Mouse tracking as a window into decision making

Mora Maldonado, Ewan Dunbar & Emmanuel Chemla

March 16, 2018

Abstract

Mouse tracking promises to be an efficient method to investigate the dynamics of cognitive processes: it is easier to deploy than eye-tracking, and yet it is in principle much more fine-grained than looking at response times. We investigate its claimed benefits directly, asking how features of decision processes, and notably decision changes, may be captured in mouse movements. We ran two experiments, one in which we explicitly manipulate whether our stimuli trigger a flip in decision, and one in which we replicate more ecological, classical mouse tracking results on linguistic negation (Dale & Duran, 2011). We conclude, first, that spatial information (mouse path) is more important than temporal information (speed and acceleration) for detecting decision changes, and we offer a comparison of the sensitivity of various typical measures used in analyses of mouse tracking (area under the trajectory, direction-flips, and others). We do so using an 'optimal' analysis of our data (a linear discriminant analysis) explicitly trained to classify trajectories), and see what type of data (position, speed, acceleration) it capitalizes on. We quantify how its results compare with those based on more standard measures.

1 Introduction

In the past ten years, mouse tracking has become a popular method for studying the dynamics of cognitive processes in different domains, ranging from phonetic competition (Cranford & Moss, 2017; Spivey, Grosjean, & Knoblich, 2005), and syntactic, semantic and pragmatic processing (Dale & Duran, 2011; Farmer, Cargill, Hindy, Dale, & Spivey, 2007; Sauerland, Tamura, Koizumi, & Tomlinson, 2015; Tomlinson, Bailey, & Bott, 2013; Xiao & Yamauchi, 2014, 2017, among others), to social cognition (Freeman & Ambady, 2010; Freeman, Dale, & Farmer, 2011; Freeman & Johnson, 2016). While response times can reveal whether a decision process is fast or slow (Donders, 1969), and analyses of response time distributions can give insight into how the decision process unfolds (Ratcliff & McKoon, 2008; Usher & McClelland, 2001, among others), mouse movements promise a more direct window onto the dynamics of cognitive processes, under the assumption that motor responses are planned and executed in parallel to the decisions they reflect (Freeman & Ambady, 2010; Hehman, Stolier, & Freeman, 2014; Song & Nakayama, 2006, 2009; Spivey & Dale, 2006).

Concretely, if a response is entered by clicking on a button, one may measure the time needed to click on that button and use it as a reflection for the complexity of the decision, roughly. But depending on whether participants are decided from the start, hesitate, or undergo a radical change of decision, the path to that button may take different trajectories (see Figure 1, Wojnowicz et al., 2009). Accordingly, researchers have studied the shape and dynamics of mouse paths to document aspects of numerous types of decision processes. Dale and Duran's (2011) approach to negation processing is an example of this. Linguistic negation has been traditionally understood as an operator that reverses sentence truth conditions, inducing an extra "step," or "mental operation," in online processing (Wason, 1965; Wason & Johnson-Laird, 1972; see review in Tian & Breheny, 2016). Dale and Duran tracked mouse trajectories as participants performed a truth-value judgment task, where they had to verify the truth of general statements such as *Cars have (no) wings*. The authors found that mouse trajectories gave rise to more shifts towards the alternative response when evaluating negative than affirmative true sentences. This was interpreted as evidence for a "two-step" processing of negation, where truth conditions for the positive content are first derived and negated only as a second step.[1] To do so, one can extract several measures from the mouse paths (e.g., maximal deviation point, number of direction changes, etc.) and argue that the deviation of these measures from what they would be for an optimal, straight trajectory reflects the relevant decision change.

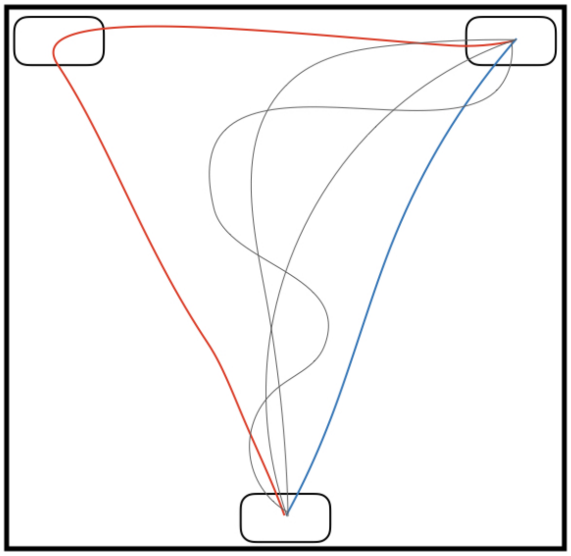


Figure 1: **Shape of trajectories underlying distinct decision processes.** One single cognitive process is expected to be mapped onto one smooth trajectory (blue line), whereas a change of mind would be reflected by a two-step path (red line). Intermediate cases are represented in gray.

Our goal is to explicitly document this method and the connection between cognition (decision making) and action (mouse trajectories): What in a decision process is reflected in mouse movements—decision changes, hesitations, or other properties?— and how —in changes in acceleration, changes in direction, or other aspects of the trajectory? We will tackle this question by asking what features of mouse trajectories distinguish *straightforward* decisions, based on a single initial commitment, and *switched* decisions, which involve a change of mind in the course of the process.

[1] Several studies have suggested that the positive argument plays an important role in negation processing (Kaup, Yaxley, Madden, Zwaan, & Lüdtke, 2007; Lüdtke, Friedrich, De Filippis, & Kaup, 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 & Kuperberg, 2008; Tian, Breheny, & Ferguson, 2010). How to explain this pattern of results has been at the center of the debate in the negation processing literature (see Tian & Breheny, 2016 for review). We will not explore this here.

First, we present a validation experiment where we directly manipulate whether the stimuli trigger a flip in what the appropriate response is in the course of a trial. We show that the mouse paths do indeed reflect these changes (Section 2). An analysis of this data using linear discriminant analysis (henceforth, LDA), confirms that the two types of decision, straightforward and switched, can be distinguished objectively (Section 3). We then compare the performance of the LDA classifier to other traditionally used mouse tracking measures (Section 4). Finally, the LDA classifier trained on the validation data is further tested with new, more "ecological" data, obtained from a replication of Dale and Duran's (2011) experiment on the processing of negation mentioned above (Section 5). If there is a change of decision triggered by negation, trajectories corresponding to negative trials should be classified together with trajectories underlying changes of decision in the validation experiment.

Data and code for all the analyses developed in this paper are provided at https://osf.io/rbx3m/?view_only=7d557aa8931c4a0886e7ce2442a77895.

2 Validation Experiment: presentation and qualitative analysis

Participants were asked to perform a two-alternative forced choice task. Each trial was triggered by clicking on a start button at the bottom of the screen. A frame surrounding the screen would then appear and the participants' task was to indicate whether the frame was blue or red by clicking on the appropriate "blue" or "red" buttons at the top left or top right of the screen, respectively. On most trials, the color of the frame remained stable throughout the trials, but in crucial cases it changed during the trial. In the first case, the initial choice was the correct response (*straightforward* trials. In the second case, participants were forced to change their answer (*switched* trials). The switched trials are meant to mimic natural decision changes. We take these to be a reasonable stand-in for changes of decision, even though there are obvious differences: in natural changes of decision, alternative responses are weighted as the pieces of information are integrated, whereas in our experiment the sensory information changes in time. We return to the question of how ecological these decisions are in Section 5. The procedure is illustrated in Figure 2.

2.1 Participants

We recruited 54 participants (F=27) 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 took approximately 5 minutes.

2.2 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. In *switched* trials, the color switched once (from red to blue or from blue to red) during the trial, forcing a revision of the initial choice. The change on *switched* trials was triggered by the cursor reaching a certain position on the y-axis, which could be at various relative heights (POINT OF CHANGE: early, at 40% of the screen, middle, at 70%, or late, at 90%). The design is schematized in Table 1.

