Behavior Research Methods (2019) 51:1085–1101
https://doi.org/10.3758/s13428-018-01194-x

Mouse tracking as a window into decision making

Mora Maldonado[1] · Ewan Dunbar[2] · Emmanuel Chemla[3]

Published online: 12 February 2019

Abstract

Mouse tracking promises to be an efficient method to investigate the dynamics of cognitive processes: It is easier to deploy than eyetracking, yet in principle it is much more fine-grained than looking at response times. We investigated these claimed benefits directly, asking how the features of decision processes—notably, decision changes—might be captured in mouse movements. We ran two experiments, one in which we explicitly manipulated whether our stimuli triggered a flip in decision, and one in which we replicated more ecological, classical mouse-tracking results on linguistic negation (Dale & Duran, Cognitive Science, 35, 983–996, 2011). We concluded, 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 sensitivities of various typical measures used in analyses of mouse tracking (area under the trajectory curve, direction flips, etc.). 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, or acceleration) it capitalizes on. We also quantify how its results compare with those based on more standard measures.

Keywords Mouse tracking · Decision making · Negation processing · LDA · Sentence verification

Introduction

In the past 10 years, mouse tracking has become a popular method for studying the dynamics of cognitive processes in different domains, ranging from general decision making (Koop, 2013; Koop & Johnson, 2013; McKinstry, Dale, & Spivey, 2008) and social cognition (Freeman & Ambady, 2010; Freeman, Dale, & Farmer, 2011; Freeman & Johnson, 2016; Freeman, Pauker, & Sanchez, 2016) to 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, 2017; Tomlinson, Bailey, & Bott, 2013; and Xiao & Yamauchi, 2014, 2017, among others).

Although 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 with the decisions they reflect (Freeman & Ambady, 2010; Hehman, Stolier, & Freeman, 2014; Resulaj, Kiani, Wolpert, & Shadlen, 2009; Song & Nakayama, 2006, 2009; Spivey & Dale, 2006; Spivey, Dale, Knoblich, & Grosjean, 2010; Wojnowicz et al., 2009). 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 (Fig. 1, see Wojnowicz et al., 2009).

Accordingly, researchers have studied the shape and dynamics of mouse paths to document aspects of numerous types of decision processes (see a review in Freeman, 2018). 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;

✉ Mora Maldonado
Mora.Maldonado@ed.ac.uk

[1] Center for Language Evolution, PPLS, University of Edinburgh, Dugald Stewart Building, Edinburgh EH8 9AD, United Kingdom

[2] Laboratoire de Linguistique Formelle, Université Paris Diderot, Sorbonne Paris Cité, CNRS, Paris, France

[3] Laboratoire de Sciences Cognitives et Psycholinguistique, PSL Research University, CNRS, EHESS École Normale Supérieure, Paris, France

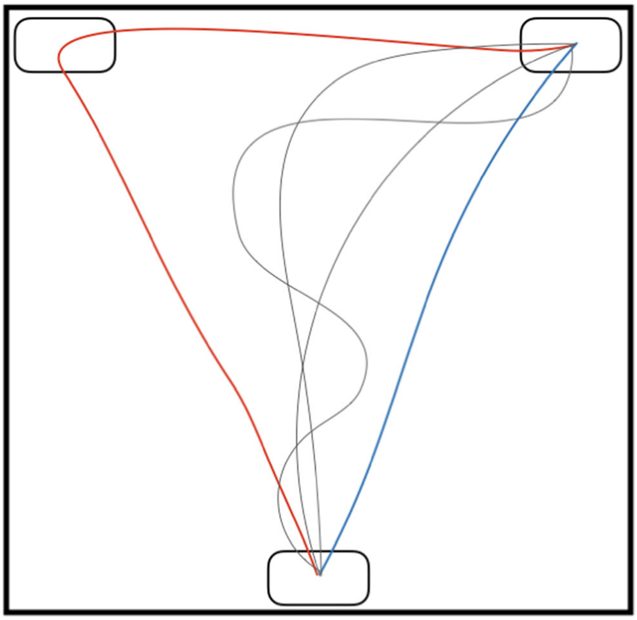


Fig. 1 Shape of the trajectories underlying distinct decision processes. One single cognitive process is expected to be mapped onto one smooth trajectory (blue line), whereas a change of mind would be reflected by a two-step path (red line). Intermediate cases are represented in gray

see the review in Tian & Breheny, 2016). Dale and Duran tracked mouse trajectories as participants performed a truth-value judgment task, in which 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 toward the alternative response when evaluating negative than affirmative true sentences out of context. This was interpreted as evidence for a "two-step" processing of negation, in which the truth conditions for the positive content are initially derived and are negated only later, as a second step (see Kaup, Yaxley, Madden, Zwaan, & Lüdtke, 2007, for details on the two-step simulation account).[1]

More generally, one can extract several measures from the mouse paths (e.g., maximal deviation point, number of direction changes, etc.) and argue, for instance, that the deviation of these measures from what they would be for an optimal, straight trajectory reflects the relevant decision change. This raises a fundamental question: what exactly does it mean for a mouse path to be deviated? That is the main topic of this article.

Our goal is to explicitly document the mouse-tracking 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 is it reflected—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.[2]

The distinction made here between *switched* and *straightforward* decisions should not be taken to rely on any specific account of decision making (e.g., serial vs. parallel). Originally, the idea of a "deviated" or "switched" mouse path came from an intuition about serial processing: The mouse path goes to one alternative, then switches to the other, because a hard decision is taken, and then a hard change of mind is made. Indeed, this is the kind of claim traditionally made about how linguistic negation is processed (Clark & Chase, 1972). There are, however, many other ways to interpret deviated mouse paths, both in general and in the specific case of negation. Decisions could, for instance, involve parallel competition between alternatives, with different degree of commitment toward each option (Spivey et al., 2010). In the specific case of negation, the deviation in mouse trajectories might be driven by task effects that might not tell us anything about negation processing per se (see note 1 and Orenes et al., 2014; Tian & Breheny, 2016).

In the present study, we just aim to identify how these two ends of the decision spectrum are reflected in mouse trajectories, without discussing the underlying mechanisms in the decision-making process. Although we will not examine the many interpretations for our experiments, we do think some of these bigger issues can be addressed with a methodology similar to the one we use here. We will come back to this in the discussion.

This article is organized as follows: First, we present a validation experiment in which we directly manipulate whether the

[1] In line with Dale and Duran (2011), several studies have shown that, at an early processing stage, negation is often ignored, and the positive argument of a negative sentence (e.g., "the door is open" for *The door is not open*) is represented (Hasson & Glucksberg, 2006; Kaup et al., 2007; Lüdtke, Friedrich, De Filippis, & Kaup, 2008, among others). This pattern of results, however, depends on a number of factors, including the amount of contextual support given for the sentence and the availability and type of alternatives at play. Indeed, the "two-step" processing of negation 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 (Dale & Duran, 2011; Nieuwland & Kuperberg, 2008; Tian, Breheny, & Ferguson, 2010). Similarly, the positive argument seems *only* to play a role for negation processing when the positive alternative is fully available (e.g., binary predicates). When there is more than one alternative or the alternative(s) are not available, negative sentences are processed straightforwardly (Orenes, Beltrán, & Santamaría, 2014). How to interpret these different processing patterns has been a center of debate in the negation literature (see Tian & Breheny, 2016, for a review), where the "two-step" strategy is often considered to be rather marginal. In the present article, however, we will precisely focus on cases that are predicted to trigger a "two-step" derivation, without discussing further examples.

[2] In a recent book, Wulff et al. (2019) propose a clustering analysis of mouse-tracking data, in order to detect different types of trajectories within one experimental condition. As pointed out by a reviewer, this approach is related to ours. The Mousetrap package developed by Kieslich and Henninger (2017) also provides a method to perform a classification of trajectories (based on the distance to prototypical trajectories). This approach differs conceptually from the approach we take here: Although it is possible to use the LDA measure we train to do classification, (zero represents an optimal threshold), we are interested primarily in extracting a continuous measure of the degree of deviation in mouse paths.

stimuli trigger a flip in what the appropriate response is in the course of a trial (see Study 3 in Farmer et al., 2007, for a similar validation experiment). We show that the mouse paths do indeed reflect these changes (Validation experiment: Presentation and qualitative analysis). An analysis of these data using linear discriminant analysis (henceforth, LDA) confirms that the two types of decision, straightforward and switched, can be distinguished objectively (Validation experiment: Classifying decision processes with LDA). We then compare the performance of the LDA classifier trained on the results of the validation experiment to traditionally used mouse tracking measures (Validation experiment: LDA versus traditional mouse tracking analyses). 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 (Extension to linguistic data). If a change of decision is triggered by negation, the trajectories corresponding to negative trials should be classified together with the trajectories underlying changes of decision in the validation experiment.[3]

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 pieces of information are integrated, whereas in our experiment the sensory information changed in time. We will return to the question of how ecological these decisions are in "Extension to linguistic data" section. The procedure is illustrated in Fig. 2.

