Bachelor Thesis

to attain the academic degree Bachelor of Science

Are Introspective Confidence Judgements Reflected in the Motor Dynamics of Response Selection?

Faculty: Human Sciences
Supervisor: Prof. Dr. Michael Franke
Second Examiner: Dr. Timo Röttger
Submission Date: June 2020

Author

Name: Tamara Keßler
Matriculation Number: 970082
Course of Studies: Cognitive Science
Semester: 8

Table of Contents

1. Introduction

Cognitive Psychology is the science of mental processes, how they are realised in the brain and how they affect human behaviour (Rosenbaum, 2005; Anderson, 2014). One of the leading research endeavours in the field is to understand how decisions are enacted. Surprisingly however, motor control has long been neglected in research as motor responses have been widely regarded as "uninteresting end-results of cognitive processes" (Rosenbaum, 2005; Freeman, 2011). By now, through advances in the field of neuroscience, cognition and motor control are both considered to be continuous processes that are dynamically influenced by each other, as well as by other related cognitive operations. This in turn implies that certain motor actions or performance errors might offer valuable insights to the mental processes underlying them (Rosenbaum, 2005; Spivey, 2006; Hehman, 2014).

In order to investigate these continuous processes, it is not enough to merely consider their outcome or summary measures related to them, such as the total time needed to execute the related motion. Therefore, traditional paradigms on their own, such as response times or error measurements, are not sufficient to research such matters. In an attempt to overcome these shortcomings, so-called process tracing methods have been developed, which rely on measuring continuous streams of motor output with a high temporal sensitivity, as it is thought that this allows to capture minor changes within mental processes (Freeman, 2011). This group of methods has proven itself to be extremely valuable, especially in the field of perceptual decision making. Mouse tracking, which is a more recently developed process tracing method, enables researchers to track how committed participants are to various choice alternatives over time by recording spatial and temporal information related to how they move a computer mouse while selecting a response option that is displayed on a screen (Hehman, 2014; Freeman, 2018; Stillman, 2018).

An alternative approach to assessing commitment to a given decision is asking participants to rate their subjective confidence. It has been shown that when participants report to feel more confident about a decision they made, this decision also tends to be more likely to be correct and to be executed faster. However, self-reported confidence in general is not considered to be an accurate measure, due to it being highly subjective and dependent on the participants (Sandberg, 2010; Grimaldi, 2015; Norman, 2015). According to recent research, it might be possible to indirectly infer subjective confidence from movement dynamics as recorded through the tracking mouse movements. Some possible measures that might be well-suited to reflect confidence are the movement's deviation from an ideal path or temporal dynamics, such as the point in time at which the highest velocity was reached (Dotan, 2018 & 2019). The present study aims to investigate to which degree such mouse tracking measures covary with self-reported confidence judgements.

As this topic has not been researched in much detail yet, this study might serve as a potential pilot-study for following more in-depth research. More research in this topic could open up new ways to investigate subjective confidence: As mouse tracking has a high temporal sensitivity, it offers a way of not only assessing confidence in real-time, but also how it changes over the course of a trial. This might also be applicable to the research on other forms of metacognition.

2. Background of the Study

2.1. Reflection of Cognitive Processes in Motor Dynamics

Psychology is commonly defined as the study of the human mind and behaviour (e.g.: Merriam-Webster), and one of its central goals is to analyse the series of cognitive processing stages that determine higher-order mental operations, such as decision making amongst others (Dotan, 2019). Following this definition, it would seem feasible that the interactions between cognition and motion would be of major research interest to understand how decisions are enacted. In reality, however, only very little research in psychology concerns itself with motor control (Rosenbaum, 2005). This could be explained by the fact that for a long time motor responses were solely regarded as an end-result of a one-way processing route starting with human perception, before undergoing cognition, followed by action planning and execution. Even further, cognition and motor control were regarded as nearabout functionally independent. As a result, motor responses were thought as being unable to disclose anything worthwhile about the cognitive processes generating them (Freeman, 2010 & 2011).

Since in research only little can be observed directly when it comes to perception and cognition, in most cases one must make inferences about the relationship between input and behaviour (Magnuson, 2005). The first time that a connection between cognition and motor responses has been verified in a scientific laboratory setting was in 1868 by the Dutch physiologist Frans Cornelis Donders. He is best known for introducing the measurement of response times (RTs) in his studies on mental chronometry to make inferences about the time course of cognitive processes in perceptual-motor tasks (Posner, 2005; Freeman, 2018; Dotan, 2019).

By now, continuous motor responses are no longer merely seen as endpoints of our sensory and cognitive processes, but rather as an integrative part of the dynamics underlying perception and cognition. Even though one's own thoughts might seem to consist of individual elements, especially if one only takes discrete behavioural outcomes as evidence for these thoughts, it is highly likely that mental activity is also happening in between seemingly discrete thoughts. Therefore, real-time cognition is best imagined as a dynamic and continuously changing pattern of neuronal activity, from which it follows that mental states exhibit in-between cases of describable states of mind (Spivey, 2006). Neuronal activity, such as in the premotor cortex, occurs in a continuously dynamic manner and is stimulated by ongoing cognitive operations. This temporal continuity of higher-level cognitive processes is closely tied to continuous motor movements as well (Spivey, 2006; Hehman, 2014). As a result, it seems plausible that cognition as a whole should not just be analysed as a series of discrete events, but as a continuous and dynamic process (Spivey, 2006).

Especially in the context of perceptual decisions, such as the categorisation of visual stimuli, this new conception of the interaction between cognition and action is crucial. Opposed to what was formerly believed, evidence from electrophysiological studies suggests that there is a two-way interaction between higher-level cognitive processes and goal-directed motor control (Hehman, 2014). Goal-directed motor control includes several aspects, such as movement selection, initiation, revision, and adjustment of said movements and finally, optimisation of the movement sequence to maximise task performance. Especially the intermediate revision and re-adjustment of the movement have long been neglected, even though they form an important part in the interaction between cognition and motor execution (Dotan, 2019). For example, during the categorisation of visual stimuli intermediate processing results are immediately shared with the (pre-)motor cortices in order to guide, for example, hand movements continuously over time. This means that certain aspects of the movement during response selection could be analysed to gain insights into the psychological processes underlying and generating the movement. Some examples for this would be acceleration, deviation from a hypothetically ideal path or changes in direction (Hehman, 2014). This means that if one were able to sample such continuous movements fast enough, they would well reflect on-line cognitive processes and how they evolve over time (Freeman, 2010 & 2011). Hence, they could be used to gain valuable insights into the time-course of perceptual cognitive processing (Freeman, 2010; Hehman, 2014). Even before ultimately deciding on a response option and the according movement plan to enact it, humans are able of representing several potential movement plans and initiate their movement based on the average of the represented plans. This is crucial, as measuring continuous movements before the decision-making process is completed allows researchers to examine the intermediate processing stages of high-level cognitive processes (Dotan, 2019).

To uncover the temporal dynamics of the processing stages of higher cognitive operations, it is crucial to use measurements with the highest possible temporal resolution – or ideally even in real-time (Dotan, 2019). Following from that, in many cases it is considered insufficient to merely analyse discrete measures, such as RTs or error rates, when investigating the development of cognitive processes over time. Since response times and error rates are only summary measures that contain conflated information from all mental processes that occur throughout an experimental trial, their ability to discern data about the absolute timing and order of the underlying cognitive operations is extremely limited (Freeman, 2011; Dotan, 2019). Largely responsible for that is that RTs, for example, are considered to be some kind of a "black box" in the sense that it is still unknow which factors exactly and to which extent contribute to how long a response takes (Freeman, 2011). Accordingly, methods that exhibit higher temporal resolutions have been introduced and increasingly used to research the temporal evolution of behavioural responses in real time.

Previously, only a few methods in cognitive psychology were able to do so, namely the brain imaging methods electroencephalography (EEG) and magnetoencephalography (MEG), and the process tracing method eye-tracking. Both EEG and MEG continuously measure brain activity with a high temporal resolution with millisecond precision. They allow for directly measuring neural mechanisms and activity patterns underlying cognitive processes before, during and in the absence of behavioural responses. Eye-Tracking is able to measure gaze shifts, fixation times and locations, as well as pupil dilations, which can be used to index both cognitive operations, as well as associated mental effort. Again, this method exhibits a high temporal resolution, so that it is able to reflect cognitive changes well (Dotan, 2019). However, these methodologies still only allow for an indirect inference of cognitive operations and so far, more direct measures examining how responses evolve in real-time have been lacking (Freeman, 2018).

2.2. Mouse Tracking

Since its establishment by Spivey et al. in 2005, mouse tracking has become a popular method with application in many fields, as it has been able to provide insights into a wide variety of psychological phenomena (Hehman, 2014; Freeman, 2018). It offers accessible real-time insights into cognitive processes that are related to human decision making, such as categorisation and judgement (Stillman, 2018). Researchers are able to track how committed participants are to various choice alternatives over time by recording how they move a computer mouse while selecting a response option that is displayed on a screen (Hehman, 2014; Freeman, 2018). By tracking mouse movements, it is possible to measure various features of the underlying cognitive processes related to perceptual decision making. Unlike with classic outcome-based measures, which are commonly used in this field of research, the mental processes of coming to and enacting a decision are not only inferred from RTs or error rates, but also through numerous measures associated with the movement trajectory, which enables capturing the continuous aspects of cognitive processes as well (Spivey, 2005; Quétard, 2016). Since mouse tracking has a comparatively high sampling rate, it is able to continuously track how committed participants are to individual response alternatives on a millisecond scale (Freeman, 2011).

So far, due to its ability to measure covert response activations of implicit decisions, mouse tracking has found usage in multiple areas of research across psychological and linguistic science: After its initial application during the study of spoken-word recognition and speech perception as in Spivey et al.'s 2005 study, it has also been successfully used in investigations of referential communication, as by Brennan et al. in 2005, categorisation as with Dale et al.'s 2007 study, and studies on social cognition, like Freeman et al.'s 2010 study on stereotyping (Farmer, 2007; Schulte-Mecklenbeck, 2011). But also other topics of psychological research have used the methodology, such as self-control, theory of mind or moral cognition, to only name a few (Freeman, 2018).

This wide applicability is most likely due to the methodology's key paradigm being flexible enough to accommodate a wide variety of tasks. Typically, mouse tracking makes use of the so-called two-alternative forced choice paradigm. In this case, participants click a button in the bottom centre of the screen to begin a trial (Fischer, 2014; Freeman, 2018; Stillman, 2018). Afterwards, they are presented with a stimulus, which can be either a letter string, an image, a sound, a video or even a combination of those (Hehman, 2014; Freeman, 2018). Following that, participants have to move the mouse cursor to one of the two available response options, which are displayed in the top left and right corners of the screen, respectively. The response alternatives may be presented before or simultaneously with the stimulus, depending on the given task (Hehman, 2014; Freeman, 2018; Stillman, 2018).

While moving the mouse to the appropriate response option, the x- and y-coordinates of the mouse trajectory are recorded 40 to 120 times per second on average (Freeman, 2011; Hehman, 2014; Stillman, 2018). After the data collection is completed, trajectories are aggregated before undergoing data analysis (Hehman, 2014). For correct trials, researchers generally differentiate between two distinct types of trajectories: Trajectories that stay in their respective correct hemispace of the screen and solely differ in the magnitude of their curvature – and trajectories that briefly cross over the central y-axis into the incorrect hemispace before returning to correct hemispace of the screen, again with varying strengths in curvature. Both movement patterns and curvature index the attraction effects that a competing stimulus exercises on (the motor execution of) the decision (Fischer, 2014). Thus, mouse tracking provides a temporally sensitive real-time measure that is able to assess when a response got partially activated, as well as the extent to which it attracted a participant's consideration, even if it ultimately remained unselected (Hehman, 2014; Freeman, 2018).

Several approaches to characterising and analysing trajectory components have been developed so far. Trajectories usually are averaged and visualised, but they can also be processed further, and a number of derivative measures can be computed (Freeman, 2010). Especially measures of spatial attraction, which can be interpreted as a measure of bias toward the unselected response option are frequently analysed. A common choice for this is curvature, as approximated through the so-called area under the curve (AUC), which is estimated as the geometric area between the recorded trajectory and the ideal, most direct trajectory from the starting point to the correct response option (Freeman, 2010 & 2011; Quétard, 2016). Further, measures regarding the trajectory's speed are also of interest. For that reason, certain time-points of acceleration might be investigated, or the overall temporal development of velocity and acceleration may be described as a function of time. Besides these measures, also many others are used in research, such as the time-course of the cursor position, and changes in movement direction, which can be interpreted as a measure of complexity (Freeman, 2010, 2011 & 2018; Quétard, 2016).

(See appendix A for a comprehensive overview of measures relevant to this study)

2.2.1. Advantages

What some view as the key benefit of mouse tracking is its ability to gauge the temporal dynamics of mental processes in real-time through the simple indexing of hand motions. Due to the method's continuous streaming of motion output data, it enjoys a high temporal sensitivity and offers millisecond-by-millisecond information (Freeman, 2018; Stillman, 2018). Therefore, mouse tracking offers a highly sensitive real-time measure of how a decision emerges and is resolved temporally (Stillman, 2018). As even single trials yield continuous time course information, the methodology is sensitive enough to provide information about cognitive changes that occur mid-trial (Magnuson, 2005). This is due to the fact that arm movements naturally proceed in a continuous manner, which means that they can exhibit in-between states between multiple responses (Freeman, 2011 & 2018; Dotan, 2019). Thus, they are able to display graded attraction effects through the curvature of the recorded trajectory even in trials in which the distractor was not chosen (Spivey, 2005 & 2006; Dale, 2007). These measurements can then be used to evaluate how committed participants are to a given choice during the evolution/development of the tracked behavioural response (Freeman, 2011; Schulte-Mecklenbeck, 2011).

This high level of temporal sensitivity can be attributed to the data richness associated with mouse tracking data, as it on average yields 30 – 60 data points per trial. Comparable set-ups with eye-tracking only generate 3 - 4 data points in the form of saccades and the paradigm of self-paced reading only affords 2-3 data points in the form of button presses per trial (Spivey, 2005; Farmer, 2007). In the case of outcome-based measures, many factors contribute to a response and are conflated to only a single value. As a result, these measures usually require to be subjected to complex modelling to isolate the various components contributing to them, which makes the interpretation of the received values difficult (Stillman, 2018). Therefore, mouse tracking may facilitate understanding the interaction of cognition and motor responses in a more dynamical sense (Magnuson, 2005) and by extension, can potentially answer research questions which could not be answered with more traditional technologies so far (Farmer, 2007).

Another strong asset is its wide applicability, as it allows for recording real-time cognitive processes in children at an age where the acquisition of this kind of on-line data is rather difficult: A study conducted by Calvert et al. in 2005 found that the mean age of first autonomous computer usage is 3.7 years, which suggests that children already aged 3.5 to 4 years are eligible to partake in mouse tracking studies (Farmer, 2007; Freeman, 2011). Further, it has been suggested that trajectory tracking methods should also be applicable for animal experiments (Dotan, 2019).

Even within human adults, mouse tracking poses fewer restrictions on potential participants than other methodologies that might be used to examine similar means: Unlike it is often the case with remote eye-tracking set-ups, there is no need to exclude participants who wear glasses due to hardware limitations when using mouse tracking.

However, while some authors only recommend the inclusion of right-handed participants (Hehman, 2014), there has been no conclusive answer as to how far handedness influences the user's performance in a given mouse tracking task. As a result, it is generally recommended to assess the participants' handedness and the hand used for administering the mouse (Kieslich, 2017).

Further, the accessibility of mouse tracking is a considerable advantage. Since computer and corresponding mouse environments are omnipresent in all research settings nowadays, setting up mouse tracking is easy, as it does not require any additional hardware (Freeman, 2011 & 2018; Hehman, 2014). All that is needed to create and launch an experiment is a freely available software program[1], which functions within user-friendly modern experiment building tools, such as PsychoPy or OpenSesame (Schulte-Mecklenbeck, 2011; Stillman, 2018). As a result, the difficulty of programming a standard mouse tracking experiment and subsequently analysing the acquired data is low by way of comparison to other methods, such as eye-tracking, EEG or functional magnetic resonance imaging (fMRI) (Magnuson, 2005; Stillman, 2018).

However, what is important to note is that mouse tracking does not aim to replace the aforementioned methods in any case, but that it rather offers an opportunity to compensate for some of their short-comings, as in many cases it would be possible to employ mouse tracking in a complementary and simultaneous manner with a number of other methods. So it would, for example, be especially easy to combine mouse and eye-tracking, as these process tracing methods often share similar experimental set-ups. To begin with, the comparison of the combined time-course data would present new possibilities to study the interactions between numerous high-level cognitive processes, such as cognition and associated motor responses (Magnuson, 2005; Farmer, 2007; Schulte-Mecklenbeck, 2011). Further, the combination with other methods, such as EEG, but also fMRI and transcranial magnetic stimulation (TMS) would be possible. A multi-method approach of combining mouse tracking, eye-tracking and EEG could offer insights into a more complete understanding of the temporal dynamics and interactions of perception, cognition, and action (Freeman, 2011; Schulte-Mecklenbeck, 2011).