To prevent participants from developing a strategy whereby they simply drag the cursor along the center line rather than moving the mouse toward their current choice of answer, the proportion of trials was adjusted so that there were a majority of 64 straightforward trials (32 repetitions per

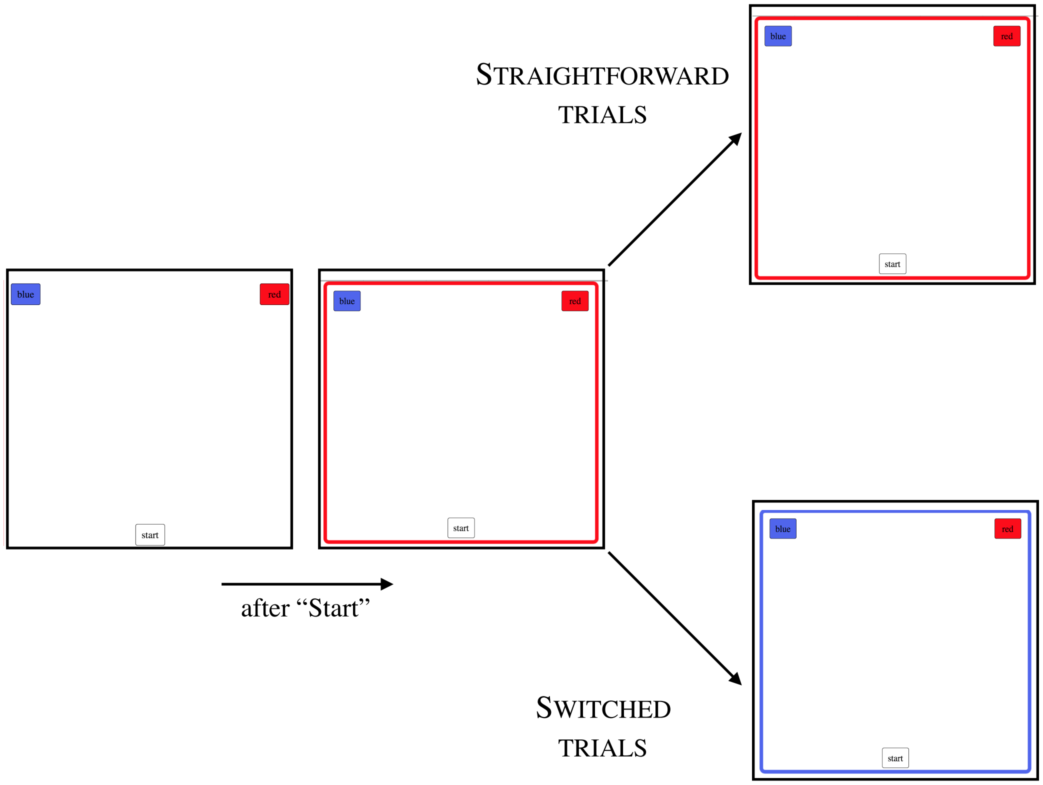


Figure 2: **Procedure in Validation Experiment.** Subjects were instructed to click the "start" button in order to see the colored frame. Response boxes were on the top left or top right. Depending on the trial condition, the frame color either did, or did not, change (once) during the trial.

Decision Pattern	Frame color	Point of change
Straightforward	Blue Red	*never*
Switched	Blue → Red Red → Blue	early (y=40%) middle (y=70%) late (y=90%)

Table 1: **Design in Validation Experiment**

frame color), while there was only 24 switched trials (4 repetitions per final frame color and change point).

2.3 Interface

The web interface was programmed using JavaScript. Mouse movements triggered the extraction of (x, y)-pixel coordinates (there was thus no constant sample rate). Three buttons were displayed during the experiment ("start" and response buttons). 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. On each trial, between start-clicks and response-clicks, mouse movements triggered the recording of the (x, y)-pixel coordinates of the cursor together with the time.

2.4 Data treatment

To allow comparisons between participants, the (x, y)-coordinates were normalized according to participants' window size: the center of the start button was mapped onto $(0, 0)$ point, the "blue" button onto $(-1, 1)$ and the "red" button onto $(1, 1)$. Variations both in response times and in the sensitivity and sampling rate of our participants' input devices imply that different trials have different numbers of (x, y) positions per trial, making comparisons difficult. We therefore normalized the time course into 101 proportional time steps by linear interpolation. That is, we reduced all time points to 101 equally distant time steps, including the first and the last positions.

2.5 Overall performance

Inaccurate responses (4% of the data) were removed from the analyses. Mean trajectories for each DECISION PATTERN and POINT OF CHANGE 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.

3 Validation Experiment: Classifying decision processes with LDA

Different decisions (that is, DECISION PATTERNS) have a different impact on mouse trajectories (Figure 3). To identify the features characteristic of each class (*switched* vs. *straightforward*), we use a linear discriminant analysis for classification.

3.1 Description of LDA classifier

The LDA is a supervised algorithm that finds a linear function of the predictors onto a single real number, such that zero represents the midpoint between the two classes to be learned, and the separation between the two classes is maximal. This linear combination of predictors can thus be used to form a decision rule to classify objects of one class (negative) or the other (positive).

The two classes here were the multi-dimensional data coming from *switched* and *straightforward* trials. The dimensions taken into account were: all the x,y coordinates, the Euclidean-distance based velocity and the Euclidean-distance based acceleration (both of which are non-linear with respect to the original (x, y) coordinates). The coordinates provide absolute spatiotemporal information about where the cursor was at what point, and velocity and acceleration provide information about how it arrived there. To avoid collinearity (which causes problems for LDA), we applied a principal component analysis (PCA) to identify 13 principal components for these predictors, and fitted and applied the LDA to 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. The procedure is schematized in Figure 4.

3.2 Performance of the LDA classifier

Figure 5 illustrates the result of applying the procedure in Figure 4 to the trajectories. To evaluate the overall performance of the classifier, we calculated the area under the ROC curve (AUC), a standard method for evaluating classifiers (Hastie, Tibshirani, & Friedman, 2009). Intuitively, the AUC gives the degree to which the histograms resulting from the classifier's continuous output (for example, Figure 5) are non-overlapping in the correct direction (in this case, *switched* more systematically in the positive direction on the classification axis than *straightforward*).