Participants

We recruited 54 participants (27 female, 27 male) using Amazon Mechanical Turk. Two participants 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 min.

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 dragged 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 a majority of the trials (64) were straightforward (32 repetitions per frame color), as compared to only 24 switched trials (four repetitions per final frame color and change point).

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 the start-clicks and response-clicks, mouse movements triggered the recording of the (x, y) pixel coordinates of the cursor together with the time.

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 the (0, 0) point, the "blue" button onto (− 1, 1), and the "red" button onto (1, 1). Variations in response times and in the sensitivity and sampling rate of our participants' input devices implied that different trials would 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.

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

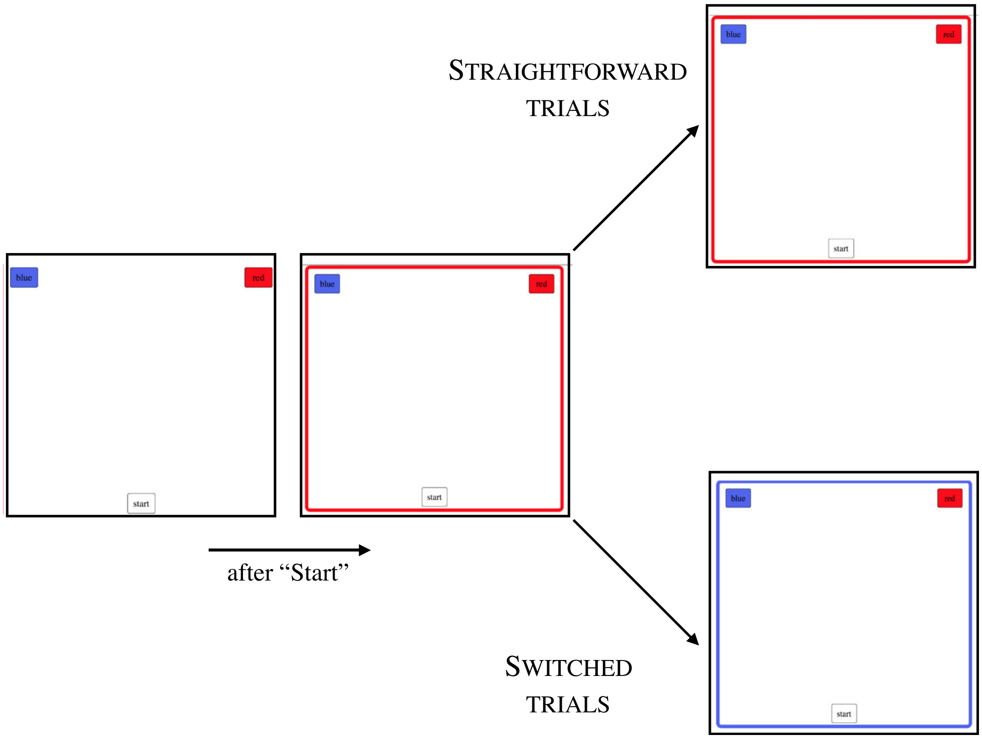


Fig. 2 Procedure of the validation experiment. Participants were instructed to click the "start" button in order to see the colored frame. Response boxes were on the top left and top right. Depending on the trial condition, the frame color either did or did not change (once) during the trial

Overall performance

Inaccurate responses (4% of the data) were removed from the analyses. The mean trajectories for each decision pattern and point of change are illustrated in Fig. 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 the mouse trajectories.

Validation experiment: Classifying decision processes with LDA

Different decisions (i.e., decision patterns) have a different impacts on mouse trajectories (Fig. 3). To identify the features characteristic of each class (switched vs. straightforward), we used a linear discriminant analysis for classification.

Table 1 Design in validation experiment

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%)

Description of the LDA classifier

The LDA is a supervised algorithm that finds a linear function projecting the predictors onto a line, giving a single real number, where zero represents the midpoint between the two classes to be learned and the separation between the two classes on this dimension 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 multidimensional data coming from *switched* and *straightforward* trials. The dimensions taken into account were all x, y coordinates, the Euclidean-distance-based velocity, and the Euclidean-distance-based acceleration (both of which are nonlinear 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 we 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. Negative LDA values correspond to trajectories than can be classified as *straightforward*, whereas positive values are associated with *switched* trials. The procedure is schematized in Fig. 4.

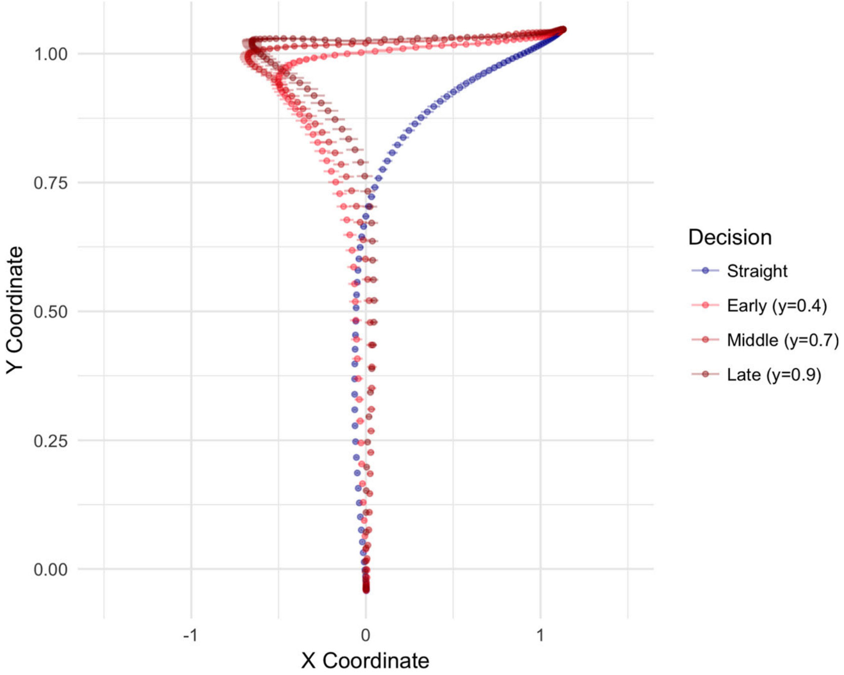


Fig. 3 Mean trajectories for different decision patterns in the validation experiment. Error bars represent the standard errors of the *x*-coordinates

Performance of the LDA classifier

Figure 5 illustrates the result of applying the procedure in Fig. 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 (e.g., Fig. 5) are nonoverlapping in the correct direction (in this case, *switched* systematically in a more positive direction on the classification axis than *straightforward*).

To properly evaluate the classifier's performance at separating trials following the distribution in the experiments, the AUC measure was cross-validated. That is, the validation data were partitioned into ten bins that kept the proportions 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 an 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 data set. The mean AUC values for the LDA, the baseline, and the topline in each bin are given in Fig. 6a.

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 the 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 a one-tailed test on the mean AUC differences. As expected, our original LDA was significantly better than a random classifier at categorizing trajectories into *straightforward* and *switched*. Conversely, there was no significant difference between the performance of our LDA and the topline; the classifier's performance was not significantly different from the best an LDA could possibly give on these data.

Meaningful features and optimal predictors

Our original LDA classifier took as predictors both absolute and relative spatiotemporal 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 gathered information about which features of mouse trajectories are most relevant to decision processes.

