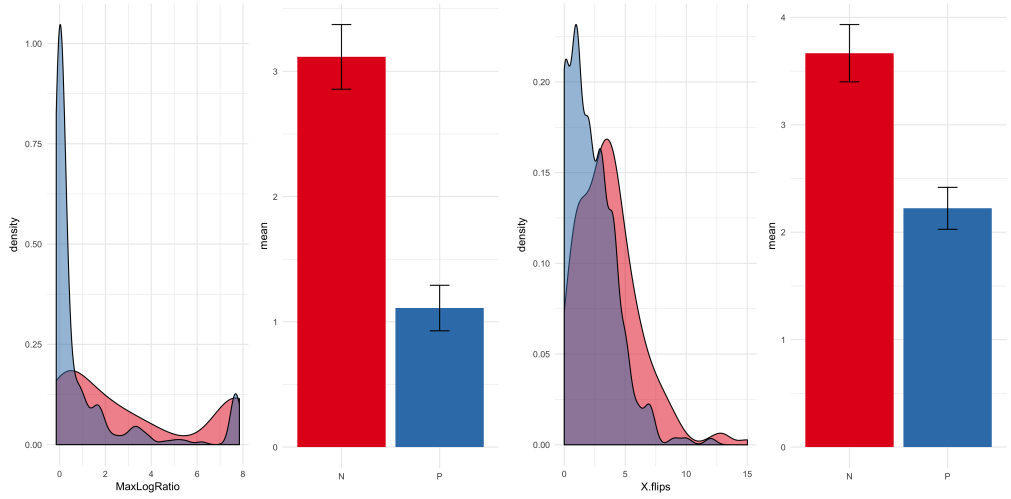
- The LDA applied to the linguistic data is not as good as it was when applied to the validation data. This difference might be related with the fact that the two sets of data (training and testing sets) correspond to two different tasks, involving different decision processes. The decision processes involved in sentence verification tasks are much more complex than the ones targeted in our validation experiment. Consequently, the LDA model might not be as detailed/complex as it would be required in order to improve the classification of negation data (i.e. it has not been trained to pay attention to important aspects of decision dynamics). The fact that negation data are *a priori* more complex additionally explains the overall drop in performance across different measures.

- From a practical point of view, our findings indicate that the contrast between processing/verifying negative and affirmative sentences can be detected in mouse trajectories by just using certain traditional measures such as the maximal deviation. In other words, it's is not required to use a LDA classifier.



