The effect of atypical exemplars on the manual response in a categorization task - a replication study

LUISA DRESCHER, HANNAH GAUDITZ, LILIAN MENZNER, LILITH OKONNEK

University of Osnabrueck

Institute of Cognitive Science

Seminar:

Experimental Psychology Lab, Prof. Dr. Michael Franke

Summer Term 2020

Original Study:

‘Graded motor responses in the time course of categorizing atypical exemplars’

by Rick Dale, Caitlin Kehoe and Michael J. Spivey

Cornell University, Ithaca, New York

published in ‘Memory & Cognition’, 2007

Abstract

The presented study is based upon parts of the study ‘Graded motor responses in the time course of categorizing atypical exemplars’ by Dale et al., published in ‘Memory & Cognition’ in 2007. More precisely, the first experiment was replicated.

For this purpose, the effect of atypical exemplars compared to typical ones on a manual response was investigated, using a categorization task and the mouse tracking method. On average, the response was faster and more accurate when the participants were presented with typical exemplars, whereas when showing atypical instances the response happened to be slower and more errors occurred. When facing atypical exemplars, the participants’ mouse movements showed evidence of competition between the two categories, most noticeable in a bias toward the competing category. The mouse trajectories gravitated to the alternative (incorrect) category more when presented with an atypical exemplar than when given a typical instance.

In summary, the manual output from the categorization task measured with mouse tracking reflects the cognitive dynamics taking place in our heads when performing categorization. The findings suggest that the cognitive mapping of a certain exemplar to its category proceeds nonlinearly over time.

In addition to the replicated hypotheses, a comparison between the performance of left- and right-handed participants was made as exploratory examination. There were no significant differences found when comparing the two groups.

Keywords: categorization task, cognitive information processing, motor output, typicality

Introduction

The phenomenon of categorization has always been of great interest in research. Especially in the last few decades, many aspects of categorization were further investigated. The focus of scientific beliefs changed from classical set theoretic accounts where cognition and categorization are discretely bounded with unique membership to nuanced theories where fuzzy categories exist that interact among each other.

In general, categorization describes the mental process of sorting objects into categories in order to facilitate the understanding and organization of things and objects in the world. It is based on distinguishing the members of a category from nonmembers by recognizing their features.

Thus, the concept of typicality is of great importance here. Some instances seem to be more natural and representative for their category because of certain qualities, traits or characteristics they possess. In contrast, other instances may not show the common traits of the category and are therefore seen as atypical. These atypical exemplars may even present features that suggest another category and therefore compete with the accurate assignment.

For example, the lion serves as typical exemplar for the category mammal. It shows typical properties like being a vertebrate animal, having a skin with fur and a constant body temperature and nourishing their babies with special mammary glands of the mothers. In contrast, the whale shows typical aspects of being a fish like having fins and living underwater but nevertheless belongs to the category of mammals. So the whale represents an atypical exemplar for the class.

In this study, the cognitive categorization process of human beings is investigated by means of comparing the effect of categorizing such typical and atypical exemplars on the manual response.

To record the graded motor responses for each categorization process, the method of mouse tracking was used. The participants were tasked to correctly assign a given animal stimulus to one of two given categories by moving the mouse cursor to the respective label. The mouse movement was tracked and with comparing the trajectories and reaction times of typical and atypical exemplars, reliable predictions about the mental categorization process can be made.

The investigated hypotheses stated that the participants' mouse trajectories gravitate to the incorrect category more when the participants are presented with an atypical exemplar than when presented with a typical instance. Furthermore, more typical category members were expected to be recognized on average more quickly than atypical exemplars. Regarding the correctness of categorization, category members which are more typical instances were assumed to be recognized more accurate than atypical exemplars. When facing atypical exemplars, the participants' mouse movements should show evidence of competition between the two categories, most noticeable in a bias toward the competing category. It was expected to find results that resemble the outcomes obtained in the original study by Dale et al.

Method

Participants:

The experiment was executed by 57 participants in total.

They were contacted via Email and social media with an uniform message (see Attachment 1 for recruitment message). The message roughly explained the requirements and instructions for participation in the online experiment. Furthermore, the invitation message contained an online link to execute the experiment. The link was operative for five days in total to collect data from the participants.

