Mouse tracking as a window into decision making

Mora Maldonado, Ewan Dunbar & Emmanuel Chemla

January 29, 2018

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 et al., 2005, Cranford and Moss, 2017], and syntactic, semantic and pragmatic processing [Farmer et al., 2007, Dale and Duran, 2011, Tomlinson et al., 2013, Xiao and Yamauchi, 2014, Sauerland et al., 2015, Xiao and Yamauchi, 2017, among others], to social cognition [Freeman and Ambady, 2010, Freeman et al., 2011, Freeman and Johnson, 2016]. All these studies have worked 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 and Nakayama, 2006, 2009, Freeman and Ambady, 2010, Spivey and Dale, 2006, Hehman et al., 2014]. As a result, features in mouse trajectories could be indicators of specific decision processes, revealing their dynamics with fine-grained temporal resolution[1]. Mouse-tracking studies typically present participants with a *two-alternative forced choice*, where they have to make a choice using the options appearing in the top left or right corner of the screen. Whenever a decision involves two independent processes –such as a change of mind–, 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 (see Figure 1, Wojnowicz et al., 2009). Of course, one could still imagine other types of decision, such as a single but late commitment, or a decision made after uncertainty or doubt. These might have a different reflection on mouse trajectories.

The existence of intuitions about how decision processes should be mapped into mouse paths (i.e. linking hypotheses) has allowed researches to draw conclusions about the cognitive processes underlying their experimental manipulations. Dale and Duran's (2011) approach to negation processing is an example of this. Negation has been traditionally understood as an operator that reverses the truth conditions of the sentence, inducing in "extra-step, or mental operation" in online processing (Wason, 1965, Wason and Johnson-Laird, 1972; see review in Tian and Breheny, 2016). To test the dynamics of negation integration, Dale and Duran tracked mouse trajectories as participants performed a Truth-Value Judgment Task (TVJT), where they had to verify the truth of general statements such as *Cars have (no) wings*. The authors found that mouse trajectories presented more shifts towards the alternative response (i.e. they were less straightforward) when evaluating negative than affirmative true sentences. These results were interpreted as evidence for a 'two-step' processing of negation, where truth conditions for the positive content are first derived and negated only afterwards, as a second step[2].

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

Our main goal here 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?

[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.

[2] Several studies have suggested that the positive argument might play an important role in negation processing [Kaup et al., 2007, Lüdtke et al., 2008, among others]. This pattern of results, however, seem to depend 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.

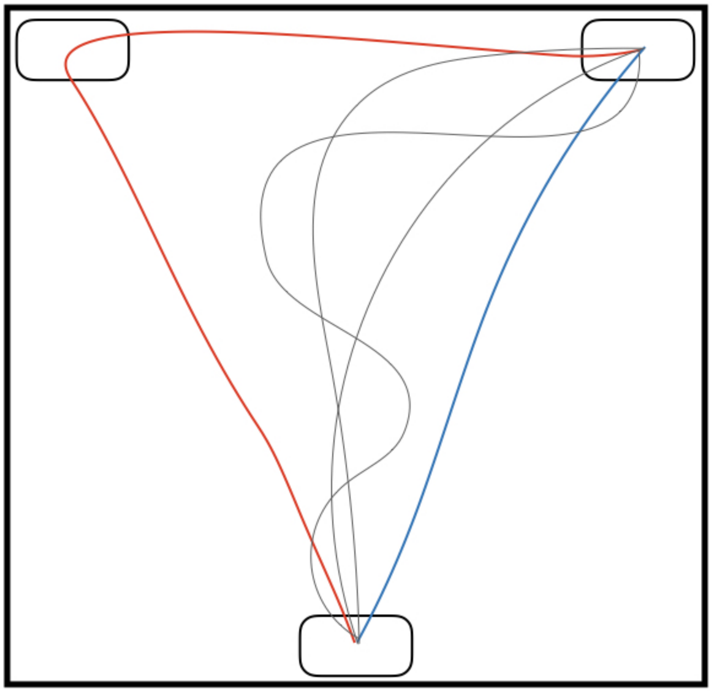


Figure 1: Shape of trajectories underlying distinct decision processes. One single cognitive process is expected to be mapped onto one smooth movement (blue line), whereas a change of mind would be reflected by two movements (red line). Intermediate cases are represented in gray.

One could easily imagine a situation where two different trajectory shapes do not correspond to two different decision mechanisms, but to a single one (due to uncertainty, noise, etc.). Differences in trajectories can therefore underlie something other than decision shift.

In this paper, we address these questions by identifying the features in mouse-trajectories corresponding to two different types of decisions: *straightforward* decisions (i.e. single commitment) and *shifted* decisions (i.e. a change of mind). First, we present a *validation* experiment where, instead of taking mouse trajectories as indicators of cognitive processes, we *manipulate* whether or not our stimuli trigger a flip in decision (Section 2). The data from this validation experiment (i.e. two groups of 'quasi-decisions') is then used to fed a *linear discriminant analysis* (henceforth, LDA), trained to classify trajectories depending on the underlying decision (Section 3). After comparing the performance of LDA classifier to other traditionally used mouse-tracking measures (Section 4), the LDA classifier will be further tested with new data, obtained from a replication of Dale and Duran's (2011) experiment on negation processing (Section 5). If there is a change of decision triggered by negation, trajectories corresponding to negative trials should be classified together with trajectories underlying decision change in the validation experiment.

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 colored 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 mirror decision processing, we manipulated whether the color of the frame remained stable or changed at some point during the trial. While in the latter case the initial choice will be the accurate response (*straightforward* trials), in the former, participants were forced to swap their answer (*switched* trials), mimicking a change of decision. Note that, since we are only *mirroring* decision making, we refer to these decision processes as *'quasi-decisions'*. An illustration of the procedure is provided in Figure 2.