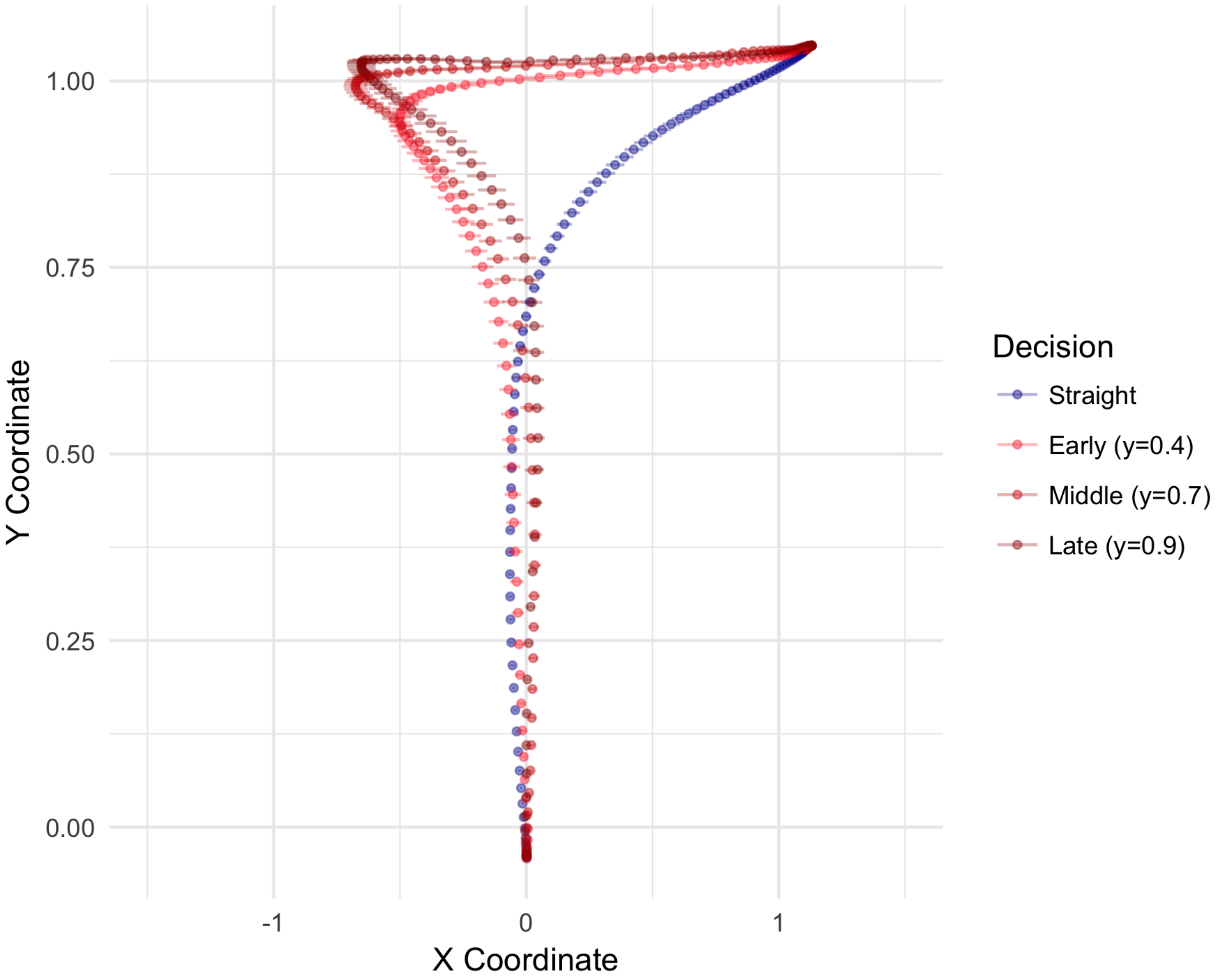


Figure 3: **Mean trajectories for different decision patterns in the validation experiment.** Error bars represent the standard error of x-coordinates.

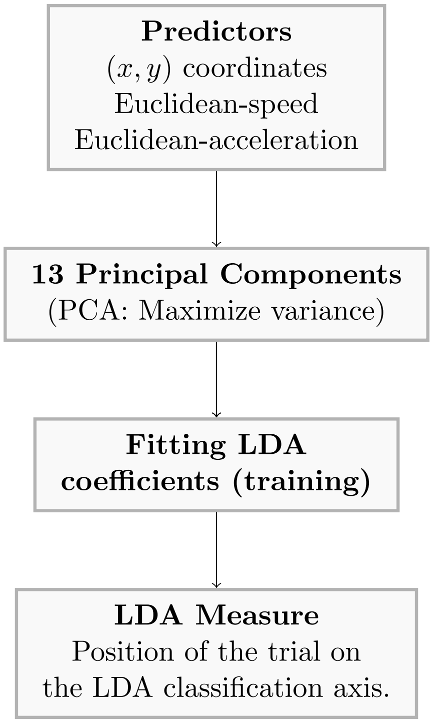


Figure 4: **Diagram of classification procedure.**

To properly evaluate the classifier's performance 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 remaining 90%) to train the classifier. The data contained in the bin were used as a test set to diagnose the classifier performance. We thus obtained one AUC score for each of the ten test bins. The performance of the LDA classifier was compared to a baseline, equivalent to the worst possible outcome, and a topline, which was what we would expect from a LDA under the best possible conditions. For the baseline, we used a random classifier that assigned labels by sampling from a beta distribution centered at the probability of straightforward trials; the topline was computed by testing and training the original LDA classifier on the same set of data. The mean AUC values for the LDA, the baseline and the topline in each bin are given in Figure 9a.

To assess whether the performance of the LDA classifier was statistically different from baseline (or topline) performance, we tested the groups of ten scores with regard to how likely it would be to obtain the attested differences in scores under the null hypothesis that the LDA classifier performance was the same as baseline (or topline) performance. The difference in the mean AUC between each of these two pairs of classifiers was calculated as a test statistic. 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-tailed test on the mean AUC differences. As expected, our original LDA is significantly better than a random classifier at categorizing trajectories into *straightforward* and *switched.* Conversely, there is no significant difference between the performance of our LDA and the topline; the classifier's performance is not significantly different from the best an LDA could possibly give on this data.

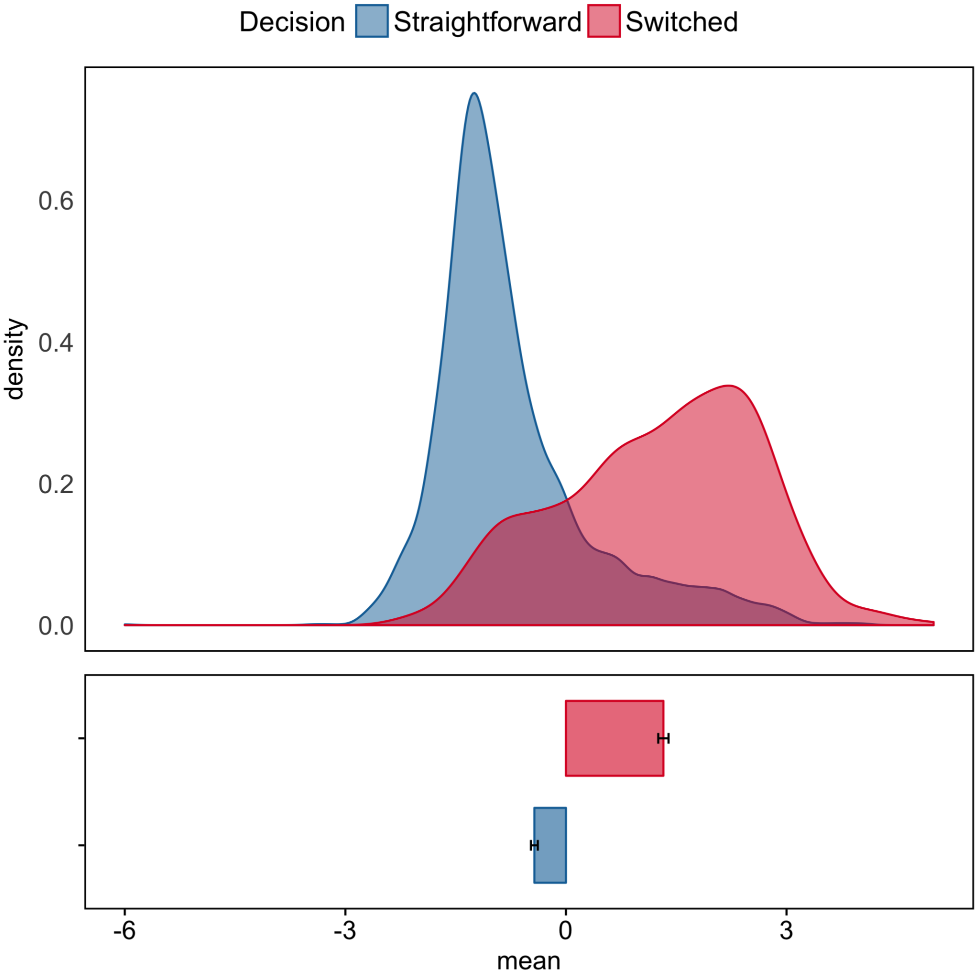


Figure 5: **Distribution and mean LDA-based measure for each class.** Classifier performance when applied to the whole validation data set. Error bars represent the mean standard error.