[1] Some examples are "MouseTracker" by Freeman and Ambady (2010), Kieslich and Henninger's "Mousetrap" (2017) or "TrajTracker" (2019) by Dotan et al.

2.2.2. Disadvantages and Limitations

As already mentioned, mouse tracking is able to be combined with a number of other well-established methods. However, it is important to note, that in some cases this should not only be regarded as a beneficial opportunity, but it could rather become a necessity.
Not in all cases are non-ballistic movements, such as hand movements, more informative than ballistic movements, such as saccades, when it comes to perception and cognition. While the former are more spatially sensitive to continuous competition and consequential attraction effects of multiple response options, the latter enjoy a higher temporal sensitivity regarding those partially active representations (Spivey, 2005). This could be explained by the fact that in general eye-tracking is more sensitive towards pre-attentive processes that occur before the initiation of movements (Quétard, 2016; Freeman, 2018), since for the most part saccades are executed subconsciously, while hand and arm movements are under conscious control (Magnuson, 2005). As a result, saccades tend to be initiated earlier than arm movements, which gives eye tracking a clear advantage over mouse tracking with respect to immediacy (Spivey, 2005). Further, ambiguity effects tend to prevail for longer in mouse tracking data than in the case of eye tracking (Magnuson, 2005). For these reasons, a combination of both methods would prove to be very valuable in order to counterbalance each other's weaknesses (Spivey, 2005; Freeman, 2018).

Another limiting aspect of mouse tracking is that the quality of the sampled data is highly dependent on the hardware used. The sampling rate of the streaming x- and y-coordinates, for example, is limited by the computer's operating system. Further, the mouse cursor can be displaced in a scaled manner depending on the mouse settings. Especially dynamic acceleration, which is enabled on default, can heavily affect the trajectory data and should therefore be manually disabled. Some researchers additionally recommend manually lowering overall mouse speed settings to de-skew the trajectory data (Fischer, 2014).

Related to this, it is important to note that also the behaviour of the participants can pose limiting factors on the mouse tracking data. For instance, it is crucial that participants will initiate their hand movements as early as possible into the trial, as otherwise the decision processes and motion planning will be completed and thus, no proper inferences can be drawn anymore from the recorded motion to the generating cognitive processes (Fischer, 2014). Further, participants might develop certain action strategies that should be controlled for as those could invalidate the underlying assumptions about the interpretation of the recorded trajectories. One example of this might be the approach of first moving the mouse a small amount in a low speed to gain more time to conclude the decision process, before carrying out the associated motion directly afterwards. A way of controlling for these strategies would be to examine velocity profiles (Stillman, 2018) or individual trajectories (Kieslich, 2018).

Further, the functionality of mouse tracking limits the flexibility of available tasks that can be investigated with the methodology. For it to collect data properly, mouse tracking requires tasks that are pre-structured in a very specific way and only support limited content in terms of stimuli (Schulte-Mecklenbeck, 2011). The method is best suited for experimental paradigms that feature easily perceivable response options and texts that neither take long to read nor to comprehend. This is due to the fact that the time needed to process and/or comprehend these factors would weaken the direct relations between the underlying cognitive processes related to decision making and the execution of the motor response. Additionally, mouse tracking only supports the making of explicit choices and in most cases is restricted to choices between two opposing options. As a result, spontaneous or in-between options between two categorisations are not well examinable with mouse tracking, even though in real life categorisations are seldom purely binary (Stillman, 2018). Due to this reduced naturalness in experimental set-up, the flexibility of mouse tracking is greatly reduced when compared to, for example, verbal reports (Magnuson, 2005; Schulte-Mecklenbeck, 2011).

Lastly, one should be careful with interpreting cognition as a trajectory starting in a neural space, which begins moving towards given response options that each with their attraction forces influence the process, before the choice finally settles on one of them. This point of viewing cognition is far too simplistic and unnatural, as in real life stimuli and potential responses occur seldomly as isolated as they do in experimental mouse tracking paradigms (Spivey, 2006). It is still unclear, which specific cognitive processes can directly be indexed through hand and arm movements and the derivative measures associated with them, as well as to which other factors might influence them. Especially the role of attention on this is something that still needs to be researched more thoroughly in the future (Freeman, 2011).

2.3. Confidence as a Measure of Metacognition

So far, the focus of this thesis has mostly been on cognitive processes that are related to judgement and decisions and how they are enacted in behavioural responses. But there is also a whole other class of cognitive processes that are related to those: so-called metacognition. Metacognition can be described as the ability to think about one's own thoughts or as the knowledge one has about one's own cognitive processes (Kepecs, 2012; Grimaldi, 2015). It is a specific aspect of consciousness, which enables one to observe continuous perceptual processes and their quality in an introspective manner. As such, metacognition is an integral part not only of learning and memory, but also of planning and making decisions. Further, metacognition can either be described as prospective or retrospective. The former refers to predicting what information might be accessible from one's memory in the future, while the latter judges whether or how successful an item related to a past experience has been committed to memory. While there are many aspects of metacognition, the present study will only focus on perceptual decisions, and more specifically on confidence in those. Decisions can be defined as the commitment to a proposition or planned action, which is based on gathered supporting evidence, as well as the expected costs for the associated outcome. Over time, evidence gets accumulated until a certain threshold (known as the 'decision bound') is reached, which determines the initial decision. Even after one has come to an initial decision, more evidence is gathered and processed and used to ultimately either re-affirm or reverse the initial decision. To the most part, such reversals serve the purpose of correcting initial errors. The process of accumulating evidence over time is able to explain not only the time needed to come to a decision in the first place, but also how decision accuracy is affected by varying levels of difficulty, which are represented through varying amounts of evidence (Resulaj, 2009).

A central example of metacognition are so-called confidence judgements. In general, confidence can be defined as someone's degree of belief that a given decision, proposition or action is correct or that a piece of information, be it retrieved from either observation, prediction, or memory, is reliable given the accumulated evidence (Kepecs, 2012; Kiani, 2014; Grimaldi, 2015; Meyniel, 2015; Pouget, 2016; Dotan, 2018). Intuitively, it can be described as a "feeling-of-knowing" (Meyniel, 2015). A specific aspect of confidence is decision confidence, which can be defined as someone's subjective estimate that a given decision that they made is correct, or more formally as the ability to approximate the likelihood of a decision being correct (Kepecs, 2012; Dotan, 2018). Making confidence judgements requires a reflective cognitive process that is highly subjective, yet humans intuitively and seemingly automatically assess their confidence in what they are perceiving, deciding, and remembering on a constant basis. Further, research even suggests that not only humans, but also other animals might have this intrinsic sense of confidence in the decisions they make (Kepecs, 2012; Grimaldi, 2015; Meyniel, 2015; Dotan, 2018).

Research suggests that confidence is computed on-line during the process of making decisions while evidence is accumulated that may support the decision (Kiani, 2009; Meyniel, 2015; Dotan, 2018). Computational models, such as by Pouget et al. (2016), suggest that the human mind may be able to compute full probability distributions of a certain assumption or decision being correct, which might aid one in withholding decisions at a given time in favour of looking for more supportive information (Meyniel, 2015; Dotan, 2018). Continuing this further, confidence would act as an estimator as to how dispersed this probability distribution is. A low level of confidence would indicate a farther spread distribution, while a high level of confidence would signify that the distribution is more concentrated around the estimated value. Therefore, in more computational terms, confidence could be formalised as the inverse of variance or as the precision of a distribution (Meyniel, 2015).

2.3.1. Confidence Ratings

Already since the 1980s, confidence ratings have been used in the field of consciousness research. Besides this initial field of use, nowadays they are employed in a wide variety of research fields, such as educational research or memory research (Norman, 2015). Amongst others, they are commonly used in implicit learning and psychophysical discrimination tasks and are used to assess the degree of correspondence between one's objective performance and subjective confidence (Kepecs, 2012; Norman, 2015). Usually, confidence in perceptual decisions is measured by using retrospective self-reports. In this case, immediately after a judgement or decision has been made, participants will give their confidence rating in the form of a numerical value that is indicative of how sure they are in their decision. Therefore, the rating must be based on the participants' memory of their initial response (Kepecs, 2012; Grimaldi, 2015; Norman, 2015).

Confidence ratings are usually obtained as follows: The accuracy of participants' performance is recorded during a so-called "Type-1 task", which could be a perceptual discrimination task, for example. After a trial has been completed, participants are then presented either with a question such as "How confident are you that your response is correct?" or merely the confidence scale, where they are supposed to choose the scale value that they most identify with. This assessment of their confidence in their performance is known as the so-called "Type-2 task". It is crucial that participants receive clear instructions both as to how they are supposed to conduct the main experimental task, as well as to how they are supposed to conduct their confidence ratings (Norman, 2015). There are various options for researchers when it comes to choosing a confidence scale. The most common alternatives are either simple binary scales, for example with scale levels such as "guess" and "sure", discrete multiple-alternative scales that typically offer something between four and six alternatives, or percentage scales that may offer scale values from either 0% or 50%, which both indicate complete uncertainty in their respective scale, ranging to 100%, which signifies complete certainty (Grimaldi, 2015; Norman, 2015).

As confidence ratings reflect someone's conscious experience, the correspondence between the performance achieved in the primary task and the confidence ratings given in the secondary task can be used to deduce to which extent the performance of the primary task is consciously mediated (Norman, 2015). Generally, accuracy of the primary task is significantly correlated with self-reported confidence ratings. This means that people are more likely to be confident in their decisions when they are correct (Kahneman, 1982). In case that a multi-trial approach is possible, confidence ratings should be gathered for each individual trial, as this increases the accuracy of correspondence as compared to assessing the average confidence to the average rate of performance (Norman, 2015).

A high degree of across-trial correlation between accuracy and confidence ratings can indicate a high reliability in the confidence measurement or, in other words, a high metacognitive sensitivity (Sandberg, 2010; Kepecs, 2012; Grimaldi, 2015). In addition to performance accuracy, self-reported confidence is also well correlated with response times of choice selection (Kepecs, 2012). Here, one can find an inverse correlation between response times and confidence ratings for both correct and incorrect trials. However, for incorrect trials response times are longer than for correct trials and confidence ratings also tend to be accordingly lower (Kiani, 2014).

Nevertheless, it should be noted that confidence ratings are not guaranteed to be accurate measures of metacognition, as they can be influenced by many factors, such as social pressure for example, and heavily rely on introspection, which humans tend to be not very good at. Further, the participants' reporting style can limit the interpretation of the correlation strength. This applies both to very liberal participants, who constantly rate their confidence as very high, as well as to very conservative participants, who constantly rate their confidence as low (Grimaldi, 2015). In general, it can be problematic to ask participants to rate their own performance or confidence, as that they might use different personal criteria to come to that decision, which makes confidence ratings not well comparable in between subjects (Sandberg, 2010).

2.3.2. Confidence Measures in Mouse Tracking

Due to the aforementioned shortcomings of assessing confidence through retrospective self-reports, some alternative methods have been introduced to do it in their stead. One possibility might be to infer confidence through certain aspects of mouse trajectories.

It has been well-established in classical paradigms which use confidence ratings that decisions participants feel more confident about are associated with shorter response times (Kiani, 2009 & 2014). Building on this evidence, Dotan et al. (2018 & 2019) suggest that it might be possible to use mouse trajectory information to infer subjective confidence indirectly. According to them, the movement deviation relative to the response options should reflect the course of the ongoing decision, while the movement speed towards the ultimately chosen option should reflect the momentary build-up of the subjective confidence that the decision is correct.

Mouse tracking allows for investigating detailed velocity profiles for each trial, as not only cursor positions are tracked in real-time, but also the momentary speed and acceleration related to them (Freeman, 2010; Hehman, 2014). The continuous movement directions, that can be derived from the profile of x- and y-coordinates over time, should be a reliable indicator of the amount of the accumulated evidence in favour of the ongoing decision. In other words, as over time more evidence is accumulated in favour of a response option over another, the more the movement will deviate from the neutral starting position towards the position representing the favoured option. Further, movement speed seems to be a good indicator of confidence levels. Positive correlations have been found for both final speed and confidence reports at the end of a trial. Further evidence suggests that speed fluctuations within a single trial, so an increased number of acceleration and deceleration bursts, are a reflection of lower degrees of subjective confidence. The idea behind this is that when participants feel like they need to collect more evidence in order to come to a decision, they will feel momentarily insecure, which should be reflected as a short period of decreased movement velocity, followed by a burst in acceleration when sufficient evidence has been gathered. Conversely, fewer bursts of acceleration and deceleration reflect higher levels of confidence, which suggests an inverse relationship between these factors. Furthermore, speed and confidence seem to be modulated by the same factors. If greater amounts of evidence in favour of a decision are present, both speed and subjective confidence levels will increase (Dotan, 2018 & 2019).

There are a few advantages of using mouse tracking in order to assess confidence. To begin with, while classical self-reports only work retrospectively, mouse tracking has the ability to measure subjective confidence levels in real-time. Resulting from that, it can detect within-trial changes in confidence through the inspection of velocity profiles. This also includes levels of pre-decision confidence, so the changes in subjective confidence before a participant decides on a response option. To this point, mouse tracking is the only method that is able to offer this real-time assessment of the evolution of confidence (Dotan, 2019).

Next, mouse tracking measures confidence indirectly by inferring it from mouse speed. This can be advantageous over explicit self-reports, as it does not require any explicit instructions (Dotan, 2019). Additionally, it is independent from using any of the different confidence scales that were previously described. As a result, it avoids the issue of participants using different rating criteria in order to assess their confidence level, which means that it might prove to be a more reliable and/or comparable measure than classical confidence ratings. Lastly, Dotan et al. (2018) argue that mouse tracking's implicit way of measuring of confidence means that it can be applied to any decision task, which allows for a broad range of application. This way, confidence could even be researched in young children or in animals.

However, since mouse tracking only measures confidence implicitly through movement velocity, it can be questioned how significant the relation between mouse speed and subjective confidence is, as they only seem to be indirectly related through the amount of the present evidence (Dotan, 2018). In addition to that, one must be mindful about a potential confound between speed and deviation. Increased movement velocity may cause a larger deviation from the middle of the screen, without accurately displaying the amount of accumulated evidence in favour of the response option one is headed towards (Dotan, 2019).

2.4. Original Study by Dale et al. (2007)

In the field of Cognitive Psychology, it has been researched extensively how knowledge is represented in the human mind. Whenever one encounters an object or a living being, the mind is tasked with identifying what it is facing. During the process known as 'categorisation', one subconsciously tries to find a fitting overarching category for the present exemplar to determine its identity. Many contending theories have been developed over the past decades in an attempt to answer how categorisation ensues. A notable account was proposed for example in Rosch and Mervis' seminal study from 1975. They found that the more features an item has in common with other exemplary items of the category it belongs to, the more typical this item is regarded for said category. In this sense, typicality can be seen as measure of "goodness-of-example" and can be used to predict within-category structures (Storms, 2004). Following from this, it is often argued that in order to categorise an item, it is compared to an abstraction of the ideal or average member of the possible categories – the so-called "prototype". An alternative standpoint is for example Medin and Schaffer's Context Model from 1978. According to them, when faced with a categorisation task, people do not assess how similar the item in question is to an abstract prototypical representation of the possible categories – but rather that it is compared to the memory of previously encountered items, which belong to the possible categories (Storms, 2004). Despite these different approaches, such similarity-based accounts of categorisation agree that highly typical category members are categorised faster and with fewer errors (Dale, 2007).

While a plethora of theories regarding the representation of category knowledge have been developed and discussed, the time course of such categorisation processes has long been neglected in research. In 2007, Dale, Kehoe and Spivey investigated the time course of categorisation processes and the effect the typicality of exemplars had on them. It aimed to show that during the process of categorisation, multiple options can exert graded attraction effects that affect the decision process in real-time. To this end, they conducted a mouse tracking study following a classic two-alternative forced choice paradigm, in which participants had to categorise animal exemplars into one of two taxonomic categories (e.g. "mammal" and "fish"). The manipulated variable was the typicality of the exemplars: It featured animals that were considered typical for their category (e.g. a cat for mammal), and ones that are considered atypical, since they share features with both the correct and the distractor category (e.g. a whale closely resembling a fish on a visual basis, even though it is a mammal). Prior to the study, it was hypothesised that atypical exemplars should activate mental representations of both competing categories, while typical exemplars should only activate the correct category for the most part. This additional activation of the distractor category in the atypical trials should exercise attraction effects on the mouse movement during the categorisation process, causing it to deviate more in its direction.

Dale et al. ran four experiments as part of their study. In experiments one and two, they used lexical stimuli for the animals that had to be categorised. The difference between the two experiments was in the atypical trials: In experiment one, the animal shared similarities with both category alternatives, while in experiment two the incorrect option was selected in such a way that the animal shared no similarities with it. Experiments three and four repeated these procedures, but with pictures instead of the lexical animal names.