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

Design Each trial instantiated one of two possible Decision patterns. In *straightforward* trials, the frame color remained stable, and the decision made at the beginning of the trial did not need to be revised. Conversely, in *switched* trials, the color swapped during the trial, forcing a revision of the initial choice. The point of change in *switched* trials was determined by the position in the y axis, and it could be early, middle or late. The Frame color at the response point was also controlled: it could be red (right button) or blue (left 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

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

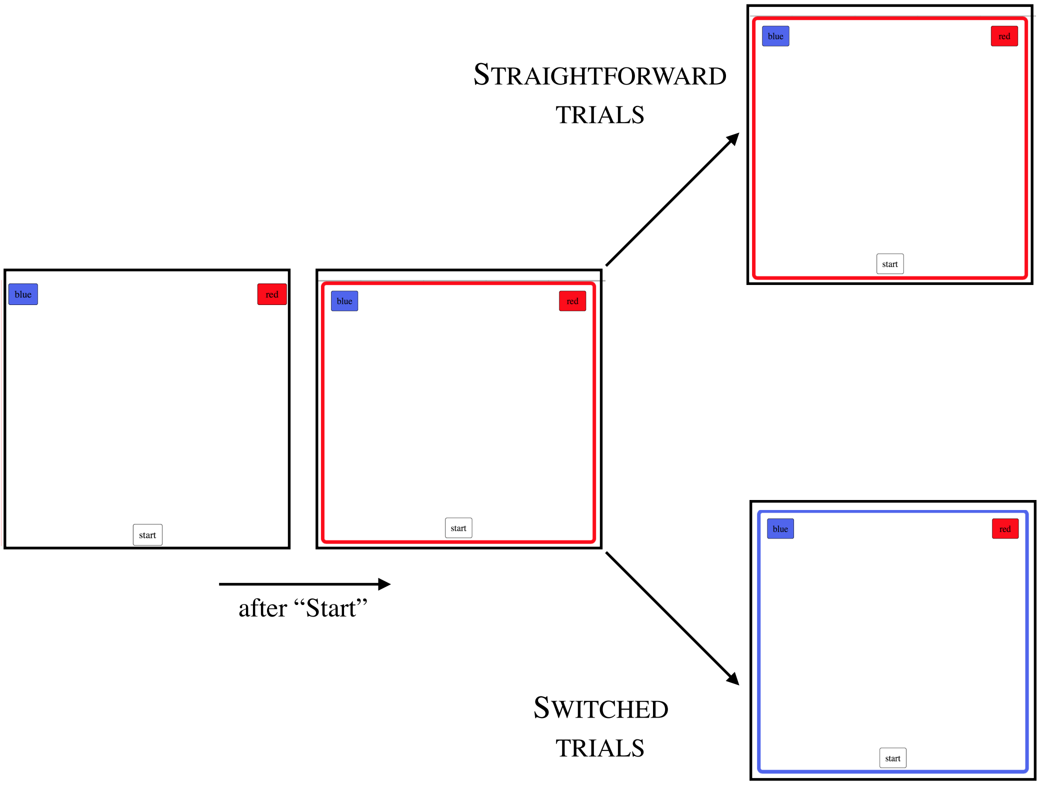


Figure 2: **Procedure in Validation Experiment.** Subjects were instructed to click the 'Start' button in order to see the colored frame. Response boxes were on the top-left or top-right. Depending on the trial condition, the frame color could change along the trial or not.

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' button 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 normalised 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.

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

3 Classifying decision processes with LDA

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

Description of LDA classifier The LDA is an optimal solution to classify continuous data –such as trajectories– into two or more classes –such as decision patterns (straightforward vs. switched quasi-decisions). In a nutshell, the LDA algorithm assumes that different classes have a common covariance 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 for the classification. It is a single number that represents the position on a line running between the two classes that maximizes their separability, with zero representing the midpoint between the two.

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

Performance of the LDA classifier The result of applying the procedure described in Figure 4 to the trajectories in the validation experiment is illustrated in Figure 5. To evaluate the overall performance of the classifier, we calculated the area under the ROC curve (AUC), a standard method for evaluating classifiers [Hastie et al., 2009]. Intuitively, the AUC gives the degree to which the histograms resulting from the