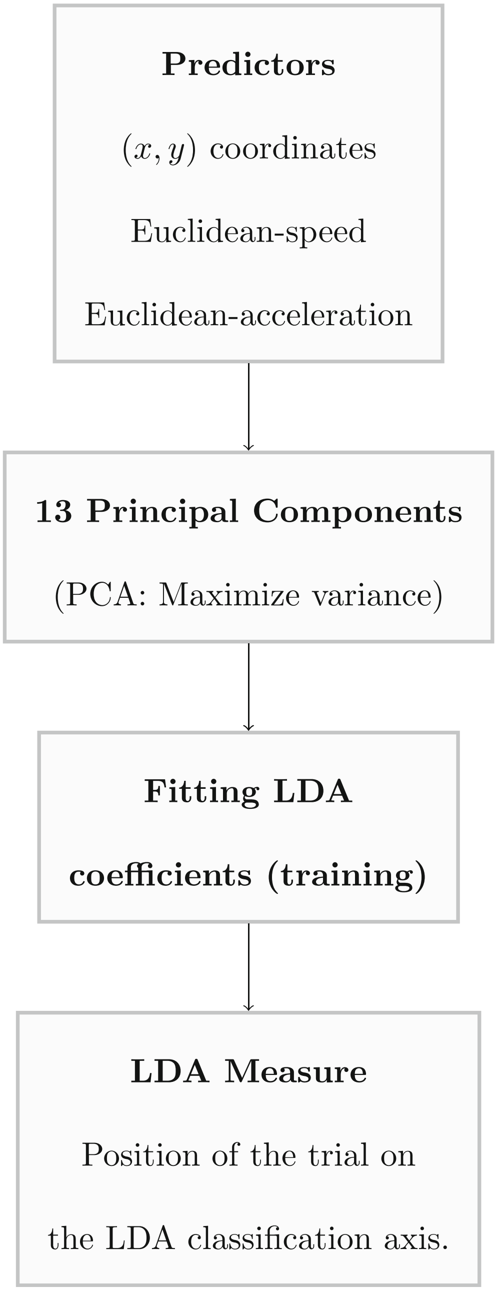


Fig. 4 Diagram of classification procedure

We trained five additional LDA classifiers obtained by subsetting the three original LDA predictors. If both absolute and relative features are required for predicting 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 ten test bins. Figure 6b 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 summarizes the comparisons between each of these classifiers and our original LDA.

The original LDA did not differ significantly from other LDA classifiers that contained the coordinates among their predictors, suggesting that the distinction between *straightforward* and *switched* decisions might be explained solely by the information contained in the (x, y) coordinates. Conversely, the original LDA was significantly better than classifiers that used only speed and acceleration as predictors. These comparisons therefore reveal that, for classifying our validation data, absolute spatiotemporal features (x,y) coordinates were 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 the effects of *true* decisions, rather than the simulated decisions tested here, might indeed have an impact on speed and acceleration. It has been suggested that speed and acceleration components can capture the level of commitment to the response, such that a change of decision (*switched* trajectories) might have associated with it a specific speed/acceleration pattern (Hehman et al., 2014). This is not visible, however, in our data.

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 curve* 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 (Fig. 7). For both measures, higher values are associated with higher trajectory deviation toward 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 Fig. 7).

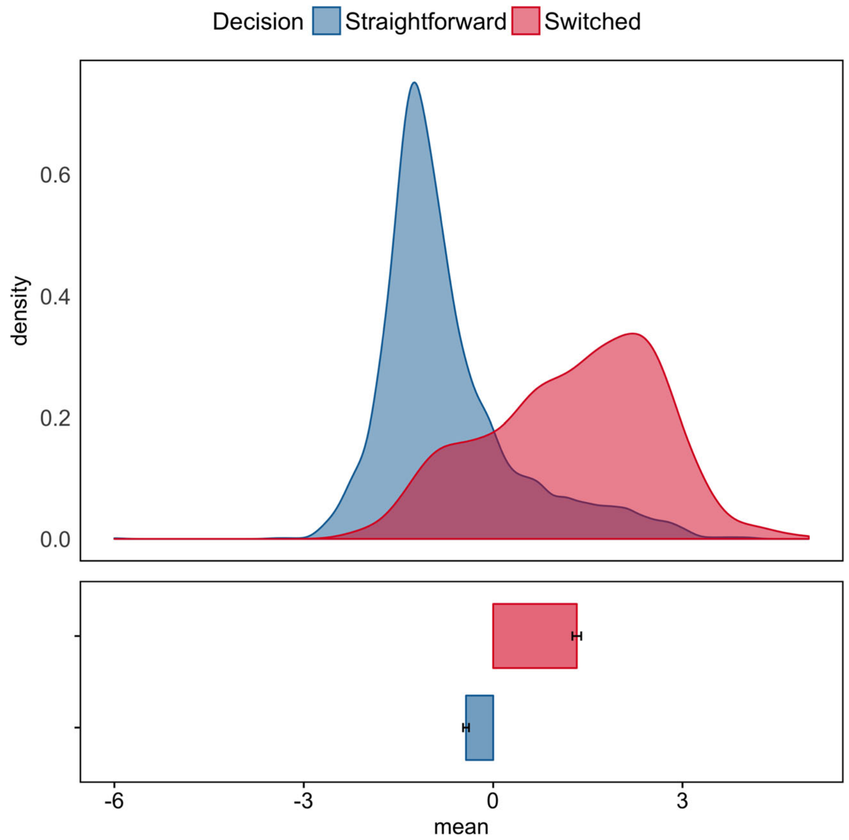


Fig. 5 Distribution and the mean LDA-based measure for each class: Classifier performance when applied to the whole validation data set. Error bars represent standard errors of the mean

Although 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) or "uncertain" (centered on the middle of the screen) trajectories.[4] To assess more directly whether mouse trajectories have a meaningful deviation toward the alternative, the distance to both the 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. Whereas ratio values closer to 1 suggest a position near the middle, higher values indicate a deviation toward the alternative response.

AUT, MD, x-coordinate flips, and the point that maximizes the log distance ratio (henceforth, the 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.[5] The mean AUC values for each of these measures are illustrated in Fig. 9. Table 3 summarizes the results 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 decision than other commonly used measures. The difference for the classifier in all cases was significant. The mean AUC values suggest that MD and the maximal log ratio are better at distinguishing decision processes than are the other alternative measures. These two measures are the only ones calculated on the basis of coordinates, and therefore give more importance to spatiotemporal information than the others do. 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 are the other measures), but weight this deviation as a function of the moment at which it