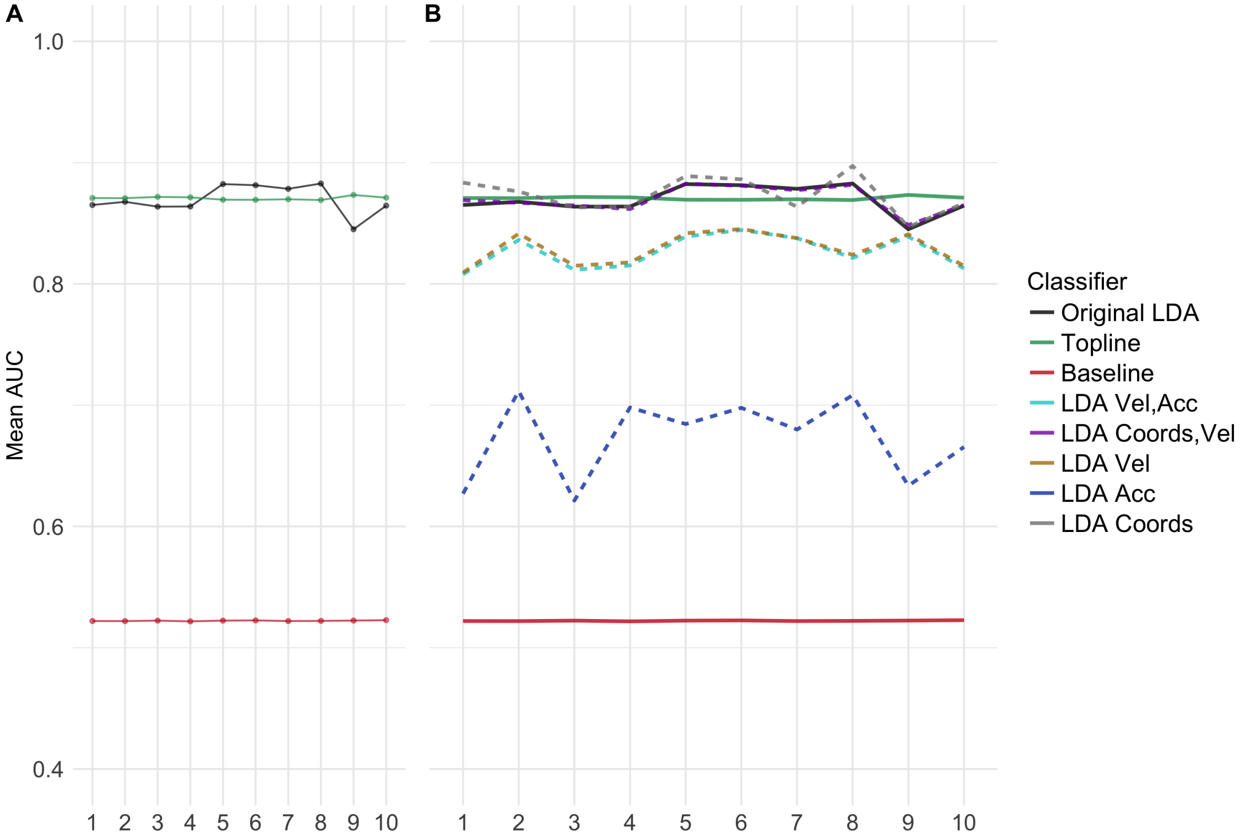


Figure 6: **Mean Area Under the ROC Curve values obtained from cross-validation.** A. Cross-validation on 10 bins for original LDA, baseline and topline. B. Comparison with values obtained for five additional classifiers obtained by subsetting the original set of predictors.

3.3 Meaningful features and optimal predictors

Our original LDA classifier takes as predictors both absolute and relative spatio-temporal features (coordinates, speed, and acceleration). 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 most relevant to decision processes.

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 9b illustrates the mean AUC values for each of these classifiers, together with the original LDA, the baseline and the topline. Pairwise comparisons with the original LDA were done by testing whether the observed mean differences would be expected under the null hypothesis of 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 other LDA classifiers that contain the coordinates among their predictors, suggesting that the distinction between *straightforward* and *switched* decisions might be solely explained by the information contained in the (x, y) coordinates. Con-

		(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) baseline and topline classifiers and (b) LDA classifiers with different predictors.

versely, 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 the mouse pointer was at a given time than to know how it got there.

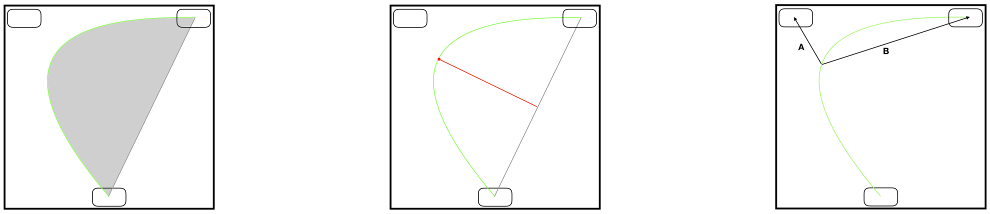
We caution that effects of *true* decisions, rather than the simulated decisions tested here, may indeed have an impact on speed and acceleration. It has been suggested that speed and acceleration components can capture the level of commitment towards the response, such that a change of decision (*swiched*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

4 Validation Experiment: LDA *versus* traditional mouse tracking analyses

The LDA classifier derives a solution to the problem of separating two kinds of mouse trajectories that is in a certain sense optimal. Previous studies have used alternative techniques to analyze mouse trajectories. In what follows, we compare the performance of our LDA to other measures commonly used in mouse tracking studies. We focus on measures that assess the spatial disorder in trajectories, 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 & Ambady, 2010). The AUT is the geometric area between the observed trajectory and an idealized straight-line trajectory drawn from the start to the end points, whereas the MD is the point that maximizes the perpendicular distance between this ideal trajectory and the observed path (Figure 7). For both measures, higher values are associated with higher trajectory deviation towards the alternative; values close to or below zero suggest a trajectory close to ideal. Another frequently used measure counts the number of times a trajectory crosses the $x-$axis (horizontal flips, Dale & Duran, 2011, as illustrated in Figure 7).

While all these measures aim to evaluate the degree of complexity of the path, they may fail to distinguish paths straight to the correct answer from "two step" (deviation to the alternative)



(a) Area Under the Trajectory

(b) Maximal Deviation

(c) Maximal Log Ratio

(d) X-coordinate 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: **Description of commonly used mouse tracking measures.**

and from "uncertain" (centered on the middle of the screen) trajectories.[2] 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 one suggest a position near the middle, higher values indicate a deviation towards the alternative response.

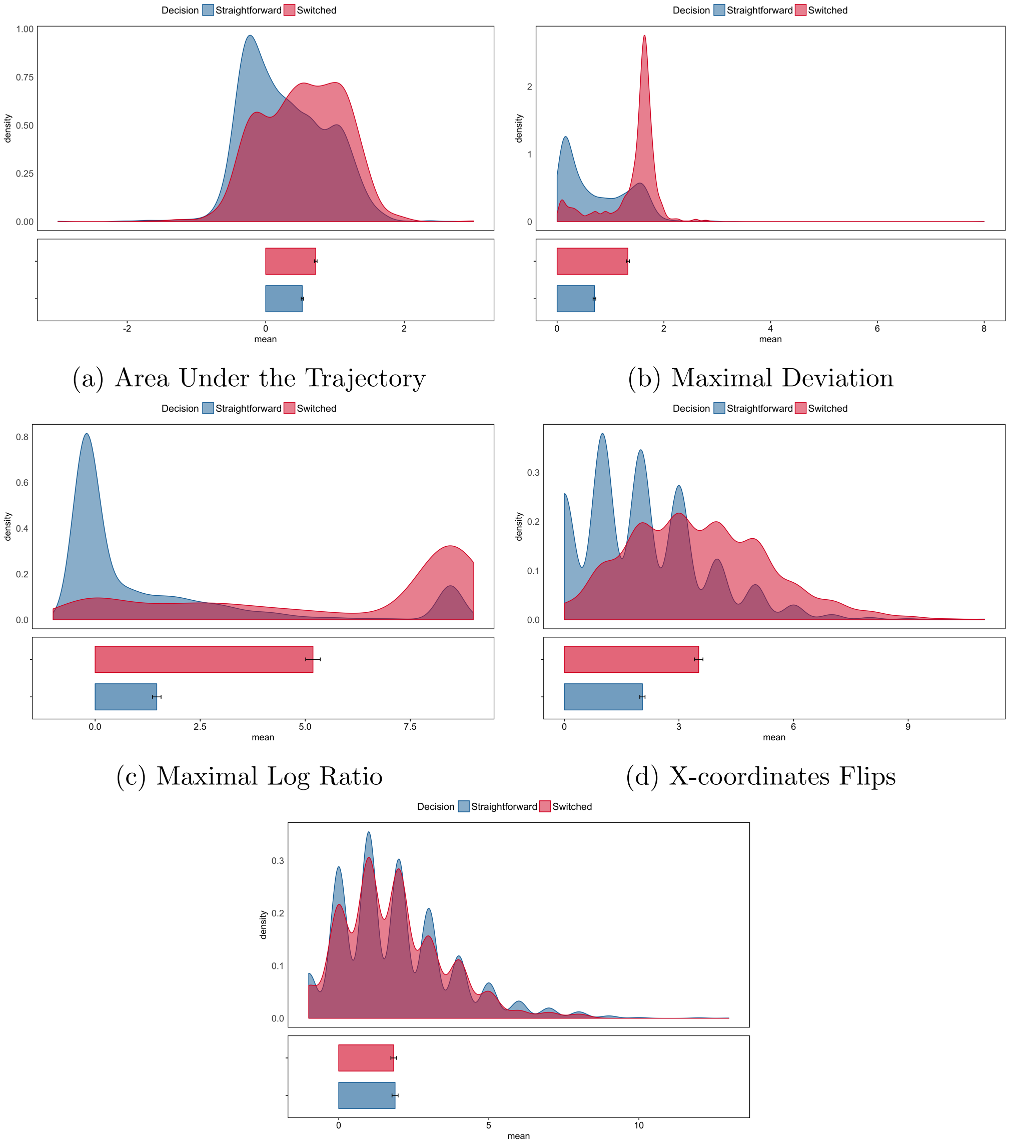
AUT, MD, x-coordinate flips, and the point that maximizes the log distance ratio (henceforth Maximal Log Ratio) were calculated for the validation data. Following Dale and Duran (2011, and other studies on error corrections), we also analyzed the *acceleration component* (AC) as a function of the number 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). Figure 8 illustrates the distribution and mean values for each decision pattern.