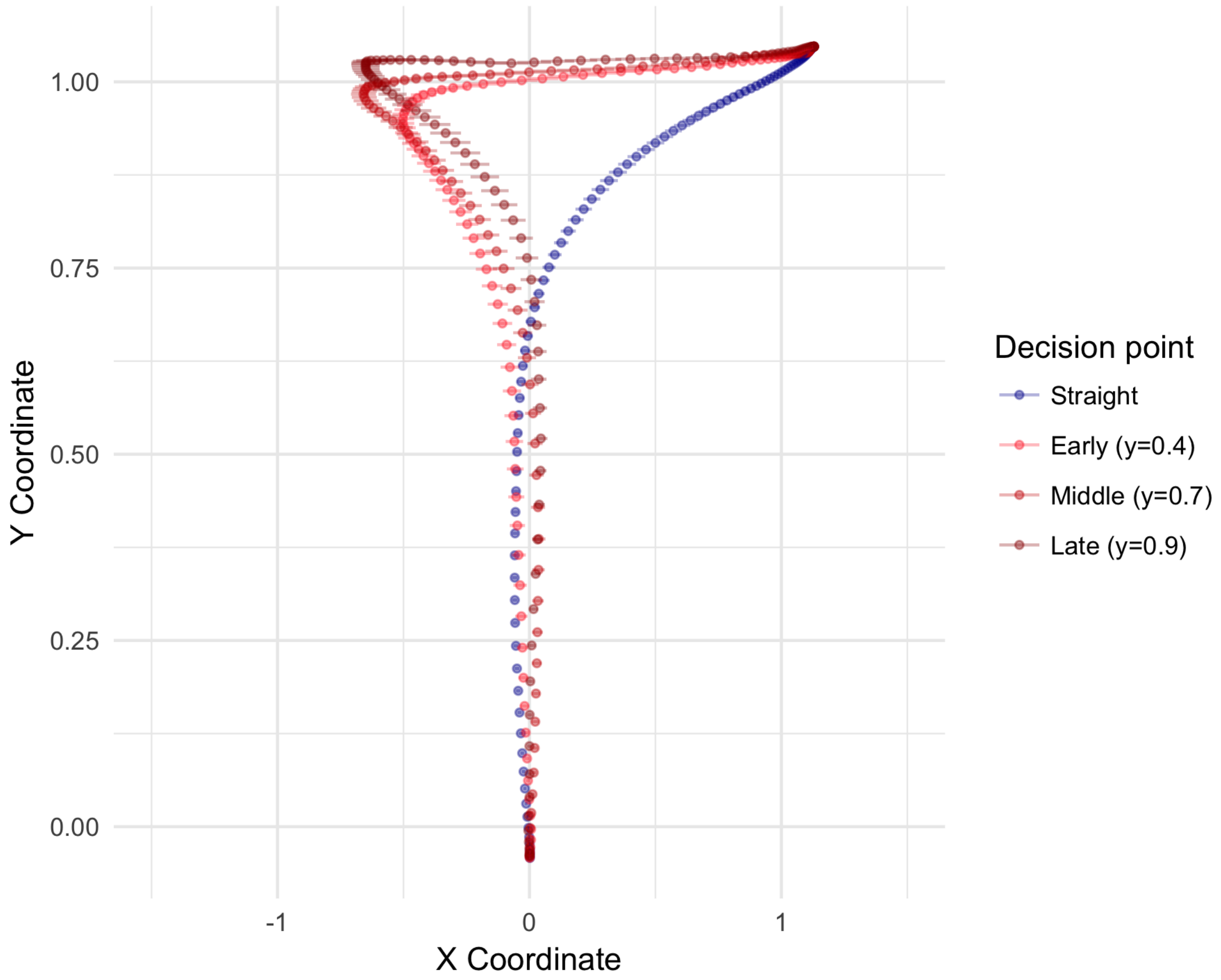


Figure 3: Mean trajectories per class

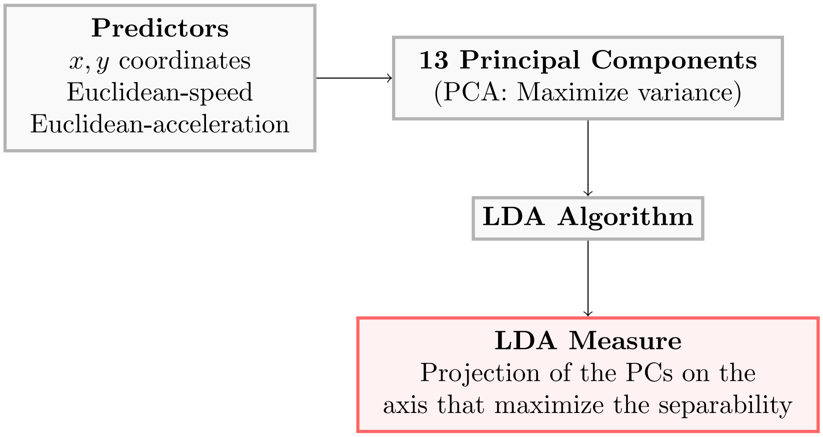


Figure 4: Diagram

classifier's continuous output (for example, Figure 5) is non-overlapping in the correct direction (in this case, "switched" more systematically in the positive direction on the classification axis than "straightforward").

To properly evaluate the classifier's perfomance at separating trials following the distribution in the experiments, the AUC measure was cross-validated. That is, calibration data were partitioned 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 were used as test set to diagnose the classifier performance. Thus, we obtained one AUC score for each test bin (ten bins). 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 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).

		(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.

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 can gather information about which features of mouse trajectories are more relevant for the distinction between decision processes (i.e. for the classification).