The data sample included 19 males, 30 females and 8 other persons as participants, with a total age span of 15 to 75 years. They represented different nationalities and education levels. Participation was voluntary and the partakers did not receive any compensation for taking part in the study.

Five additional participants took part in a previous pilot study. Their task was to test the experiment and provide feedback to the researchers before the main experimental phase was initiated. The pilot participants represented the age of 11 to 75 years.

All partakers were asked to use their right hand while performing the experiment. Furthermore, they were instructed to use a device with a mouse or mouse pad to supply useable data and to focus their full attention on the experimental task without any distractions. An additional requirement was that the participants should be able to understand basic English, as the online experiment was executed in English language.

Materials:

The categorization task was adapted from Dale et al. (2007). More precisely, it was a replication of the first experiment of the study. In total, 19 different lexical items representing different animal names were presented to the participants (see Table 1 for animal stimuli). The items were split in 13 typical exemplars that covered the most typical traits for their categories and 6 atypical exemplars. The atypical items comprised certain properties which might suggest another category label rather than the correct one and therefore the atypical instances caused competition between the labels.

Procedure:

At the beginning of each trial, the two categories appeared at the upper left and right corner of the screen. After a 2000 ms pause in which the participants could get familiar with the category names, a button marked with “Click me!” appeared at the bottom center of the screen. As soon as the participants clicked the button, the animal stimulus was presented at the same position on the screen. The participants were supposed to assign the animal to the fitting category by moving the mouse cursor to the respective corner and selecting the category name. The main phase included 19 trials in total.

Before participating in the main experiment, the partakers performed three practice trials to get familiar with the task.

The order of the animal stimuli was assigned completely random as well as the corner for the categories in each trial. This holds for both, the practice run and the target trials of the experiment.

After completing the experimental trials, the participants were asked to specify if they are left- or right-handed in daily life. This additional information was required for the exploratory part of the study. Therefore, it was obligatory for the participants to respond to this question whereas answering the following final standardized questionnaire on age, level of education, gender and native language was voluntary.

Data Analysis:

Measured variables. The overall time spent per trial onset and the selection of the category label was measured in each trial as well as the time between stimulus onset and first movement and the general movement duration. Furthermore, the distance in pixels covered by the cursor and the full trajectory of the cursor movement were recorded.

The measured trajectory was further divided into time-normalized and space-normalized data. For the time-normalized trajectories, all trajectories were normalized to 101 time steps and translated to start at the coordinates of (0,0). Differences in the cursor paths compared to a previously computed averaged tract were quantifiable by the *x*-coordinates in the set timesteps. In the so called space-normalized examination, the end coordinates of each trial were set to (0,0), in addition to the normalized starting point of (0,0). On the way between the start and end position, the real-time information were clustered into time bins of 500 ms to enable better analysis.

Statistical models. All analyses was done only on the correct responses. Incorrect responses were removed. For the analysis of the time-normalized data, two-sample t-tests for the difference between *x*-coordinate of each of the 101 time-steps to zero between the atypical and typical trials were conducted. To provide more statistically powerful results, the right- and left-ward responses were pooled. In addition, a more conservative test was performed. A *reliable divergence* as some *d* consecutive time-steps with significant *p*-values was established by performing a variation of bootstrapping of the 101 time steps. 10,000 simulations were performed with the same mean and standard deviation of *x* at each timestep for atypical and typical trials, simulated for each participant. The frequencies of significantly different consecutive time steps were recorded. A *d* s.t. $p < 0.01$ defined as the probability that this number of consecutive timesteps occurs randomly was picked, in order to discern whether the divergence occurred by chance or not. Once the sequence size was fixed, the t-tests were applied on each timestep between atypical and typical trials. If *d* consecutive timesteps were found, the trajectories were considered divergent. Here, *reliable divergence* was defined as at least seven consecutive time-steps that have significant *p*-values with significance at 0.05. Afterwards, pooled bins from the time-normalised data were computed to conduct a 2x3[1] repeated measures ANOVA. A planned comparison between trial types was then conducted within each bin to reveal portions of the data exhibiting divergence. For the space-normalized data, a 2x3[1] repeated measures ANOVA and planned comparisons were performed as described above.