Their main finding was that the cognitive competition that is due to the two competing categories exercises dynamic spatial attraction effects on the trajectories associated with the categorisation process. These arising effects are amplified when the alternative category represents a viable possibility, as in the case of experiment one, in which the atypical exemplars were featurally similar to the distractor category. However, even when a non-competing alternative category was used, its presence exerted a certain influence on the trajectories. When visual stimuli were used instead of lexical ones, the observed attraction affects became more robust. Especially in the atypical condition when the exemplars shared visual similarities with the distractor category, increased spatial attraction effects as well as longer total response times were detected, which indicates an increase in cognitive competition. These effects were even more pronounced than for the experiment that used only lexical stimuli. But also when comparing the two experiments that used less competitive categories, experiments two and four, stronger attraction effects were found for when visual stimuli were used over lexical stimuli. This means that competition affects motor dynamics more strongly when the alternative category shares more features with the exemplar (in accordance to the similarity-based accounts of category knowledge representation), as well as when these similarities are more 'obvious' due to the supportive nature of visual stimuli (Dale, 2007).

3. The Study

This mouse tracking study is based on and is a conceptual replication of the first experiment of Dale et al.'s study on atypical classification, with the addition of self-reported confidence ratings after each trial: Like in the original study, participants were shown the name of an animal along with two taxonomic classes and their task was to select the correct category for the animal. Some animals are regarded to be typical for their respective class while others are regarded atypical since they also share similarities with the distractor class. After each trial, participants were asked to rate how confident they felt in their decision on a 1 to 6 scale, where a rating of 1 indicated they felt "uncertain" and a rating of 6 indicated they felt "very certain".

The aim of this study is two-fold: Firstly, to conceptually replicate the main patterns that were found in Dale's 2007 study, namely that there are characteristic differences between the trajectories found for typical and atypical trials. And secondly, to investigate whether there is any covariance between the participants' self-reported confidence levels and the manual movement dynamics of response selection, such as response time and curvature of mouse trajectories.

The focus of this study was de facto not put on the role of typicality, but the categorisation of exemplars was chosen for the underlying task, since the investigation of confidence ratings requires a separate type-1 task to obtain these ratings. Further, this task had the additional benefit of the participants being less aware of the actual matter of investigation, namely confidence.

To this end, a number of five hypotheses has been formulated and pre-registered in advance[2] to conducting the study:

1. The classification of typical exemplars should exhibit a higher mean of confidence ratings than the classification of atypical exemplars.
2. A higher degree of confidence should covary with a higher ratio of correct answers.
3. A higher degree of confidence should covary with a lower mean of response times (RT).
4. A higher degree of confidence should covary with an earlier occurrence of the movement's maximum velocity (vel_max_time).
5. A higher degree of confidence should covary with a decreased mean curvature of the trajectory, as measured by the trajectories' area under the curve (AUC), which is the geometric area between the observed, actual trajectory and the direct path between the start and end coordinates.

[2] The full pre-registration can be found under https://osf.io/q8af9

3.2. Method

To investigate the interaction between mouse tracking measures and self-reported confidence judgements, Dale et al.'s 2007 study was used as a peg. A German translation of the original stimuli set containing thirteen typical and six atypical exemplars that was done by Kieslich et al. was used to accommodate for the native language of the participants.

3.2.1. Participants

Initially, it was aimed to recruit 15 participants for the study. This number was exceeded, as a group of 20 participants was able to be recruited among students of the University of Osnabrück, Germany. All participants were offered course credit in the form of experimental subject hours for their participation. The participants were German native speakers, aged between 18 and 34 (μ = 22.55, SD = 3.634) and all had normal or corrected-to-normal vision. Their handedness was assessed with a simple questionnaire prior to the conduction of the experiment. Four out of the 20 participants were left-handed, the remaining 16 were right-handed. They were instructed to only use their dominant hand to operate the mouse.

Participants were not asked to specify their gender as there is no indication found in previous studies that gender is an influencing variable when it comes to mouse tracking trajectories and therefore, assessing the participants' gender was regarded as non-relevant for the purposes of the present study.

3.2.2. Set-Up and Materials

Before the experiment started, participants were asked for some personal information (age, handedness, mother language, vision) and were asked to read and sign the form of informed consent. Afterwards, each participant was led into the mouse tracking lab room, in which they did the experiment.

The mouse tracking lab was equipped with a Dell UP2716D UltraSharp monitor, 27" 2560 * 1440 (59.8 cm * 33.6 cm), an Intel Xeon CPU E3-1245 v% @ 3.5GHz * 8 computer with 32GB RAM and a Dell MS116p optical mouse with 1000 dpi. This set-up ran on Ubuntu 16.04 LTS and used OpenSesame 3.2.8b1 to run the experiment, using the mousetrap-os 2.0.0 extension. Further, the mouse settings were manipulated via the 'xset m 0 0' command of the terminal, which disabled the mouse's dynamic acceleration.

The stimuli were presented using the OpenSesame software in full-screen mode. The task was to match the animal word stimulus, which was presented in a box located in the centre bottom of the screen, with the correct (taxonomic) category. The experiment used the same set of animal stimuli and categories as Kieslich et al.'s German replication of Dale et al.'s 2007 study. Accordingly, the available categories were "Säugetier", "Vogel", "Fisch", "Reptil", "Amphibie" and "Insekt" (= mammal, bird, fish, reptile, amphibian, and insect).
(For the full set of stimuli-category match-ups, see appendix B.)

3.2.3. Experimental Procedure

The experimental procedure closely resembled the one already implemented in the original study by Dale and colleagues. Like in the replication by Kieslich et al., German stimuli were used. The biggest differences in procedure to the one used by Kieslich are the addition of another set of practice trials where participants could familiarise themselves with the usage of the mouse and the general experimental set-up, and the addition of the participants' self-report of their confidence following each trial.

Before the experiment began, participants were informed that their mouse movements would be recorded, however, they received no specific instructions about how to operate or move the mouse. Further, they were neither informed about the expectations of the conditions, nor about any of the hypotheses that were to be investigated with the study.

Each trial started with a 1,000ms interval in which only a blank screen was presented. Afterwards, the top left and right corners of the screen displayed a category each within a box. The order of those categories was randomised at run time. After 2,000ms the start button was presented in the bottom centre of the screen and the trial was started after the participant clicked on it. The cursor position was not reset. Directly following that, the stimulus word that was supposed to be categorised appeared above the start button. The participants' response was indicated by hovering the mouse over the box of the chosen category without having to click on it.

First, participants were instructed to familiarise themselves with the experimental set-up and the usage of the mouse. To that end, after clicking on the "Start" button in the bottom centre of the screen, they had to move the mouse to one of the two displayed boxes in the top left or right corner, which was indicated as "Ziel" (= goal), while they were instructed to ignore the other box, which was marked "---". This was repeated three times. After that, they had three further test trials, which matched the actual experimental set-up: Now, the boxes in the top corners of the screen featured taxonomic categories, while the start button changed to the name of an animal after clicking. Then, the participants had to move the mouse to the box matching the animal's category.

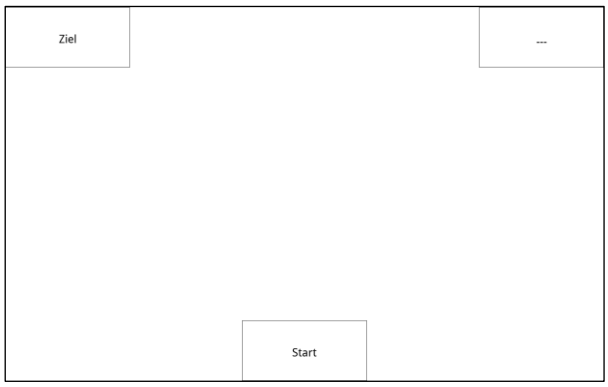


Fig. 1 *Experimental set up as seen on the screen during the first practise phase*

After completion of these trials, the experimental trials began, which shared their exact set-up with the previous test trials. There were 19 trials and the only difference was in the used exemplars, such that animals that already appeared in the practice trials did not re-appear in the experimental trials. The order in which the exemplars appeared was randomised, as well as the position of the presented categories was randomised at run time.

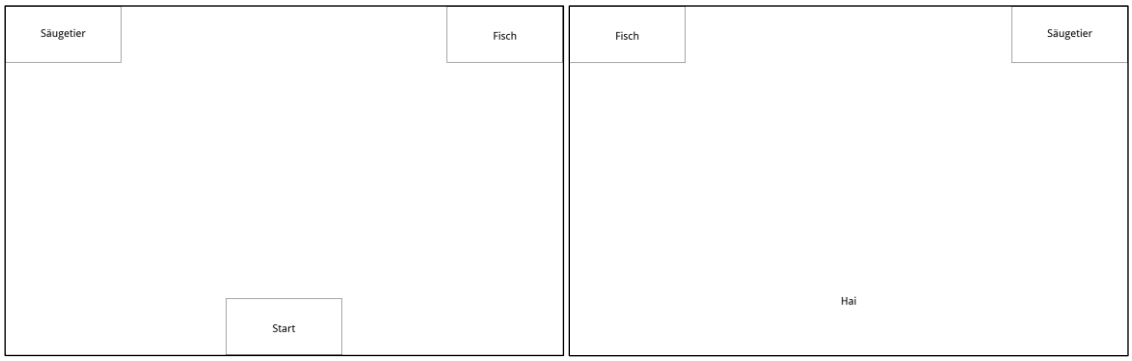


Fig. 2 *Experimental set up as seen on the screen during an experimental trial*
2a) *Before beginning the trial (by pressing the 'Start' button);* ***2b)*** *After beginning the trial*

Finally, the participants were asked to rate their subjective confidence in the correctness of their decision on a scale of 1 to 6, where the value 1 denoted feeling "Unsicher" (= not confident) and 6 denoted feeling "Sehr sicher" (= very confident).

Fig. 3 *Experimental set up as seen on the screen after completing an experimental trial during the rating of subjective confidence*

3.3. Data Analysis

For the following analyses, R version 3.6.1. was used alongside with the tidyverse version 1.2.1 as well as the mousetrap version 3.1.3. and readbulk version 1.1.2. packages by Kieslich et al.

First, the raw data was imported and transformed into a mousetrap object to utilise the various analysis functions provided by the mousetrap package. There were three main steps of data pre-processing. First, data from the practise trials and data resulting from incorrect experimental trials was removed, following the procedure as applied by both Dale et al. and Kieslich et al. Second, spatial normalisations were applied to the trajectories. For that, the trajectories were re-mapped, such that they end on the same hemispace of the screen and share the same general direction, namely diagonally up and left. Additionally, the trajectories were adjusted, such that they shared the same starting position of (0,0). These remapping steps enable an easier comparison of mouse trajectories across different conditions.

However, it should be noted that these steps are only applicable in the case of a typical two-alternative forced choice design, as the used functions assume a coordinate system that is centred on the screen centre, as well as that the response buttons are equally distant from the screen centre. Third, the trajectories were subjected to temporal normalisations by diving them into a number of equally sized time steps. By default, mouse positions are recorded every 10ms, which corresponds to a sampling rate of 100Hz. However, since trials can differ significantly in their length, this also means that they will differ in their respective number of recorded cursor positions, which impedes proper statistical comparisons. After time-normalisation all trajectories have an equal number of recorded positions[3], which allows for averaging over trajectories and comparing the cursor positions at relative time points across conditions that differ in their overall response times. Lastly, the normalised data was reshaped so it would accurately represent the additionally gathered confidence measures, which were not supported by the original data structure of the used package.

Following the original study, the manipulated variable was the typicality of exemplars. The two levels of this categorical variable were "typical" and "atypical". The measured variables were the correctness, response times in milliseconds and confidence ratings on a scale of 1 – 6 for each trial, in addition to the associated mouse trajectories that recorded x- and y-coordinates in z time steps. From that the measures of vel_max_time and AUC were derived.

[3] Following Spivey et al.'s 2005 study, a number of 101 time-steps is used as a default.

To get a first overview over the effects present in the experiment, the time-normalised trajectories were aggregated for each condition and then plotted for visual inspection. Thereafter, summary measures in the form of the mean and standard deviation were computed for total response time, the time at which the maximum velocity occurred first (vel_max_time), the area under the curve (AUC) as a measure for the trajectory curvature, and the confidence rating for the two conditions, as well as for each exemplar. Both these summary measures, as well as the raw data distributions for the aforementioned variables were plotted. Additionally, the ratio of correct answers was computed, again aggregated over exemplars. Further, covariance and Bravais-Pearson correlation were computed for the interaction between confidence ratings and correctness, RT, vel_max_time and AUC, respectively.

3.4. Results

Pre-Registered Data Analysis

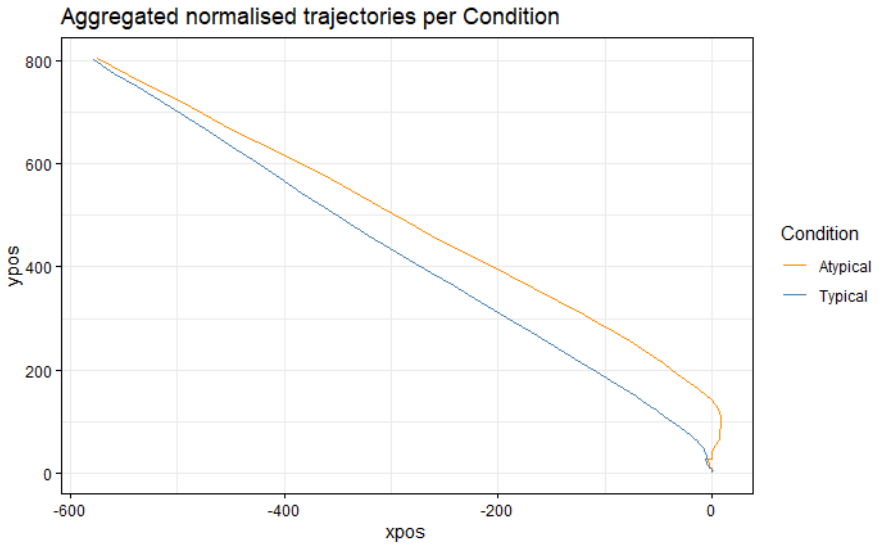


Fig. 4 *Visual inspection of x- and y-coordinates of aggregated normalised trajectories for the atypical (represented in dark orange) and typical (represented in blue) conditions.*

To get a first overview over the present condition-level effects, a visual inspection of the aggregated normalised trajectories was conducted. It is visible that the average trajectory recorded from typical trials differs from the one obtained from atypical trials: The aggregated mouse trajectory for typical exemplars seems to be very direct, or in other words seems to have a lower curvature towards the non-chosen option. Its general movement direction is left-upwards for its entire duration. For the atypical condition, however, this is not the case. Here, the average movement exhibits a brief right-upwards movement into the right hemispace of the screen, before changing to a left-upwards movement.

After the visual inspection, the hypotheses were first tested by simply comparing the means of the investigated measures.

(For plots of the data distributions of these measures, see appendix C.)

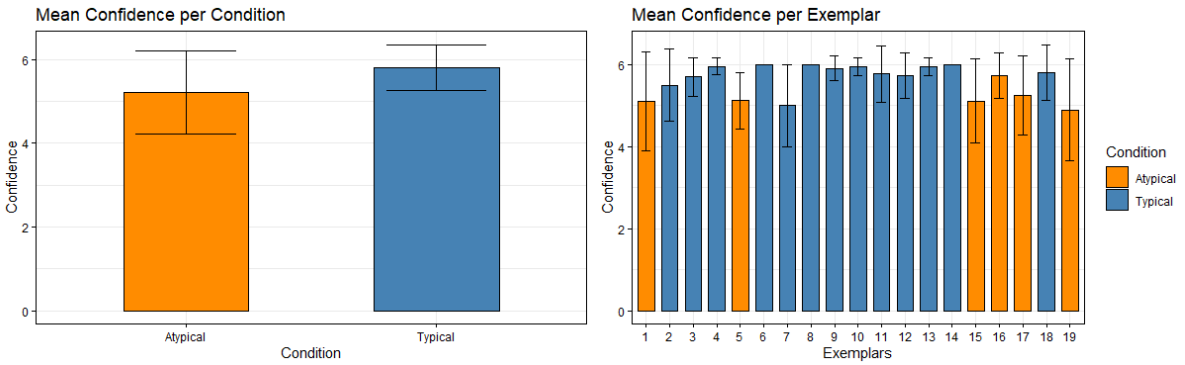


Fig. 5 *Mean confidence rating on a scale of 1 – 6 with standard deviations for the atypical (= dark orange) and typical (= blue) conditions;*
5a) *Mean confidence ratings for both conditions;* ***5b)*** *Mean confidence ratings for all exemplars. A list of which exemplar corresponds to which number can be found in* appendix B.