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

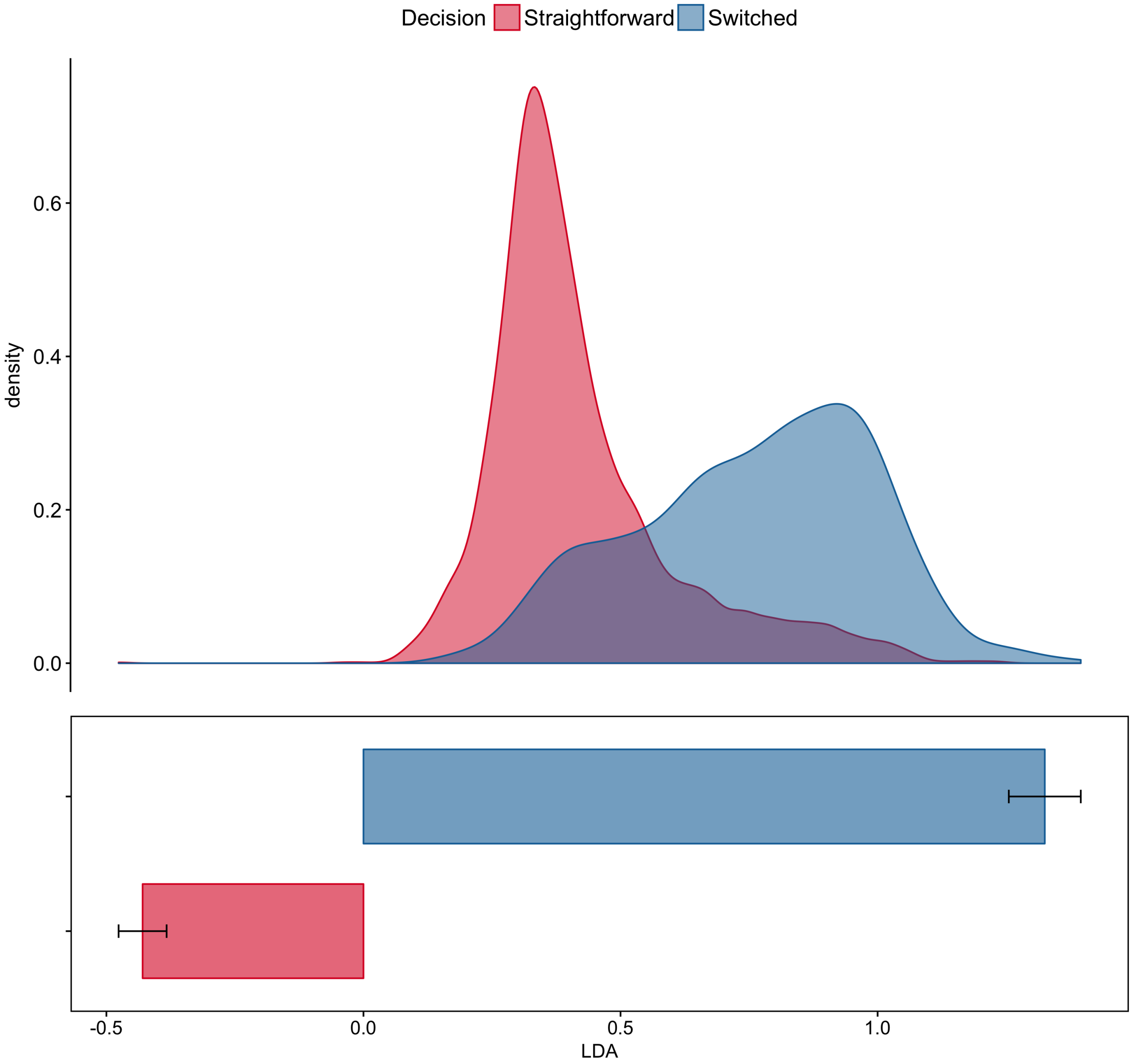


Figure 5: Classifier performance

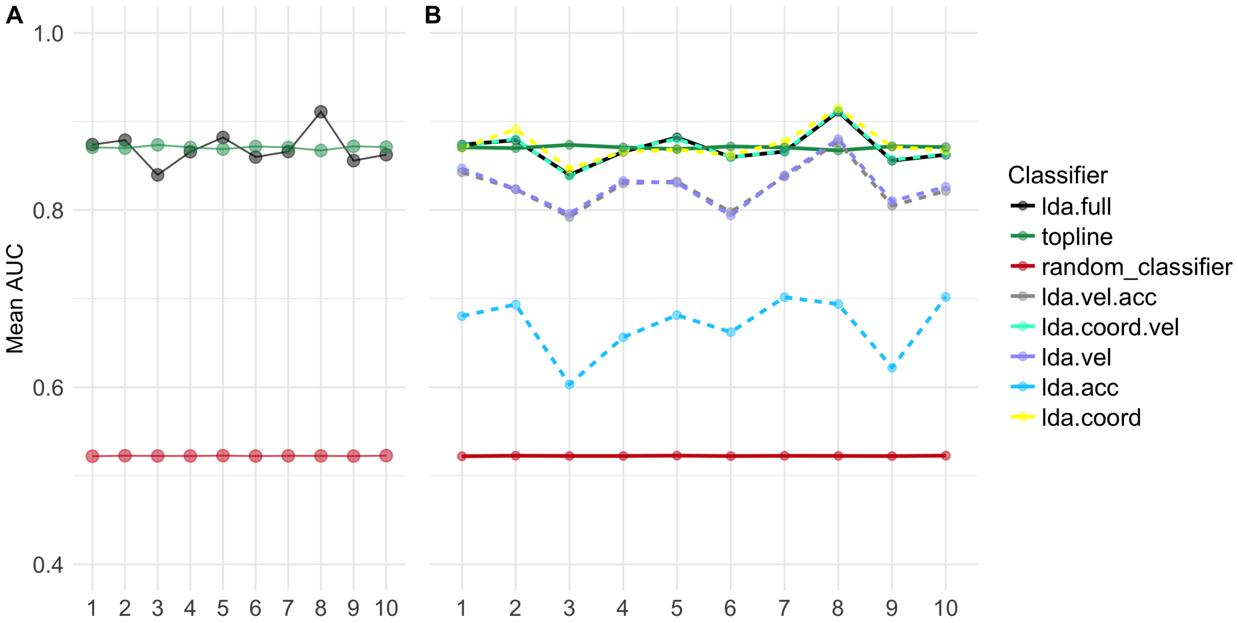


Figure 6: Mean AUC values

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 to say, it seems to be more relevant to know where are you when than how you got there.

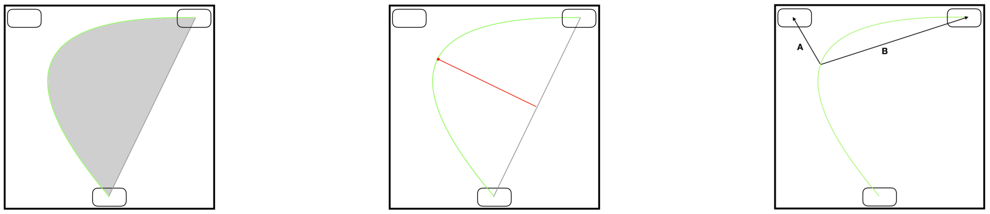
We caution that effects of true decisions, rather than the simulated quasi-decisions tested here, may indeed have an impact on speed and acceleration. It has been suggested that the speed and acceleration components can capture the level of commitment towards the response, such that a change of decision (*swiched* 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 classifier 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 mouse-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 to estimate the complexity of the trajectory 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



(a) Area Under the Trajectory (b) Maximal Deviation (c) Maximal LogRatio

(d) X-coordinates Flips

$$\sum H[(x_t - x_{t-1})(x_{t-1} - x_{t-2})]$$

(e) Acceleration flips

$$(\sum H[(a_t - a_{t-1})(a_{t-1} - a_{t-2})]) - 1$$

Figure 7: Traditional mouse-tracking measures.

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[3]. In order 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.

AUT, MD, x-coordinates flips and 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 (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). 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]. Figure 8 illustrates the distribution and mean values for each 'quasi-decision'.

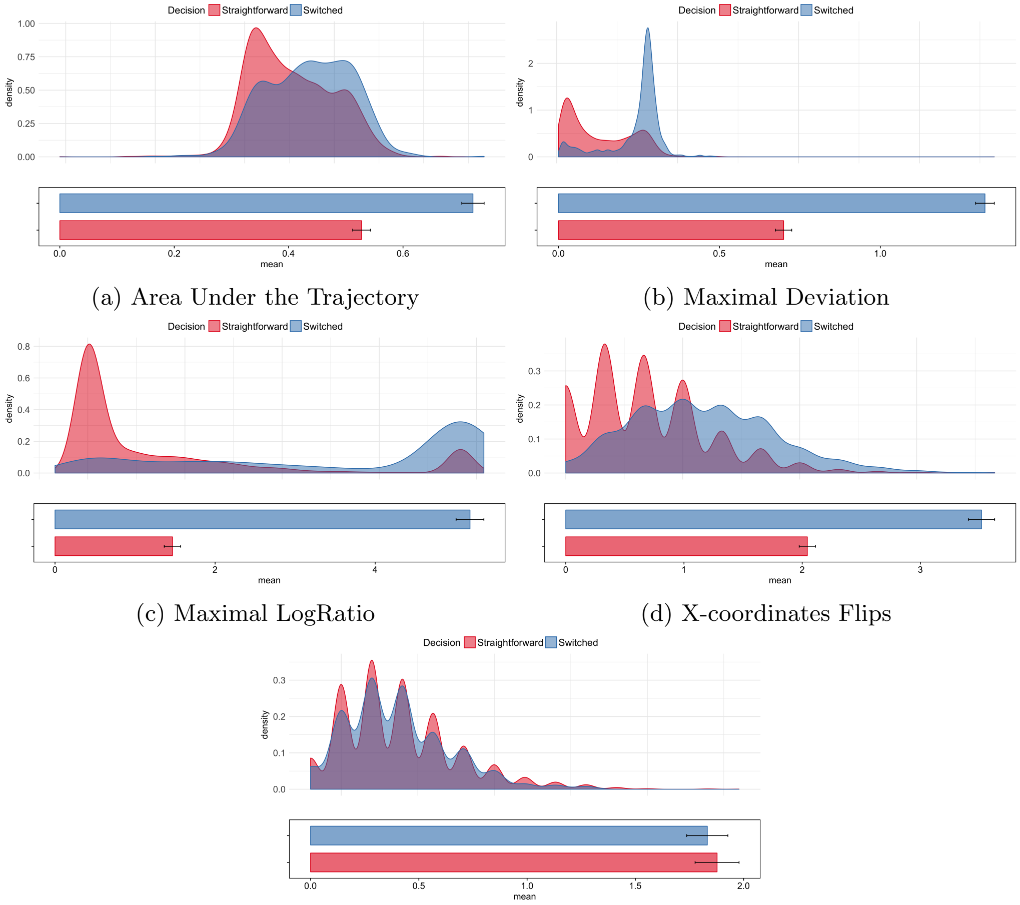
The same cross-validation procedure described in the previous section was used to diagnose the performance of each of these measures. The mean AUC values for each of these measures are illustrated in Figure 9. Table 3 summarises the result of comparing the LDA performance to the one of each alternative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

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 other measures such as the Area, the number of X-coordinate flips, and the Acceleration Component. These two measures are the only ones calculated based on coordinates, and therefore give more importance to

[3] 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 AUC might not be able to make a significant distinction between them.



