

1 Mouse tracking as a window into decision making

2 Mora Maldonado, Ewan Dunbar & Emmanuel Chemla

3 January 31, 2018

4 1 Introduction

5 In the past ten years, mouse-tracking has become a popular method to target the processes underlying
6 decision making in different domains, ranging from phonetic competition [Spivey et al., 2005, Cranford and
7 Moss, 2017], and syntactic, semantic and pragmatic processing [Farmer et al., 2007, Dale and Duran, 2011,
8 Tomlinson et al., 2013, Xiao and Yamauchi, 2014, Sauerland et al., 2015, Xiao and Yamauchi, 2017, among
9 others], to social cognition [Freeman and Ambady, 2010, Freeman et al., 2011, Freeman and Johnson, 2016].
10 All these studies have worked on the assumption that motor responses are prepared in parallel to cognitive
11 processing and performed in a cascade manner [Song and Nakayama, 2006, 2009, Freeman and Ambady, 2010,
12 Spivey and Dale, 2006, Hehman et al., 2014]. As a result, features in mouse trajectories could be indicators
13 of specific decision processes, revealing their dynamics with fine-grained temporal resolution. Mouse-tracking
14 studies typically present participants with a *two-alternative forced choice*, where they have to make a choice
15 using the options appearing in the top left or right corner of the screen. Whenever a decision involves two
16 independent processes –such as a change of mind–, mouse trajectories are expected to be displayed as two
17 movements, whereas a single smooth and graded movement would reflect a commitment with an initial choice
18 (see Figure 1, Wojnowicz et al., 2009). Of course, one could still imagine other types of decision, such as
19 a single but late commitment, or a decision made after uncertainty or doubt. These might have a different
20 reflection on mouse trajectories.

21 The existence of intuitions about how decision processes should be mapped into mouse paths (i.e. linking
22 hypotheses) has allowed researchers to draw conclusions about the cognitive processes underlying their
23 experimental manipulations. Dale and Duran's (2011) approach to negation processing is an example of
24 this. Negation has been traditionally understood as an operator that reverses the sentence truth conditions,
25 inducing in “extra-step, or mental operation” in online processing (Wason, 1965, Wason and Johnson-Laird,
26 1972; see review in Tian and Breheny, 2016). To test the dynamics of negation integration, Dale and Duran
27 tracked mouse trajectories as participants performed a Truth-Value Judgment Task (TVJT), where they
28 had to verify the truth of general statements such as *Cars have (no) wings*. The authors found that mouse
29 trajectories presented more shifts towards the alternative response when evaluating negative than affirmative
30 true sentences. These results were interpreted as evidence for a ‘two-step’ processing of negation, where truth
31 conditions for the positive content are first derived and negated only as a second step¹.

32 The impact of experimental manipulations in the shape of trajectories has been taken as evidence for
33 underlying decision patterns. This association, however, has never been explicitly tested. While no one can
34 deny that mouse trajectories are sensitive to experimental manipulations –such as negation–, it is unclear
35 whether this can be directly interpreted as reflecting decision making.

36 Our main goal is to test the connection between cognition (decision making) and action (mouse trajectories): Are decision processes always reflected in mouse trajectories? If yes, how are they reflected? One could
37 easily imagine a situation where two different trajectory shapes do not correspond to two different decision
38 mechanisms, but to a single one (due to uncertainty, noise, etc.). Differences in trajectories can therefore
40 underlie something other than decision shift.

¹Several studies have suggested that the positive argument plays an important role in negation processing [Kaup et al., 2007, Lüdtke et al., 2008, among others]. This pattern of results, however, depends on the amount of contextual support given for the sentence: ‘two-step’ negation processing seems to occur specifically for sentences presented out-of-the-blue, whereas no difference between negative and positive sentences arises when the right contextual support is provided [Nieuwland and Kuperberg, 2008, Tian et al., 2010]. How to explain this pattern of results has been at the center of the debate in the negation processing literature (see Tian and Breheny, 2016 for review), but we will not explore it here.

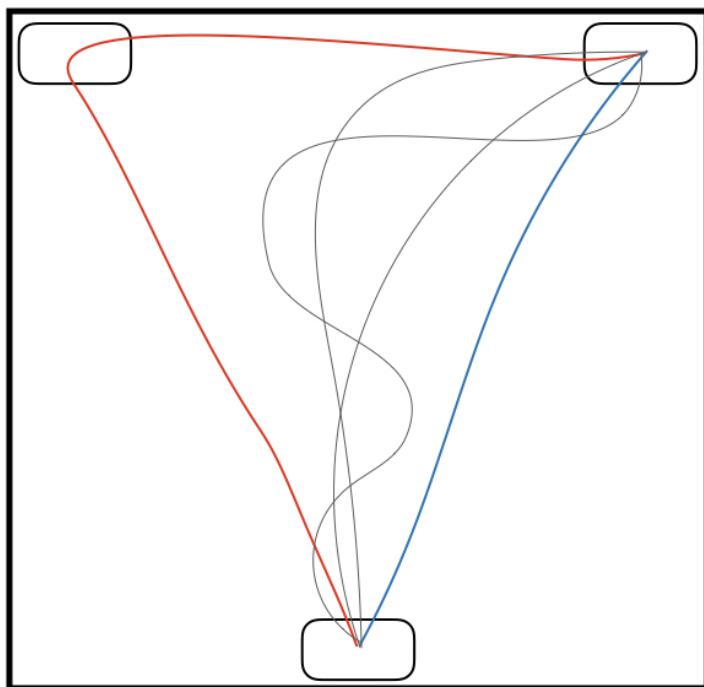


Figure 1: Shape of trajectories underlying distinct decision processes. One single cognitive process is expected to be mapped onto one smooth movement (blue line), whereas a change of mind would be reflected by two movements (red line). Intermediate cases are represented in gray.

41 In this paper, we address these questions by identifying the features in mouse-trajectories corresponding
 42 to two different types of decisions: *straightforward* decisions (i.e. single commitment) and *switched* decisions
 43 (i.e. a change of mind). First, we present a *validation* experiment where, instead of taking mouse trajectories
 44 as indicators of cognitive processes, we *manipulate* whether or not our stimuli trigger a flip in decision
 45 ([Section 2](#)). The data from this validation experiment (i.e. two groups of ‘quasi-decisions’) is then used
 46 to feed a *Linear Discriminant Analysis* (henceforth, LDA), trained to classify trajectories depending on the
 47 underlying decision ([Section 3](#)). After comparing the performance of LDA classifier to other traditionally
 48 used mouse-tracking measures ([Section 4](#)), the LDA classifier will be further tested with new data, obtained
 49 from a replication of Dale and Duran’s ([2011](#)) experiment on negation processing ([Section 5](#)). If there is
 50 a change of decision triggered by negation, trajectories corresponding to negative trials should be classified
 51 together with trajectories underlying decision change in the validation experiment.

52 2 Manipulating decision making: Validation Experiment

53 We developed an experiment where participants had to perform a *two-alternatives forced task*: at each trial,
 54 they were presented with a colored frame surrounding the screen and they had to determinate whether
 55 the frame was blue or red. Responses were made by clicking on the “blue” or “red” buttons, allowing the
 56 recording of mouse-movements during each trial. Responses were considered accurate if they described the
 57 color at the moment of the click. To mirror decision processing, we manipulated whether the color of the
 58 frame remained stable or changed at some point during the trial. While in the latter case the initial choice
 59 will be the accurate response (*straightforward* trials), in the former, participants were forced to swap their
 60 answer (*switched* trials), mimicking a change of decision. Note that, since we are only *mirroring* decision
 61 making, we refer to these decision processes as ‘*quasi-decisions*’. An illustration of the procedure is provided
 62 in [Figure 2](#).

63 **Participants** Fifty four participants ($F=27$) were recruited using Amazon Mechanical Turk. Two subjects
 64 were excluded from the analyses because they did not use a mouse to perform the experiment. All of them
 65 were compensated with 0.5 USD for their participation, which required approximately 5 minutes.

66 **Design** Each trial instantiated one of two possible DECISION PATTERNS. In *straightforward* trials, the
 67 frame color remained stable, and the decision made at the beginning of the trial did not need to be revised.
 68 In *switched* trials, the color swapped during the trial, forcing a revision of the initial choice. The POINT OF
 69 CHANGE in *switched* trials was determined by the position in the y axis, and it could be early, middle or late.
 70 The FRAME COLOR at the response point was also controlled: it could be red (right button) or blue (left
 71 button). A summary of the design is given in [Table 1](#).

DECISION PATTERN	FRAME COLOR		POINT OF CHANGE
Straightforward	Blue		—
	Red		—
Switched	Blue	Red	early ($y=.4$), middle ($y=.7$), late ($y=.9$)
	Red	Blue	early ($y=.4$), middle ($y=.7$), late ($y=.9$)

Table 1: Design in Validation Experiment

72 To prevent participants from developing a strategy based on staying on the middle of the screen, the
 73 proportion of trials was adjusted so that straightforward trials were the majority (32 repetitions per frame
 74 color), whereas switched trials had 4 repetitions per frame color and change point. The total number of trials
 75 was 88.