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

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

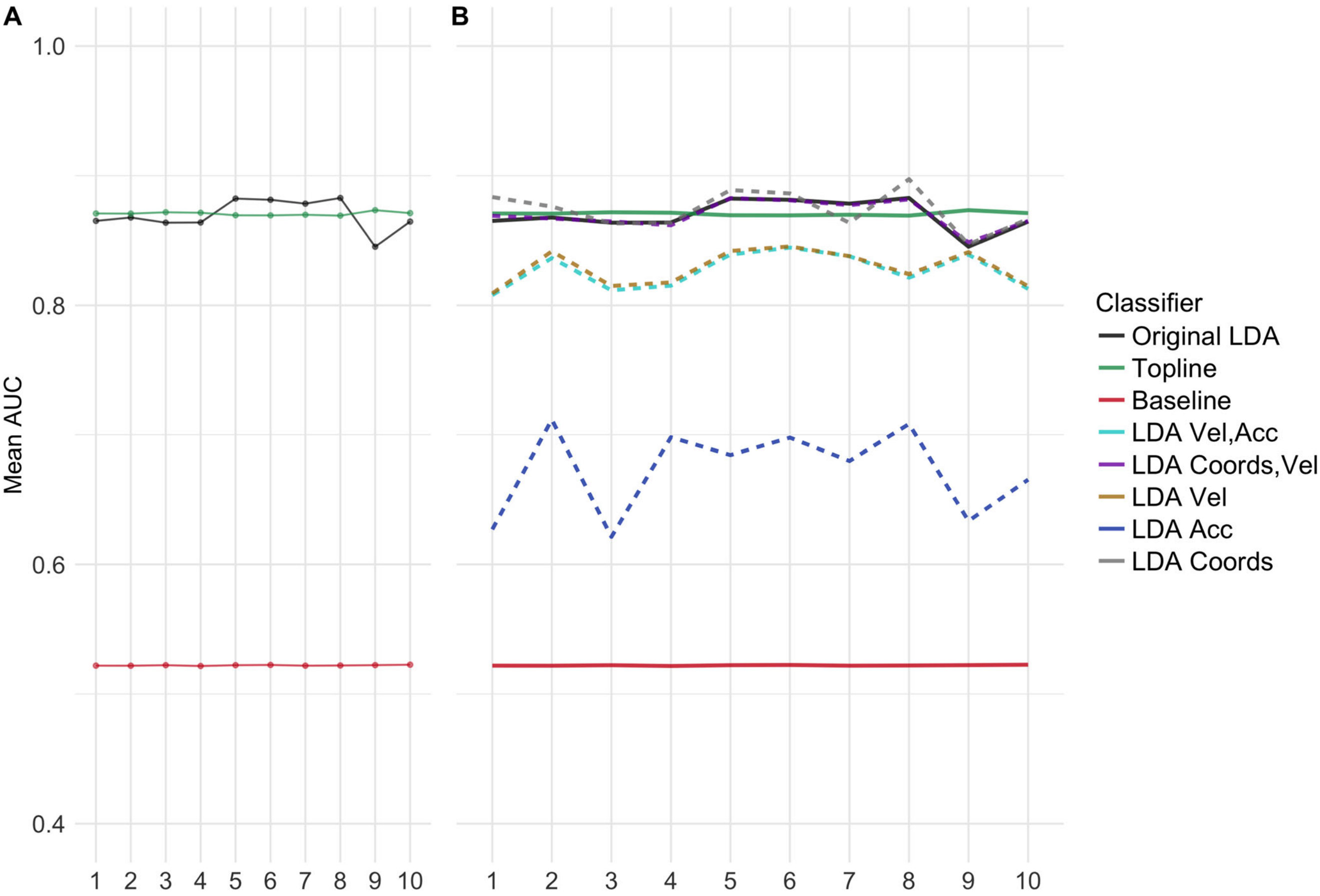


Fig. 6 Mean area under the ROC curve values obtained from cross-validation. **a** Cross-validation over ten bins for the original LDA, baseline, and topline. **b** Comparison with values obtained for five additional classifiers, obtained by subsetting the original set of predictors

occurred, assigning higher values to late deviations. This information seems to be essential for the classification, as we observed in "Validation experiment: Classifying decision processes with LDA" section.

Finally, we had previously observed that acceleration and, to a minor extent, velocity 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.

To summarize, in this section we have shown that (1) a rough manipulation of decision-making processes has a direct impact on mouse trajectories, (2) an LDA using absolute temporal information is enough to accurately distinguish these decision patterns, and (3) this LDA does a better classification than other, traditional mouse-tracking measures.

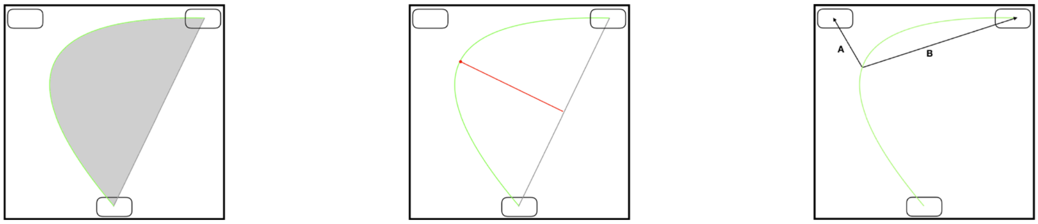
Extension to linguistic data

Mouse paths obtained from the validation experiment, in which we explicitly induced “decision changes,” were used to construct a transformation that takes mouse trajectories as input and transforms them into a single “degree of change of decision” measure (i.e., the LDA measure). This transformation can in principle be applied to new mouse trajectories in order to detect changes of decision. Can our LDA, then, help

Table 2 Cross-validation results for the LDA classifier

	Original LDA (Coords., Speed, Acc.)	(a)		(b) LDA With Different Predictions				
		Baseline	Topline	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	.20
p value	–	< .001	.58	.50	< .001	.68	< .001	< .001

The performance of the LDA was compared to that of (a) the baseline and topline classifiers and (b) the LDA classifiers with different predictors



(a) Area Under the Trajectory Curve

(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

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

Fig. 7 Description of commonly used mouse-tracking measures

characterize more complex decision processes, such as the ones involved in sentence verification tasks?

To address this question, we tested our classifier on data obtained from a replication of Dale and Duran's (2011) experiment. This experiment revealed 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 for 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 is negated only as a second step. On the other hand, if negation does not involve a change in decision—or if participants' behavior in the validation experiment is simply too different from natural changes of decision—then the LDA measure trained on validation data would not reveal systematic differences between positive and negative sentences.

Experiment

Participants had to perform a truth-value judgment task in which they had to decide whether a sentence (e.g., *Cars have wheels*) is true or false, on the basis of common world knowledge. Each sentence could either be a negated form or a nonnegated form, and could either be a true or a false statement. Unlike Dale and Duran (2011) experiment, the complete statement was presented in the middle of the screen after participants pressed "start" (i.e., no self-paced reading, see example in Fig. 10). The response buttons appeared at the top left and top right corners of the screen, as in our validation experiment. The materials and design are exemplified in Table 4, and a sample trial can be seen in Fig. 10.

Participants

In all, 53 English native speakers (29 female, 24 male) were tested using Amazon Mechanical Turk. They were compensated for their participation (1 USD). The experiment lasted approximately 10 min.

Design

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

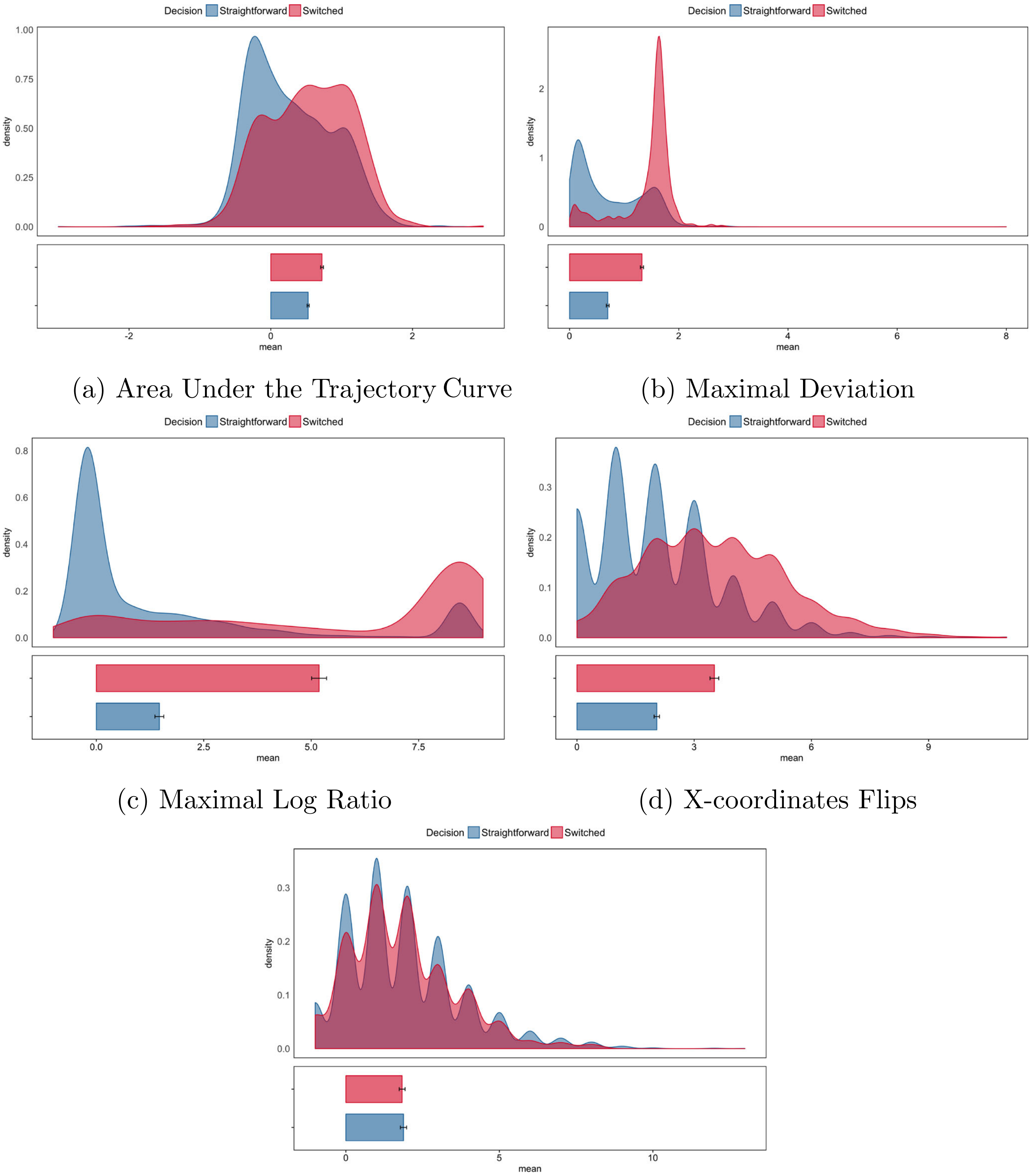
Interface and data treatment The interface and data treatment were the same as we had used in the validation experiment. The time course of mouse trajectories was again normalized into 101 time steps.