(a) Area Under the Trajectory

(b) Maximal Deviation

(c) Maximal LogRatio

(d) X-coordinates Flips

(e) Acceleration Component

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

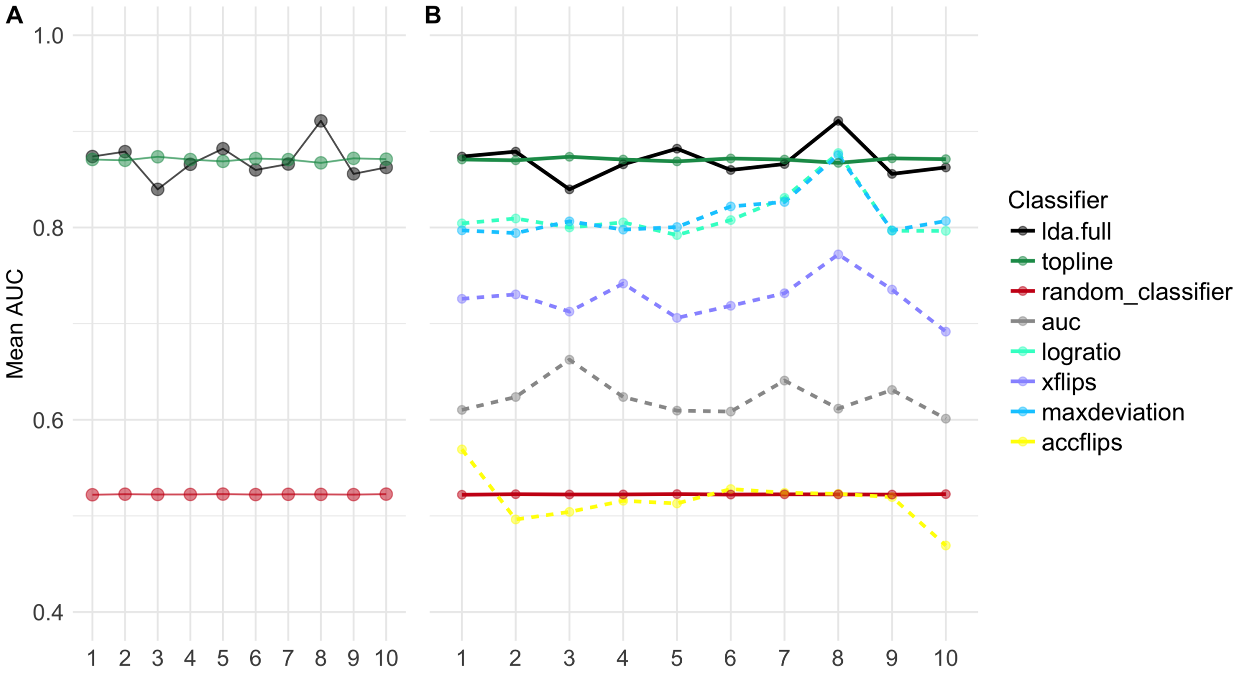


Figure 9: Mean AUC values

spatio-temporal information than the others. In other words, both the MD and the maximal 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. This information seems to be essential for the classification, as observed in Section 3.

Finally, we previously observed that velocity and acceleration were not good predictors for the LDA classifier. Indeed, the performance of the Acceleration Component overlaps here with that of the baseline (i.e. random classification), suggesting that this type of information is not helpful.

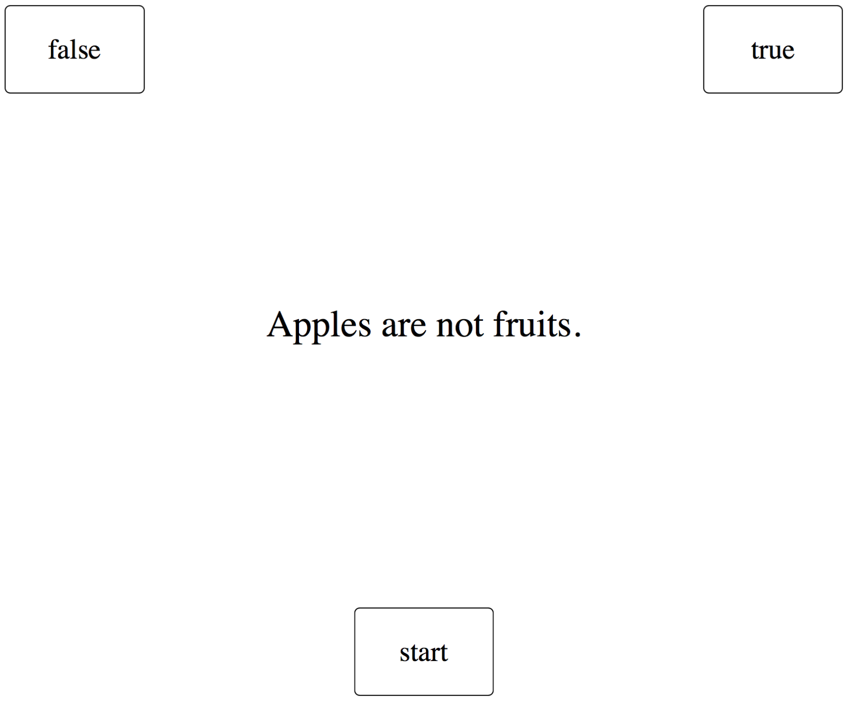
We have shown that (i) a rough manipulation of decision making processes has a direct impact on mouse trajectories; (ii) a linear discriminant analysis (LDA) using absolute-temporal information is enough to accurately distinguish these quasi-decisions; and (iii) this LDA does a better classification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

5 Extension to linguistic data

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

Dale and Duran (2011) found differences in the processing of true positive and negative sentences when people performed a simple truth-value judgment task. These results were interpreted as indicating that negation underlies an abrupt shift in cognitive dynamics (i.e. an unconscious change of decision). If this is indeed the case, we would expect mouse trajectories corresponding to the verification of negative sentences to pattern with *switched* trajectories from the validation experiment. This pattern of results would provide additional support to the hypothesis that, at least in *out-of-the-blue* contexts, processing negation does involve a ‘two-step’ derivation, where the positive argument is initially derived and negated only as a second step[4].