A comparison between the accuracy[2] of atypical and typical trials was made by using a two-sample t-test against the null hypothesis of the difference being 0. Further, the performance of two-sample t-tests between atypical and typical trial types was conducted for movement duration, total categorization time, total distance traveled and movement initiation latency. Item-based repeated measures ANOVA on these same measures were examined with exclusion of the accuracy measure.

The nature of curvature across trajectories was examined. In particular, the focus was on exploring its bimodality. If AUC exhibited bimodality, the curvature of trajectories in the atypical category towards the competitor could be explained by two distinct types of behaviour, namely a movement towards the competitor followed by a course-correction, or a relatively straight path towards the answer. The z-scores for all areas were computed and subjected to the distributional analysis. A Kolmogorov-Smirnov test was performed on the difference in distribution. Bimodality coefficients were computed for each trial type. It was investigated if the coefficients were within the unimodal range of b < 0.555.

Transformations. Analysis of the data was conducted in two ways. First, all trajectories were normalized into 101 discrete time-steps, translated s.t. they all started at (0, 0). This is called the time-normalized dataset. Second, the real-time information of the trajectories were recorded by the *x,y* coordinates as the mouse cursor travelled from 0 to 1, more precise form the origin (0,0) to the final answer (1,1). Additionally, the mouse movement initiation time was calculated as well as the movement duration, the total categorization response time, and the distance traveled in pixels. Further, the area under the curve, viz. the area in pixels between the trajectory line and a straight line between the start and finish, was computed.

Inference criteria. For the evaluation of significance, the standard criteria of $p < 0.05$ was used for all tests.

Data exclusion. Participants who noted that they used a smartphone or tablet for executing the experiment were excluded because no cursor movement could be collected from these devices. Only correctly categorized trials were included in the analysis.

If participants denoted severe technical difficulties that impacted the measured variables, their dataset was excluded from the analysis. Further, a Rosner Test was used to remove any outliers in reaction time, using the median and a threshold of 3.5.

Missing data. In case a participant did not execute the entire experiment, the partly collected data was still used in the analysis.

Results

The proportion of correct answers in the atypical (M = 0.83, SD = 0.21) and typical (M = 0.96, SD = 0.10) condition was examined as well as their significant difference, $t(79.97) = -4.03$, $p < 0.001$.

Incorrect responses were removed from all further analysis and right and left responses were pooled for greater statistical power.

Time-normalized analysis:

A series of two-samples t-tests was performed to measure the divergence of trajectories between atypical and typical trials within each of the 101 timesteps (see Figure 2). The null hypothesis stated that the difference in the *x*-coordinate within each timestep is zero. Divergence of trajectories was defined as a certain amount of consecutively divergent timesteps. In order to establish this amount, a bootstrap of 10,000 simulated experiments were conducted using the same mean and standard deviation of each timestep (see Appendix). Seven consecutive timesteps occurred very infrequently ($p < 0.01$) with a criterion of $p < 0.05$ which we consider sufficient to determine reliable divergence. 54 consecutive timesteps were found that showed significance difference in *x*-coordinates (*ps* < 0.05). More precise, these were the 31st to 84th timesteps (see Figure 1).

A repeated measures ANOVA was performed with two levels of type (typical and atypical) and three levels of bins of pooled x-coordinates for timesteps (bins for 1–33, 34–67, 68–101 timesteps). There was a significant main effect for bin ($F[1.44, 73.35] = 477.610$, $p < 0.001$), a main effect for type ($F[1,51] = 15.461$, $p < 0.001$) and a significant interaction, $F[1.65, 84.32] = 6.895$, $p < 0.01$. A planned comparison between types within each bin was performed using paired t-tests in order to see which bins contained timesteps with divergent trajectories. The first bin was significant, $t(51) = -2.41$, $p < 0.05$, the second bin was also significant, $t(51) = -3.73$, $p < 0.001$, and the final bin was significant, $t(51) = -3.63$, $p < 0.001$. This indicates that all thirds of the time-normalised trajectories contained divergent timesteps between the atypical and typical trials.