Results and discussion

Replicating Dale and Duran (2011)

All participants responded correctly more than 75% of the time. No participant was discarded on the basis of accuracy. Only accurate trials were analyzed. Figure 11 illustrates the mean trajectories for the four conditions.

To assess whether we replicated Dale and Duran's (2011) results, we calculated *x*-coordinate flips (see Validation experiment: LDA versus traditional mouse tracking analyses section) and analyzed them with a linear mixed-effect model, taking truth, polarity, and their interaction as predictors. We included random intercepts per subject and a random slope for the interaction of both factors. All *p* values were obtained by comparing the omnibus model to a reduced model in which the relevant factor was



(a) Area Under the Trajectory Curve

(b) Maximal Deviation

(c) Maximal Log Ratio

(d) X-coordinates Flips

(e) Acceleration Component

Fig. 8 Distribution and means obtained from applying different mouse-tracking measures to the validation data. Error bars represent standard errors of the mean

removed. This was 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 the mouse trajectories underlying different decision patterns in the validation experiment.

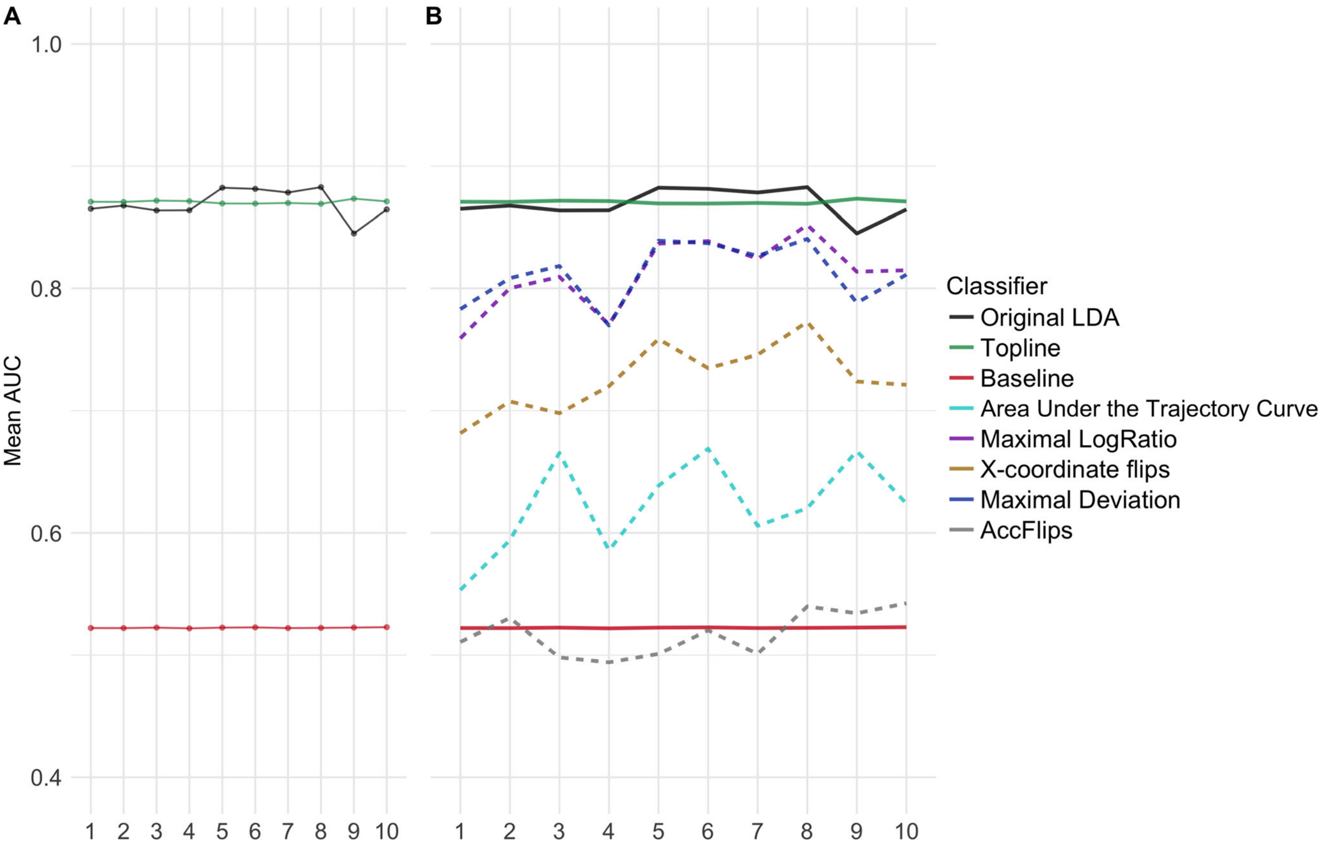


Fig. 9 Mean area under the ROC curve values obtained from cross-validation. **a** Cross-validation on ten bins for the original LDA, baseline, and topline. **b** Comparison with the values obtained for other commonly used mouse-tracking measures

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 Truth × Polarity interaction ($\chi^2 = 24.7, p < .001$), such that the difference between negative and positive sentences was 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 Dale and Duran's (2011) results.

We seem to have replicated Dale and Duran's (2011) findings: Verifying true negated sentences produces less-straightforward trajectories than do true positive sentences. The values obtained in the two experiments were 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 that could be responsible for the range difference. Moreover, Dale and Duran used a smoothing method over their trajectories, whereas we did not. This probably causes our estimates to be higher.

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? We tested these new data using two different LDA classifiers, both of them trained on mouse trajectories from the validation experiment. In other words, switched and straightforward trials from the validation were used to train the LDA algorithms, which we then used to

Table 3 Cross-validation results for the LDA classifier

	Original LDA	AUT	MD	Maximal Log Ratio	*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

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

Table 4 Design of Dale and Duran's (2011) replication

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

test new trajectories. 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. The validation results (see Validation experiment: Classifying decision processes with LDA section) 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 it can be interpreted more straightforwardly.

The relevant difference in processing between positive and negative sentences was expected to arise specifically for true statements. Consequently, we analyzed 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 1,000 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 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 that would have been obtained from applying the LDA to a set of data in which there was no difference between experimental conditions? We calculated the AUC values for a set of data in which the 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 data. Figure 13b illustrates the separability of the two classifications for each sample size.

The LDA classifier trained with the validation data seems to make a distinction between experimental conditions. This finding suggests that the contrast between negative and positive trials was 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 was proposed by Dale and Duran (2011), among others. However, although the 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 Figs. 5 and 12). This is not surprising, given that we were dealing with different cognitive processes—simulated

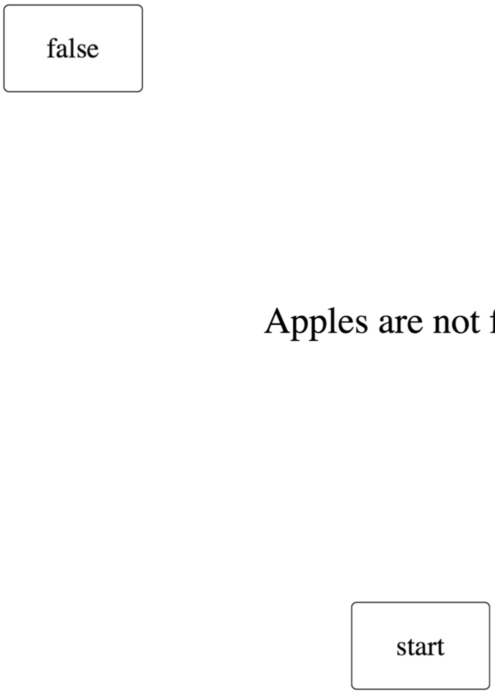


Fig. 10 Illustration of a trial in the replication of Dale and Duran (2011)