The average mean of confidence ratings on a scale of 1 to 6 was higher during the typical condition (μ= 5.806, SD = 0.548) than during atypical condition (μ = 5.216, SD = 0.994). Interestingly, there is one exemplar with a mean that deviates by even more by the standard deviation of if its condition, which is typical exemplar #7 "Hai" (μ = 5.0). With this value, it resembles rather the trends found for atypical than for typical exemplars. The exemplars with the highest mean confidence rating were #6 "Goldfisch", #8 "Hund" and #14 "Pferd" (μ = 6.0) and were all considered to be typical, while the lowest mean confidence rating was found for exemplar #19 "Wal" (μ = 4.895), which belonged to the atypical group.

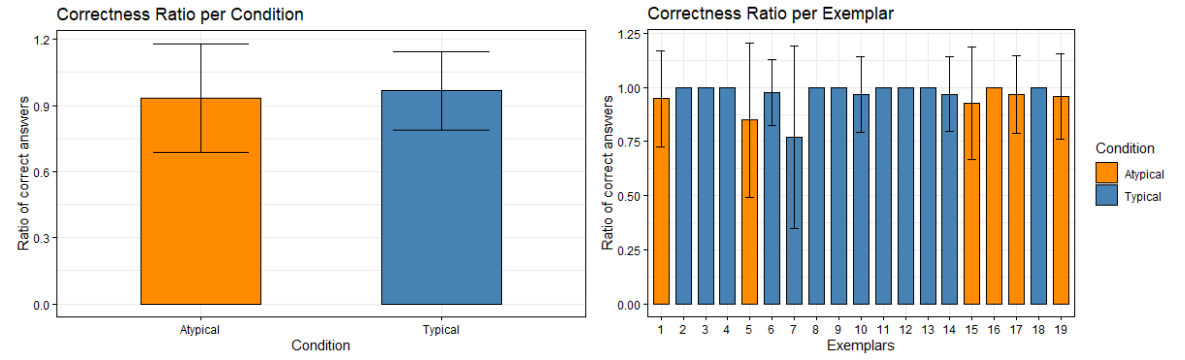


Fig. 6 *Mean ratio of correct classifications with standard deviations for the atypical (= dark orange) and typical (= blue) conditions;*
6a) *Mean correctness ratio for both conditions;* ***6b)*** *Mean correctness ratio for all exemplars. A list of which exemplar corresponds to which number can be found in* appendix B.

On average, the mean ratio of correct classifications was found to be higher for typical exemplars (μ = 0.967, SD = 0.180) than for atypical exemplars (μ = 0.935, SD = 0.247). Likewise with the mean confidence ratings, the only exemplar that differed for more than its group's standard deviation from the respective mean exemplar #7 "Hai" (μ = 0.768). With this, "Hai" achieved the overall minimum of correct classifications out of all exemplars.

Nine out of the thirteen typical exemplars[4] reached perfect scores of 100% correct classifications, while the only atypical exemplar this was achieved for was #16 "Schmetterling".

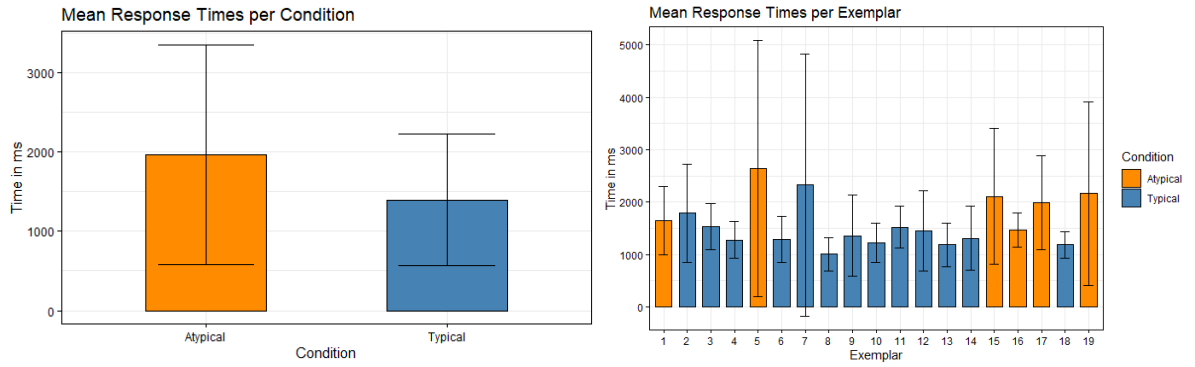


Fig. 7 *Mean response times in milliseconds with standard deviations for the atypical (= dark orange) and typical (= blue) conditions;*
7a) *Mean response times for both conditions;* ***7b)*** *Mean response times for all exemplars. A list of which exemplar corresponds to which number can be found in* appendix B.

For total response times, it was found that the mean for the typical condition (μ = 1,398.249ms, SD = 830.177) was lower than in the case of the atypical condition (μ = 1,972.966ms, SD = 1,387.260). Again, the mean RTs that were measured for exemplar #7 "Hai" (μ = 2,326ms) deviated strongly from the mean of the typical exemplars overall by exceeding it by over 900ms. The only exemplar that a longer mean RT was found for was #5 "Fledermaus" (μ = 2,637.706ms), which belonged to the atypical group. The lowest mean RTs were found for #8 "Hund" (μ = 1,007.714ms) and #16 "Schmetterling" (μ = 1,466.455ms) for the typical and atypical conditions, respectively.

[4] Those were exemplars #2 "Alligator", #3 "Chamäleon", #4 "Falke", #8 "Hund", #9 "Kaninchen", #11 "Klapperschlange", #12 "Lachs", #13 "Löwe" and #18 "Spatz".

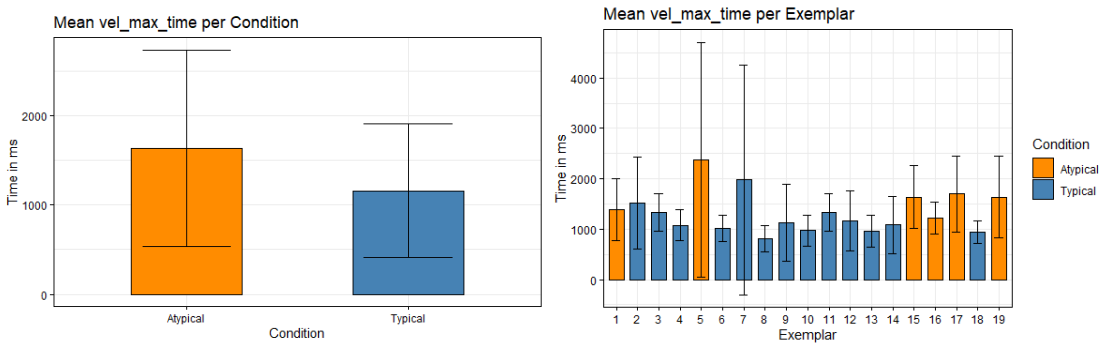


Fig. 8 *Mean vel_max_time (= time at which the maximum velocity occurred first) in milliseconds with standard deviations for the atypical (= dark orange) and typical (= blue) conditions;*
8a) *Mean vel_max_time for both conditions;* ***8b)*** *Mean vel_max_time for all exemplars. A list of which exemplar corresponds to which number can be found in* appendix B.

On average, the highest movement velocity was reached earlier during the classification of typical (μ = 1,160.506ms, SD = 750.995) than for atypical exemplars (μ = 1,634.379ms, SD = 1,104.303). The trends that were found for the mean reaction times of the exemplars are repeated here: The only exemplar that deviates more from its group mean than by the standard deviation is #7 "Hai" (μ = 1,983.334ms) and the only exemplar that exceeds this mean first occurrence of the peak velocity is the atypical #5 "Fledermaus" (μ = 2,379.471ms). Again, the lowest means were found for #8 "Hund" (μ = 815.810ms) for the typical exemplars and #16 "Schmetterling" (μ= 1,222.818ms) for the atypical exemplars.

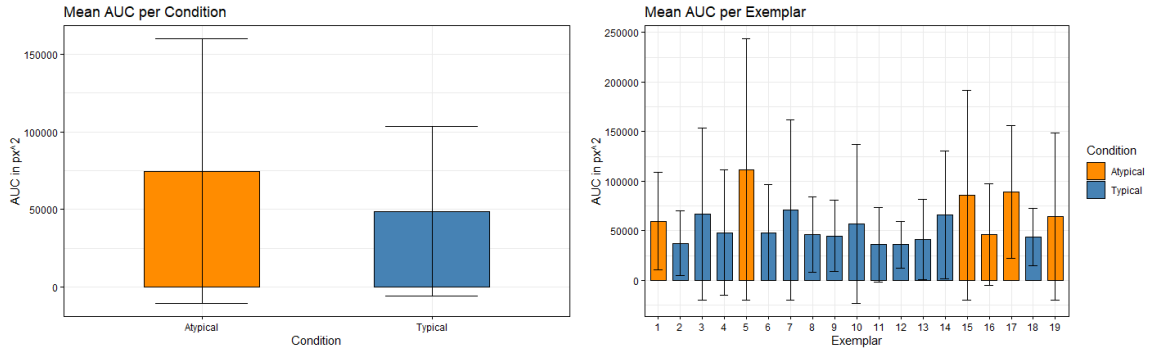


Fig. 9 *Mean absolute area under the curve (AUC) in squared pixels with standard deviations for the atypical (= dark orange) and typical (= blue) conditions;*
9a) *Mean absolute AUC for both conditions;* ***9b)*** *Mean absolute AUC for all exemplars. A list of which exemplar corresponds to which number can be found in* appendix B.

While the visual inspection of the aggregated and normalised trajectories already revealed that trajectories showed stronger attraction effects towards to the non-selected distractor category, in order to statistically test for differences in curvature the average AUC was compared across conditions.

Here, it was found that the mean absolute AUC during typical trials (μ = 48,734.451px^2, SD = 54,619.218) was lower than during atypical trials (μ = 74,642.259px^2, SD = 85,198.427). Similar to RT and vel_max_time, the highest means for their respective conditions were found during trials #5 and #7, during which are "Fledermaus" (μ = 111,783.853px^2) and "Hai" (μ = 70,689.234px^2) had to be classified. The lowest mean absolute AUC of an atypical exemplar was found for #16 "Schmetterling" (μ = 45,950.136px^2). For the typical exemplars, the lowest mean was found for #12 "Lachs" (μ = 35,919.659px^2). Further, no exemplar exceeded its group mean by more than the standard deviation with regard to this measure.

However, there were some exceptions to the previously made hypotheses and discovered trends: For example, trials associated with the classification of typical exemplar #7 "Hai" displayed a higher mean absolute AUC than those, in which the atypical exemplar #16 "Schmetterling" had to be classified (μ(Hai) = 70,689.234px^2 > μ(Schmetterling) = 45,950.136px^2). Even further, the mean AUC of "Schmetterling" was lower than the average mean AUC of the typical category (μ = 48,734.451px^2). In fact, two atypical exemplars, namely "Schmetterling" as well as exemplar #17 "Seelöwe", both exhibited means for their ratio of correct answers (μ = 1 and μ = 0.966) that were at least as high as the mean of the typical condition (μ(Typical) = 0.966). Lastly, all measures that were recorded for the typical exemplar #7 "Hai were closer to the mean of the atypical condition that of the typical condition, to which it de facto belonged.

Following the comparison of means, covariance and correlation coefficients were computed between confidence ratings and correctness ratio, RT, vel_max_time and AUC, respectively. Covariances and correlations were computed three times for each match-up of measures: independently for each of the two conditions and the weighted average of both. (The full covariance and correlation matrices can be found in appendix B.)

Without differentiating between the two levels of typicality, virtually no covariance could be determined between self-reported confidence and correctness of a trial (cov = 0.016). Both response times and vel_max_time achieve a moderately negative covariance with confidence, reaching values of -158.336 and -132.906 respectively. Of the investigated measures, curvature as approximated by AUC shows the strongest covariance with confidence with -5,040.721. This means, the higher someone's confidence was in their decision within a trial, the lower the measured response time, earliest occurrence of peak velocity and trajectory curvature were on average.

When computing covariance coefficients separately for the typical and atypical condition, similar trends are found, though less pronounced. The covariance between confidence and correctness is twice as strong (though still vanishingly low) for typical (cov = 0.014) than for atypical trials (cov = 0.007). Response time covaries with confidence by cov = -89.873 in the typical and by cov = -71.815 in the atypical condition. In case of vel_max_time, the coefficients were cov = -81.188 and cov = -50.301 for the typical and atypical condition, respectively.

The previously observed trend of stronger covariances in case of the typical condition is reversed for the measure of AUC: Here, the covariance with the confidence ratings is cov = -1,009.889 for the aggregated typical condition and cov = -2,767.135 for the atypical condition. The majority of these findings is in line with the predictions that were made before sighting the data as stated in the preregistration. The only exception is that there was no covariance to speak of found between correctness and confidence.

Since covariance is not invariant under linear transformations, it can be difficult to interpret and compare it across measures. For that reason, the Bravais-Pearson correlation (commonly abbreviated as "r") was computed for the same measures additionally. Again, it was calculated threefold: separately for each of the two conditions, as well as the weighted average of both.

When looking at the two conditions separately, three of the four measures showed stronger correlations in typical trials than in the atypical ones. The correlation coefficients between confidence and correctness were r = 0.777 for typical and r = 0.482 for atypical trials. In the typical condition, confidence correlated with response times to a degree of r = -0.953, while in the atypical condition it was only r = -0.617. For vel_max_time, correlations were found to be r = -0.939 in case of typical trials and r = -0.449 in atypical trials on average. Analogous to what was discovered in the analysis of covariances, this pattern is reversed for the relation between confidence ratings and AUC: Here, stronger correlations were detected for the classification of atypical exemplars (r = -0.410) than for typical exemplars (r = -0.298). This data is in line with the preregistered hypotheses.

Since the main research question of this study was to investigate if introspective confidence judgements are generally reflected in mouse tracking measures, from now on the main focus during further analyses and discussions will be put on the weighted average of correlation coefficients. The correlation coefficient between confidence and correctness was found to be r = 0.661, indicating a moderate positive relationship between these measures. This is interesting compared to the lack of covariance that was found for the same measures. A strong negative correlation was found between confidence and response time (r = -0.902). With a correlation coefficient of r = -0.847, another pronounced negative relationship was detected between confidence and vel_max_time. Interestingly, even though AUC and confidence were shown to covary to a great degree, their correlation is only of moderate strength (r = -0.627).

This means that the higher the reported confidence during a given trial was, the faster the appropriate response was selected and the earlier the highest movement velocity was reached on average. Additionally, a trial that was given a higher confidence rating, was also more likely to be correct and to exhibit a lower curvature of the recorded trajectory. However, confidence was more adequately reflected in response time and vel_max_time than in correctness and AUC.