Space-normalized analysis:

In this analysis, the *x*- and *y*-coordinates were normalized between (0,0) and (1,-1/1) (see Figure 2). The values were pooled into three bins, distinguished by time: 0-500, 500-1000 and 1000-1500 ms. A repeated measures ANOVA was performed, as above, containing two levels of type (atypical and typical) and three levels of bins (0-500, 500-1000, 1000-1500 ms). There was a significant main effect on bin (*F*[2, 102] = 280.854, $p < 0.001$) , a significant main effect on type (*F*[1,51] = 42.436, $p < 0.001$) and a significant interaction, *F*[2,102] = 8.360, $p < 0.001$. Additionally, a similar planned comparison was performed between the types within each bin. A statistically significant difference was found between atypical and typical trials within each bin.

Additional measures:

Two-sample t-tests were performed on the additional measures. Movement durations for atypical and typical trials were 1737 ms and 1347 ms, respectively. The two groups showed a statistical significance difference, *t*(342.148) = 5.31, $p < 0.001$. The total categorization times, more specifically the total reaction time of the participants, for atypical and typical trials were 2482 ms and 2005 ms, respectively. A statistically significant difference was found between the two groups, *t*(356.96) = 7.52, $p < 0.001$. The total distances travelled for atypical and typical trials were 839 pixels and 664 pixels. The difference in distance travelled was statistically significant, *t*(277.83) = 4.53, $p < 0.001$. Movement initialization latencies, i.e. the time it took for participants to start moving their mouse after being shown the animal word, were 745 ms and 658 ms, respectively. The difference between the two groups was not significant, *t*(354) = 1.56, $p = 0.119$.

Further, item-based repeated measures ANOVA was run on the above measures. Movement duration, total categorization time, total distance travelled and movement initiation latency, they all showed statistical significance, *F*(1,55) = 61.986, 123.637, 19.124, and 11.558, $ps < 0.05$. Mean atypical trajectories were longer in time and distance travelled.

It is possible that the divergence between atypical and typical trials could be explained by bimodality in the distribution of atypical trajectories. That is, some participants may veer of into the direction of the competitor and then correct the cursor mid-flight. This would cause a divergence in the trajectories. In order to examine whether this is the case, the bimodality of the area under the curve

(AUC) for atypical and typical trials was explored. All AUC values were converted into z-scores (see Figure 4). The difference between the atypical and typical distributions were subjected to a Kolmogorov-Smirnov test. A significant difference between the distributions was found, *D*(785) = 0.151, $p < 0.01$. Bimodality coefficients were computed for atypical and typical trials and were compared against the unimodal range of <0.555. The coefficient for atypical (0.741) and typical (0.773) trials were both outside of the unimodal range, which may indicate bimodality. These values are very similar, which implies that even if there is bimodality it may exist in both trial types. However, the result of the Kolmogorov-Smirnov test indicates that the distributions are different which is a cause for concern. It is possible that the difference in bimodality explains away the divergence in atypical trajectories.

Exploratory:

We examined the effect of which hand was dominant for the participant had on both the proportion of correct responses and the total reaction time. Two-sample t-tests were run to examine this effect. The proportion of correct responses showed no significance between the right- and left-handed participants, $t(8.71) = -0.86$, $p = 0.41$, and neither did reaction time, $t(69.65) = 1.24$, $p = 0.22$.

Conclusion

Comparison of results:

The comparison of the proportion of correct responses between atypical and typical trials showed a similar significance as in Dale et al. In the time normalised analysis, reliable divergence in the trajectories of atypical and typical trials was found as in Dale et al. Bootstrapping gave a different number of consecutive timesteps required for reliable divergence but that is a reasonable difference. The repeated measures ANOVA showed similar main effects for bin and type along with their interaction. The planned comparison in our experiment showed a significance in all bins, whereas Dale et al. found significance only in the second and third bins. They did not state the p-value for the first bin and the significance was higher for us in the first bin. It is possible that the paired t-test did now have enough statistical power in their analysis (marginally significance) but their consecutively convergent time slice did not start until the 47th timestep so it is possible that their participants simply diverged later in the atypical trials. During the space-normalised analysis we found the same significance as in Dale et al. When performing two sample t-tests on the additional measures between the atypical and typical trials, we found the same significance for movement duration, total categorization time and total distance travelled. Similarly, we found the same insignificance in the movement initialization latency. When performing item based repeated ANOVA on the same measures we found across-the-board significance, whereas Dale et al. found only marginal significance in movement duration and distance. This difference may be caused by the size of the data sample.