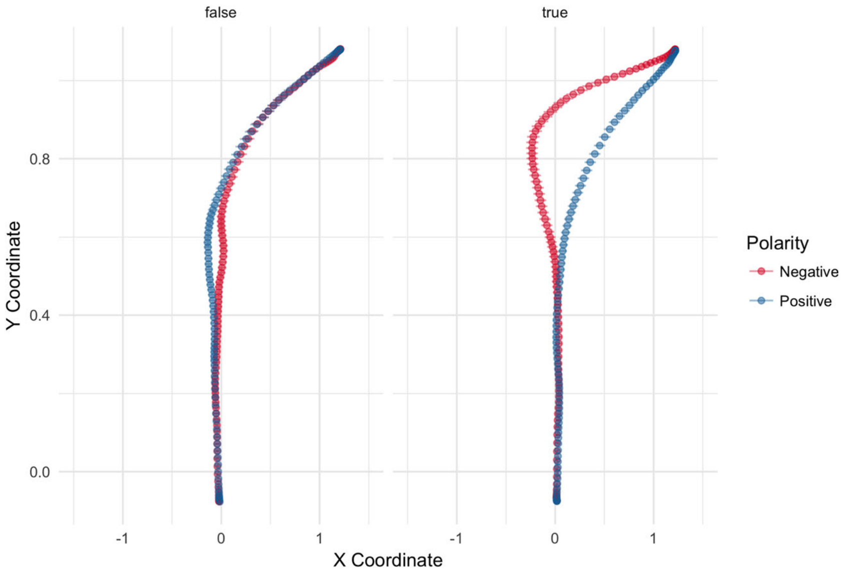


Fig. 11 Mean trajectories for accurate trials

decisions versus sentence verification—but, as we discussed above, the difference could easily also result from an idiosyncrasy of these two data sets.

Finally, although the classifier comparison in Fig. 6 indicated that *relative* spatiotemporal 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, Fig. 13 reveals that the *full* classifier—which takes all features as predictors—gives a better separation between the two experimental conditions than does the simplified one.

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 distributions 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 Fig. 15a) and contrasted these values against a null hypothesis of no difference between experimental conditions (see Fig. 15b).

Table 5 Mean and effect estimates for Dale and Duran’s (2011; D&D) original experiment and our replication

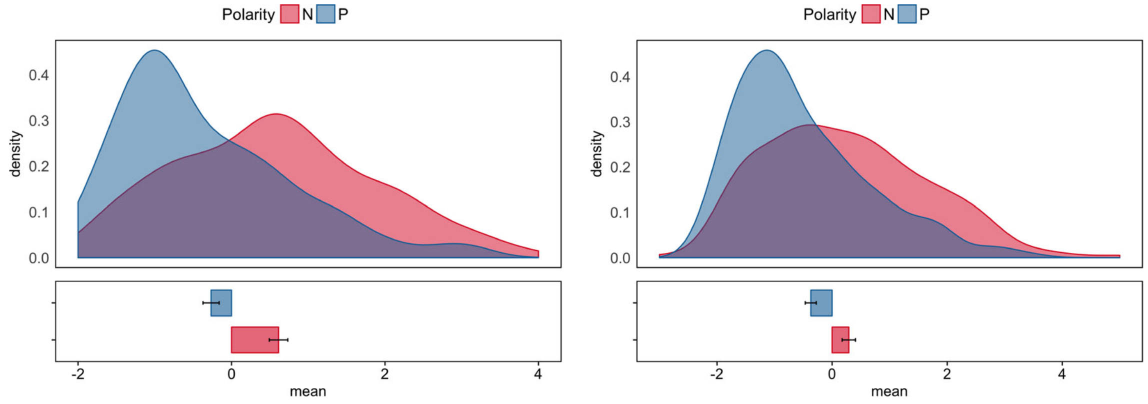
Condition	*x*-Flips	*x*-Flips in D&D
T/no negation	2.22	1.13
T/negation	3.67	1.71
F/no negation	2.82	1.24
F/negation	2.9	1.34
Estimate Polarity	0.76	0.35
Estimate Truth	0.07	0.13
Estimate Truth × Polarity	1.35	0.47

The results in Fig. 15a suggest that most measures performed less well here than on the validation data (cf. Fig. 9). Since a decrease in performance was attested across the board and not only for the classifiers trained with the 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 might 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, since it does not make any specific assumptions about how the change of decision should be reflected by mouse trajectories beyond the observed ones.

Baseline

A linear classifier trained on simulated decisions can separate the two experimental conditions of the replication of the previous study by Dale and Duran (2011). We interpreted this result as suggesting that 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 the mouse paths that happened 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.



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

Fig. 12 Two LDA classifiers applied to *true* trials (negative vs. affirmative). Error bars represent standard errors of the mean

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 that contained only the positive trials from the replication of the experiment by Dale and Duran (2011). 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 to the response, whereas longer response times would reflect a late commitment. As is illustrated by Fig. 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 Fig. 16b. The performance was evaluated following the same procedure applied above (see the blue line in Fig. 15).

The classification on *early* versus *late* categories is less accurate than the one performed on 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 look similar to *switched*