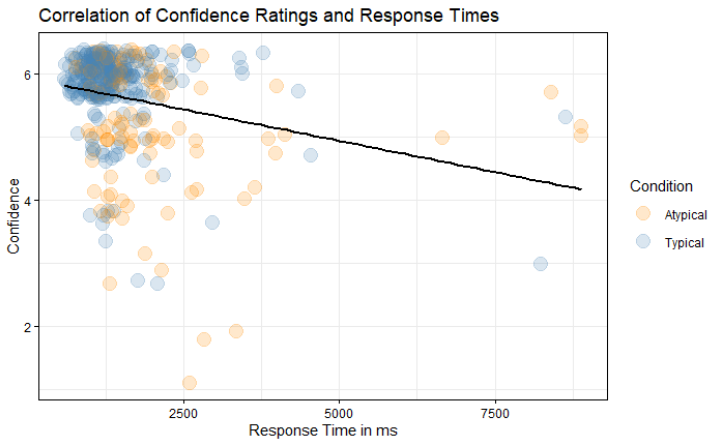


Fig. 10 *Relation between response times and confidence ratings. Data points from atypical trials are depicted in dark orange, data from typical trials in blue. The black line shows the Bravais-Pearson correlation between the two measures.*

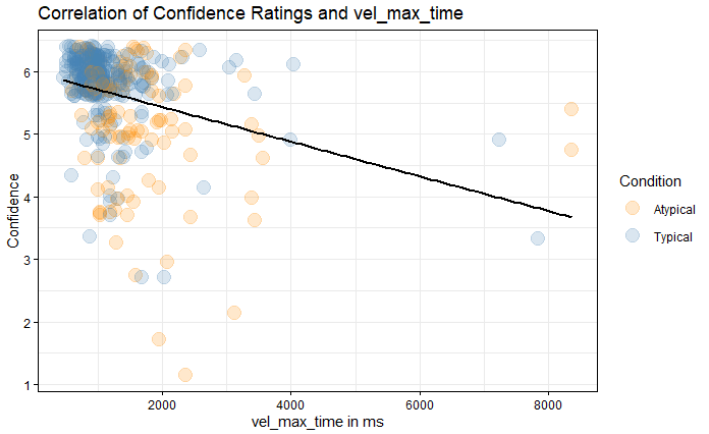


Fig. 11 *Relation between vel_max_time and confidence ratings. Data points from atypical trials are depicted in dark orange, data from typical trials in blue. The black line shows the Bravais-Pearson correlation between the two measures.*

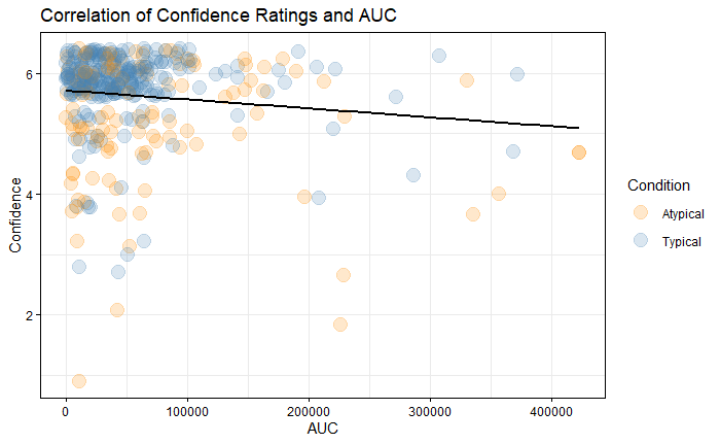


Fig. 12 *Relation between the area under the curve and confidence ratings. Data points from atypical trials are depicted in dark orange, data from typical trials in blue. The black line shows the Bravais-Pearson correlation between the two measures.*

Exploratory Analyses

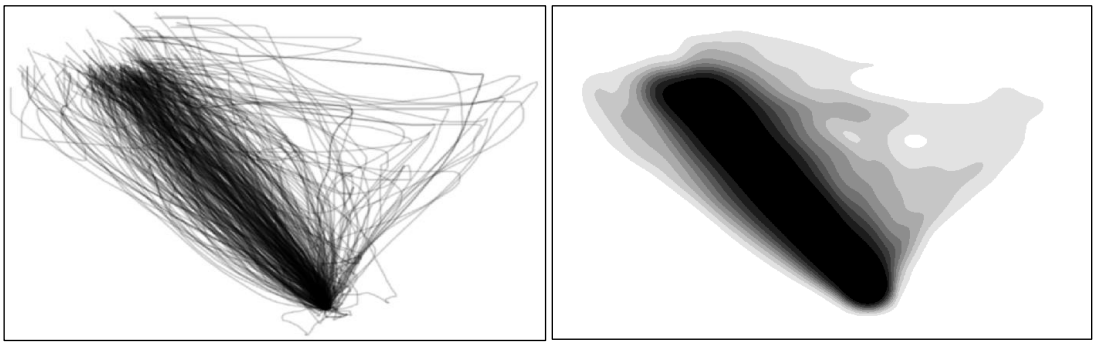


Fig. 13 *Heatmaps of all recorded trajectories;*
13a*) Plot of all individual normalised trajectories, not differentiating between conditions;* ***13b)*** *Smoothed heatmap of individual trajectories, not differentiating between conditions*

When inspecting the entirety of the correct trajectories, without differentiating between typical and atypical trials, it appears that the majority of individual trajectories follow a direct path from the starting position to the selected option. Most trajectories that diverge from this direct path do so in the direction of the non-selected option, so towards the upper right. Only few trajectories seem to deviate toward the bottom left. These patterns get especially clear when looking at the smoothed heatmap: The majority of trajectories is concentrated along the most direct path.

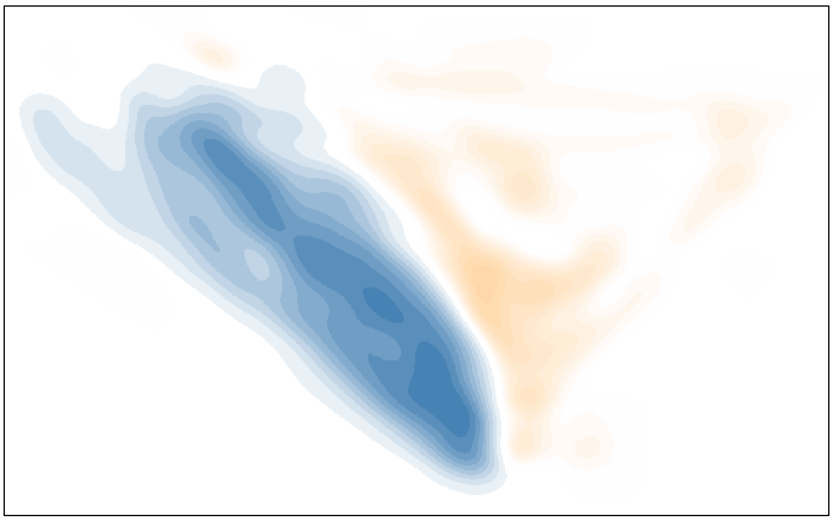


Fig. 14 *Smoothed difference heatmap, in which the density of a trajectory type per condition determines the colouring. Dark orange indicates higher densities within the atypical condition, blue indicates higher densities within the typical condition and white indicates comparable densities.*

The difference heatmap allows to compare trajectory densities between conditions, as it reveals the relative occurrence of trajectory types per condition. The darker the colouring, the higher is the density of that particular trajectory type in the respective condition: Orange indicates the atypical condition, blue the typical condition and white signifies a comparable density in both conditions.

The majority of trajectories that are concentrated along the direct line connecting the starting point in bottom centre and end point in the top left were recorded in the typical condition , which is visualised by a large number of dark blue clusters along that line. Further, typical trajectories were not the most prevalent type of trajectories at any locations that are not closely adjacent to the most direct path between the start and end points, which can be deduced from a lack of blue coloured clusters in the rest of the heatmap. However, that does not mean that no typical trajectories were detected in other locations, but merely that atypical trajectories occurred there at least as often. Locations coloured in white were part of trajectories of both conditions with equal probability densities, while the orange coloured clusters were predominantly for atypical trajectories. Like the plot depicting the aggregated normalised trajectories for both conditions, the difference heatmap shows that atypical trials exhibited stronger attraction effects towards the upper right: The direct path connecting the start point and the non-selected response options exhibits a small accumulation of orange clusters, and the remaining clusters indicating atypical trials are located between this and the blue clusters that are concentrated along the direct path towards the correct response option.

4. Discussion

The present study investigated the covariance between subjective confidence judgements and various mouse tracking measures related to the motor dynamics of response selection in a classic two-alternative forced choice paradigm. The underlying experiment was based on Dale et al.'s 2007 study on the effect of exemplar typicality on categorisation processes. As a result, the aim of this study was two-fold: Examining if the same patterns of systematic differences of trajectories can be found between typical and atypical trials that were discovered in the original study, and inspecting if self-reported confidence levels are reflected in the movement dynamics during the respective decision process.

The conceptual replication of Dale's original study is based on the idea that the categorisation of typical exemplars should proceed in a faster manner and with fewer errors. The present study found the same general trends of results from the original study, namely that there are significant differences between the trajectories found for typical and atypical trials. Generally, trajectories that were recorded during the classification of atypical exemplars demonstrated stronger attraction effects towards the unselected distractor category than trajectories of the typical condition. As mentioned beforehand, already the initial visual inspection of the aggregate normalised trajectories revealed a higher degree of curvature for the atypical condition, which can be interpreted as an indicator for higher attraction effects, towards the unselected response option. This impression is further supported by the generated difference heatmap. Apart from the strong concentration of typical trajectories along the most direct path between the start and end points, a high density of atypical trajectories was detected with a lower distance to the unselected response option compared to trajectories associated with typical trials. This is expressed by more locations that lie either along the central axis or even to the right of it exhibiting a dark orange colour, which signifies a higher overall density of present trajectories that were recorded during the atypical condition.

This tendency was also present in the statistical analysis: The atypical condition has a higher absolute mean AUC on average when compared to the typical condition. When comparing the absolute means of the AUC between exemplars across both conditions, this also generally holds. Curvature in general, and therefore in extension also AUC, is assumed to be modulated by the differences in activation between the chosen and non-chosen response option. If one response option is strongly activated, while the other only exhibits a mild or moderate activation, the associated trajectory tends to be a mostly direct straight line towards the strongly activated option. In contrast to that, if there is only a small difference in activations, this would lead to a trajectory with a stronger curvature (Spivey, 2010; Kieslich, 2018). Therefore, it is possible to assess the degree of attraction towards an unselected response through AUC and thus, it can serve as an index for the magnitude of activation of response options over time (Freeman, 2010; Hehman, 2014; Stillman, 2018; Maldonado, 2019).

Higher values for the atypical than for the typical condition and exemplars were also found for the measures of RTs and vel_max_time, while it is reversed for the correctness ratio. All of this is in line with the pre-registered hypotheses and the trends of findings from Dale's study.

Following similarity-based accounts of category knowledge representation (e.g.: Rosch & Mervis, 1975; Dale, 2007), exemplars that exhibit a higher typicality for their respective category should be categorised faster than exemplars that are considered to be atypical. This is most likely due to a decreased amount of time that is needed to gather sufficient evidence during the deliberation process. In addition to that, the categorisation of highly typical exemplars should be more likely to be correct – or in other words, typical trials should have a higher ratio of correct answers as when compared to atypical trials. The findings of this study are in line with this: On average, typical trials had lower RTs and a higher ratio of correct answers compared to atypical trials.

Further, the movement was found to reach its maximum velocity at a later point in time for atypical trials than for typical trials. Velocity peaks can be seen as indicators for both the point in time, as well as the degree to which a participant commits to a response. Competing response alternatives inhibit each other during early stages of the processing related to response selection. Upon the resolution of this competition (at which one decides in favour of one of the options), this initial inhibition is relieved, which leads to a rapid increase of activation of the previously inhibited option. This means that a strong competition between response options goes along with a decreased initial velocity, which is only increased when one comes to decision and the inhibition is alleviated (Hehman, 2014).

Summing up, as predicted on the basis of Dale's 2007 study, typical trials were found to feature higher ratios of correct answers, as well as lower RTs, vel_max_time and AUC when compared to atypical trials on average. However, this study also investigated a factor that was not present in the original study: subjective confidence and in how far it is reflected in the aforementioned mouse tracking measures.

Before the study was conducted, it was hypothesised that the average mean of confidence ratings should be higher for typical than for atypical trials. Again, this was based on similarity-based accounts of knowledge representation, as proposed for example by Rosch and Mervis (1975). This prediction was met by the acquired data. Further, it was hypothesised that confidence should covary with correctness ratios, RTs, vel_max_time and AUC, such that trials participants felt highly confident about should covary with higher ratios of correct answers, as well as lower values for RTs, vel_max_time and AUC when compared to trials that were rated lower on the confidence scale.

During the investigation of covariances, the predictions were only partly met: The confidence ratings and correctness associated with a trial show effectively no covariance. Both RTs and vel_max_time possess moderate negative covariances with confidence, with it being more pronounced in the case of RTs. AUC was found to have a very strong, and again negative, covariance with the subjective confidence ratings. The fact that the predictions were only met for RTs, vel_max_time and AUC, but not for correctness, is most likely due to covariance not being a normalised measure and it not being invariant under linear transformations. This means that the magnitude of covariance is dependent on the magnitudes of the variables that are being put into relation. Since covariances were always calculated between confidence and another measure, the range of values the latter can take should strongly influence the covariance overall. By definition, the correctness ratio can only take values from 0 to 1 (as it represents the percentage of correct trials in decimal form), while by definition the absolute AUC ranges between values from 0 to a maximum of 320,000px^2. RTs and vel_max_time both show moderate values of covariance to confidence ratings that are quite comparable to each other in magnitude since both of these measures operate on a similar scale of values.

In brief, in the present study correctness only negligibly covaries with confidence, which is opposed to what was hypothesised prior to conduction. However, the remaining predictions were met as negative covariances were found between confidence and RTs, vel_max_time, as well as AUC. This means, a trial that received a high mean of confidence ratings generally was executed faster, the maximum velocity of the movement was reached earlier, and its associated trajectories possessed a decreased curvature.

As already mentioned in the results section, the Brevais-Pearson correlation coefficient was computed additionally, due to covariances not being invariant under linear transformation and thereby impeding comparison across exemplars and conditions. Here, the data matched the predicted trends: Correctness was found to be positively correlated to confidence rating with moderate strength, while RT, vel_max_time and AUC all were negatively correlated to confidence with varying magnitudes. Like correctness, AUC was only moderately correlated with confidence, while both RT and vel_max_time showed strong correlation coefficients. Therefore, trials that received a higher confidence rating also had a higher ratio of correct answers, while their associated RTs, vel_max_time and AUC were lower on average. So, trials were completed faster, the maximum velocity was achieved earlier in the trial, and their trajectories were less curved.

When evaluating in how far confidence is reflected in mouse tracking measures, not only the existence and direction of the correlation should be taken into account, but also its magnitude: Vel_max_time and response times were both strongly correlated with subjective confidence ratings, with coefficients of $r = -0.847$ and $r = -0.902$, respectively. Therefore, in the present study these two measures were found to be well reflective of the participants' subjective confidence ratings in an inverse relation.

The correlation coefficients that were detected for correctness and AUC, namely r = 0.661 and r = -0.673, indicate that for the present data set these two measures did not reflect confidence as well as the two time-based measures. However, these results would still be considered to be highly correlated by the standards of many researchers in the behavioural sciences. Following Cohen (1988), correlation coefficients larger than r = 0.5 can be interpreted to be large in size, while a survey conducted by Hemphill (2003) found that most active researchers in the field of psychology started considering a correlation to be large if it was at least in the order of magnitude of r = 0.6. Therefore, one could argue that even AUC and correctness reflect subjective confidence reasonably well, despite them only yielding comparatively low correlation coefficients. Even so, this effect might be due to AUC and correctness themselves being correlated to RTs. Across various studies, response times and AUC have been found to be correlated with a magnitude of approximately r = 0.4 (Stillman, 2018), while in the present study these two measures exhibited correlation coefficients of r = 0.883 for atypical trials, and r = 0.349 for typical trials, which resulted in a weighted average of r = 0.756. This large difference in correlation strength is most likely due to the large differences in magnitude of present competition in the two conditions. Therefore, the correlation coefficient between AUC and RTs that was calculated for typical trials could be best suited for making more general statements, as not all mouse tracking studies feature as strong cognitive competition as in the categorisation of atypical exemplars and as it is the closest to the value reported by Stillman et al. (2018). Response times and trial accuracy have also long been known to be related through the so-called "speed accuracy trade-off (SAT)" (Schouten and Bekker, 1967; Wickelgren, 1977). In previous mouse tracking studies, it was found that on average RTs were increased for erroneous responses and that this is also directly related to subjective confidence, as within incorrect trials participants felt more confident about faster decisions than about slower decisions (Kiani, 2009 & 2014). Consequently, one should exercise caution when it comes to considering AUC and correctness as good indexes for subjective confidence. Overall, the results of the present study should generally be regarded cautiously, as a merely descriptive approach to statistics was applied and the study does not aim to make any universally valid claims about the relations of the investigated measures.

As discussed before, mouse tracking can be used to infer information about the ongoing decision, as well as about the dynamic evolution of subjective confidence. It has been established that the former is represented by the movement direction relative to the available response options and that in classical two-alternative forced-choice tasks a larger deviation from the ideal straight line towards the correct response - or in other words, an increased curvature towards the alternative - can be seen as an indicator for the degree of competition that is present between the response options (Stillman, 2018; Dotan, 2019; Maldonado, 2019). So far, the reflection of introspective confidence in the motor dynamics of response selection has not been studied in depth.

The best established relation is the inverse relation between response times and subjective confidence ratings, as trials that feature lower RTs generally receive higher confidence ratings (Kiani, 2014; Dotan, 2019). Further, Dotan et al. demonstrated in their 2018 study that the final speed measured in a trial is correlated with the retrospective confidence ratings that were given after completion of said trial. Therefore, the (momentary) speed towards the chosen option has been considered to represent introspective confidence.

However, the true nature of the relation between RTs and confidence is still a disputed topic: Kiani et al. (2014) argue that confidence is not only informed by the amount of accumulated evidence in favour of the decision as was classically believed, but that also the time taken to form the decision directly influences subjective confidence. Following this, it would mean that longer response times actively decrease subjective confidence. They maintain that "[the] coupling between decision accuracy and decision time suggests that the latter might inform a judgment of [confidence]. Longer decision times are often associated with weaker sensory evidence and higher error rates. Thus, the brain may learn, by association, to use decision time or some function of it as a proxy for stimulus strength and [confidence] judgment" (Kiani, 2014).