[4] Note that the data used to train the classifier correspond to *quasi*-decisions; namely, the training set is only an approximation



Apples are not fruits.

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

5.1 Experiment

Participants were asked to perform a truth-judgment task, where they had to decide whether a sentence (e.g. *Cars have wheels*) is true or false according to common knowledge. Each of the sentences could be either a true or a false statement in its negated or non-negated form. 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. 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. The experiment lasted approximately 10 minutes.

Design The experimental design consisted of two fully crossed factors: TRUTH VALUE (true, false) and SENTENCE 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.
	Negative	Cars have no wings.
False	Positive	Cars have wings.
	Negative	Cars have no wheels.

Table 4: Design

of what should happen during an unconscious change of decision, such as the one expected for negation processing. As a result, we expect some aspects of the decision processes on sentence-verification data not to be captured by the LDA.

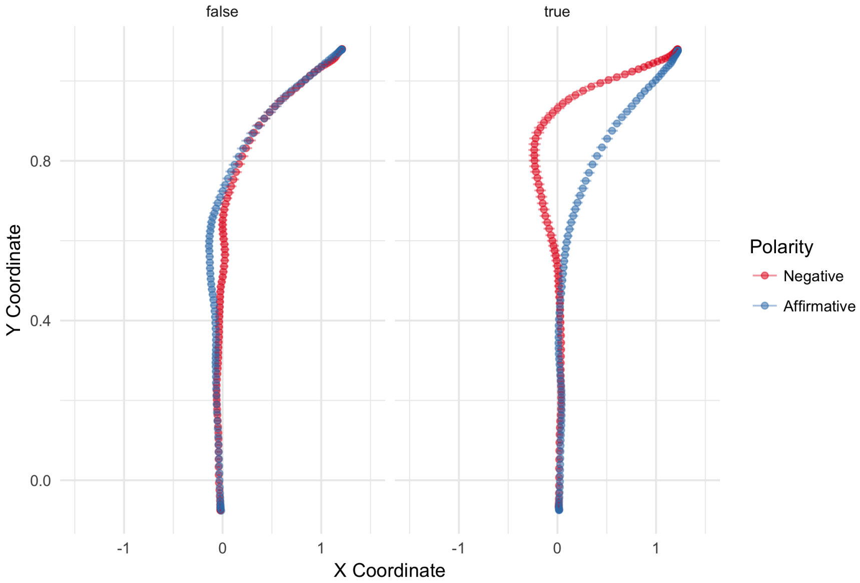


Figure 11: Mean trajectories for accurate trials

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.

5.2 Results and discussion

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 analyses. Figure 11 illustrates mean trajectories for the four possible trial conditions.

To assess whether we replicate Dale and Duran's results, we calculated the x-coordinate flips (see Section 4) and analysed them with a linear mixed-effects model [Baayen et al., 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 Dale and Duran.

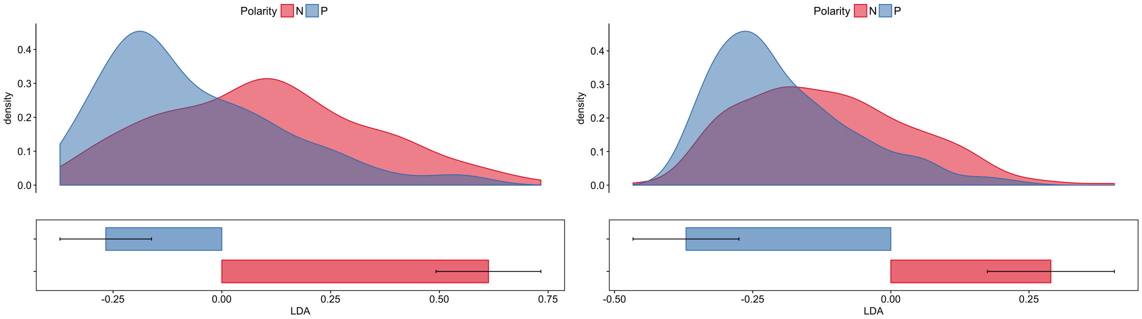
Unlike Dale and Duran, we did not perform statistical analyses based on the acceleration component (i.e. acceleration flips). In our validation experiment, this quantitative measure was unable to distinguish mouse trajectories underlying different "quasi-decisions". The origin of this inadequacy is hard to determinate: it could be a property of the kind of decisions were are manipulating, or just a consequence of noisy data. We reasoned that if the different decision processes involved in a rather simple task were not captured by the acceleration component, this measure might also be unable to classify more complex processes, such as the ones at play in a sentence verification task.

The model of x-coordinate flips revealed a main effect of Polarity, such that negation significantly increases flips in the x-coordinate by 0.76 ($\chi^2 = 10.11; p = .0014$), and a significant interaction Truth × Polarity ($\chi^2 = 22.7; p < .001$), such that the difference between negative and positive sentences is bigger for the true than for the false statements. There was no significant effect of Truth ($\chi^2 < 1; p = .5$). Table 5 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 (i.e. negation gives rise to more 'curvy'/rough? trajectories). The values obtained in the two experiments, however, are slightly different; namely, our results present higher range of values (see Table 5). Note that, in our experiments, the mouse-position was not sampled at a fixed rate (see Interfase and data treatment), creating additional noise which could be responsible for the range differencemaybe foonote?.

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



(a) LDA (coords, speed and acceleration)

(b) LDA only with coordinates

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

Our findings therefore pattern with a broader set of psycholinguistic studies, which, using different techniques, have revealed that the positive argument plays a role in negation processing: verifying negative sentences might involve computing the positive content at an early processing stage [Tian and Breheny, 2016].

Classifier performance Two different LDA classifiers, trained with data from the validation experiment, were applied to the new experimental data. The first classifier was our original LDA, which had as predictors x, y coordinates as well as distance-based velocity and acceleration. The second LDA had only x, y coordinates as predictors. Validation results (see Section 3) suggest that the simpler model, which only relies on absolute information, might be sufficient to classify the two basic kinds of decision-making processes. That is to say, the simple model might fit the data as well as a more complex model, and be interpreted more straightforwardly.