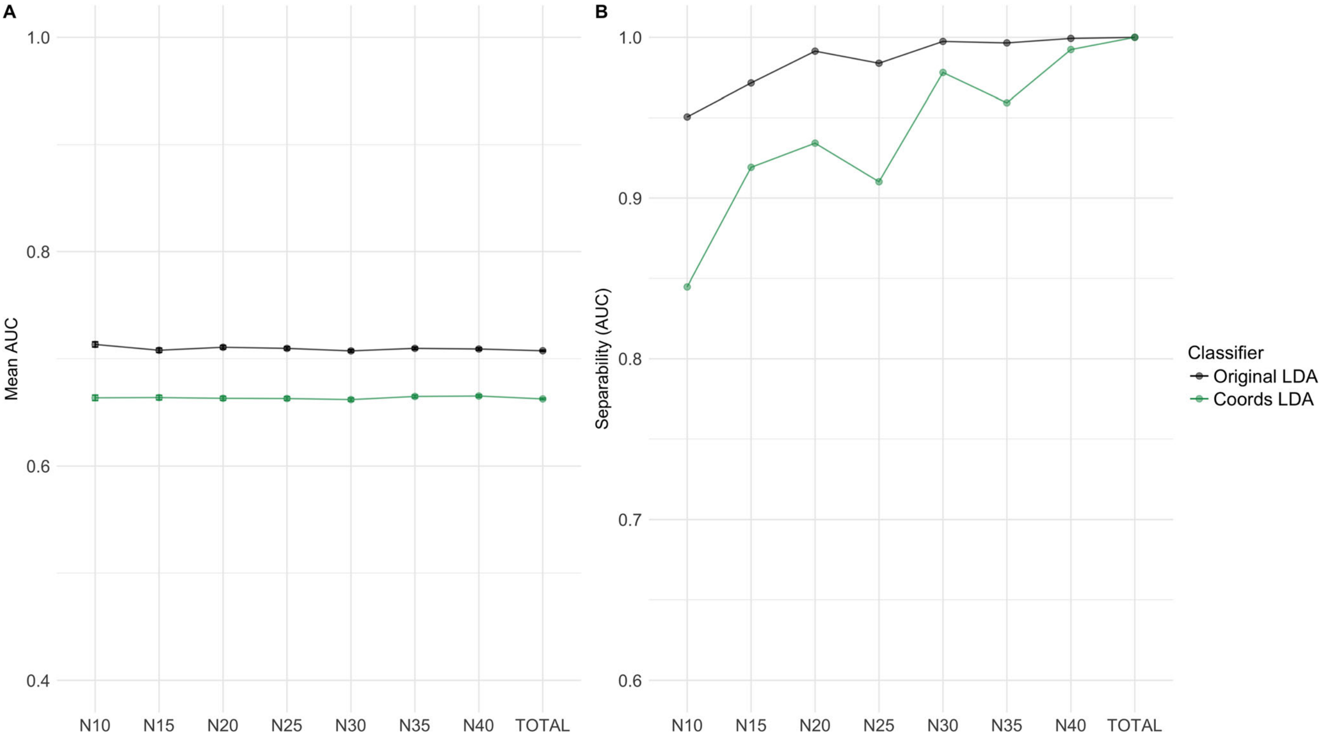
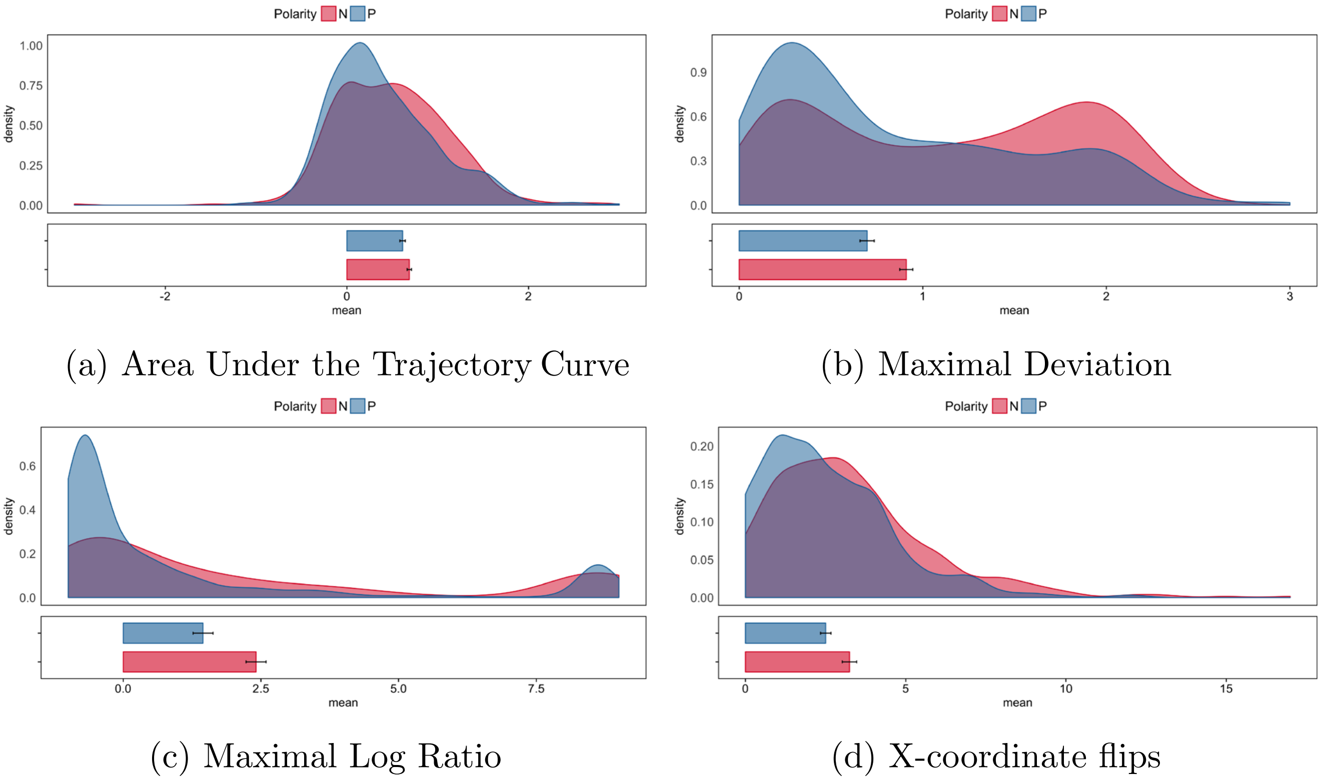


Fig. 13 Performance of the LDA classifiers. **a** Mean AUC values over bootstrapped data (iterations = 1,000) for different sample sizes. **b** Difference in classifier performance when applied to scrambled versus the original data



(a) Area Under the Trajectory Curve

(b) Maximal Deviation

(c) Maximal Log Ratio

(d) X-coordinate flips

Fig. 14 Distributions and means of negative and positive *true* trials obtained from applying different mouse-tracking measures to negation data. Error bars represent standard errors of the mean

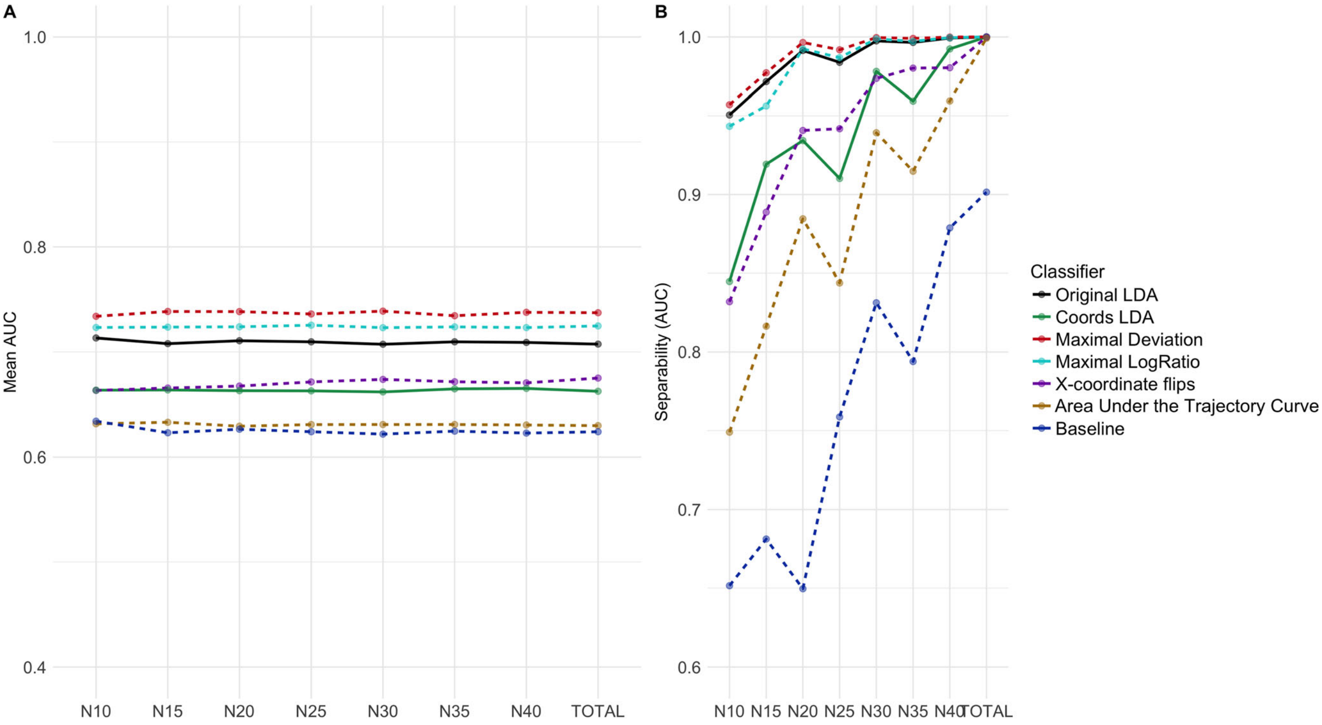
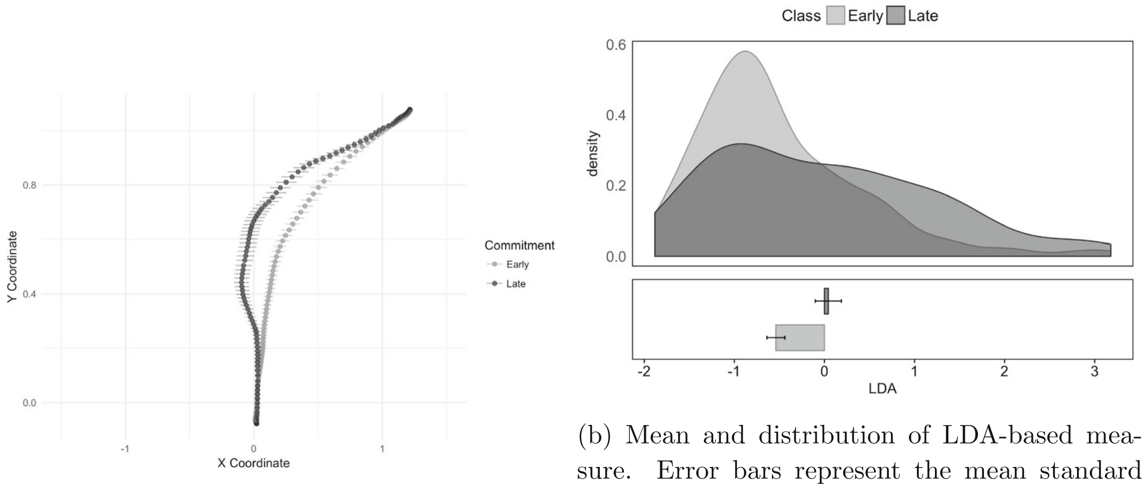


Fig. 15 Performance of other mouse-tracking measures. **a** Mean AUC values over bootstrapped data (iterations = 1,000) for different sample sizes. **b** Difference in measure performance when measures were applied to scrambled versus the original data



(a) Mean Trajectories

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

Fig. 16 Analyses performed on baseline data set (early vs. late decisions)

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 the replication experiment, the similarities appear to be more than accidental.