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

Overall, these comparisons reveal that the LDA trained on the validation data is significantly better at classifying this type of decisions than other commonly used measures. The difference with the classifier is in all the cases significant. Mean AUC values suggest that MD and the Maximal Log Ratio 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, the MD and the Maximal Log Ratio are not only sensitive to whether or not there was a deviation from the ideal trajectory (as the other measures), but weight this deviation as a function of the moment at which it occurred, assigning higher values to late deviations. This information seems to be essential for the classification, as observed in Section 3.

[2] A late medium-size deviation towards the alternative might underlie a "two-step" decision, whereas an early, but large, deviation towards the alternative might very well be considered noise. Measures such as the AUT might not be able to make a distinction between these.

[3] Note that these measures do not need training; we simply applied the measure to the same ten test subsets as before to make the results comparable.



(a) Area Under the Trajectory

(b) Maximal Deviation

(c) Maximal Log Ratio

(d) X-coordinates Flips

(e) Acceleration Component

Figure 8: **Distribution and means obtained from applying different mouse tracking measures to validation data.** Error bars represent the mean standard error.

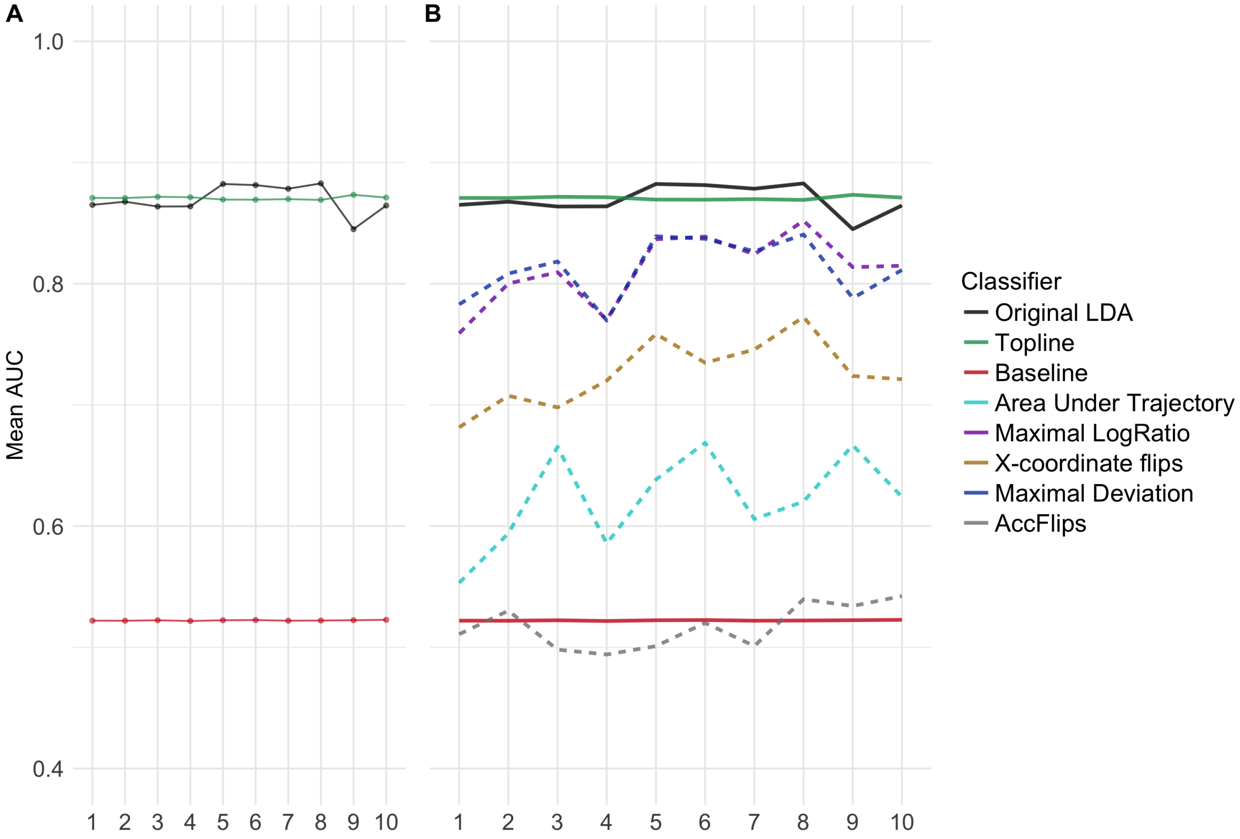


Figure 9: **Mean Area Under the ROC Curve values obtained from cross-validation.** A. Cross-validation on 10 bins for original LDA, baseline and topline. B. Comparison with values obtained for other commonly used mouse tracking measures.

	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 LDA was compared to each of five commonly used measures in mouse tracking studies.

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

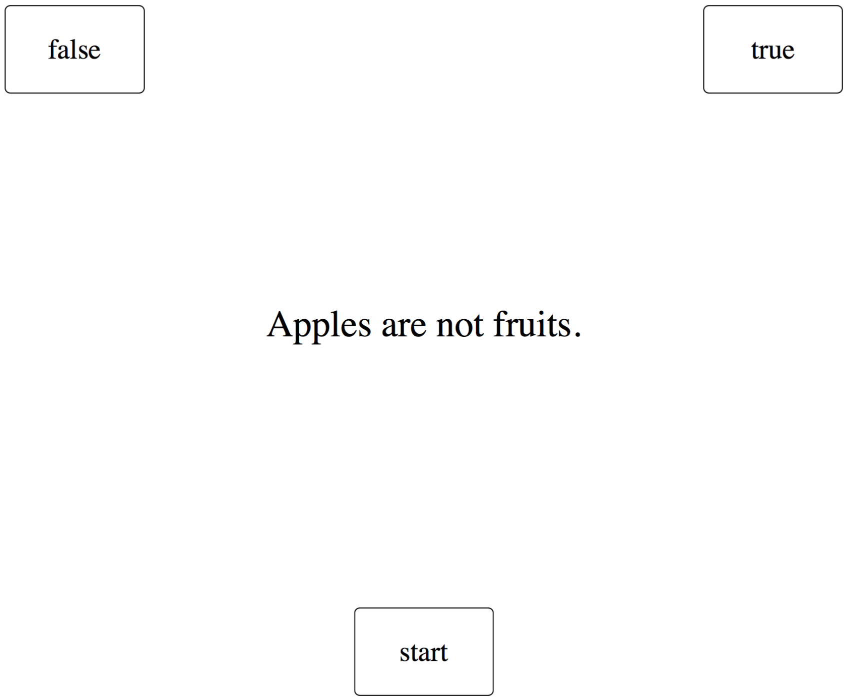
5 Extension to linguistic data

So far, we have shown that (i) a rough manipulation of decision making processes has a direct impact on mouse trajectories; (ii) an LDA using absolute temporal information is enough to accurately distinguish these decision patterns; and (iii) this LDA does a better classification than other traditional mouse tracking measures. Can our LDA help characterize more complex decision processes, such as the ones involved in sentence verification tasks?