In contrast to that, Dotan et al. (2019) merely state that decision speed and confidence are affected by the same factors, primarily the amount of evidence: Larger quantities of momentary evidence increase both speed and confidence. Even so, they emphasise that this could be due to a possible confound, as movement velocity is affected by both confidence (represented by movement along the y-axis) and the ongoing decision conflict (represented by the movement along the x-axis). Put more concretely, higher speed may cause larger deviations from the ideal path[5]. Additionally, Dotan et al. argue in their 2018 study that participants use slowing down while momentarily unconfident as an adaptive mechanism in order to gain more time to enhance their decision - thus, directly opposing the causality as suggested by Kiani et al. (2014).

To summarise, more research on the true relation between RTs and confidence is still needed, as there is evidence for both opposing possibilities, which claim that either higher RTs may decrease subjective confidence due to learned associations (Kiani, 2014) or that an increase in RTs may serve as an compensatory mechanism in times of increased uncertainty (Dotan, 2018).

[5] Therefore, they advise to control for this potential relation, for example by using momentary curvature as a covariate during the data analysis (Dotan, 2019).

5. Shortcomings of the Present Study and Considerations for Future Studies

Mouse tracking studies generate extraordinarily rich data, which allows for a plethora of statistic investigations. The present study has only exhausted a very small fraction of what is possible in terms of available measures, as well as approaches of analysis.

All of the investigated measures have been found to correlate with subjective confidence at least with moderately strength, with especially high correlation coefficients in the case of RTs and vel_max_time. Taken together with the findings from Dotan et al. (2018), this suggests that a more in-depth analysis of how well the temporal aspects of motor dynamics reflect introspective confidence during response selection. To that end, it suggests itself to not just investigate related summary measures (as in the case of this study), but to exhaust the full capacity of the method. Only the analysis of trajectories as a continuous series of timepoints, as well as the full velocity profiles can reveal certain patterns within the data (Dotan, 2019). For example, it might be interesting to study if and how movement initiation latency (cf. Dale, 2007) and final speed of trial (cf. Dotan, 2018) differ in their abilities to reflect subjective confidence. Furthermore, one should consider adding momentary curvature as a covariate for such future analyses as propositioned by Dotan et al. (2019) due to the likely confound between speed and spatial deviations.

In addition to this, trajectory types should be analysed as well, since an aggregate-level approach of trajectory features such as curvature (e.g. through mean AUC) does not reflect the high movement variability that is found on a trial-level: Average trajectories that show a continuous curvature could potentially arise from a combination of trajectories that exhibit a strong curvature towards the non-chosen response option, as well as of trajectories that just follow the straight direct path to the correct response option. In that case, it would be possible that not a single trajectory would match the shape and location of the aggregate trajectory. While such purely bimodal distributions of trajectory types are rather rare, the majority of data sets from published mouse tracking studies has been revealed to feature multiple types of trajectories (Wulff, 2018; Maldonado, 2019). The analysis of trajectory types can also be used to exclude trajectories that show irrelevant movements that might be due to the participant trying to overcome a physical obstacle while administering the mouse. Such trajectories might feature erratic movements with a comparatively high number of up- and down-movements that are untypical for the context of the study and thus might bias aggregate measures (Kieslich, 2018). Further, the trial-level investigation of trajectory types can reveal adaptive strategies that participants develop, which could invalidate the underlying assumptions about how to interpret the trajectories (*c.f. chapter 2.2.2.: Disadvantages and Limitations*), and such trajectories could be excluded from further analyses.

As this study for the most part investigated aggregate measures on the condition-level, a stronger emphasis on trial-level analyses in the future might bring more insights to the examined relations.

Equally important would be to also investigate the relation between introspective confidence judgements and mouse tracking measures while employing a different underlying task, since the categorisation of typical and atypical exemplars has some inherent flaws that became apparent during the analysis. Its biggest fault is that typicality is treated as a binary variable for the sake of simplicity. Resulting from that the binary distinction between "typical" and "atypical" exemplars does not seem to fit well for all animals, even though only a relatively small stimulus set was chosen to begin with (Dale, 2007). So for example, while exemplar #7 "Hai" was deemed to be a typical member of the category "Fisch", the results from the present study suggest that participants did not share this sentiment. Even besides exceptions such as this one, the two conditions exhibited large differences in magnitude of cognitive competition. Consequently, the recorded measures and calculated covariances and correlation coefficients differ drastically in between conditions. Therefore, it would be useful to compare the results obtained in this study with ones from a study that used another underlying task, which does not share this stark disparity in competition between its conditions.

Lastly, another small shortcoming of the present study was that handedness should be assessed in a more sophisticated manner in the future than it was done now. One sensible option would be to use the Edinburgh Handedness Inventory, as it was also used in Kieslich et al.'s (2017) replication of Dale et al.'s (2007) original study.

6. Conclusion

The present study investigated if introspective confidence judgements are reflected in the motor dynamics of response selection based on Dale et al.'s 2007 study on taxonomic categorisation of typical and atypical exemplars. Following that original study, it was hypothesised that there are characteristic differences in the trajectories that are recorded for typical and atypical trials. Further, it was predicted that the investigated mouse tracking measures should covary or correlate to a higher degree to subjective confidence ratings in the typical than in atypical condition.

These hypotheses were generally met. Based on the calculated correlation coefficients, the investigated temporal measures (RTs and vel_max_time) seem to reflect confidence better than spatial measures (AUC) or the correctness ratio in this study. These results, however, should not necessarily be interpreted in such a way that introspective confidence judgements are generally better reflected in temporal measures. Rather, they should be seen such that these results in conjunction with the most relevant literature of this field (c.f. Kiani, 2014; Stillman, 2018; Dotan, 2019) suggest that more research for the relationship between introspective confidence and motor dynamics is still needed in the future and even more so that that it would possibly be sensible to focus that research more on temporal rather than on spatial measures.

References

Anderson, J. R. (2014). *Cognitive Psychology and Its Implications* (8th ed.) New York, United States: Macmillan Publishers.

Brennan, S. E. (2005). *How Conversation Is Shaped by Visual and Spoken Evidence.* In J. C. Trueswell & M. K. Tanenhaus (Eds.), Approaches to studying world-situated language use: Bridging the language-as-product and language-as-action traditions (pp. 95–130). Retrieved from http://www.psychology.sunysb.edu/sbrennan-/papers/brennan2004.pdf

Calvert, S. L., Rideout, V. J., Woolard, J. L., Barr, R. F., & Strouse, G. A. (2005). *Age, Ethnicity, and Socioeconomic Patterns in Early Computer Use.* American Behavioral Scientist, 48(5), 590–607. https://doi.org/10.1177/0002764204271508

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences (*2nd ed.). Hillsdale, NJ, USA: Lawrence Erlbaum Associates.

Dale, R., Kehoe, C., & Spivey, M. J. (2007). *Graded motor responses in the time course of categorizing atypical exemplars.* Memory & Cognition, 35(1), 15–28. https://doi.org/10.3758/bf03195938

Dotan, D., Meyniel, F., & Dehaene, S. (2018). *On-line confidence monitoring during decision making.* Cognition, 171, 112–121.
https://doi.org/10.1016/j.cognition.2017.11.001

Dotan, D., Pinheiro-Chagas, P., Al Roumi, F., & Dehaene, S. (2019). *Track It to Crack It: Dissecting Processing Stages with Finger Tracking.* Trends in Cognitive Sciences, 23(12), 1058–1070. https://doi.org/10.1016/j.tics.2019.10.002

Farmer, T., Cargill, S., Hindy, N., Dale, R., & Spivey, M. (2007). *Tracking the Continuity of Language Comprehension: Computer Mouse Trajectories Suggest Parallel Syntactic Processing.* Cognitive Science: A Multidisciplinary Journal, 31(5), 889–909. https://doi.org/10.1080/03640210701530797

Fischer, M. H., & Hartmann, M. (2014). *Pushing forward in embodied cognition: may we mouse the mathematical mind?* Frontiers in Psychology, 5, 1–4. https://doi.org/10.3389/fpsyg.2014.01315

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.
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, 1–6.
https://doi.org/10.3389/fpsyg.2011.00059

Freeman, J. B. (2018). *Doing Psychological Science by Hand.* Current Directions in Psychological Science, 27(5), 315–323.
https://doi.org/10.1177/0963721417746793

Grimaldi, P., Lau, H., & Basso, M. A. (2015). *There are things that we know that we know, and there are things that we do not know we do not know: Confidence in decision-making.* Neuroscience & Biobehavioral Reviews, 55, 88–97. https://doi.org/10.1016/j.neubiorev.2015.04.006

Hehman, E., Stolier, R. M., & Freeman, J. B. (2014). *Advanced mouse-tracking analytic techniques for enhancing psychological science.* Group Processes & Intergroup Relations, 18(3), 384–401.
https://doi.org/10.1177/1368430214538325

Hemphill, J. F. (2003). *Interpreting the magnitudes of correlation coefficients*. American Psychologist, 58(1), 78–79.
https://doi.org/10.1037/0003-066x.58.1.78

Kahneman, D., & Tversky, A. (1982). *Variants of uncertainty.* Cognition, 11(2), 143–157. https://doi.org/10.1016/0010-0277(82)90023-3

Kepecs, A., & Mainen, Z. F. (2012). *A computational framework for the study of confidence in humans and animals.* Philosophical Transactions of the Royal Society B: Biological Sciences, 367(1594), 1322–1337.
https://doi.org/10.1098/rstb.2012.0037

Kiani, R., & Shadlen, M. N. (2009). *Representation of Confidence Associated with a Decision by Neurons in the Parietal Cortex.* Science, 324(5928), 759–764. https://doi.org/10.1126/science.1169405

Kiani, Roozbeh, Corthell, L., & Shadlen, M. N. (2014). *Choice Certainty Is Informed by Both Evidence and Decision Time.* Neuron, 84(6), 1329–1342. https://doi.org/10.1016/j.neuron.2014.12.015

Kieslich, P. J., & Henninger, F. (2017). *Mousetrap: An integrated, open-source mouse-tracking package.* Behavior Research Methods, 49(5), 1652–1667.
https://doi.org/10.3758/s13428-017-0900-z

Kieslich, P. J., Henninger, F., Wulff, D. U., Haslbeck, J., & Schulte-Mecklenbeck, M. (2018). *Mouse-tracking: A practical guide to implementation and analysis.* PsyArXiv Preprints. https://doi.org/10.31234/osf.io/zuvqa

Magnuson, J. S. (2005). *Moving hand reveals dynamics of thought.* Proceedings of the National Academy of Sciences, 102(29), 9995–9996.
https://doi.org/10.1073/pnas.0504413102

Maldonado, M., Dunbar, E., & Chemla, E. (2019). *Mouse tracking as a window into decision making*. Behavior Research Methods, 51(3), 1085–1101.
https://doi.org/10.3758/s13428-018-01194-x

Medin, D. L., & Schaffer, M. M. (1978). *Context theory of classification learning.* Psychological Review, 85(3), 207–238.
https://doi.org/10.1037/0033-295x.85.3.207

Meyniel, F., Schlunegger, D., & Dehaene, S. (2015). *The Sense of Confidence during Probabilistic Learning: A Normative Account.* PLOS Computational Biology, 11(6), e1004305. https://doi.org/10.1371/journal.pcbi.1004305

Norman, E., & Price, M. C. (2015). *Measuring consciousness with confidence ratings.* Behavioral Methods in Consciousness Research, 159–180. https://doi.org/10.1093/acprof:oso/9780199688890.003.0010

Posner, M. I. (2005). *Timing the Brain: Mental Chronometry as a Tool in Neuroscience.* PLoS Biology, 3(2), e51.
https://doi.org/10.1371/journal.pbio.0030051

Pouget, A., Drugowitsch, J., & Kepecs, A. (2016). *Confidence and certainty: distinct probabilistic quantities for different goals.* Nature Neuroscience, 19(3), 366–374. https://doi.org/10.1038/nn.4240

Quétard, B., Quinton, J. C., Mermillod, M., Barca, L., Pezzulo, G., Colomb, M., & Izaute, M. (2016). *Differential effects of visual uncertainty and contextual guidance on perceptual decisions: Evidence from eye and mouse tracking in visual search.* Journal of Vision, 16(11), 28. https://doi.org/10.1167/16.11.28

Resulaj, A., Kiani, R., Wolpert, D. M., & Shadlen, M. N. (2009). *Changes of mind in decision-making.* Nature, 461(7261), 263–266.
https://doi.org/10.1038/nature08275

Rosch, E., & Mervis, C. B. (1975). *Family resemblances: Studies in the internal structure of categories.* Cognitive Psychology, 7(4), 573–605.
https://doi.org/10.1016/0010-0285(75)90024-9

Rosenbaum, D. A. (2005). *The Cinderella of Psychology: The Neglect of Motor Control in the Science of Mental Life and Behavior.* American Psychologist, 60(4), 308–317. https://doi.org/10.1037/0003-066x.60.4.308

Sandberg, K., Timmermans, B., Overgaard, M., & Cleeremans, A. (2010). *Measuring consciousness: Is one measure better than the other?* Consciousness and Cognition, 19(4), 1069–1078.
https://doi.org/10.1016/j.concog.2009.12.013

Schouten, J. F., & Bekker, J. A. M. (1967). *Reaction Time and Accuracy*. Acta Psychologica, 27, 143–153.
Retrieved from https://bit.ly/378a7Ld

Schulte-Mecklenbeck, M., Kühberger, A., & Ranyard, R. (2011). *The role of process data in the development and testing of process models of judgment and decision making.* Judgment and Decision Making, 6(8), 733–739.
Retrieved from http://journal.sjdm.org/11/m36/m36.pdf

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

Spivey, Michael J., & Dale, R. (2006). *Continuous Dynamics in Real-Time Cognition.* Current Directions in Psychological Science, 15(5), 207–211. https://doi.org/10.1111/j.1467-8721.2006.00437.x

Spivey, Michael 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(1), 251–254. https://doi.org/10.1037/a0017170

Stillman, P. E., Shen, X., & Ferguson, M. J. (2018). *How Mouse-tracking Can Advance Social Cognitive Theory.* Trends in Cognitive Sciences, 22(6), 531–543. https://doi.org/10.1016/j.tics.2018.03.012

Storms, G. (2004). *Exemplar Models in the Study of Natural Language Concepts.* In B. H. Ross (Ed.)*, The Psychology of Learning and Motivation: Advances in Research and Theory* (pp. 1–35). Retrieved from https://bit.ly/3ctdA8a

Wickelgren, W. E. (1977). *Speed–accuracy tradeoff and information processing Dynamics*. Acta Psychologica, 41, 67–85. Retrieved from http://www.columbia.edu/~nvg1/Wickelgren/papers/1977bWAW.pdf

Wulff, D. U., Haslbeck, J. M. B., Kieslich, P. J., Henninger, F., & Schulte-Mecklenbeck, M. (2018). *Mouse-tracking: Detecting Types in Movement Trajectories.* PsyArXiv Preprints, 1–28. https://doi.org/10.31234/osf.io/6edca

Declaration of Authorship

I hereby certify that the work presented here is, to the best of my knowledge and belief, original and the result of my own investigations, except as acknowledged, and has not been submitted, either in part or whole, for a degree at this or any other university.

_____

signature

_____

city, date

Appendix A

Overview of measures relevant to this study

Correctness Ratio

The ratio of correct to incorrect trials was assessed per exemplar and condition prior to filtering the data. For all subsequent analyses, only correct trials were included to guarantee that the other measures could be consistently interpreted across trials.

Response times (RTs)

Response times are a measure of the total time between the start of a trial until a response was given. In the present study, this is the period that begins with the participant clicking on the "Start" button and ends as soon as the cursor reaches one of the two response boxes.

vel_max_time

vel_max_time is the point in time at which the maximum velocity occurred first during a given trial. This measure of velocity/acceleration is able to reveal the degree and onset of a commitment to a specific response and reflects the magnitude of activation of the available response options. Even further, it can be used to index the extent of competition between response options (Hehman, 2014; Kieslich, Mousetrap Documentation).

Area under the curve (AUC)

The area under the curve can be used a measure of trajectory curvature. It is the geometric area between the recorded trajectory and an idealised direct path in the form of a straight line between the start and end points. Trajectories that follow the idealised direct path would have an associated AUC of 0, while trajectories that follow an extreme triangular path (see Figure x) feature a maximum AUC of 320,000px^2. Deviations above the direct path cause the AUC to become positive, while AUCs associated with trajectories that deviate below the direct path receive a negative sign during calculation. For the purposes of this study, only absolute values were used for improved comparisons. The AUC is commonly used to assess the degree of attraction or bias towards the unselected alternative response option, as it indexes the magnitude of activation of the available response options. Consequently, it can be seen as a quantifier of cognitive conflict between the response options (Freeman, 2010; Hehman, 2014; Quétard, 2016; Kieslich, 2017 & 2018; Stillman, 2018; Maldonado, 2019).