76 **Interface** The interface was programmed using JavaScript. Mouse movements triggered the extraction of
 77 x , y -pixel coordinates (i.e., no constant sample rate). The software was adapted proportionally to the window
 78 of the participant’s browser, forming a rectangle. Three buttons were displayed during the experiment (‘start’
 79 and response buttons). The ‘start’ button was placed at the bottom center of the screen. The two response

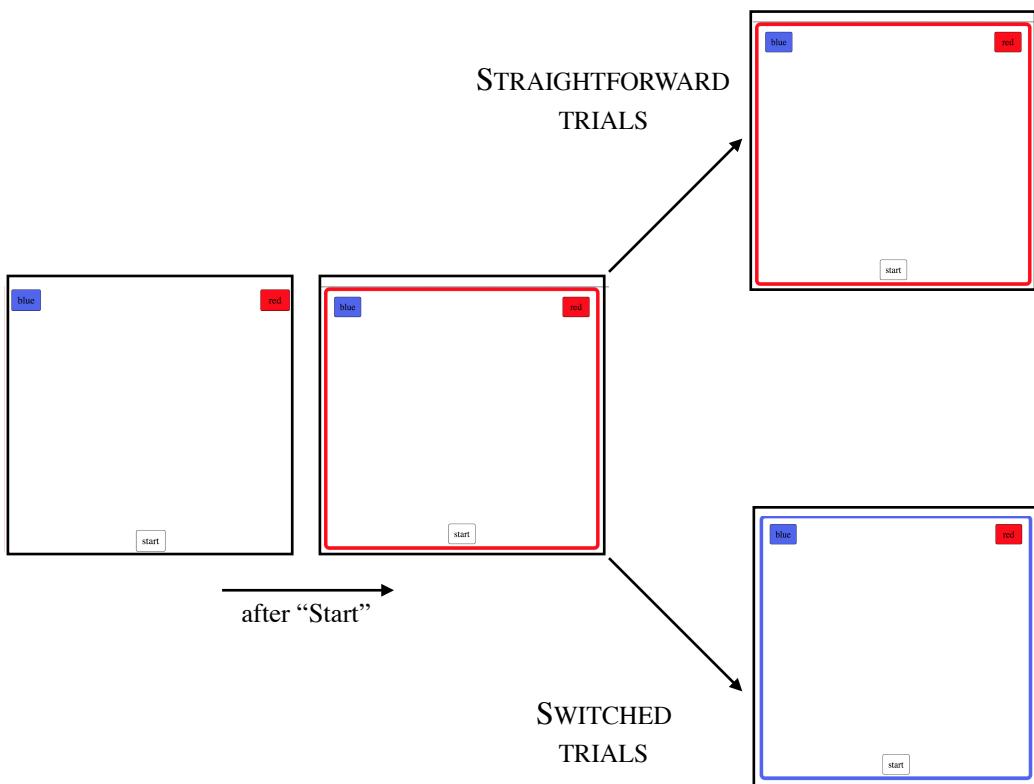


Figure 2: **Procedure in Validation Experiment.** Subjects were instructed to click the ‘Start’ button in order to see the colored frame. Response boxes were on the top-left or top-right. Depending on the trial condition, the frame color could change along the trial or not.

80 boxes were located at the top left ('blue') and top right ('red') corners of window. This location was constant
81 across participants, and handedness was controlled. In each trial, mouse movements were recorded between
82 start-clicks and response-clicks. The x, y -pixel trajectory was saved together with its raw time. Afterwards,
83 the positions were normalised according to participants' window size, to allow comparisons between subjects.
84 The normalization was done by considering the start button at the [0,0] point, the 'blue' button corner at
85 [-1,1] and the 'red' button at [1,1].

86 **Data treatment** Mouse-tracking data are particularly variable trial to trial. Variations in response times
87 imply different quantity of x, y positions per trial, making difficult the comparisons between items. Moreover,
88 in our design, positions are extracted based on mouse movements, and devices with different sensibility could
89 influence the number of samples taken during the trial. In order to compare mouse trajectories, we normalized
90 the time course into 101 proportional times steps (percentage of trial duration).

91 **Overall performance** Inaccurate responses, corresponding to 4% of the data, were removed from the
92 analyses. Mean trajectories for each DECISION PATTERN and POINT OF CHANGE are illustrated in [Figure 3](#).
93 These trajectories suggest that participants made a decision as soon as they were presented with the color
94 frame, and revised this decision if needed. When they were forced to change their choice, this switch was
95 reflected in mouse trajectories. [MM: Do we want to include other graphs?](#)

96 3 Classifying decision processes with LDA

97 Different 'quasi-decisions' (i.e. DECISION PATTERNS) have a different impact on mouse trajectories, as ob-
98 served in [Figure 3](#). To identify the features characteristic of each class (*switched* vs. *straightforward*), we use
99 a Linear Discriminant Analysis method for classification.

100 **Description of LDA classifier** The LDA is an optimal solution to classify continuous data –such as
101 trajectories– into two or more classes –such as decision patterns (i.e. quasi-decisions). In a nutshell, the LDA
102 algorithm assumes that different classes have a common covariance matrix, and finds the linear combination
103 of predictors that gives maximum separation between the classes. This linear combination of predictors
104 is obtained as a linear coefficient and it can be used to form a decision rule for the classification. It is a
105 single number that represents the position on a line running between the two classes that maximizes their
106 separability, with zero representing the midpoint between the two.

107 The predictors used by the classification algorithm were: (a) the x, y coordinates, (b) Euclidean-based
108 velocity, and (c) Euclidean-based acceleration (both of which are non-linear with respect to the original x, y
109 cooordinates). The coordinates provide absolute spatio-temporal information about where the cursor was
110 when, and velocity and acceleration provide information about how did it arrived there. To avoid collinearity
111 (which causes problems for LDA), we applied a Principle Component Analysis (PCA) to identify the 13
112 principal components on these predictors, and performed the LDA on these principal components. We thus
113 obtained an *LDA measure* for each trial, the single number giving the position of the trial on the LDA
114 classification axis. The procedure is schematized in [Figure 4](#).

115 **Performance of the LDA classifier** The result of applying the procedure described in [Figure 4](#) to the
116 trajectories in the validation experiment is illustrated in [Figure 5](#). To evaluate the overall performance of the
117 classifier, we calculated the area under the ROC curve (AUC), a standard method for evaluating classifiers
118 [[Hastie et al., 2009](#)]. Intuitively, the AUC gives the degree to which the histograms resulting from the
119 classifier's continuous output (for example, [Figure 5](#)) is non-overlapping in the correct direction (in this case,
120 *switched* more systematically in the positive direction on the classification axis than *straightforward*).

121 To properly evaluate the classifier's perfomance at separating trials following the distribution in the
122 experiments, the AUC measure was cross-validated. That is, calibration data were partitioned into 10 bins
123 that kept the proportion of *straightforward* and *switched* trajectories constant (75/25 proportion). For each
124 bin, we took the complementary set of data (the rest 90%) to train the classifier. The data contained in
125 the bin were used as test set to diagnose the classifier performance. Thus, we obtained one AUC score for
126 each test bin (ten bins). The performance of the LDA classifier was compared to *baseline*, equivalent the