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

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

Could the observed performance be expected if negative and positive trials were actually not different from each other? Are these AUC values significantly different from the ones one would have obtained from applying the LDA to a set of data where there is no difference between experimental conditions (i.e. *null hypothesis*)? We calculated the AUC values for a set of data where experimental labels (positive, negative) were scrambled. The distribution of AUC values under the *null hypothesis* was compared to the performance

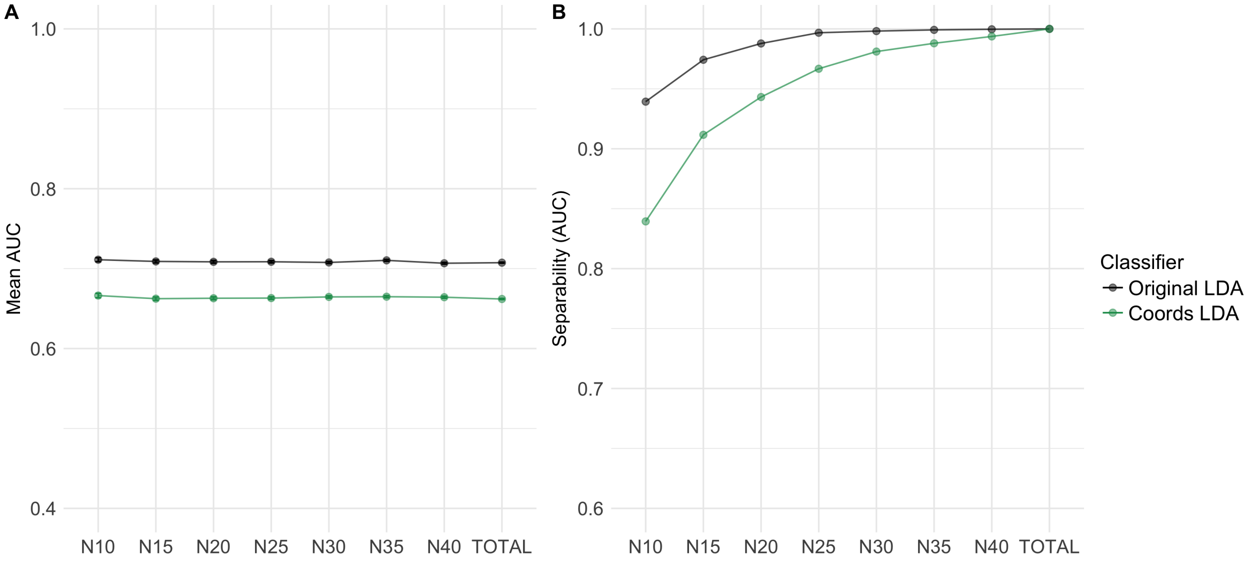


Figure 13: **Performance of LDA classifiers.** A. Mean AUC values over bootstrapped data (iterations=1000) for different sample sizes; B. Difference of classifier performance when applied to scrambled vs. original set of data.

observed for the original set of data. Figure 13B illustrates the separability of the two classifications for each sample size.

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

A last note should be made about the comparison between the two classifiers. While the classifiers' comparison in Figure 6 indicated that relative spatio-temporal features, such as distance-based acceleration and speed, were not essential for the classification of 'quasi-decisions', these features do seem to play a role in the classification of sentence-verification data. Indeed, Figure 13 reveals that the *full* classifier –which takes all features as predictors– makes a better distinction than the simplified one.

Other mouse-tracking measures In the validation experiment, the performance of the LDA classifier was shown to be significantly better than the one of other mouse-tracking measures. Does this difference remain when these measures are applied to the new experimental data (and hence to slightly different decision processes)? While it's true the LDA trained on validation data can make a distinction between negative and positive trials, it might not be the *best* possible strategy for classification.

We address the question of whether different measures differ on their ability to find the observed effect by applying the same procedure as before: we calculated the mean area under the ROC curve for different sample sizes (cf. Figure 14A), and contrasted these values against the null hypothesis (i.e. the values we would have obtained if there had been no difference between the experimental conditions; Figure 14B)[5].

The results in Figure 14A suggest that most measures perform a worse classification than the one observed

[5] Interestingly, the 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.