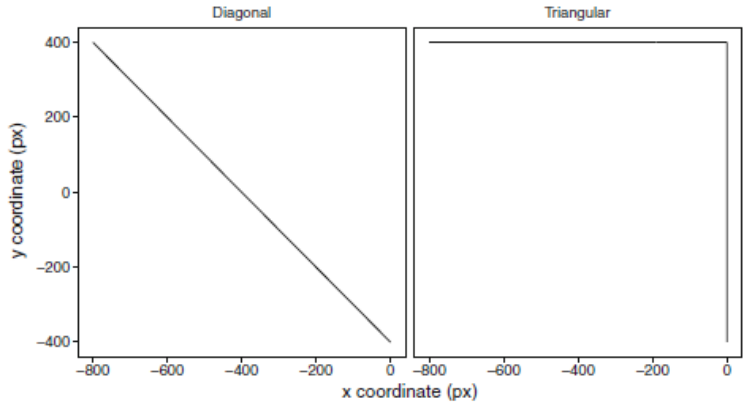


Fig. 6 Plot of all raw trajectories for each simulation. All trajectories started at the bottom center of the screen and ended at the top left

Add.-Fig. 1 *Adopted from Kieslich et al. (2017)*

Confidence ratings

Participants were instructed[6] to retrospectively rate how confident they felt in their decision on a scale of 1 – 6. Here, a rating of 1 represented feeling "unsicher" (= uncertain) and a rating of 6 symbolised feeling "sehr sicher" (= very certain). No further lexical analogies were given for the remaining values. A scale with an even number of values was chosen, so that it would not include a neutral middle point.

[6] The exact wording was „Wie überzeugt waren Sie von der Richtigkeit Ihrer Antwort während des Beurteilungsvorgangs?" (= How confident were you of your answer's correctness during the judgement process?)

Appendix B

Supplementary Tables

German Stimuli (as used in the present study)

Trial	Exemplar	Kondition	Korrekte Kategorie	Distraktor-Kategorie
-	Giraffe	Übung	Säugetier	Amphibie
-	Taube	Übung	Vogel	Reptil
-	Biene	Übung	Insekt	Fisch
1	Aal	Atypisch	Fisch	Reptil
2	Alligator	Typisch	Reptil	Säugetier
3	Chamäleon	Typisch	Reptil	Insekt
4	Falke	Typisch	Vogel	Reptil
5	Fledermaus	Atypisch	Säugetier	Vogel
6	Goldfisch	Typisch	Fisch	Amphibie
7	Hai	Typisch	Fisch	Säugetier
8	Hund	Typisch	Säugetier	Insekt
9	Kaninchen	Typisch	Säugetier	Reptil
10	Katze	Typisch	Säugetier	Reptil
11	Klapperschlange	Typisch	Reptil	Amphibie
12	Lachs	Typisch	Fisch	Säugetier
13	Löwe	Typisch	Säugetier	Fisch
14	Pferd	Typisch	Säugetier	Vogel
15	Pinguin	Atypisch	Vogel	Fisch
16	Schmetterling	Atypisch	Insekt	Vogel
17	Seelöwe	Atypisch	Säugetier	Fisch
18	Spatz	Typisch	Vogel	Säugetier
19	Wal	Atypisch	Säugetier	Fisch

English Stimuli (as used in Dale's 2007 study)

Trial	Exemplar	Condition	Correct Category	Distractor Category
-	Giraffe	Practise	Mammal	Amphibian
-	Pigeon	Practise	Bird	Reptile
-	Bee	Practise	Insect	Fish
1	Eel	Atypical	Fish	Reptile
2	Alligator	Typical	Reptile	Mammal
3	Chameleon	Typical	Reptile	Insect
4	Hawk	Typical	Bird	Reptile
5	Bat	Atypical	Mammal	Bird
6	Goldfish	Typical	Fish	Amphibian
7	Shark	Typical	Fish	Mammal
8	Dog	Typical	Mammal	Insect
9	Rabbit	Typical	Mammal	Reptile
10	Cat	Typical	Mammal	Reptile
11	Rattlesnake	Typical	Reptile	Amphibian
12	Salmon	Typical	Fish	Mammal
13	Lion	Typical	Mammal	Fish
14	Horse	Typical	Mammal	Bird
15	Penguin	Atypical	Bird	Fish
16	Butterfly	Atypical	Insect	Bird
17	See lion	Atypical	Mammal	Fish
18	Sparrow	Typical	Bird	Mammal
19	Whale	Atypical	Mammal	Fish

Summary Measures

Condition-level

	Atypical	Typical
Correctness	μ = 0.935 SD = 0.247	μ = 0.966 SD = 0.180
RT	μ = 1,972.966 SD = 1,387.260	μ = 1,398.249 SD = 830.177
vel_max_time	μ = 1,634.379 SD = 1,104.303	μ = 1,160.506 SD = 750.995
AUC	μ = 74,642.259 SD = 85,198.427	μ = 48,734.451 SD = 54,619.218
Confidence	μ = 5.216 SD = 0.994	μ = 5.808 SD = 0.548

Exemplar-level

Exemplar (Condition)	Correctness	RT	vel_max_ time	AUC	Confidence
#1 Aal (Atypical)	μ = 0.946 SD = 0.223	μ = 1,651.25 SD = 652.340	μ = 1,383 SD = 616.732	μ = 59,568 SD = 49,204.808	μ = 5.1 SD = 1.210
#2 Alligator (Typical)	μ = 1 SD = 0	μ = 1,784.75 SD = 940.353	μ = 1,517.15 SD = 916.676	μ = 37,182 SD = 32,629.112	μ = 5.5 SD = 0.889
#3 Chamäleon (Typical)	μ = 1 SD = 0	μ = 1,533 SD = 437.626	μ = 1,331.5 SD = 372.873	μ = 67,039.65 SD = 86,718.921	μ = 5.7 SD = 0.470
#4 Falke (Typical)	μ = 1 SD = 0	μ = 1,274.826 SD = 348.086	μ = 1,077.87 SD = 310.319	μ = 48,172.87 SD = 62,974.13	μ = 5.957 SD = 0.209
#5 Fledermaus (Atypical)	μ = 0.849 SD = 0.358	μ = 2,637.706 SD = 2,441.790	μ = 2,379.471 SD = 2,333.586	μ = 111,783.853 SD = 131,684.576	μ = 5.118 SD = 0.697
#6 Goldfisch (Typical)	μ = 0.976 SD = 0.153	μ = 1,288.1 SD = 433.441	μ = 1,018.1 SD = 256.471	μ = 48,273.525 SD = 48,289.168	μ = 6 SD = 0
#7 Hai (Typical)	μ = 0.768 SD = 0.422	μ = 2,326 SD = 2,502.173	μ = 1,983.334 SD = 2,281.618	μ = 70,689.234 SD = 90,820.463	μ = 5 SD = 1

#8 Hund (Typical)	μ = 1 SD = 0	μ = 1,007.714 SD = 316.167	μ = 815.809 SD = 264.104	μ = 46,160.595 SD = 38,111.552	μ = 6 SD = 0
#9 Kaninchen (Typical)	μ = 1 SD = 0	μ = 1,360.476 SD = 773.508	μ = 1,132.381 SD =761.82	μ = 44,959.381 SD = 35,770.297	μ = 5.905 SD = 0.301
#10 Katze (Typical)	μ = 0.968 SD = 0.177	μ = 1,219.5 SD = 376.150	μ = 970 SD = 313.553	μ = 56,830.725 SD = 79,875.270	μ = 5.95 SD = 0.223
#11 Klapper-schlange (Typical)	μ = 1 SD = 0	μ = 1,522.364 SD = 403.906	μ = 1,332.818 SD = 379.751	μ = 36,061.5 SD = 37,526.199	μ = 5.773 SD = 0.685
#12 Lachs (Typical)	μ = 1 SD = 0	μ = 1,450.909 SD = 772.799	μ = 1,158.182 SD = 594.896	μ = 35,919.659 SD = 23,815.785	μ = 5.727 SD = 0.550
#13 Löwe (Typical)	μ =1 SD = 0	μ = 1,182.55 SD =421.977	μ = 958.55 SD = 310.485	μ = 41,239.8 SD = 40,377.110	μ = 5.95 SD = 0.223
#14 Pferd (Typical)	μ = 0.969 SD = 0.174	μ = 1,310.1 SD = 616.207	μ = 1,083.6 SD = 568.716	μ = 66,213.25 SD = 64,506.927	μ = 6 SD = 0
#15 Pinguin (Atypical)	μ = 0.927 SD = 0.260	μ = 2,108.444 SD = 1,300.315	μ = 1,636.722 SD = 625.765	μ = 85,890.75 SD = 105,643.743	μ = 5.111 SD = 1.023
#16 Schmetter-ling (Atypical)	μ = 1 SD = 0	μ = 1,466.455 SD = 327.327	μ = 1,222.818 SD = 317.126	μ = 45,950.136 SD = 51,172.672	μ = 5.727 SD = 0.550
#17 See-löwe (Atypical)	μ = 0.966 SD = 0.179	μ = 1,984.55 SD = 896.167	μ = 1,699.05 SD = 751.184	μ = 89,273.725 SD = 66,889.007	μ = 5.25 SD = 0.967
#18 Spatz (Typical)	μ = 1 SD = 0	μ = 1,181.952 SD = 245.550	μ = 938.143 SD = 225.824	μ = 43,506.238 SD = 43,506.238	μ = 5.81 SD = 0.68
#19 Wal (Atypical)	μ = 0.959 SD = 0.198	μ = 2,162.789 SD = 1,753.357	μ = 1,638.579 SD = 812.466	μ = 64,442.395 SD = 83,974.487	μ = 4.895 SD = 1.243

Covariance Matrices

Aggregated over both Conditions

	Correct-ness Mean	RT Mean	vel_max_ time Mean	AUC Mean	Confidence Mean
Correctness Mean	0.001	-9.09	-7.495	-409.758	0.009
RT Mean	-9.09	165,149.505	136,171.493	7444821.766	-170.125
vel_max_ time Mean	-7.495	136,171.493	112,278.118	6,138,513.693	-140.274
AUC Mean	-409.758	7,444,821.766	6,138,513.693	335,607,249.322	-7,669.099
Confidence Mean	0.009	-170.125	-140.274	-7,669.099	0.175

Aggregated over Exemplars

Atypical Exemplars

	Correct-ness Mean	RT Mean	vel_max_ time Mean	AUC Mean	Confidence Mean
Correctness Mean	0.003	-18.306	-18.604	-1,038.140	0.007
RT Mean	-18.306	17,0288.907	156,411.326	8,713,513.469	-71.815
vel_max_ time Mean	-18.604	156,411.326	157,604.093	8,794,016.528	-50.301
AUC Mean	-1,038.140	8,713,513.469	8,794,016.528	572,189,667.889	-2,767.135
Confidence Mean	0.007	-71.815	-50.301	-2,767.135	0.077

Typical Exemplars

	Correct-ness Mean	RT Mean	vel_max_ time Mean	AUC Mean	Confidence Mean
Correctness Mean	0.004	-16.444	-14.420	-472.473	0.014
RT Mean	-16.444	112,948.341	102,906.271	1,414,370.437	-89.873
vel_max_ time Mean	-14.420	102,906.271	94,852.461	1,305,124.296	-81.188
AUC Mean	-472.473	1,414,370.437	1,305,124.296	145,459,766.976	-1,009.889
Confidence Mean	0.014	-89.873	-81.188	-1,009.889	0.079

Weighted Average

	Correct-ness Mean	RT Mean	vel_max_ time Mean	AUC Mean	Confidence Mean
Correctness Mean	0.004	-20.556	-18.505	-810.122	0.016
RT Mean	-20.556	200,181.553	176,123.192	6,921,260.21	-158.336
vel_max_ time Mean	-18.505	176,123.192	159,928.780	6,251,262.287	-132.906
AUC Mean	-810.122	6,921,260.21	6,251,262.287	419,085,781.706	-5,040.721
Confidence Mean	0.016	-158.336	-132.906	-5,040.721	0.154

Correlation Matrices

Aggregated over both Conditions

	Correctness Mean	RT Mean	vel_max_time Mean	AUC Mean	Confidence Mean
Correctness Mean	1	-1	-1	-1	1
RT Mean	-1	1	1	1	-1
vel_max_time Mean	-1	1	1	1	-1
AUC Mean	-1	1	1	1	-1
Confidence Mean	1	-1	-1	-1	1

Aggregated over Exemplars

Atypical Exemplars

	Correctness Mean	RT Mean	vel_max_time Mean	AUC Mean	Confidence Mean
Correctness Mean	1	-0.862	-0.911	-0.844	0.482
RT Mean	-0.862	1	0.955	0.883	-0.617
vel_max_time Mean	-0.911	0.955	1	0.926	-0.449
AUC Mean	-0.844	0.883	0.926	1	-0.41
Confidence Mean	0.482	-0.617	-0.449	-0.41	1

Typical Exemplars

	Correctness Mean	RT Mean	vel_max_time Mean	AUC Mean	Confidence Mean
Correctness Mean	1	-0.77	-0.737	-0.616	0.777
RT Mean	-0.77	1	0.994	0.349	-0.953
vel_max_time Mean	-0.737	0.994	1	0.351	-0.939
AUC Mean	-0.616	0.349	0.351	1	-0.298
Confidence Mean	0.777	-0.953	-0.939	-0.298	1

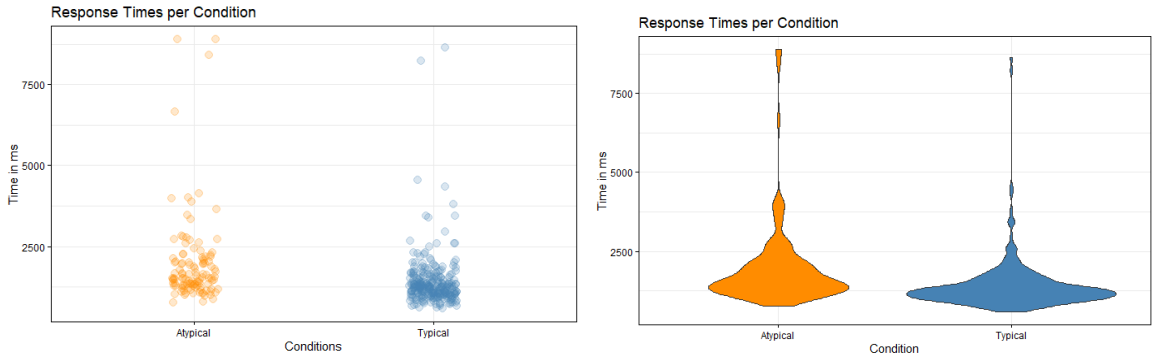
Weighted Average

	Correctness Mean	RT Mean	vel_max_time Mean	AUC Mean	Confidence Mean
Correctness Mean	1	-0.756	-0.762	-0.652	0.661
RT Mean	-0.756	1	0.984	0.756	-0.902
vel_max_time Mean	-0.762	0.984	1	0.764	-0.847
AUC Mean	-0.652	0.756	0.764	1	-0.628
Confidence Mean	0.661	-0.902	-0.847	-0.627	1