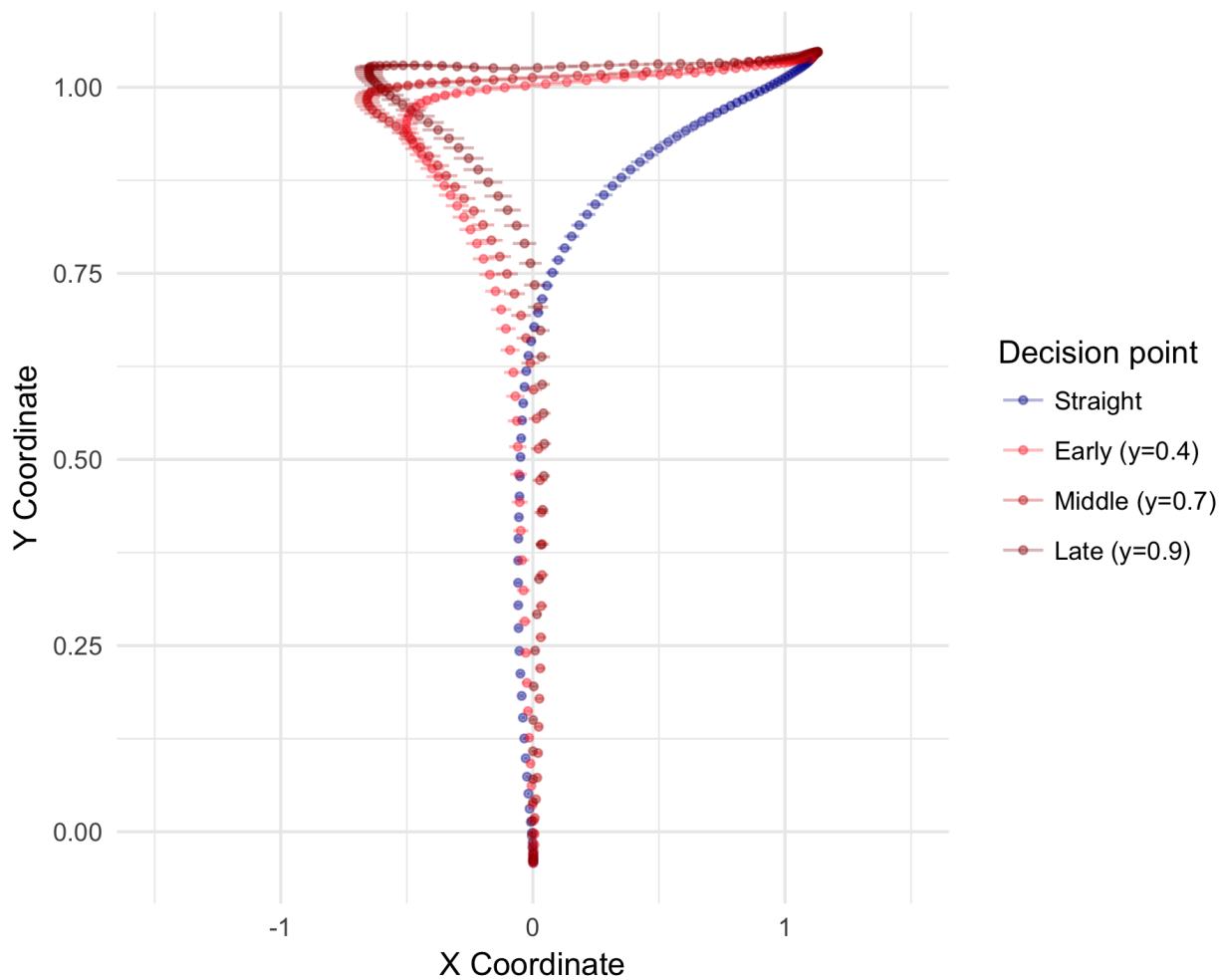


Figure 3: Mean trajectories per class

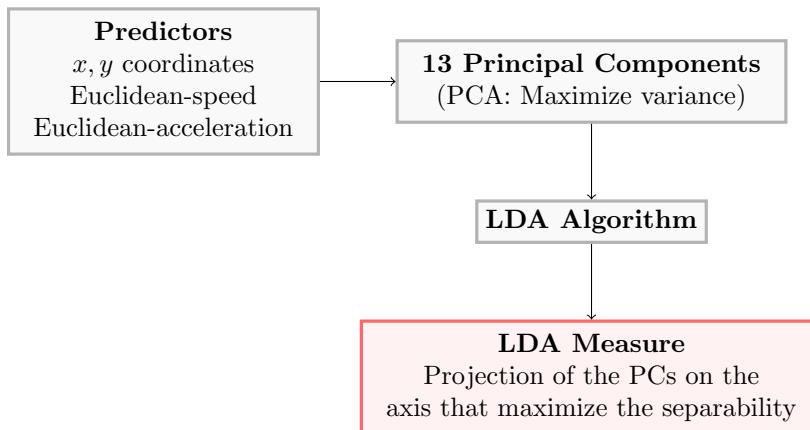


Figure 4: Diagram

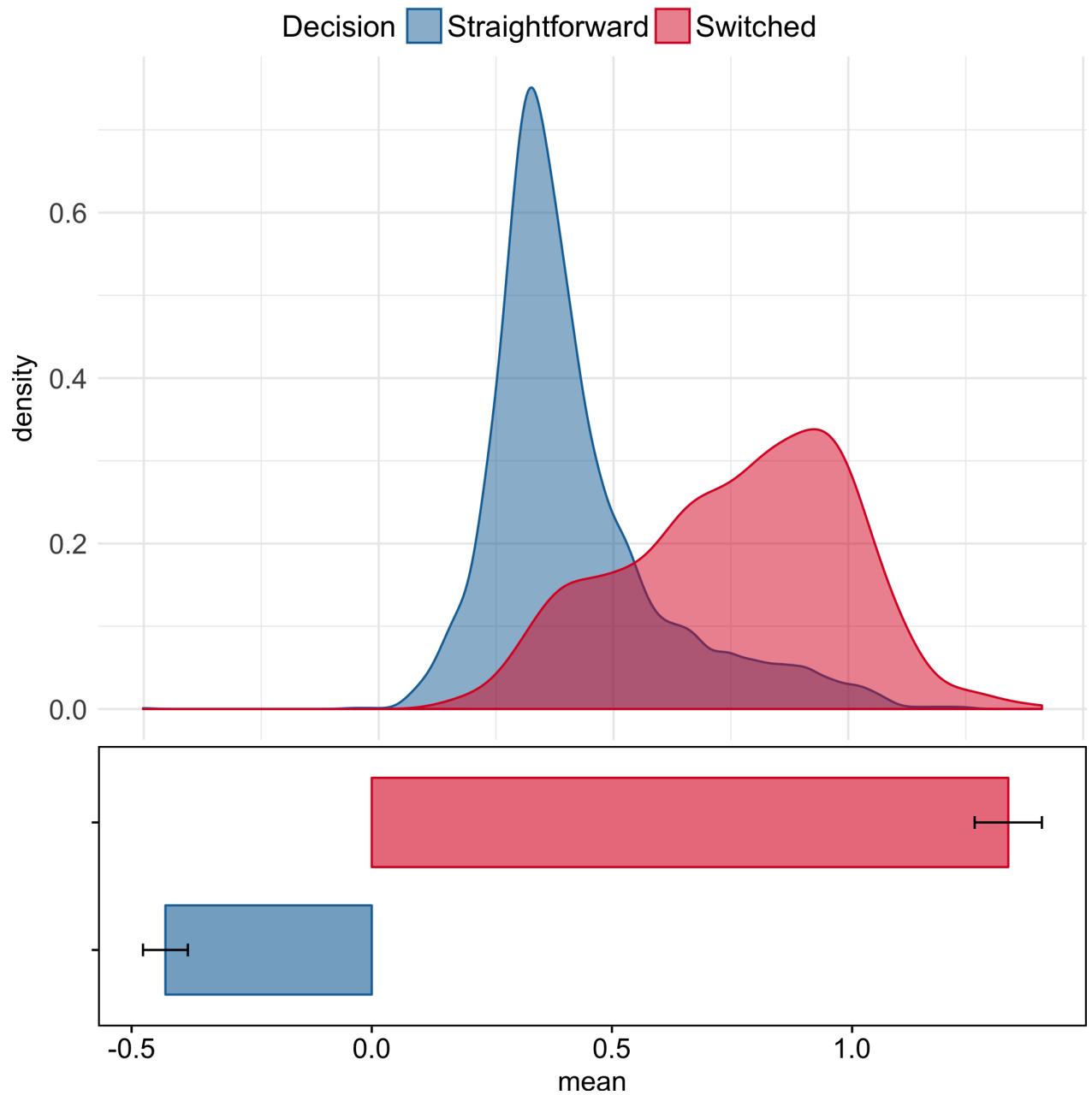


Figure 5: Classifier performance

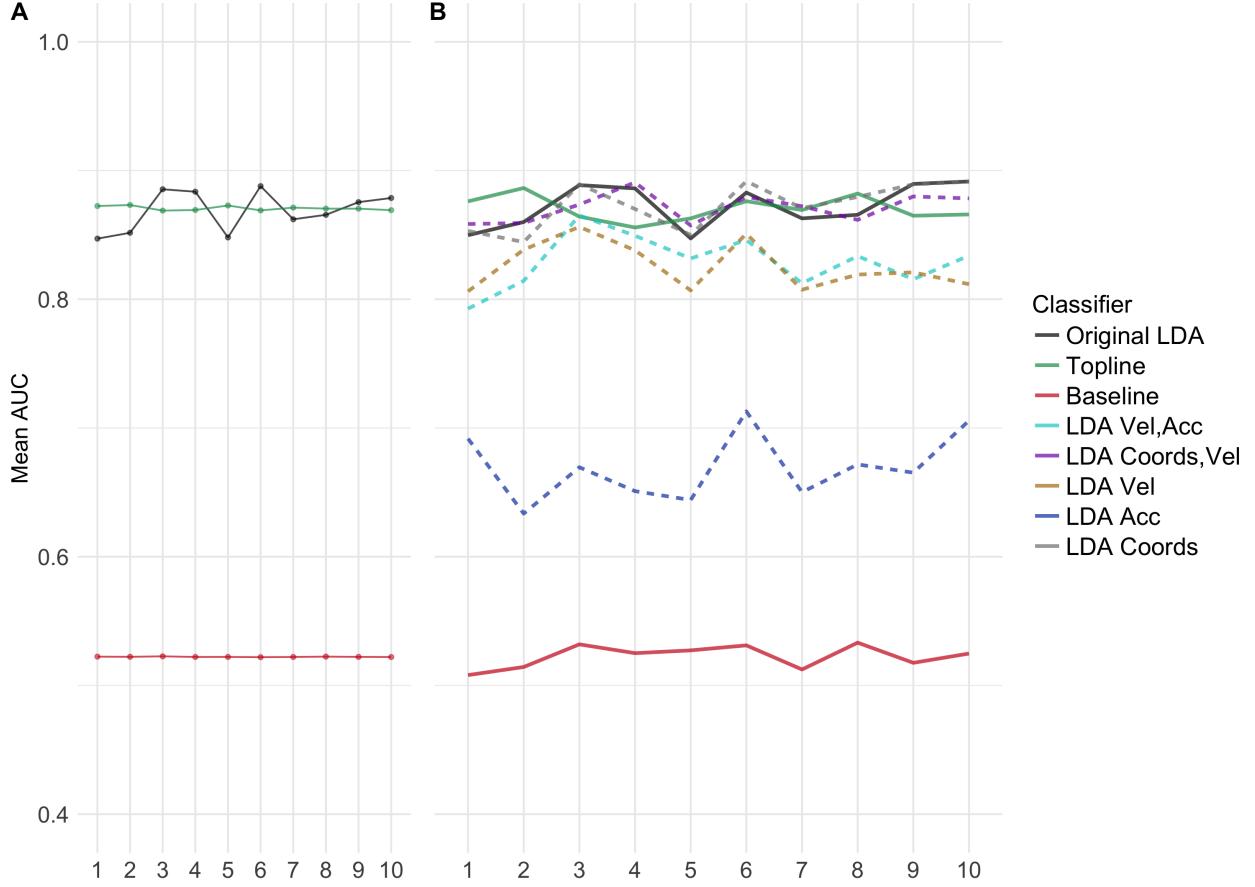


Figure 6: Mean AUC values

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 probability of straightforward trials; the topline was computed by testing and training the original LDA classifier with the same set of data. The mean AUC values for the LDA, the baseline and the topline in each bin are given in Figure 6a.