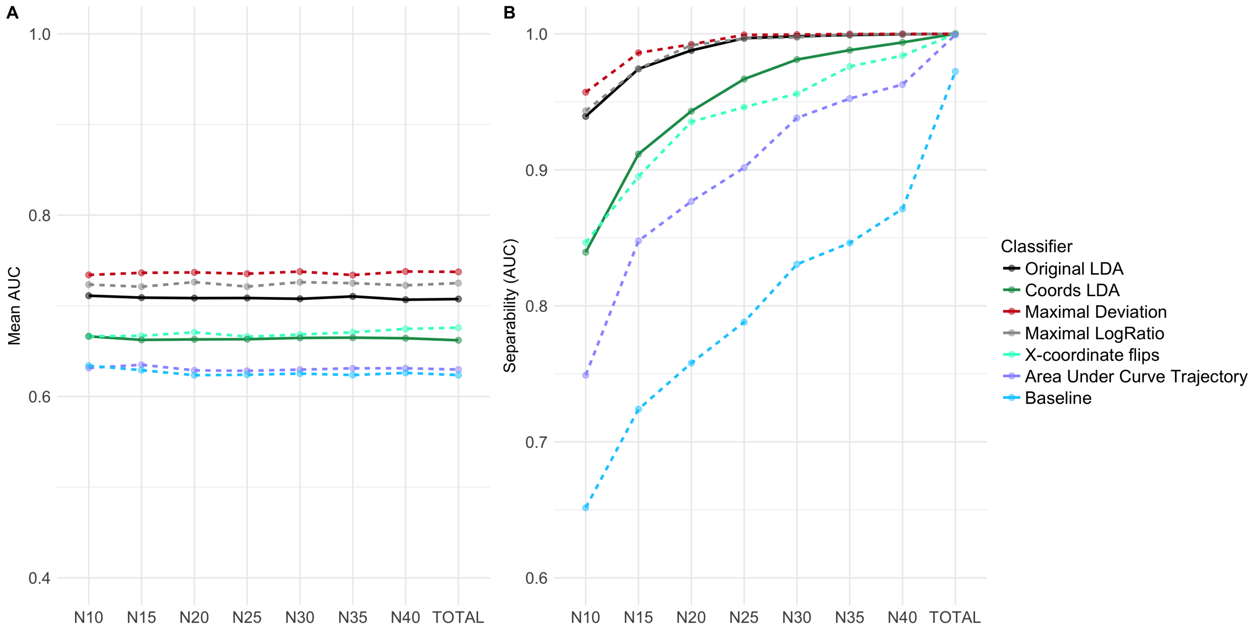


Figure 14: Performance of other measures

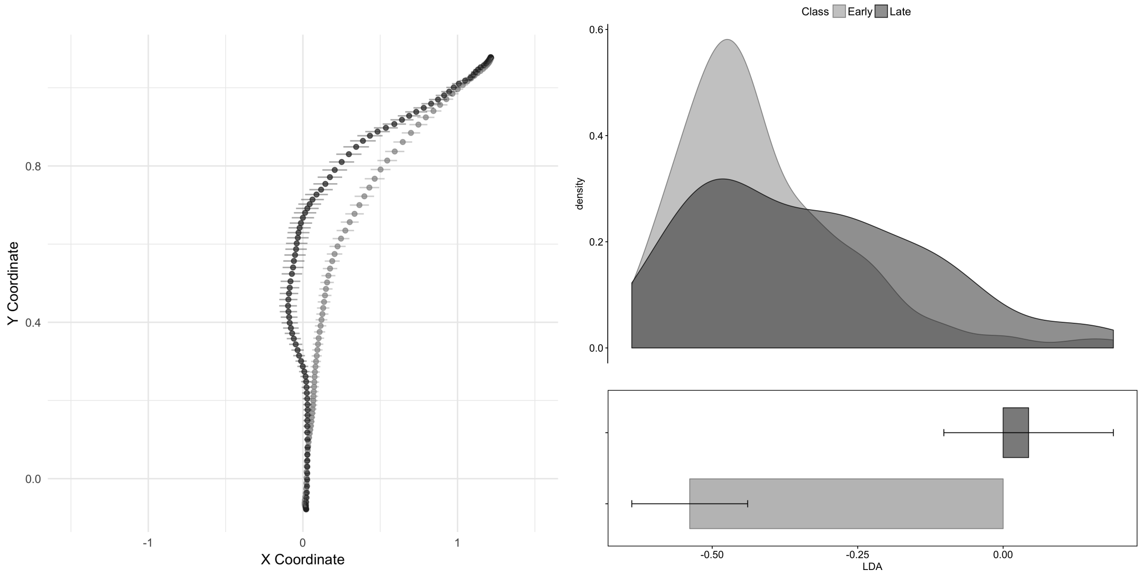
for the validation data (compare with Figure 9). Given that a decrease in performance is attested across the board and not only for the classifiers trained with validation data, this difference must be driven by properties of the new data set. Specifically, the sentence-verification data might be more variable, such that both negative and positive trials might underlie instances of different decision processes.

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

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

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

As illustrated by mean trajectories in Figure 15a, the two classes in the baseline data have slightly different trajectory shapes. The distribution of the LDA measure after testing the classifier on the new data set is shown in Figure 15b. The performance was evaluated following the same procedure applied above (see blue line in Figure 14).



(a) Mean Trajectories

(b) Mean and distribution of LDA measure

Figure 15: Baseline data set

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

6 Conclusion/General discussion

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

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

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

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

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

7 Supplementary Materials

References

R. H. Baayen, D. J. Davidson, and D. M. Bates. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, 59(4):390–412, 2008.

E. A. Cranford and J. Moss. Mouse-tracking evidence for parallel anticipatory option evaluation. *Cognitive processing*, pages 1–24, 2017.

R. Dale and N. D. Duran. The Cognitive Dynamics of Negated Sentence Verification. *Cognitive Science*, 35: 983–996, 2011. ISSN 03640213. doi: 10.1111/j.1551-6709.2010.01164.x.

T. a. Farmer, S. a. Cargill, N. C. Hindy, R. Dale, and M. J. Spivey. Tracking the continuity of language comprehension: computer mouse trajectories suggest parallel syntactic processing. *Cognitive science*, 31: 889–909, 2007. ISSN 0364-0213. doi: 10.1080/03640210701530797.

J. B. Freeman and N. Ambady. MouseTracker: software for studying real-time mental processing using a computer mouse-tracking method. *Behavior research methods*, 42(1):226–241, 2010. ISSN 1554-351X. doi: 10.3758/BRM.42.1.226.

J. B. Freeman and K. L. Johnson. More than meets the eye: split-second social perception. *Trends in cognitive sciences*, 20(5):362–374, 2016.

J. B. Freeman, R. Dale, and T. A. Farmer. Hand in motion reveals mind in motion. *Frontiers in Psychology*, 2(APR):1–6, 2011. ISSN 16641078. doi: 10.3389/fpsyg.2011.00059.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, New York, second edition edition, 2009.

E. Hehman, R. M. Stolier, and J. B. Freeman. Advanced mouse-tracking analytic techniques for enhancing psychological science. *Group Processes & Intergroup Relations*, pages 1–18, 2014. ISSN 1368-4302. doi: 10.1177/1368430214538325.

B. Kaup, R. H. Yaxley, C. J. Madden, R. A. Zwaan, and J. Lüdtke. Experiential simulations of negated text information. *The Quarterly Journal of Experimental Psychology*, 60(7):976–990, 2007.

J. Lüdtke, C. K. Friedrich, M. De Filippis, and B. Kaup. Event-related potential correlates of negation in a sentence–picture verification paradigm. *Journal of Cognitive Neuroscience*, 20(8):1355–1370, 2008.

M. S. Nieuwland and G. R. Kuperberg. When the truth is not too hard to handle: An event-related potential study on the pragmatics of negation. *Psychological Science*, 19(12):1213–1218, 2008.

U. Sauerland, A. Tamura, M. Koizumi, and J. M. Tomlinson. Tracking down disjunction. In *JSAI International Symposium on Artificial Intelligence*, pages 109–121. Springer, 2015.

J.-H. Song and K. Nakayama. Role of focal attention on latencies and trajectories of visually guided manual pointing. *Journal of Vision*, 6(9):11, 2006.

J. H. Song and K. Nakayama. Hidden cognitive states revealed in choice reaching tasks. *Trends in Cognitive Sciences*, 13(8):360–366, 2009. ISSN 13646613. doi: 10.1016/j.tics.2009.04.009.

M. J. Spivey and R. Dale. Continuous dynamics in real-time cognition. *Current Directions in Psychological Science*, 15(5):207–211, 2006.

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. ISSN 0027-8424. doi: 10.1073/pnas.0503903102.

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 https://doi.org/10.1007/978-3-319-17464-8_2.

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.

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.

P. C. Wason. The contexts of plausible denial. *Journal of verbal learning and verbal behavior*, 4(1):7–11, 1965.

P. C. Wason and P. N. Johnson-Laird. *Psychology of reasoning: Structure and content*, volume 86. Harvard University Press, 1972.

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.

K. Xiao and T. Yamauchi. Semantic priming revealed by mouse movement trajectories. *Consciousness and cognition*, 27:42–52, 2014.

K. Xiao and T. Yamauchi. The role of attention in subliminal semantic processing: A mouse tracking study. *PloS one*, 12(6):e0178740, 2017.