The largest difference between the replicated findings and the findings by Dale et al. occurred during the examination of bimodality in the area under the curve distribution. Between the atypical and typical trajectories, Dale et al. found no difference in the distribution using a Kolmogorov-Smirnov test, while in the presented study there was a significant difference. Further, the examination of the bimodality coefficient indicates potential bimodality in the replication study, but does not do so in the original experiment. The source of this difference in the results is not clear. The density of AUC is much greater around zero for the typical condition, while the atypical displays a different mode further into the positive region as can be seen in Figure 3.

Discussion

Replication:

The results of the study enable further insights into the nature of categorization. Different categories are competing during cognitive categorization and the participants' manual response is reflective of these mental dynamic processes. One disadvantage, which was also criticized by Dale et al. in the original paper, is the relatively small stimulus set of the presented experiment. As already described in great detail earlier, the presented stimuli comprised of only 19 different instances. But there exist many more suitable exemplars and there even may be other perfectly fitting atypical and typical instances for each category. Perhaps a more extensive stimuli set would enable even more insights into the human cognitive processes and would allow the researchers to dive even deeper into the analyses. A possible follow-up work with a larger stimuli set could aid the researchers to conclude more precise statements and reliable information about the categorization processes.

Another concern is that the original study from Dale et al. was performed in English language with English native speakers as participants. The currently presented experiment was executed in English language too, but unlike the original study the data was collected primarily from German native speakers. This difference in language might lead to differences in the degree of understanding and therefore could cause an unintentional bias in the results. The question arises if the difference of native speakers was of such great importance that the study could not be correctly titled as replication further on.

General:

The general question arises to what extent a study can correctly be titled as replication study. The distinction between direct and systematic replication is already very common, indicating if the original study gets simulated nearly perfectly or if few aspects are modified. But it is not transparent how much freedom the researchers own in order to vary different experimental features and still guarantee a replication. There might be a point where the variation is too daring and hence leads to too many other influences on the data in comparison to the original study. The results might then be biased to such a high degree that the label “replication” becomes unfitting. But in practice, it is very doubtful that something like an uniform limit could be specified to control the degree of variation.

Another critical aspect is the online experiment as instrument for data collection. It is indeed a fast and uncomplicated way to gather a great amount of data. But on the other side, the control options for the researchers are minimal. Although the participants in this study were explicitly asked to focus their full attention on performing the experimental task, they still might ignore this request. The researchers are not able to retrace if the participants perhaps used their phone while working on the experiment, asked a friend or the internet to find out the correct answers or if they simply were distracted by whatever caught their attention. These influences may bias the measured variables, most important the reaction time and the accuracy, and therefore could easily lead to wrong results. The researchers are able to control the mentioned effects better with other methods, e.g. when inviting the participants to a laboratory study and supervise them while performing the experiment.

These criticized formal aspects are important to remember to further improve research and strive towards maximal validity and generalizability of science.

References

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.*

American Psychological Association. *APA-style sixth edition ressources: Sample One-Experiment Paper.* Retrieved from https://apastyle.apa.org/6th-edition-resources/sample-experiment-paper-1.pdf

Drescher, L., Gauditz, H., Menzner, L., Okonnek, L. (2020). *The effect of atypical exemplars on the manual response in a categorization task - a replication study.* Retrieved from https://github.com/lokonnek/The-effect-of-atypical-exemplars-on-the-manual-response-in-a-categorization-task-REPLICATION

Footnotes

[1] type typical and atypical by the bins, 1-33, 34-67, 68-101

[2] Accuracy of atypical and typical trials here means how often participants answered correctly.

Attachment 1.

Recruitment message.

‘Hallo,

wir studieren Cognitive Science an der Uni Osnabrück und suchen im Rahmen eines Seminarprojektes Versuchsteilnehmer für ein Online-Experiment. Es wäre super wenn du 10 min Zeit hättest und unser Experiment machen könntest, klick dazu einfach auf den Link.

Die Datenerhebung läuft bis einschließlich Freitag, den 24.07.2020.

Die Studie ist anonym und die Daten werden nur für unsere Seminararbeit genutzt und danach wieder gelöscht.

Das Experiment ist auf Englisch, du solltest dich mit englischen Tiernamen etwas auskennen.