To assess whether the performance of the LDA classifier was statistically different from baseline (or topline performance), we tested how likely it would be to obtain the attested differences under the null hypothesis that the LDA classifier performance was the same as the baseline (or topline) performance. The difference in AUC means between each pair of these two pairs of classifiers was calculated, and the sampling distribution under the null hypothesis was estimated by randomly shuffling the labels indicating which classifier the score came from.

In Table 2a, we report the results of performing a one-tail test on these mean differences (observed vs. under null hypothesis). As expected, our original LDA is significantly better than a random classifier at categorising trajectories into the two classes. Conversely, there is no significant difference between the performance of our LDA and the topline (i.e., the LDA is not significantly different from the best possible classification).

Meaningful features and optimal predictors Our original LDA classifier takes as predictors both absolute and relative spatio-temporal features (coordinates, speed and acceleration). Some of these features, however, might not be relevant for the classification. By comparing classifiers trained with different predictors, we gather information about which features of mouse trajectories are more relevant for the distinction between decision processes (i.e. for the classification).

	ORIGINAL LDA (coords, speed, acc)	(a)		(b)		
		BASELINE	TOPLINE	LDA WITH DIFFERENT PREDICTORS		
				coords, vel	vel, acc	coords vel
AUC (mean)	.87	.52	.87	.87	.83	.87
Mean Difference	—	.35	-.002	-.0004	.04	-.006
p value	—	<.001	0.58	.5	<.001	.68
				<.001	.04	<.001

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

We trained five additional LDA 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 “full” original LDA classifier to be better than any other classifier that takes only a subset of these original predictors. 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. Figure 6b illustrates the mean AUC values for each of these classifiers, together with the original LDA, the baseline and topline. Pair-wise comparisons with the original LDA were done by testing whether the observed mean differences would be expected under the null hypothesis (i.e. no difference in performance between classifiers). Table 2b 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* ‘quasi-decisions’ might be solely explained by the information contained in the x, y coordinates. In contrast, the original LDA is significantly better than classifiers that use only speed and acceleration as predictors. These comparisons therefore reveal that, for classifying our validation data, absolute spatio-temporal features (x, y coordinates) are generally better predictors than relative features (speed and acceleration). That is, it seems to be more relevant to know where are you when than how you got there.

We caution that effects of true decisions, rather than the simulated quasi-decisions tested here, may indeed have an impact on speed and acceleration. It has been suggested that the speed and acceleration components can capture the level of commitment towards the response, such that a change of decision (*switched* trajectories) might have associated a specific speed/acceleration pattern [Hehman et al., 2014]. This is not visible, however, in our data.

4 LDA *versus* traditional mouse-tracking analyses

The LDA classifier derives a solution to separating two kinds of mouse trajectories that is in a certain sense optimal. Previous studies have used alternative techniques to analyse mouse trajectories. In what follows, we will compare the performance of our LDA to the one of other measures commonly used in mouse tracking studies. We focus on measures that mainly assess the spatial disorder in trajectories, which is typically taken to be indicative of unpredictability and complexity in response dynamics [Hehman et al., 2014].

Two of the most commonly used methods of mouse tracking **spatial analysis** are the *Area under the trajectory* and the *Maximal deviation* (henceforth, AUT and MD respectively) (see Freeman and Ambady, 2010) The AUT is the geometric area between the observed trajectory and an idealised straight-line trajectory drawn from the start to the end points, whereas the MD is the point that maximises the perpendicular distance between this ideal trajectory and the observed path (Figure 7). For both measures, higher values are associated with higher deviation peaks towards the alternative; values closer to zero (or below) suggest trajectory close to ideal. Another frequently used measure is based on quantifying the number of times the trajectories goes back and forth along the x -axis (horizontal flips, Dale and Duran, 2011, as illustrated in Figure 7).

While all these measures serve to evaluate the degree of complexity of the path, they might fail to distinguish between ‘two-step’ and ‘uncertain’ decision processes –i.e., trajectories with a true deviation to

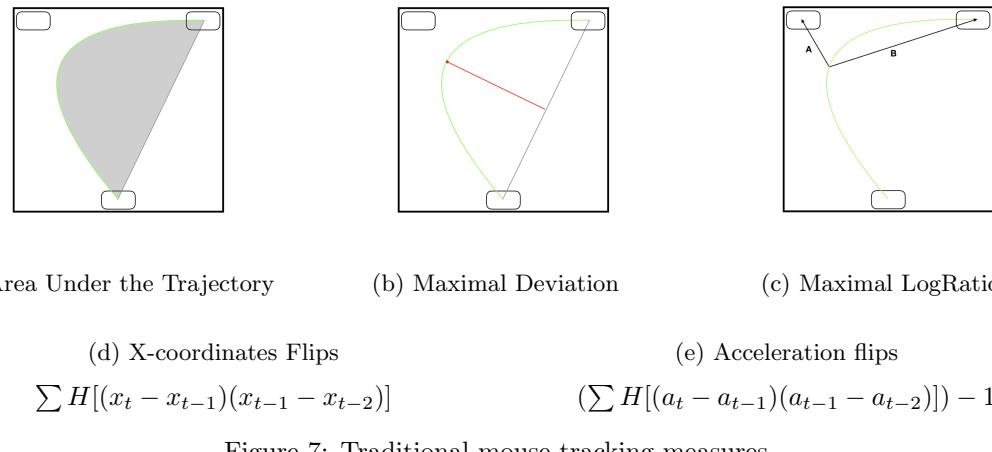


Figure 7: Traditional mouse-tracking measures.

the alternative or centred on the middle of the screen². To assess more directly whether mouse trajectories have a meaningful deviation towards the alternative, the distance to both target and alternative responses should be taken into account. For instance, the *ratio of the target distance to the alternative distance* can be calculated 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.

¹⁹¹ AUT, MD, x-coordinates flips and the point that maximises the log-ratio (Maximal LogRatio, henceforth) were calculated for the validation data. Following Dale and Duran (2011, and other studies on error corrections), we also analysed the *acceleration component* (AC) as a function of the number of changes in acceleration. Since stronger competition between alternative responses is typically translated into steeper acceleration peaks, changes in acceleration can be interpreted as decision points [Hehman et al., 2014]. ?? illustrates the distribution and mean values for each ‘quasi-decision’.

197 The same cross-validation procedure described in the previous section was used to diagnose the performance
198 of each of these measures. The mean AUC values for each of these measures are illustrated in Figure 9.
199 Table 3 summarises the result of comparing the LDA performance to the one of each alternative measure.

	ORIGINAL LDA	AUT	MD	MAXIMAL LOGRATIO	X-COORD. FLIPS	AC
AUC (mean)	.87	.62	.81	.81	.73	.53
Mean Difference	–	.24	.06	.06	.14	.34
<i>p</i> 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.

Overall, these comparisons reveal that the LDA is significantly better at classifying validation data than other commonly used measures. The difference with the classifier is in all the cases significant. Mean AUC values suggest that the MD and the Maximal LogRatio are better at distinguishing decision processes than the other alternative measures. These two measures are the only ones calculated based on coordinates, and therefore give more importance to spatio-temporal information than the others. In other words, both the MD and the Maximal LogRatio give different weight to positions depending the moment when they occurred, and therefore are more sensitive to the moment at which deviation occurred. This information seems to be essential for the classification, as observed in Section 3.

²A 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. Measures such as the AUT might not be able to make a significant distinction between them.