To address this question, we test our classifier on data obtained from a replication of Dale and Duran's experiment (2011). This experiment found differences in the processing of true positive and negative sentences when people performed a truth-value judgment task. These results were interpreted as indicating that negation gives rise to an abrupt shift in cognitive dynamics (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 two steps, in which the positive value is initially derived and negated only as a second step. On the other hand, if negation does not involve a change in decision—or if subjects' behavior in the validation experiment is simply too different from natural changes of decision—then the LDA measure trained on validation data will not reveal systematic differences between positive and negative sentences.

5.1 Experiment

Participants had to perform a truth-value judgment task in which they had to decide whether a sentence (for example, *Cars have wheels*) is true or false, based on common world knowledge. Each sentence could either be a negated form or a non-negated form, and could either be a true or a false statement. Unlike Dale and Duran's experiment, the complete statement was presented in the middle of the screen after participants pressed "start" (that is, no self-paced reading). The response buttons appeared at the top left or top right corners of the screen, as in our validation experiment. Materials and design are exemplified in Table 4.



Apples are not fruits.

Figure 10: **Illustration of a trial in the replication of Dale & Duran.**

5.1.1 Participants

53 English native speakers (F=29) were tested using Amazon Mechanical Turk. They were compensated for their participation (1 USD). The experiment lasted approximately 10 minutes.

5.1.2 Design

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

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

Table 4: **Design of Dale and Duran's replication.**

5.1.3 Interface and data treatment

The interface and data treatment were the same as those used in the validation experiment. The time course of mouse trajectories was normalized into 101 time steps.

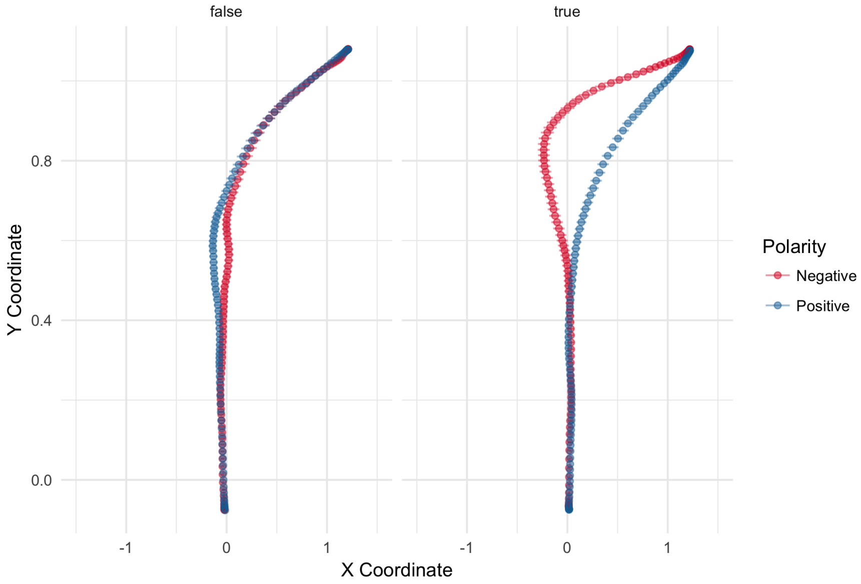


Figure 11: Mean trajectories for accurate trials

5.2 Results and discussion

5.2.1 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 analyzed. Figure 11 illustrates mean trajectories for the four conditions.

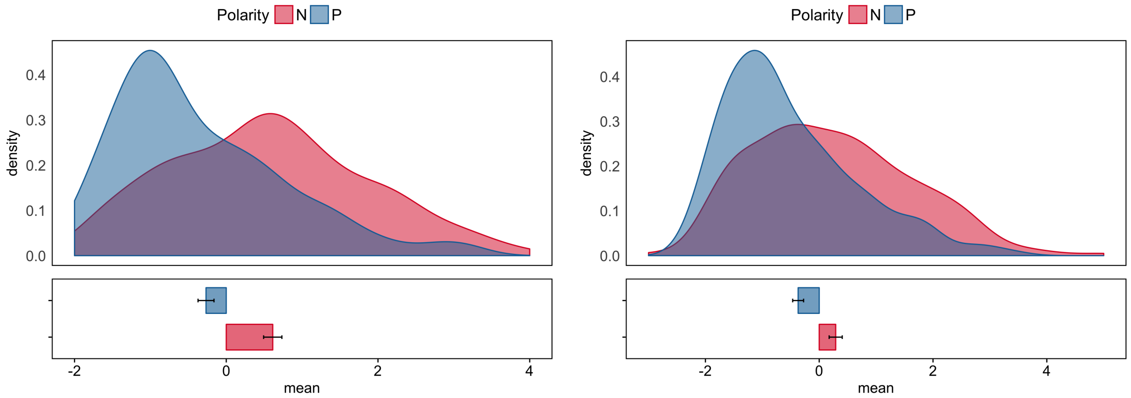
To assess whether we replicate Dale and Duran's results, we calculated the x-coordinate flips (see Section 4) and analyzed them with a linear mixed-effects model, taking Truth, Polarity and their interaction as predictors. We included random intercepts per subject and a random slope with the interaction of both factors. P-values were obtained by comparing the omnibus model to a reduced model where the relevant factor was removed. This is the analysis done by Dale and Duran. Unlike Dale and Duran, we did not perform statistical analyses based on the acceleration component, since this quantitative measure was unable to distinguish mouse trajectories underlying different decision patterns in the validation experiment.

The model for x-coordinate flips revealed a main effect of Polarity, such that negation increased the number of flips by an estimated of 0.76 ($\chi^2 = 21.7; p < .001$), 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 summarizes our and of Dale and Duran's results.

We seem to replicate Dale and Duran's findings: verifying true negated sentences produces less straightforward trajectories than true positive sentences. The values obtained in the two experiments are slightly different; our results present a higher range of values (see Table 5). In our experiments, the mouse position was not sampled at a fixed rate, creating additional noise which could be responsible for the range difference.

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 for Dale & Duran original experiment and our replication.**



(a) LDA (coordinates, speed, and acceleration)

(b) LDA only with coordinates

Figure 12: **Two LDA classifiers applied to *true* trials (negative vs. affirmative).** Error bars represent the mean standard error.

5.2.2 Classifier performance

How well does our LDA classify new trajectories underlain by cognitive processes that might, or might not, involve different decision patterns across conditions? Two different LDA classifiers, trained with the 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 fits the data just as well as a more complex model, and can be interpreted more straightforwardly.