Es ist außerdem wichtig, dass du das Experiment auf einem Gerät mit Maus oder Mauspad durchführst, Daten von einem Smartphone können wir leider nicht benutzen.

Schicke den Link gerne weiter, wir freuen uns über jeden Teilnehmer!

Vielen Dank für deine Hilfe!

English Version:

Hi,

we are studying Cognitive Science at the University Osnabrueck and we are currently looking for participants for an online experiment as part of a seminar project. It would be great if you could take 10 min of your time and do our experiment, just click on the link. Deadline for collecting the data is Friday, 24/07/2020 at midnight.

The study is anonymous and the data is only used for our seminar work and will be deleted afterwards.

The experiment is in English, you should be familiar with english animal names.

It is also important that you perform the experiment on a device with mouse or mouse pad, unfortunately we cannot use data from a smartphone.

Feel free to further share the link, we are happy about each participant!

Thank you for your help!

Luisa, Lilian, Hannah & Lilith’

Table 1.

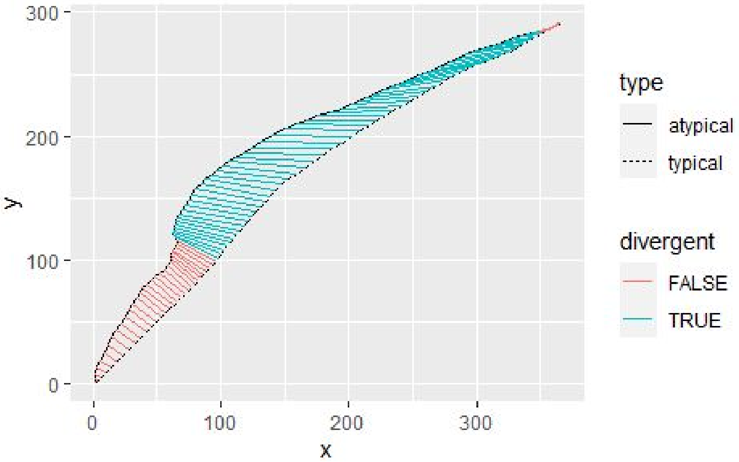
Atypical and typical animal words in the experiment.

Atypical	**Typical**
Eel (*fish*; reptile) Whale (*mammal*; fish) Sea lion (*mammal*; fish) Penguin (*bird*; fish) Butterfly (*insect*; bird) Bat (*mammal*; bird)	Hawk (*bird*; reptile) Dog (*mammal*; insect) Horse (*mammal*; bird) Shark (*fish*; mammal) Alligator (*reptile*; mammal) Rabbit (*mammal*; reptile) Chameleon (*reptile*; insect) Cat (*mammal*; reptile) Sparrow (*bird*; mammal) Goldfish (*fish*; amphibian) Salmon (*fish*; mammal) Rattlesnake (*reptile*; amphibian) Lion (*mammal*; fish)

Note: The response options presented to the participants are given in parentheses, the correct category is written in italian.

Figure 1.

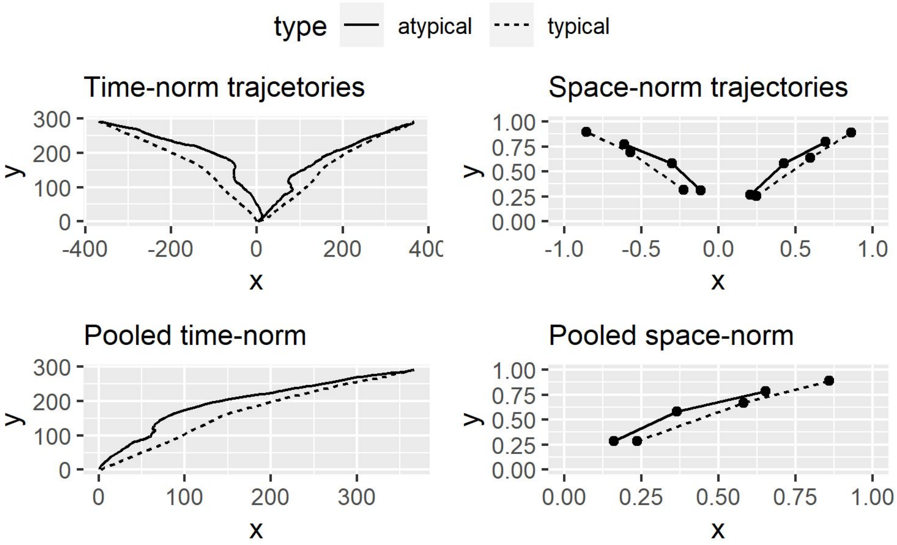
Pooled time-normalized data.