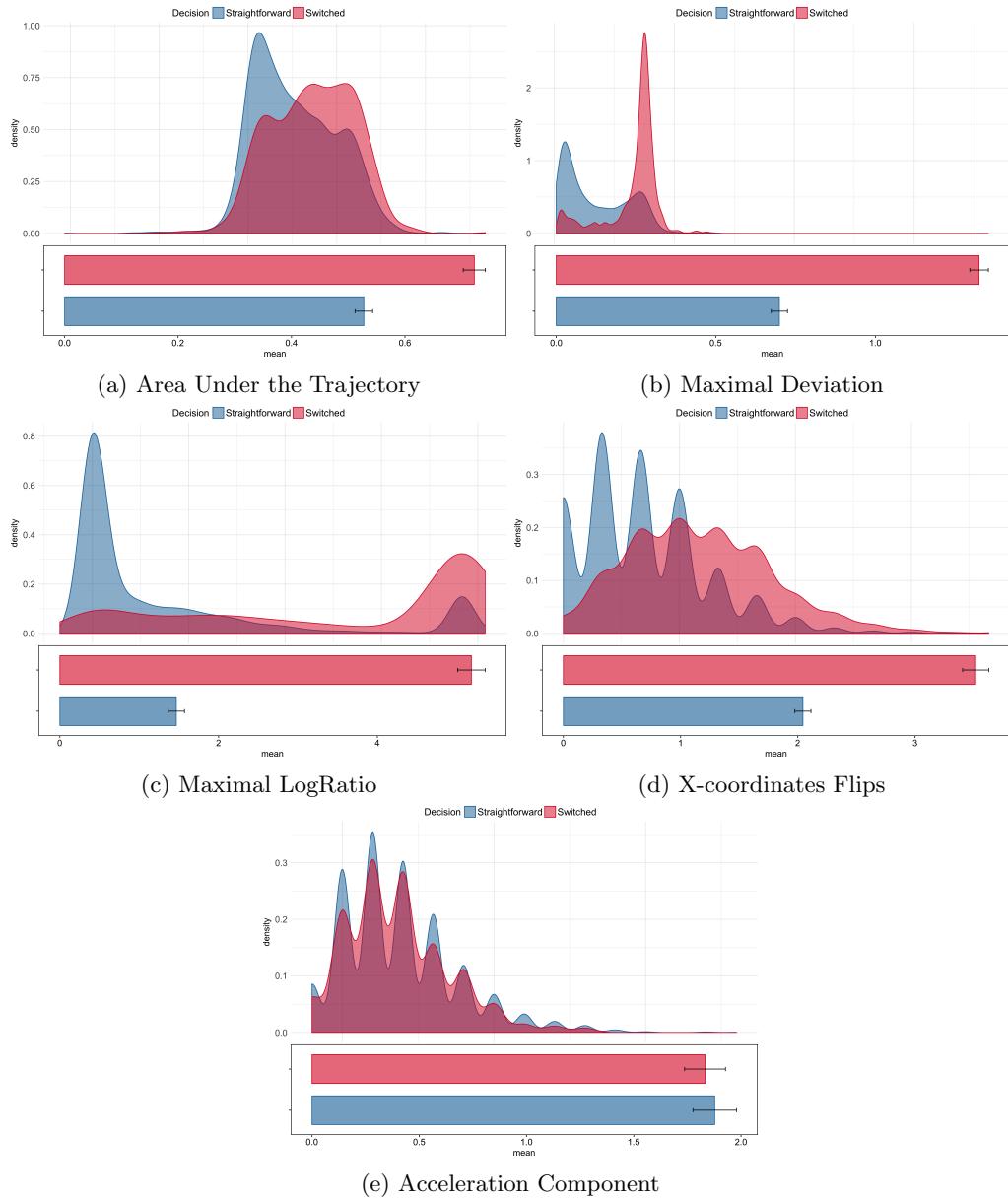


Figure 8: Different measures applied to validation data.

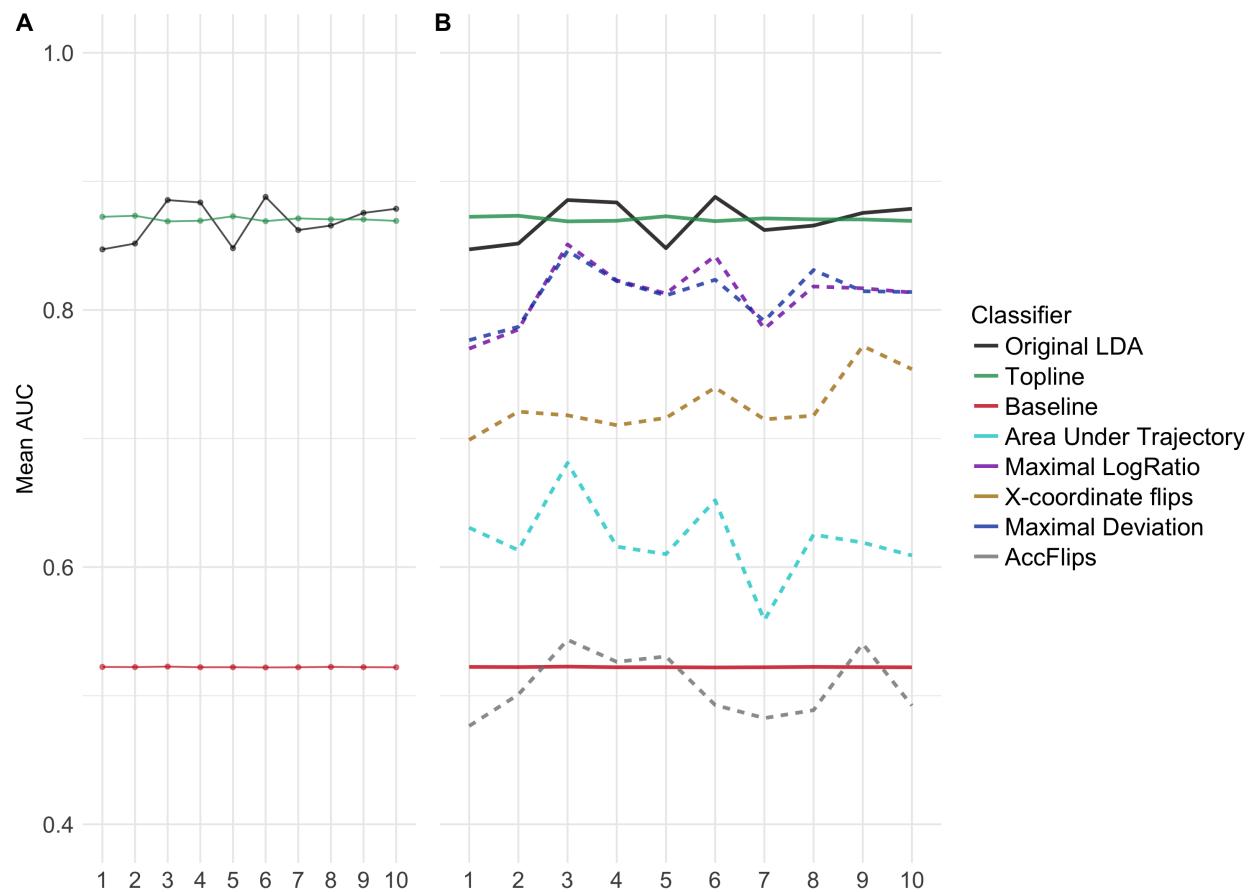


Figure 9: Mean AUC values

208 Finally, we previously observed that velocity and acceleration were not good predictors for the LDA
209 classifier. Indeed, the performance of the Acceleration Component overlaps here with that of the Baseline,
210 suggesting that this type of information is not helpful.

211 We have shown that (i) a rough manipulation of decision making processes has a direct impact on mouse
212 trajectories; (ii) a LDA using absolute-temporal information is enough to accurately distinguish these quasi-
213 decisions; and (iii) this LDA does a better classification than other traditional mouse-tracking measures. But,
214 can our LDA can classify more complex decision processes, such as the ones involved in sentence verification
215 tasks?

216 5 Extension to linguistic data

217 How well does our LDA, trained on “quasi-decisions”, classify new trajectories, which underly cognitive
218 processes that might or not correspond to different decision patterns? To address this question, we test our
219 classifier on data obtained from a replication of Dale and Duran’s (2011) experiment.

220 Dale and Duran (2011) found differences in the processing of true positive and negative sentences when
221 people performed a simple truth-value judgment task. These results were interpreted as indicating that
222 negation underlies an abrupt shift in cognitive dynamics (i.e. an unconscious change of decision). If this is
223 indeed the case, we would expect mouse trajectories corresponding to the verification of negative sentences
224 to pattern with *switched* trajectories from the validation experiment. This pattern of results would provide
225 additional support to the hypothesis that, at least in out-of-the-blue contexts, processing negation does
226 involve a ‘two-step’ derivation, where the positive argument is initially derived and negated only as a second
227 step³.

228 5.1 Experiment

229 Participants were asked to perform a truth-judgment task, where they had to decide whether a sentence (e.g.
230 *Cars have wheels*) is true or false according to common knowledge. Each of sentence could be either a true
231 or a false statement in its negated or non-negated form. Unlike Dale and Duran’s experiment, the complete
232 statement was presented in the middle of the screen after participants pressed “start” (i.e. no self-paced
233 reading). The response buttons appear at the top-left or right corners of the screen, as in our validation
234 experiment. Materials and design are exemplified in Table 4.

235 **Participants** 53 English native speakers were tested using Amazon Mechanical Turk. They were rewarded
236 for their participation (1USD). The experiment lasted approximately 10 minutes.

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

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

Table 4: Design

240 **Interface and data treatment** The interface and data treatment were similar to the ones used for in the
241 validation experiment. Mouse trajectories’ time course was normalised into 101 time steps.

³The data used to train the classifier correspond to *quasi*-decisions; namely, the training set is only an approximation of what should happen during an unconscious change of decision, such as the one expected for negation processing. We then expect some aspects of the decision processes on sentence-verification data not to be captured by the LDA.



Apples are not fruits.

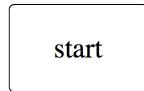


Figure 10: Illustration of a Trial in Dale & Duran’s Replication

242 5.2 Results and discussion

243 **Replicating Dale and Duran (2011)** All participants responded correctly more than 75% of the time.
 244 No participant was discarded based on accuracy. Only accurate trials were analysed. [Figure 11](#) illustrates
 245 mean trajectories for the four conditions.

246 To assess whether we replicate Dale and Duran’s results, we calculated the x -coordinate flips (see [Section 4](#))
 247 and analysed them with a linear mixed-effects model [Baayen et al., 2008], taking TRUTH, POLARITY and
 248 their interaction as predictors. We included random intercepts per subject and a random slope with the
 249 interaction of both factors. P -values were obtained by comparing the omnibus model to a reduced version
 250 of itself, where the relevant factor was removed. This pipeline mirrors the model performed by Dale and
 251 Duran. Note that, unlike Dale and Duran, we did not perform statistical analyses based on the acceleration
 252 component, since this quantitative measure was unable to distinguish mouse trajectories underlying different
 253 ‘quasi-decisions’ in the validation experiment.

254 The model of x -coordinate flips revealed a main effect of POLARITY, such that negation significantly
 255 increases flips in the x -coordinate by 0.76 ($\chi^2 = 10.11; p = .0014$), and a significant interaction TRUTH \times
 256 POLARITY ($\chi^2 = 22.7; p < .001$), such that the difference between negative and positive sentences is bigger
 257 for the true than for the false statements. There was no significant effect of TRUTH ($\chi^2 < 1; p = .5$). [Table 5](#)
 258 summarises the results of ours and Dale and Duran’s experiments.