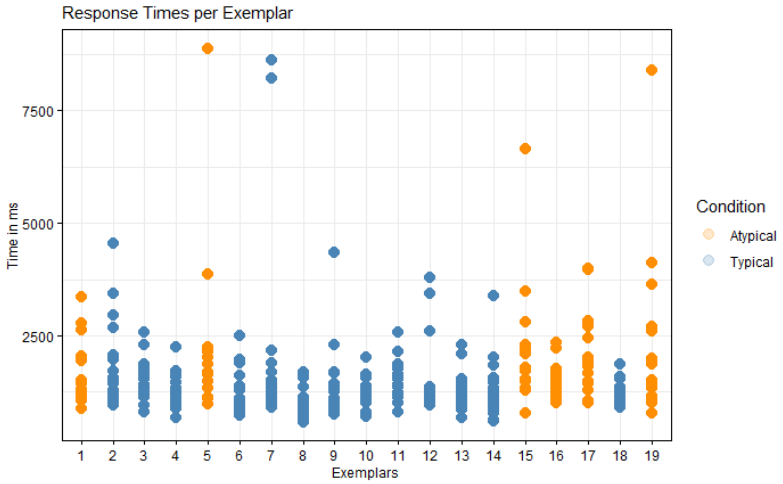
Appendix C

Supplementary Plots

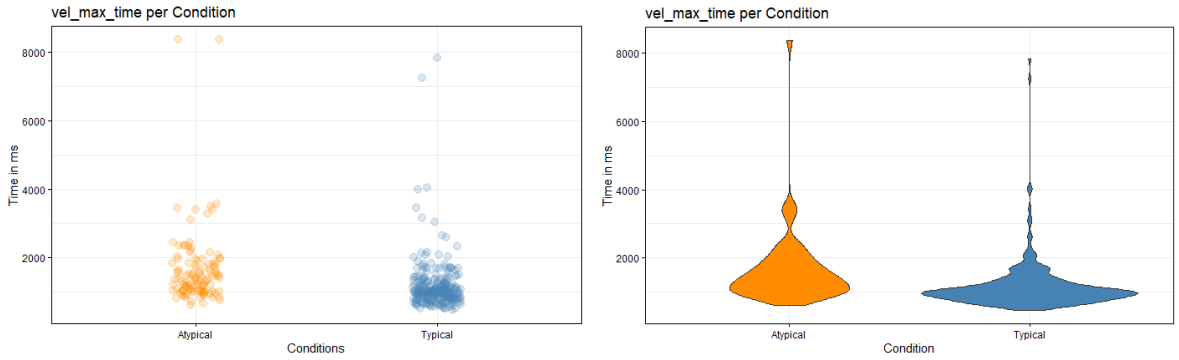
Data Distribution Plots



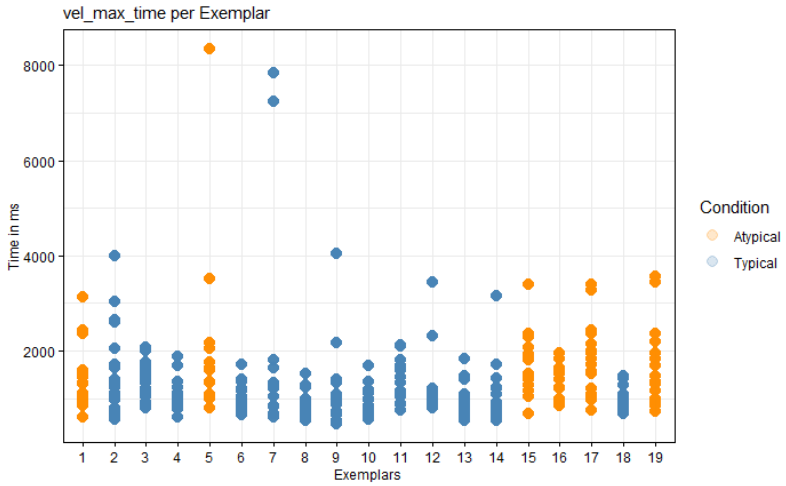
Add.-Fig. 2 *Response Times in milliseconds for the atypical (depicted in dark orange) and typical (depicted in blue) conditions; Visualised as jitter plots* ***(A-2a)*** *and violin plots* ***(A-2b)***



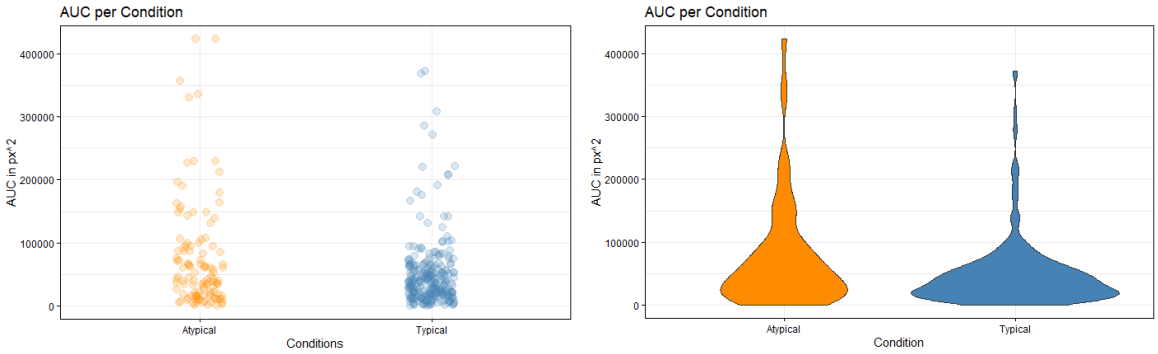
Add.-Fig. 3 *Response Times in milliseconds for all exemplars, visualised as a scatter plot. Atypical exemplars are depicted in dark orange, typical exemplars are depicted in blue. A list of which exemplar corresponds to which number can be found in* appendix B.



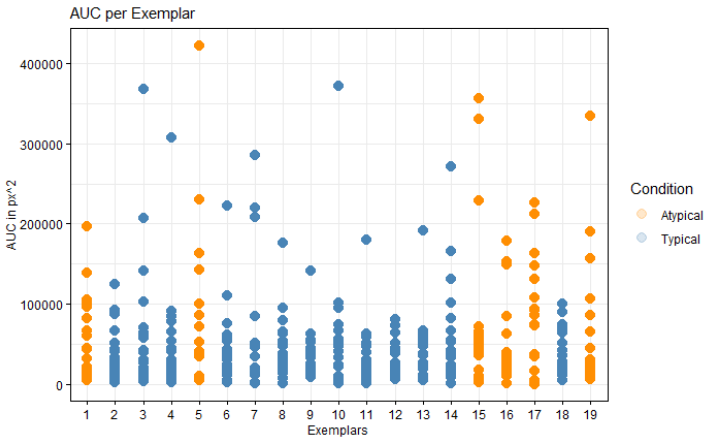
Add.-Fig. 4 *vel_max_time in milliseconds for the atypical (depicted in dark orange) and typical (depicted in blue) conditions; Visualised as jitter plots* ***(A-4a)*** *and violin plots* ***(A-4b)***



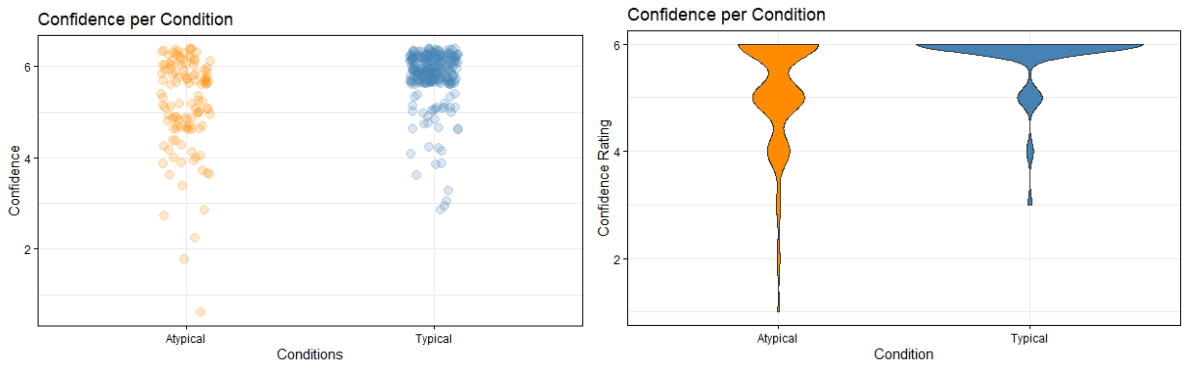
Add.-Fig. 5 *vel_max_time in milliseconds for all exemplars, visualised as a scatter plot. Atypical exemplars are depicted in dark orange, typical exemplars are depicted in blue. A list of which exemplar corresponds to which number can be found in* appendix B.



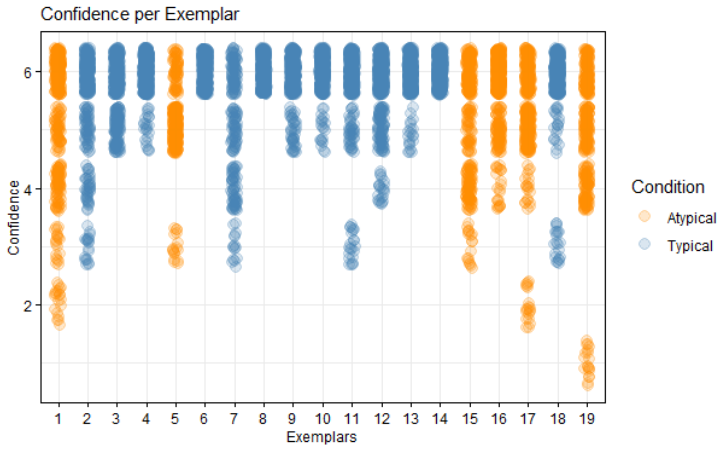
Add.-Fig. 6 *AUC in squared pixels for the atypical (depicted in dark orange) and typical (depicted in blue) conditions; Visualised as jitter plots* ***(A-6a)*** *and violin plots* ***(A-6b)***



Add.-Fig. 7 *AUC in squared pixels for all exemplars, visualised as a scatter plot. Atypical exemplars are depicted in dark orange, typical exemplars are depicted in blue. A list of which exemplar corresponds to which number can be found in* appendix B.

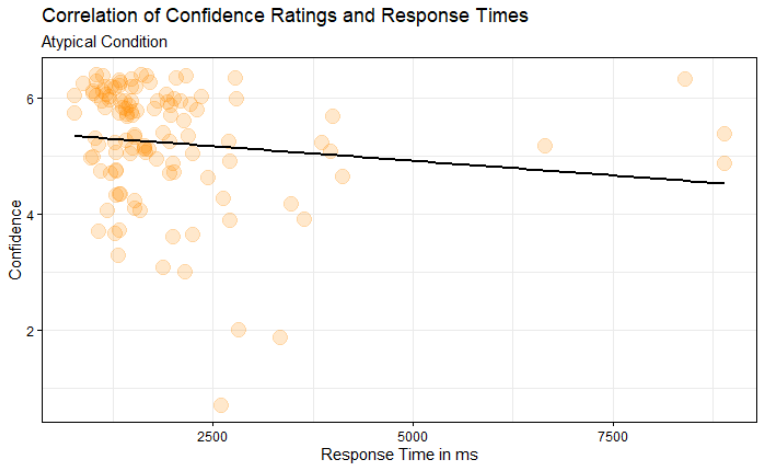


Add.-Fig. 8 *Confidence ratings on a scale of 1 - 6 for the atypical (depicted in dark orange) and typical (depicted in blue) conditions; Visualised as jitter plots* ***(A-8a)*** *and violin plots* ***(A-8b)***

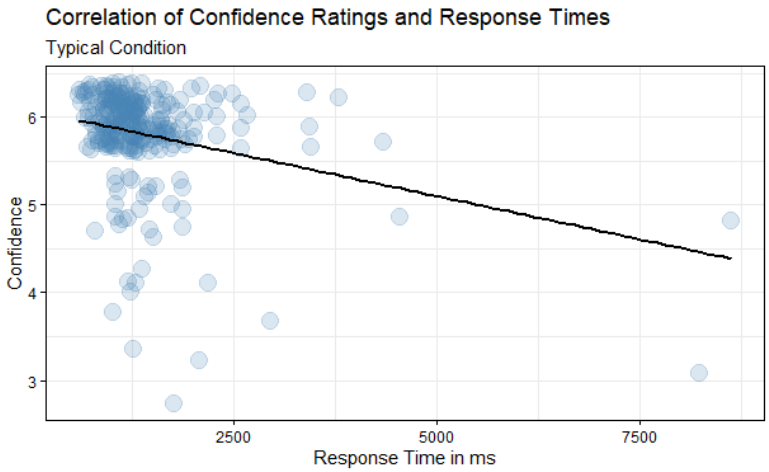


Add.-Fig. 9 *Confidence ratings on a scale of 1 - 6 for all exemplars, visualised as a jitter plot. Atypical exemplars are depicted in dark orange, typical exemplars are depicted in blue. A list of which exemplar corresponds to which number can be found in* appendix B.

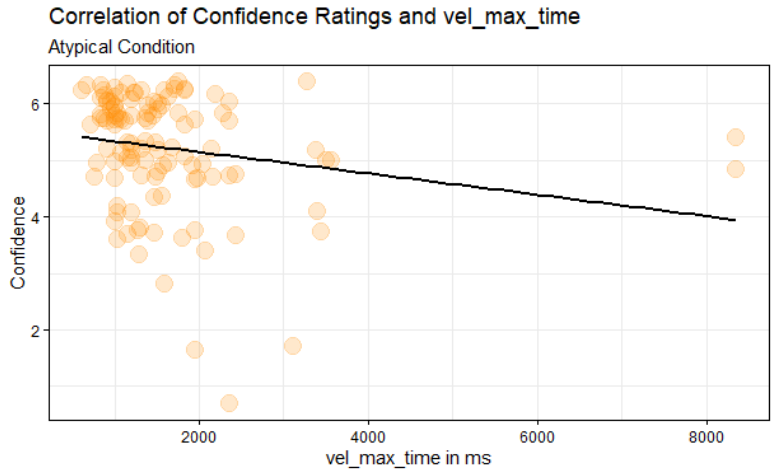
Correlation Plots



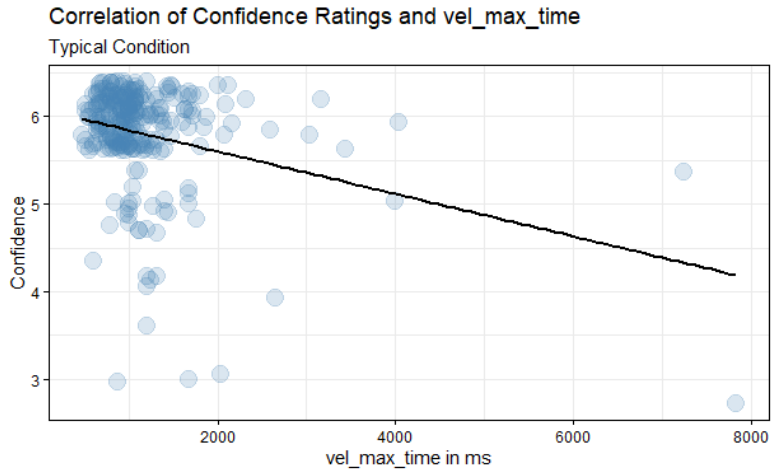
Add.-Fig. 10 *Relation between response times and confidence ratings in atypical trials. The black line shows the Bravais-Pearson correlation between the two measures.*



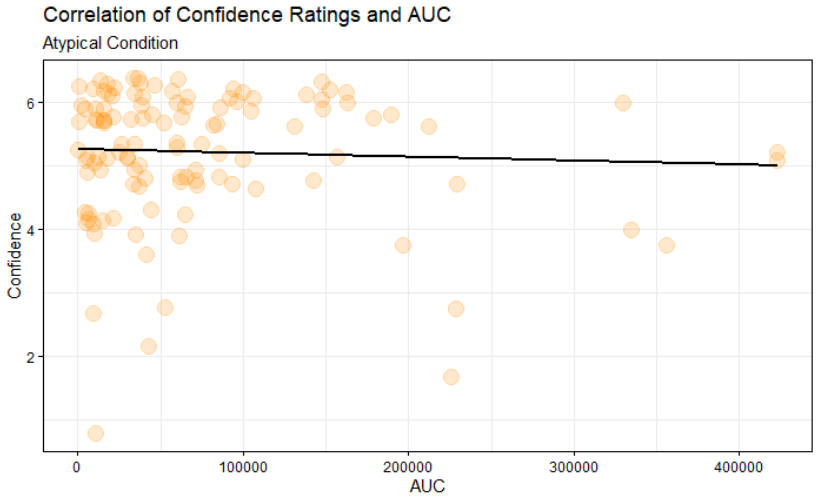
Add.-Fig. 11 *Relation between response times and confidence ratings in typical trials. The black line shows the Bravais-Pearson correlation between the two measures.*



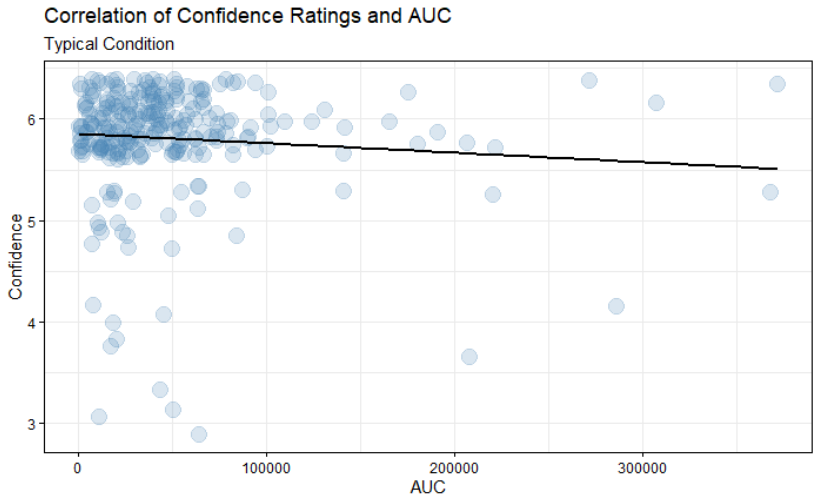
Add.-Fig. 12 *Relation between vel_max_time and confidence ratings in atypical trials. The black line shows the Bravais-Pearson correlation between the two measures.*



Add.-Fig. 13 *Relation between vel_max_time and confidence ratings in typical trials. The black line shows the Bravais-Pearson correlation between the two measures.*



Add.-Fig. 14 *Relation between the area under the curve and confidence ratings in atypical trials. The black line shows the Bravais-Pearson correlation between the two measures.*



Add.-Fig. 15 *Relation between the area under the curve and confidence ratings in typical trials. The black line shows the Bravais-Pearson correlation between the two measures.*