Note: A line between the points for atypical and typical trials within each timestep is drawn and coloured depending on whether it is divergent or not.

Figure 2.

Trajectories of mouse movements.



Note: Top-left is the time-normalised trajectories of mouse movements with separate left and right responses. Bottom-left is the pooled time-normalised trajectories for both directions. Top-right is the space-normalised trajectories, transformed s.t. the mouse travels from (0, 0) to (-1/1, 1). Bottom-right is the pooled space-normalised data.

Figure 3.

Density of AUC values for atypical and typical trials.

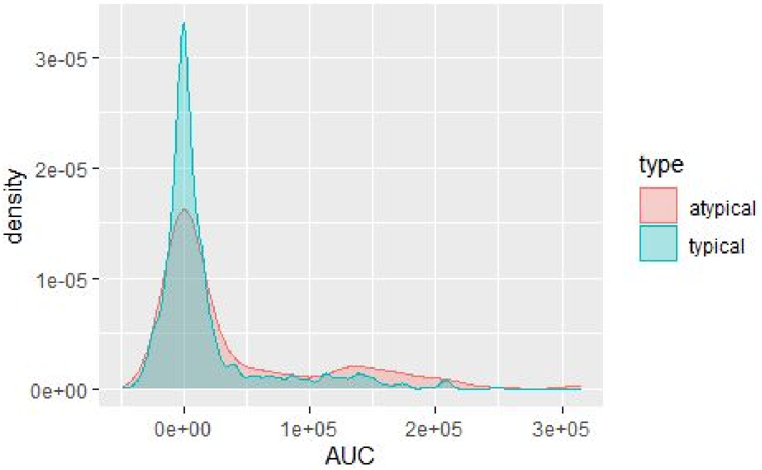
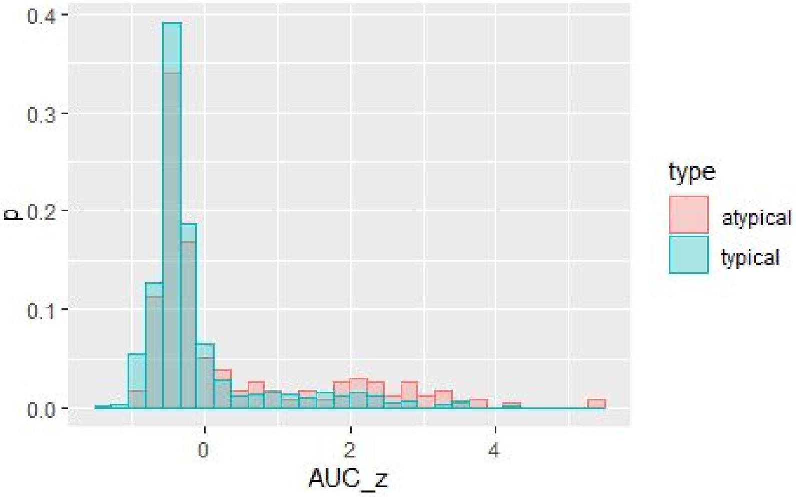


Figure 4.

Percentage histogram for the distribution of AUC z-scores.



Note: Distribution analysis showed significant differences between the typical and atypical type.

Appendix

During examination of the time-normalised data, we wished to answer the question of whether one group diverges in trajectory from the other. A certain number of consecutive divergences at the timestep level was considered sufficient evidence of divergence. Choosing this number was done by bootstrapping. For each timestep, we recorded the mean and standard deviation. We simulated 10,000 experiments for each 101 timestep and each participant (N = 57). Divergences were recorded in each experiment, and their frequencies averaged out for each sequence size. The result can be seen in Table A below. A threshold of $p < 0.01$ was used, which occurs at sequence size 7.

Table A

Sequence size	p	Frequency (%)
3		50.93
4		19.25
5		6.14
6	< 0.05	1.69
7	< 0.01	0.66
8		0.19
9	< 0.001	0.09