259 These results seem to replicate Dale and Duran’s findings: Verifying true negated sentences produces
 260 less straightforward trajectories than true positive sentences. The values obtained in the two experiments,
 261 however, are slightly different; namely, our results present higher range of values (see [Table 5](#)). In our
 262 experiments, the mouse-position was not sampled at a fixed rate, creating additional noise which could be
 263 responsible for the range difference.

264 **Classifier performance** Two different LDA classifiers, trained with data from the validation experiment,
 265 were applied to the new experimental data. The first classifier was our original LDA, which had as predictors
 266 x, y coordinates as well as distance-based velocity and acceleration. The second LDA had only x, y coordinates
 267 as predictors. Validation results (see [Section 3](#)) suggest that the simpler model, which only relies on absolute

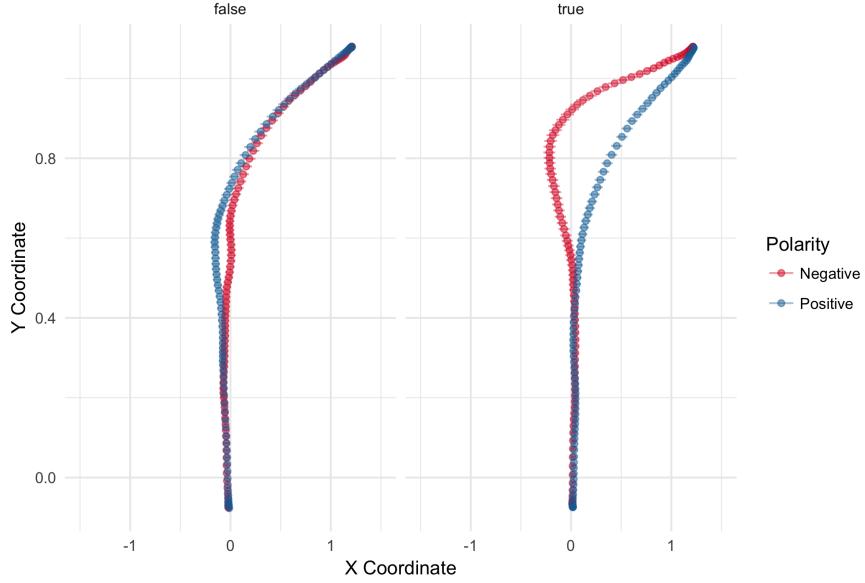


Figure 11: Mean trajectories for accurate trials

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 5: Mean and effect estimates

268 information, might be sufficient to classify the two basic kinds of decision-making processes. That is to say, the
269 simple model might fit the data as well as a more complex model, and be interpreted more straightforwardly.

270 The relevant difference in processing between positive and negative sentences is expected to arise specifically
271 for *true* statements. Consequently, we analyse the performance of both classifiers when applied to true
272 trials. Figure 12 illustrates the distribution and means of the resulting *LDA measure*.

273 To assess how well these classifiers separate positive from negative trials, we bootstrapped new samples
274 from the original set of data (iterations=1000) and calculated the area under the ROC curve for the classifi-
275 cation of each of these samples. To estimate the classification power, we evaluated the performance after
276 reducing the sample size. Figure 13A shows the mean AUC values obtained after applying the same procedure
277 to different sample sizes. Note that these values are generally lower than the ones obtained in the validation
278 experiment. This is not surprising given that the classifier is being trained and tested with different sets of
279 data, which target different cognitive processes.

280 Could the observed performance be expected if negative and positive trials were actually not different
281 from each other? Are these AUC values significantly different from the ones one would have obtained from
282 applying the LDA to a set of data where there is no difference between experimental conditions (i.e. *null*
283 *hypothesis*)? We calculated the AUC values for a set of data where experimental labels (positive, negative)
284 were scrambled. The distribution of AUC values under the *null hypothesis* was compared to the performance
285 observed for the original set of data. Figure 13B illustrates the separability of the two classifications for each
286 sample size.

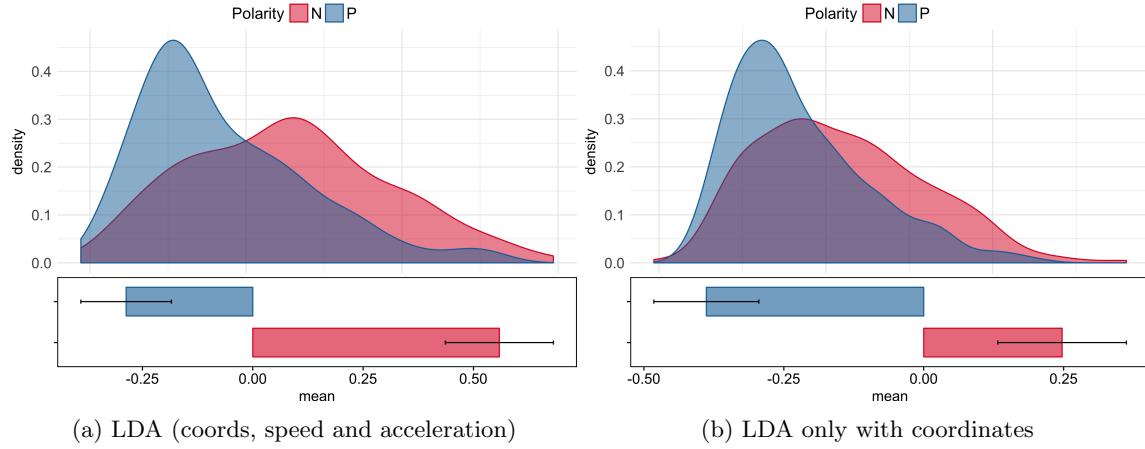


Figure 12: Two LDA classifiers applied to *true* trials.

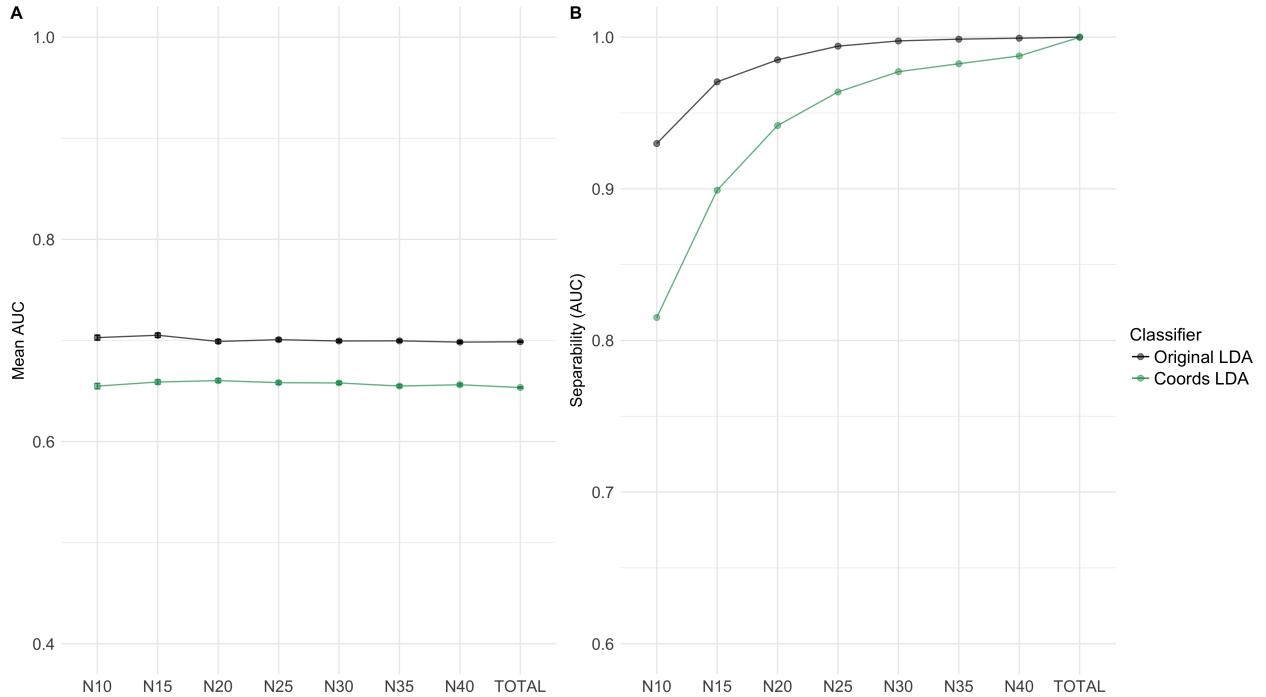


Figure 13: **Performance of LDA classifiers.** A. Mean AUC values over bootstrapped data (iterations=1000) for different sample sizes; B. Difference of classifier performance when applied to scrambled vs. original set of data.

287 The LDA classifier trained with “quasi-decisions” seems to make a relevant distinction between exper-
288 imental conditions. This finding suggest that the contrast between negative and positive trials has something
289 in common with the contrast in the validation experiment. The fact that negation has similar properties as
290 pseudo “switched-decisions” indicates that verifying negative sentences might underlie a change of decision,
291 as proposed by Dale & Duran (2011), among others. However, while mouse trajectories corresponding to
292 negative and ‘switched’ trials do share basic properties (e.g. shape), they seem to differ on how they are
293 placed in the “change of decision” spectrum; namely, they occupy different parts of the quasi-decision-based
294 LDA continuum (compare Figure 5 and Figure 12). This is not surprising given that we are dealing with
295 different cognitive processes –very simple ‘quasi-decisions’ vs. sentence-verification based decision.

296 Finally, note that while the classifiers’ comparison in Figure 6 indicated that relative spatio-temporal
297 features, such as distance-based acceleration and speed, were not essential for the classification of ‘quasi-
298 decisions’, these features do seem to play a role in the classification of sentence-verification data. Indeed,
299 Figure 13 reveals that the *full* classifier –which takes all features as predictors– makes a better distinction
300 than the simplified one.