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 rough change of decision, we showed directly, for the first time, how mouse trajectories are impacted by decision processes: 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 this sort of decision shift. This classifier, freely available online, accurately classifies not only the paths corresponding to quasi-decisions, but also the paths underlying a more complex cognitive process, such as the verification of negative sentences.

The approach developed here in this sense makes an important contribution to all lines of research that may rely on mouse-tracking data to investigate cognitive processing. Our results not only replicate previous findings but, more importantly, show that the LDA classifier performs at least as well as the best of the other commonly used mouse-tracking measures. This comparison of performance raises the question of whether we should abandon traditional mouse-tracking measures, adopting our LDA classifier instead.

On the one hand, we have established that the maximal deviation and maximal log ratio measures as comparable alternatives to the LDA analysis in terms of performance. These measures are in principle easier to deploy than our classifier and have been used successfully in a number of studies.

However, unlike these other measures, the performance of the LDA classifier is contingent on the characteristics of the training data set—in our case, the one coming from the validation experiment. Although, as it is now, our validation experiment is just a first, very simple approximation to a decision switch, it can potentially be refined and adapted in order to test new hypotheses. That is, if one has clearer hypotheses about the mechanisms at play during decision making or sentence verification, one could build more-representative validation experiments. This would in turn serve to identify prototypical mouse path patterns for different types of cognitive processes. Indeed, we believe this refinement will be a necessary step and is only made possible by our classifier, making LDA analysis conceptually more powerful than alternative measures.

Moreover, the LDA classifier has the unique advantage of not relying on any specific assumption about how switched trajectories should look like. Because it is assumption-free, our approach can be applied to other processing measures in order to perform a classification of decisions that goes beyond the specific mouse-tracking methodology.

Author note The research leading to these results received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013)/ERC Grant Agreement No. 313610, and was also supported by Grants ANR-10-IDEX-0001-02 PSL* and ANR-10-LABX-0087 IEC.

References

Clark, H. H., & Chase, W. G. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology*, *3*, 472–517.

Cranford, E. A., & Moss, J. (2017). Mouse-tracking evidence for parallel anticipatory option evaluation. *Cognitive Processing*, *19*, 327–350. https://doi.org/10.1007/s10339-017-0851-4

Dale, R., & Duran, N. D. (2011). The cognitive dynamics of negated sentence verification. *Cognitive Science*, *35*, 983–996. https://doi.org/10.1111/j.1551-6709.2010.01164.x

Donders, F. C. (1969). On the speed of mental processes. *Acta Psychologica*, *30*, 412–431. (Original work published 1868) https://doi.org/10.1016/0001-6918(69)90065-1

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. https://doi.org/10.1080/03640210701530797

Freeman, J. B. (2018). Doing psychological science by hand. *Current Directions in Psychological Science*, *27*, 315–323. https://doi.org/10.1177/0963721417746793

Freeman, J. B., & Ambady, N. (2010). MouseTracker: software for studying real-time mental processing using a computer mouse-tracking method. *Behavior Research Methods*, *42*, 226–241. https://doi.org/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*, 59:1–6. https://doi.org/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*, 362–374.

Freeman, J. B., Pauker, K., & Sanchez, D. T. (2016). A perceptual pathway to bias: Interracial exposure reduces abrupt shifts in real-time race perception that predict mixed-race bias. *Psychological Science*, *27*, 502–517.

Hasson, U., & Glucksberg, S. (2006). Does understanding negation entail affirmation? An examination of negated metaphors *Journal of Pragmatics*, *38*, 1015–1032.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference and prediction (2nd). New York, NY: Springer.

Hehman, E., Stolier, R. M., & Freeman, J. B. (2014). Advanced mouse-tracking analytic techniques for enhancing psychological science. *Group Processes & Intergroup Relations*, *18*, 384–401. https://doi.org/10.1177/1368430214538325

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

Kieslich, P. J., & Henninger, F. (2017). Mousetrap: An integrated, open-source mouse-tracking package. *Behavior Research Methods*, *49*, 1652–1667. https://doi.org/10.3758/s13428-017-0900-z

Koop, G. J. (2013). An assessment of the temporal dynamics of moral decisions. *Judgment and Decision Making*, *8*, 527.

Koop, G. J., & Johnson, J. G. (2013). The response dynamics of preferential choice. *Cognitive Psychology*, *67*, 151–185.

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*, 1355–1370. https://doi.org/10.1162/jocn.2008.20093

McKinstry, C., Dale, R., & Spivey, M. J. (2008). Action dynamics reveal parallel competition in decision making. *Psychological Science*, *19*, 22–24.

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*, 1213–1218. https://doi.org/10.1111/j.1467-9280.2008.02226.x

Orenes, I., Beltrán, D., & Santamaría, C. (2014). How negation is understood: Evidence from the visual world paradigm. *Journal of Memory and Language*, *74*, 36–45.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*, 873–922. https://doi.org/10.1162/neco.2008.12-06-420

Resulaj, A., Kiani, R., Wolpert, D. M., & Shadlen, M. N. (2009). Changes of mind in decision-making. *Nature*, *461*, 263–266. https://doi.org/10.1038/nature08275

Sauerland, U., Tamura, A., Koizumi, M., & Tomlinson, J. M. (2017). Tracking down disjunction. In M. Otake, S. Kurahashi, Y. Ota, K. Satoh, & D. Bekki (Eds.), New Frontiers in Artificial Intelligence: JSAI-isAI 2015 Workshops (pp. 109–121). Cham: Springer.

Schulte-Mecklenbeck, M., Kühberger, A., & Johnson, J. (Eds.). (2019). A handbook of process tracing methods (2nd). New York: Psychology Press.

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

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. https://doi.org/10.1167/6.9.11

Spivey, M. J., & Dale, R. (2006). Continuous dynamics in real-time cognition. *Current Directions in Psychological Science*, *15*, 207–211. https://doi.org/10.1111/j.1467-8721.2006.00437.x

Spivey, M. J., Dale, R., Knoblich, G., & Grosjean, M. (2010). Do curved reaching movements emerge from competing perceptions? A reply to van der Wel et al. (2009). *Journal of Experimental Psychology: Human Perception and Performance*, *36*, 251–254. https://doi.org/10.1037/a0017170

Spivey, M. J., Grosjean, M., & Knoblich, G. (2005). Continuous attraction toward phonological competitors. *Proceedings of the National Academy of Sciences*, *102*, 10393–10398. https://doi.org/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. https://doi.org/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. *Quarterly Journal of Experimental Psychology*, *63*, 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*, 18–35.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, *108*, 550–592. https://doi.org/10.1037/0033-295X.111.3.757

Wason, P. C. (1965). The contexts of plausible denial. *Journal of Verbal Learning and Verbal Behavior*, *4*, 7–11. https://doi.org/10.1016/S0022-5371(65)80060-3

Wason, P. C., & Johnson-Laird, P. N. (1972). Psychology of reasoning: Structure and content. Cambridge: 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. *Psychological Science*, *20*, 1428–1435. https://doi.org/10.1111/j.1467-9280.2009.02448.x

Wulff, D. U., Haslbeck, J. M. B., Kieslich, P. J., Henninger, F., Schulte-Mecklenbeck, M. (to appear). Mousetracking: Detecting types in movement trajectories. In M. Schulte-Mecklenbeck, A. Kuehberger, & J. G. Johnson (Ed.), A handbook of process tracing methods (2. ed.). Psychology Press

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, e0178740. https://doi.org/10.1371/journal.pone.0178740