The relevant difference in processing between positive and negative sentences is expected to arise specifically for *true* statements. Consequently, we analyze 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 1000 new samples of various different sample sizes from the data from the replication experiment and calculated the area under the ROC curve for the classification of each one. Figure 13A shows the mean AUC values obtained after applying the classification procedure across these various samples

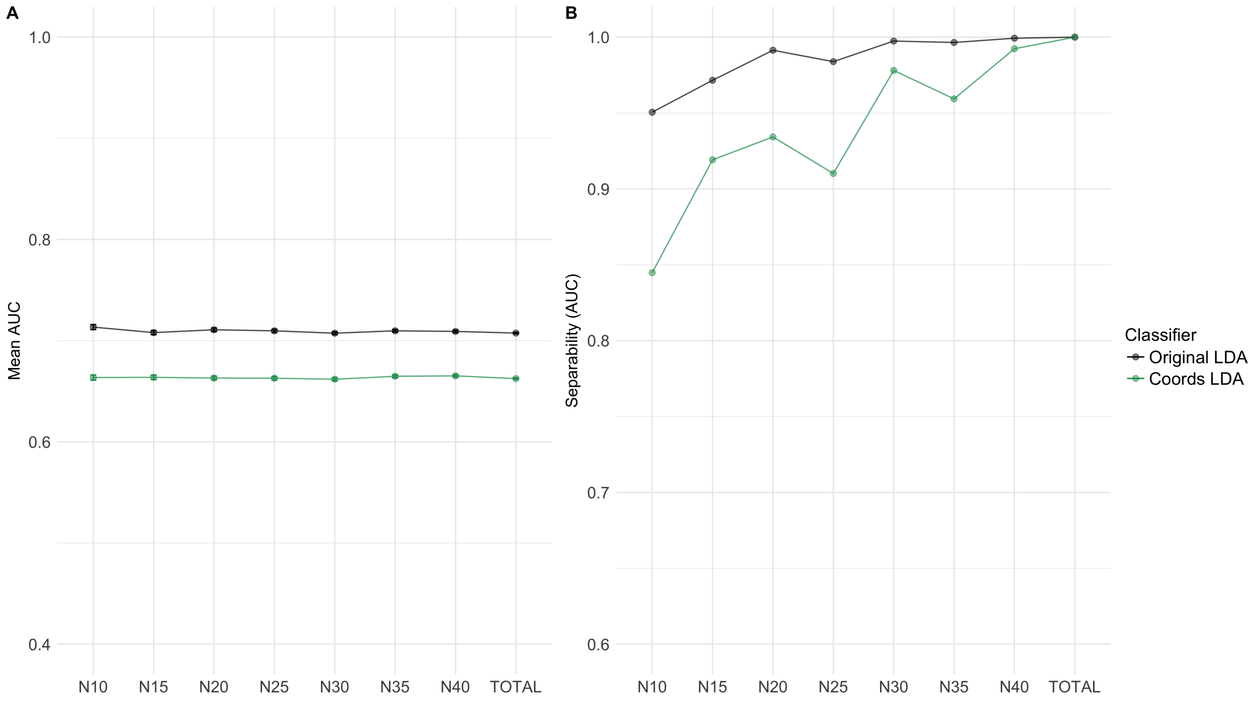


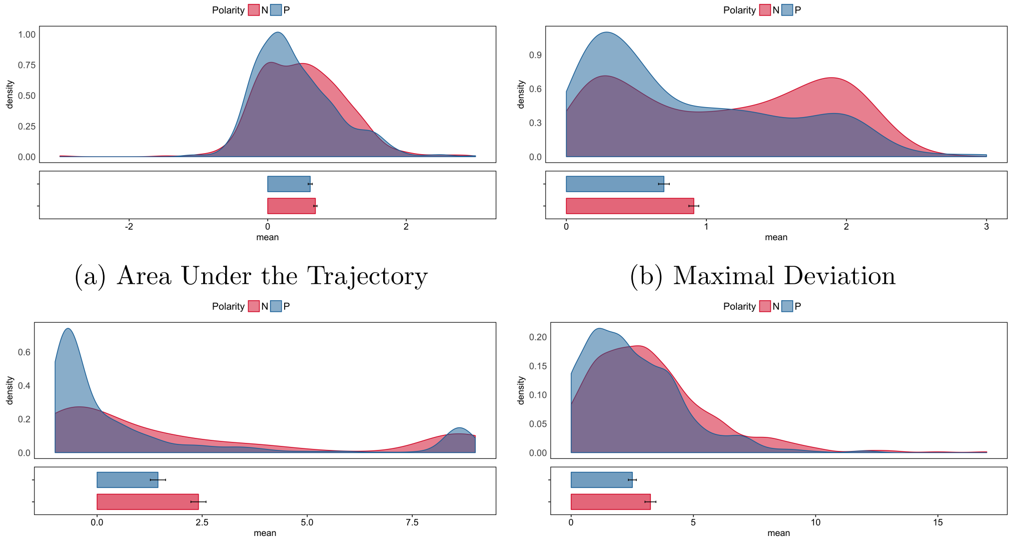
Figure 13: **Performance of LDA classifiers.** A. Mean AUC values over bootstrapped data (iterations=1000) for different sample sizes; B. Difference of classifier performance when applied to scrambled vs. original set of data.

of different sizes. The values are generally lower that the ones obtained in the validation experiment. This could be due to the fact that the tasks were different; or it could simply reveal idiosyncrasies of the original validation experiment data, or of this replication experiment.

Might the observed performance be expected, even if negative and positive trials were actually not systematically different? 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? We calculated the AUC values for a set of data where experimental labels (positive, negative) were scrambled (once per sample). The distribution of AUC values under this null hypothesis was compared to the performance 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 validation data seems to make a distinction between experimental conditions. This finding suggests that the contrast between negative and positive trials is similar to the contrast in the validation experiment. The fact that negation has similar properties to *switched* decisions indicates that verifying negative sentences might give rise to a change of decision, as proposed by Dale & Duran (2011), among others. However, while mouse trajectories corresponding to negative and to *switched* trials do share basic properties, they seem to differ on how they are placed on the "change of decision" spectrum: they occupy different parts of the decision-based LDA continuum (compare Figure 5 and Figure 12). This is not surprising, given that we are dealing with different cognitive processes—simulated decisions versus sentence verification—but, as discussed above, could easily also be an idiosyncrasy of these two data sets.

Finally, while the classifiers' comparison in Figure 9 indicated that *relative* spatio-temporal features, such as acceleration and speed, were not essential for the classification of simple decisions, these features do seem to play a role in the classification of sentence verification data. Indeed,



(a) Area Under the Trajectory

(b) Maximal Deviation

(c) Maximal Log Ratio

(d) X-coordinate flips

Figure 14: **Distribution and means of negative and positive *true* trials obtained from applying different mouse tracking measures to negation data.** Error bars represent the mean standard error.

Figure 13 reveals that the *full* classifier—which takes all features as predictors— gives a better separation between the two experimental conditions than the simplified one.

5.3 Other mouse tracking measures

Does the difference in performance between the LDA and other mouse tracking measures remain when these are applied to the new experimental data? Figure 14 illustrates the distribution of each measure. The question of whether different measures differ in their ability to separate the experimental conditions was addressed by applying the same procedure as before: we calculated the mean area under the ROC curve for different sample sizes (see Figure 15A), and contrasted these values against a null hypothesis of no difference between experimental conditions (see Figure 15B).

The results in Figure 15A suggest that most measures perform less well here than on the validation data (cf. Figure 9). Since 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. The sentence verification data might be more variable, such that both negative and positive trials may underlie instances of different decision processes.

The LDA classifier seems here to be roughly as powerful as other traditional mouse tracking measures, such as the Maximal Deviation and the Maximal Log Ratio. In contrast with the validation results, this opens the possibility of using these alternative measures to analyze mouse tracking data from sentence verification tasks. The classifier is still a better choice from a conceptual point of view, as it does not make any specific assumptions about how the change of decision should be reflected by mouse trajectories beyond the observed.

5.3.1 Baseline

A linear classifier trained on simulated decisions can separate the two experimental conditions of the replication of a previous study by Dale & Duran's. We have interpreted this result as suggesting that