301 **Other mouse-tracking measures** Does the difference in performance between the LDA and other mouse-
302 tracking measures remain when these are applied to the new experimental data? While it’s true the LDA
303 trained on validation data can make a distinction between negative and positive trials, it might not be the
304 *best* possible strategy for classification.

305 We address the question of whether different measures differ on their ability to find the observed effect
306 by applying the same procedure as before: we calculated the mean area under the ROC curve for different
307 sample sizes (cf. Figure 15A), and contrasted these values against the null hypothesis (i.e. the values we
308 would have obtained if there had been no difference between the experimental conditions; Figure 15B)⁴.

309 The results in Figure 15A suggest that most measures perform a worse classification than the one observed
310 for the validation data (cf. Figure 9). Since a decrease in performance is attested across the board and not
311 only for the classifiers trained with validation data, this difference must be driven by properties of the new
312 data set. Specifically, the sentence-verification data might be more variable, such that both negative and
313 positive trials might underlie instances of different decision processes.

314 Moreover, the LDA classifier seems here to be as powerful as other traditional mouse-tracking measures,
315 such as the Maximal Deviation and the Maximal LogRatio. In contrast with the results of the validation
316 experiment, this opens the possibility of using any of these alternative measures to analyse mouse-tracking
317 data from sentence verification tasks. Importantly, the classifier is still a better choice from a conceptual
318 perspective, as long as it does not make any specific assumption about how the change of decision should be
319 reflected by mouse-trajectories.

320 **Baseline** Our LDA, trained to classify ‘quasi-decisions’, can separate the two experimental conditions of
321 Dale & Duran’s replication. We have interpreted this result by suggesting that the LDA is distinguishing
322 mouse trajectories that underlie two different decision processes. Alternatively, one could argue that the
323 classification made by the LDA is not based on decision processes, but on some other feature of mouse paths,
324 which happens to be partially shared between conditions in both experiments. For example, the LDA might
325 not be sensitive to decision shift but to cognitive cost, and the contrast between straightforward and switched
326 trials, on one hand, and positive and negative trials, on the other, might have that in common.

327 To disentangle these possibilities, we asked how our LDA classifies trajectories that might have different
328 shapes but do not underlie two different decision processes (straightforward vs. switched), but a single one.
329 To this end, we constructed a *baseline* set of data, which contained only positive trials from the original
330 data set. These trials were further categorised into two classes depending on whether their response time
331 was above or below the subject mean. We reasoned that shorter response times would correspond to an
332 ‘early commitment’ towards the answer, whereas longer response times would reflect a ‘late commitment’.
333 Importantly, no trial in the *baseline* data set was assumed to underlie a decision shift; thus, the LDA was
334 expected to perform a poor classification.

335 As illustrated by in Figure 16a, the two classes in the baseline data have slightly different trajectory
336 shapes. The distribution of the LDA measure after testing the classifier on the new data set is shown in

⁴The ranking based on power does not correspond exactly to the one based on the AUC mean values: how good is each measure at detecting the effect is not necessarily equivalent to its absolute performance.

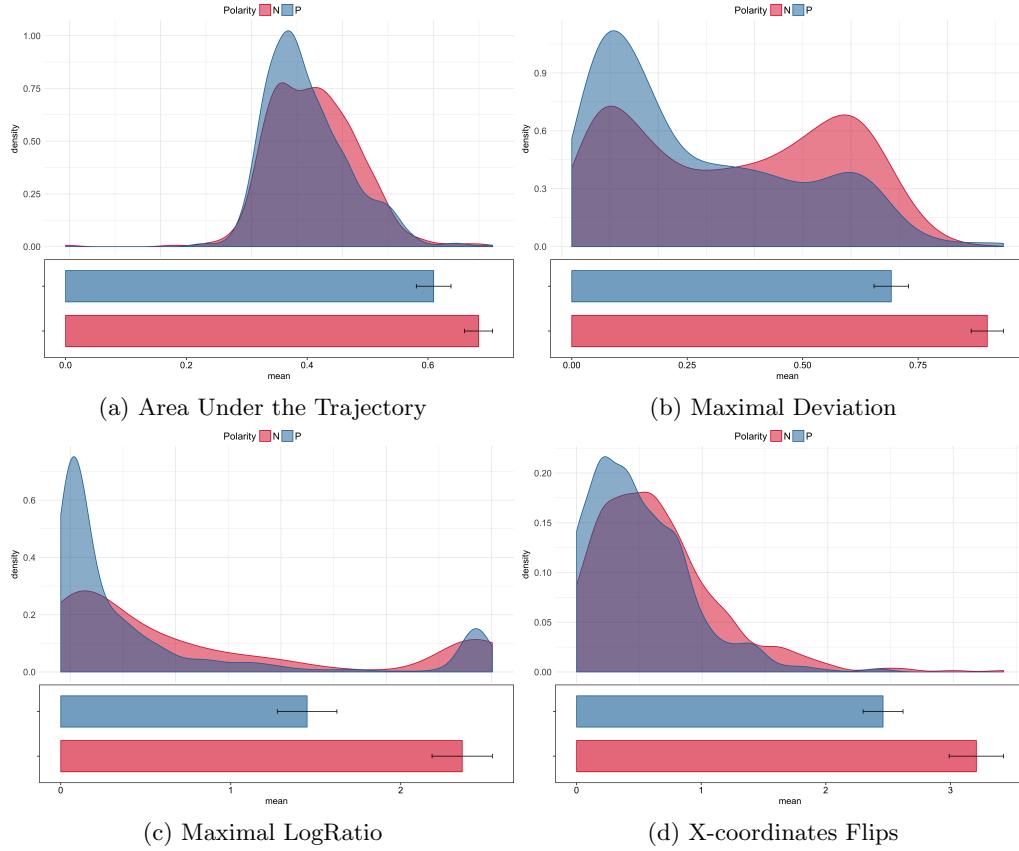


Figure 14: Different measures applied to Dale & Duran replication.