(a) Area Under the Curve

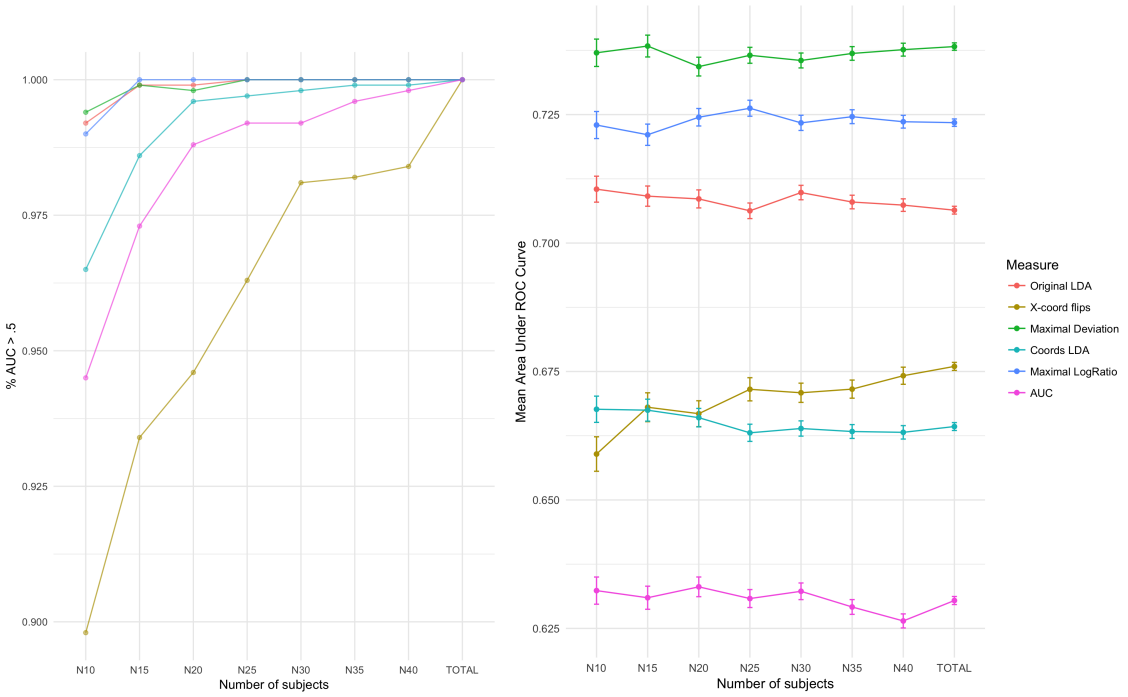
(b) Maximal Deviation



(c) Maximal LogRatio

(d) X-coordinates Flips

Figure 8: Different measures applied to Dale & Duran replication.



(a) Power (b) Means

Figure 9: Measure performance for different sample sizes

6 General Discussion

- This study aimed to test the mapping from decision making to mouse movements. Our results make the following contributions:

- By manipulating whether the stimulus triggered or not a change of decision, we have shown that mouse trajectories capture differences in decision dynamics. That is to say, when participants are forced to switch their decision, this change has a direct repercussion in their hand movements.

- We have demonstrated that mouse trajectories underlying each type of decision process can be distinguished by a LDA classifier. In order to make such distinction, this classifier can rely only on absolute temporal and spatial information about the position of the mouse at each time step, without taking into account relative information about how the mouse got there (speed and acceleration). Our LDA classifier not only has a better performance than more traditional mouse-tracking measures but also has the advantage of being able to make a distinction without making any specific assumption about how trajectories should look like.

- We replicated Dale and Duran findings regarding negation processing. The result of applying our LDA classifier, trained on validation data, to data obtained from a sentence verification task suggest that processing negative sentences involves a change of decision, absent during the verification of positive sentences.

- Although the LDA manages to dissociate between these two sentence verifications, its performance is overall lower than in the validation case, and not significantly different from the one of other traditional measures.

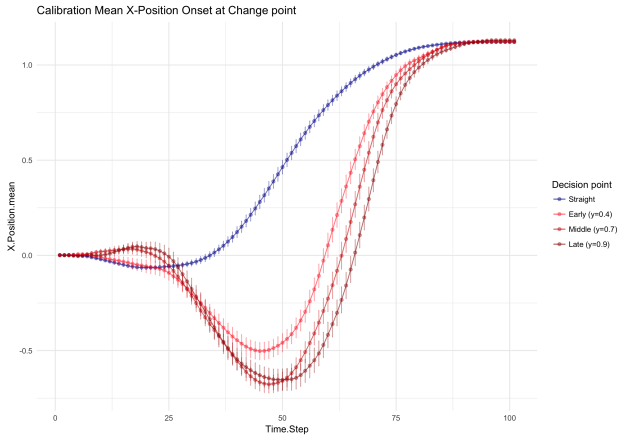


Figure 10

- The difference LDA performance across experiments reveals that the two data sets are capturing slightly different decision processes. Ideally, we should be able to make our validation experiment slightly more complex in order to capture subtleties.

7 Supplementary materials

- Time information for validation experiment
- Other LDA with different predictors: Another two classifiers were trained with a different set of predictors. A first one had as predictors x, y coordinates and partial derivatives: vertical velocity and acceleration (y-based) and horizontal velocity and acceleration ($x-$based). Spatial and temporal features for each axis might predict the type of decision to a different extent (both axes might be relevant but to different extent). For instance, the movement on the horizontal axis might have stronger relevance for the classification than the movement on the vertical axis. A second classifier had as predictor the ratio between the euclidean distance to the alternative response and the distance to the target response. The ratio contains spatial information, but it is sensitive to how close the point was to the target response and to the alternative. Add results