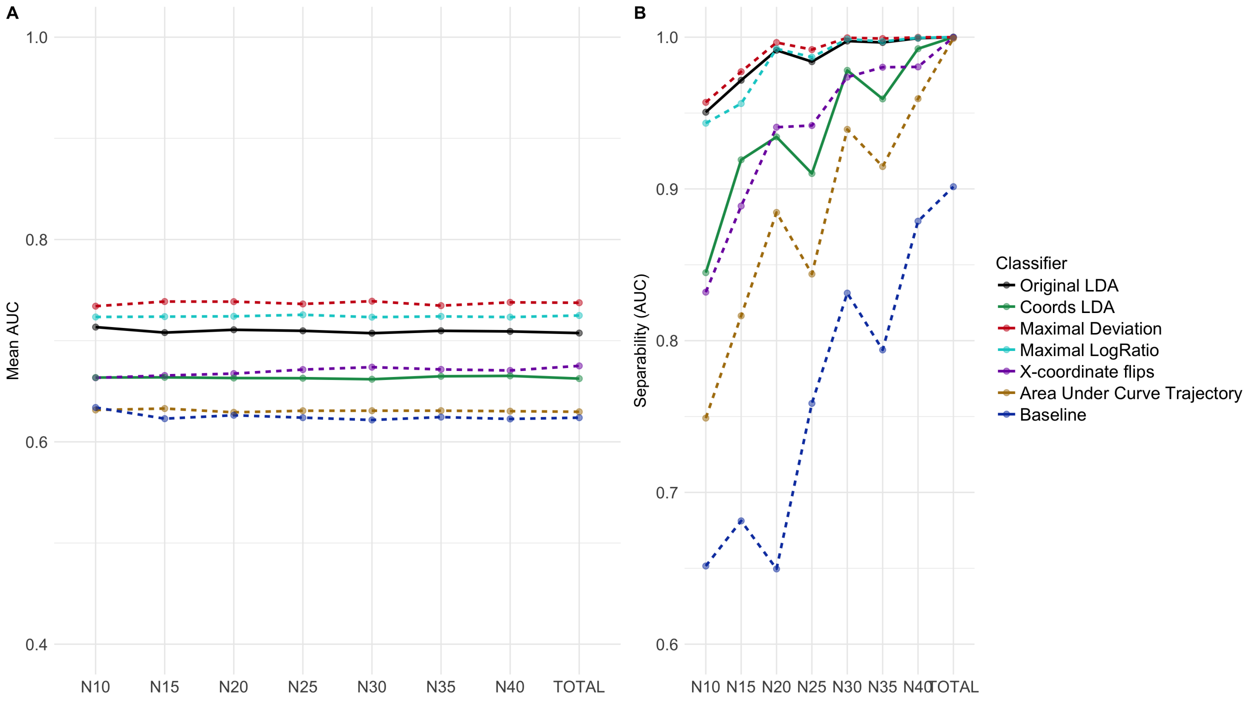


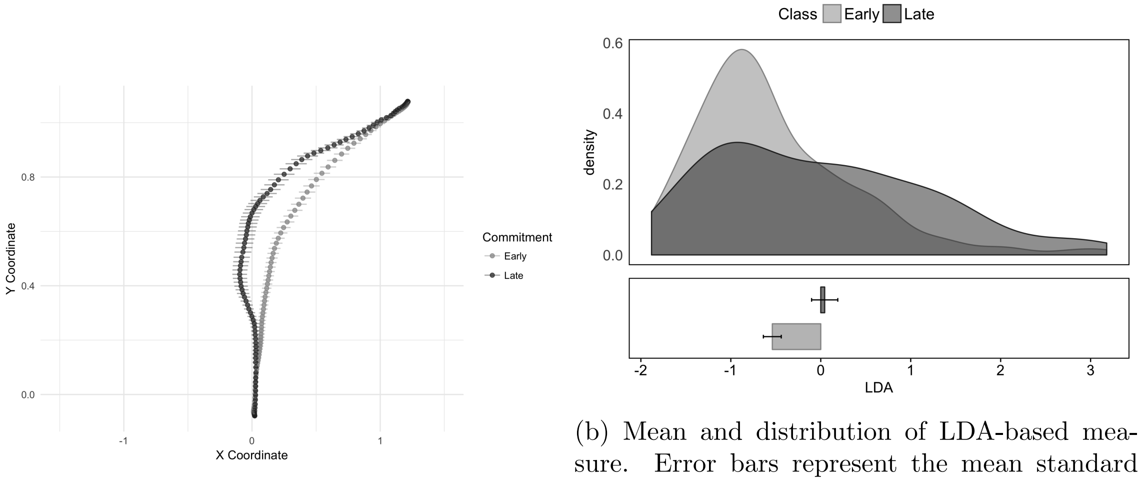
Figure 15: **Performance of other mouse tracking measures.** A. Mean AUC values over bootstrapped data (iterations=1000) for different sample sizes; B. Difference of measure performance when applied to scrambled vs. original set of data.

the key features being extracted reflect two different decision processes. It could instead be argued that the classification is not based on properties related to decision processes, but on some other feature of mouse paths which happen to be partially shared between conditions in both experiments. For example, the LDA might be sensitive not to decision shift but to differences in cognitive cost, something both experiments might have in common.

To disentangle these possibilities, we asked how the classifier trained on simulated decisions classifies trajectories that have different shapes but ought not to be related to differing decision processes. We constructed a set of baseline data which contained only positive trials from the replication of the experiment by Dale and Duran. The trials were classified as to whether their response time was above or below the subject mean. We reasoned that shorter response times would correspond to early commitment towards the response, whereas longer response times would reflect a late commitment. As illustrated by in Figure 16a, the two classes in the baseline data have slightly different trajectory shapes. Importantly, however, nothing about this split implies that these shapes correspond to a change of decision. Thus, the classifier trained on *straightforward* versus *switched* trials was expected to perform poorly.

The distribution of the LDA measure after testing the classifier on the new data set is shown in Figure 16b. The performance was evaluated following the same procedure applied above (see blue line in Figure 15).

The classification on *early* versus *late* categories is less accurate than the one performed 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 have look similar to *switched* and *negation* trials are not taken to be underlying a change of decision. Thus, despite the differences between the experimental conditions in the validation experiment and in the replication



(a) Mean Trajectories

(b) Mean and distribution of LDA-based measure. Error bars represent the mean standard error.

Figure 16: **Analyses performed on Baseline data set (early vs. late decision).**

experiment, the similarities appear to be more than accidental.

6 Conclusion

We investigated the correspondence between some types of decision processes and mouse movements. By manipulating whether a stimulus triggered, or did not trigger, a change of decision, we showed directly, for the first time, how mouse trajectories are impacted by decision changes: a forced switch in decision has an impact on mouse movements, which is for the most part observable in the spatial information (the path), and not so much in the timing of the trajectory.

We trained a classifier on the mouse trajectories underlying these simulated decisions to predict whether or not a given trial involved a decision shift. The classifier is freely available online. It accurately classifies not only paths corresponding to quasi-decisions, but also paths underlying a more complex process: the verification of negative sentences. Our results replicate previous findings but, more importantly, the LDA classifier performs at least as well as the best of the other commonly used mouse tracking measures. It also has the unique advantage of not relying on assumptions about what change-of-decision trajectories should look like. We also established the Maximal Deviation and Maximal LogRatio measures as comparable alternatives to the LDA analysis.

References

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

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

Donders, F. C. (1969). On the speed of mental processes. *Acta psychologica*, *30*, 412–431.

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

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

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

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

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference and prediction* (Second Edition ed.). New York: Springer.

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

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

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

Nieuwland, M. S., & Kuperberg, G. R. (2008). 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.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural computation*, *20*(4), 873–922.

Sauerland, U., Tamura, A., Koizumi, M., & Tomlinson, J. M. (2015). Tracking down disjunction. In *Jsai international symposium on artificial intelligence* (pp. 109–121).

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

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

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

Spivey, M. J., Grosjean, M., & Knoblich, G. (2005). Continuous attraction toward phonological competitors. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(29), 10393–10398. doi: 10.1073/pnas.0503903102

Tian, Y., & Breheny, R. (2016). Dynamic pragmatic view of negation processing. In P. Larrivée & C. Lee (Eds.), *Negation and polarity: Experimental perspectives* (pp. 21–43). Cham: Springer International Publishing. Retrieved from https://doi.org/10.1007/978-3-319-17464-8_2 doi: 10.1007/978-3-319-17464-8_2

Tian, Y., Breheny, R., & Ferguson, H. J. (2010). Why we simulate negated information: A dynamic pragmatic account. *The Quarterly Journal of Experimental Psychology*, *63*(12), 2305–2312.

Tomlinson, J. M., Bailey, T. M., & Bott, L. (2013). Possibly all of that and then some: Scalar implicatures are understood in two steps. *Journal of memory and language*, *69*(1), 18–35.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: the leaky, competing accumulator model. *Psychological review*, *108*(3), 550.

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

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

Wojnowicz, M., Ferguson, M. J., Spivey, M., Wojnowicz, M. T., Ferguson, M. J., Dale, R., & Spivey, M. J. (2009). The Self-Organization of Explicit Attitudes. , *20*(July 2017), 1428–1435. doi: 10.1111/j.1467-9280.2009.02448.x

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

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