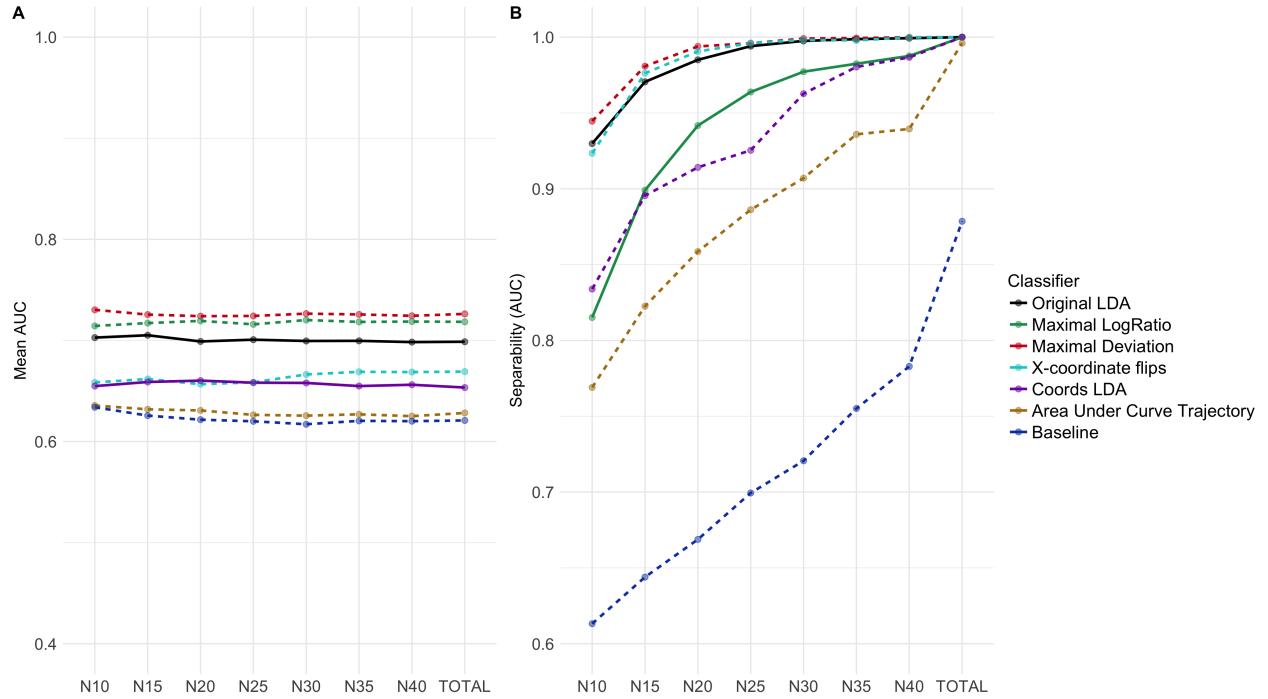


Figure 15: Performance of other measures

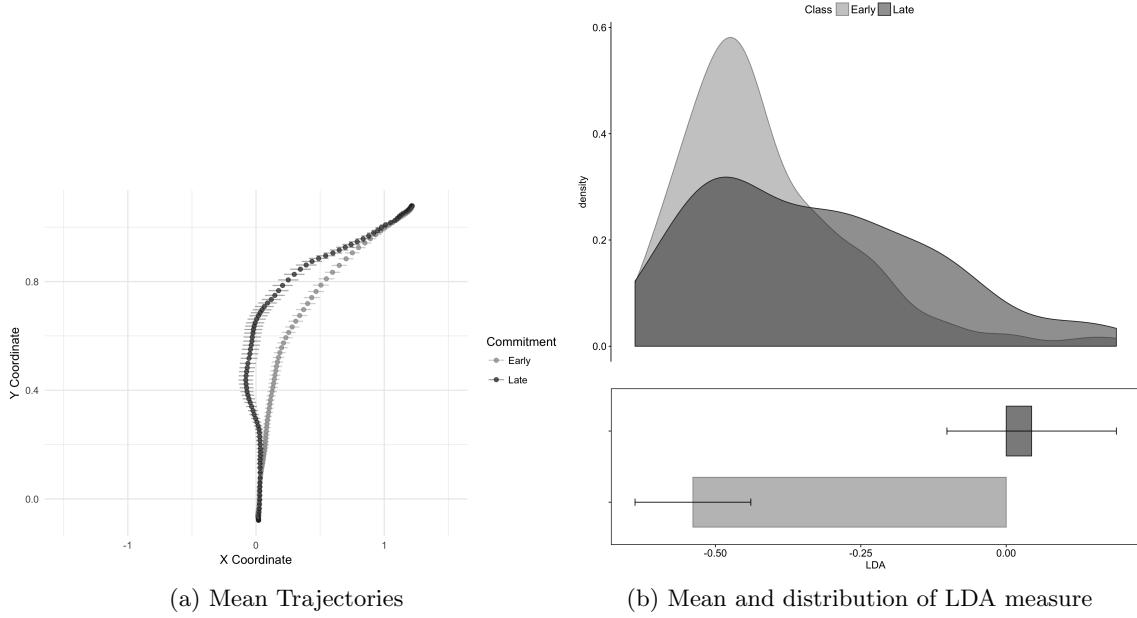


Figure 16: Baseline data set

337 [Figure 16b](#). The performance was evaluated following the same procedure applied above (see blue line in
 338 [Figure 15](#)).

339 The classification on ‘early’ and ‘late’ categories is less accurate than the one performed in the original data
 340 set, to separate negative and positive trials. Differences in trajectories that are not due to the experimental
 341 manipulation are poorly captured by the LDA measure: even trajectories that do have some properties in
 342 common with *switched* and *negation* trials are not taken to be underlying a change of decision. These findings
 343 suggest that our classifier is not just tapping onto trajectory similarity but onto decision processes reflected
 344 on mouse trajectories.

345 6 Conclusion/General discussion

346 We aimed to investigate the connection between action and cognition by testing one of its specific instances:
 347 the mapping of decision making processes into mouse movements. Our findings make three main contributions
 348 on this point.

349 First, by manipulating whether the stimulus triggered or not a change of decision, we have shown –for
 350 the first time in a direct way– that mouse trajectories reflect basic decision processing: When participants
 351 were forced to change their answer, this switch had a systematic/direct mapping/impact on hand movements
 352 ([Section 2](#)).

353 Second, we trained a LDA classifier with mouse-trajectories underlying these ‘quasi-decisions’ to determine
 354 whether or not a given trial involved a decision shift. This LDA has been proven to accurately classify
 355 not only paths corresponding to other quasi-decisions’, but also mouse-trajectories underlying more complex
 356 decision processes, such as sentence verification. While the performance of the classifier –at this stage– might
 357 be as good as the one of other commonly used mouse-tracking measures (i.e. Maximal Deviation), it has the
 358 unique advantage of not relying on any specific assumption about how trajectories should look like. Indeed,
 359 we demonstrated that the LDA classifier is not just sensitive to superficial similarity between trajectories,
 360 but to the underlying cognitive processes.

361 Lastly, our results also contribute to the research in negation processing. Besides replicating Dale & Duran’
 362 experiment, the classification performed by the LDA suggests that verifying negative sentences involves a
 363 decision shift, similar to the one used for the training. We then provide new evidence to the hypothesis that
 364 processing negated sentences –at least in out-of-the-blue contexts– involves a two-step derivation, where the
 365 positive argument is initially computed.

366 To conclude, we should mention that the differences in the LDA performance across the two experiments
367 can be well understood by noticing that two data sets capture slightly different decision processes. In order to
368 capture more subtle contrasts in decision making, the training data should contain more variation. Further
369 research will explore this possibility.

370 **References**

- 371 R. H. Baayen, D. J. Davidson, and D. M. Bates. Mixed-effects modeling with crossed random effects for
372 subjects and items. *Journal of memory and language*, 59(4):390–412, 2008.
- 373 E. A. Cranford and J. Moss. Mouse-tracking evidence for parallel anticipatory option evaluation. *Cognitive
374 processing*, pages 1–24, 2017.
- 375 R. Dale and N. D. Duran. The Cognitive Dynamics of Negated Sentence Verification. *Cognitive Science*, 35:
376 983–996, 2011. ISSN 03640213. doi: 10.1111/j.1551-6709.2010.01164.x.
- 377 T. a. Farmer, S. a. Cargill, N. C. Hindy, R. Dale, and M. J. Spivey. Tracking the continuity of language
378 comprehension: computer mouse trajectories suggest parallel syntactic processing. *Cognitive science*, 31:
379 889–909, 2007. ISSN 0364-0213. doi: 10.1080/03640210701530797.
- 380 J. B. Freeman and N. Ambady. MouseTracker: software for studying real-time mental processing using a
381 computer mouse-tracking method. *Behavior research methods*, 42(1):226–241, 2010. ISSN 1554-351X. doi:
382 10.3758/BRM.42.1.226.
- 383 J. B. Freeman and K. L. Johnson. More than meets the eye: split-second social perception. *Trends in
384 cognitive sciences*, 20(5):362–374, 2016.
- 385 J. B. Freeman, R. Dale, and T. A. Farmer. Hand in motion reveals mind in motion. *Frontiers in Psychology*,
386 2(APR):1–6, 2011. ISSN 16641078. doi: 10.3389/fpsyg.2011.00059.
- 387 T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference
388 and Prediction*. Springer, New York, second edition edition, 2009.
- 389 E. Hehman, R. M. Stolier, and J. B. Freeman. Advanced mouse-tracking analytic techniques for enhancing
390 psychological science. *Group Processes & Intergroup Relations*, pages 1–18, 2014. ISSN 1368-4302. doi:
391 10.1177/1368430214538325.
- 392 B. Kaup, R. H. Yaxley, C. J. Madden, R. A. Zwaan, and J. Lüdtke. Experiential simulations of negated text
393 information. *The Quarterly Journal of Experimental Psychology*, 60(7):976–990, 2007.
- 394 J. Lüdtke, C. K. Friedrich, M. De Filippis, and B. Kaup. Event-related potential correlates of negation in a
395 sentence–picture verification paradigm. *Journal of Cognitive Neuroscience*, 20(8):1355–1370, 2008.
- 396 M. S. Nieuwland and G. R. Kuperberg. When the truth is not too hard to handle: An event-related potential
397 study on the pragmatics of negation. *Psychological Science*, 19(12):1213–1218, 2008.
- 398 U. Sauerland, A. Tamura, M. Koizumi, and J. M. Tomlinson. Tracking down disjunction. In *JSAI Interna-
399 tional Symposium on Artificial Intelligence*, pages 109–121. Springer, 2015.
- 400 J.-H. Song and K. Nakayama. Role of focal attention on latencies and trajectories of visually guided manual
401 pointing. *Journal of Vision*, 6(9):11, 2006.
- 402 J. H. Song and K. Nakayama. Hidden cognitive states revealed in choice reaching tasks. *Trends in Cognitive
403 Sciences*, 13(8):360–366, 2009. ISSN 13646613. doi: 10.1016/j.tics.2009.04.009.
- 404 M. J. Spivey and R. Dale. Continuous dynamics in real-time cognition. *Current Directions in Psychological
405 Science*, 15(5):207–211, 2006.

- 406 M. J. Spivey, M. Grosjean, and G. Knoblich. Continuous attraction toward phonological competitors. *Proceedings of the National Academy of Sciences of the United States of America*, 102(29):10393–10398, 2005.
407
408 ISSN 0027-8424. doi: 10.1073/pnas.0503903102.
- 409 Y. Tian and R. Breheny. *Dynamic Pragmatic View of Negation Processing*, pages 21–43. Springer International Publishing, Cham, 2016. ISBN 978-3-319-17464-8. doi: 10.1007/978-3-319-17464-8_2. URL
410 https://doi.org/10.1007/978-3-319-17464-8_2.
- 411
412 Y. Tian, R. Breheny, and H. J. Ferguson. Why we simulate negated information: A dynamic pragmatic account. *The Quarterly Journal of Experimental Psychology*, 63(12):2305–2312, 2010.
- 413
414 J. M. Tomlinson, T. M. Bailey, and L. Bott. Possibly all of that and then some: Scalar implicatures are understood in two steps. *Journal of memory and language*, 69(1):18–35, 2013.
- 415
416 P. C. Wason. The contexts of plausible denial. *Journal of verbal learning and verbal behavior*, 4(1):7–11, 1965.
- 417
418 P. C. Wason and P. N. Johnson-Laird. *Psychology of reasoning: Structure and content*, volume 86. Harvard University Press, 1972.
- 419
420 M. Wojnowicz, M. J. Ferguson, M. Spivey, M. T. Wojnowicz, M. J. Ferguson, R. Dale, and M. J. Spivey. The Self-Organization of Explicit Attitudes. 20(July 2017):1428–1435, 2009. doi: 10.1111/j.1467-9280.2009.02448.x.
- 421
422
423 K. Xiao and T. Yamauchi. Semantic priming revealed by mouse movement trajectories. *Consciousness and cognition*, 27:42–52, 2014.
- 424
425 K. Xiao and T. Yamauchi. The role of attention in subliminal semantic processing: A mouse tracking study. *PloS one*, 12(6):e0178740, 2017.
- 426