UiT The Arctic University of Norway

Faculty of health science, Department of Psychology

The Role of Eye-Movements in the Representation of Value in a Reinforcement Learning Context: A Web-Based Study

Pål Ovanger Stensland

Thesis, cand.psychol., (PSY-2901) February 2022, Tromsø

The Role of Eye-Movements in the Representation of Value in a Reinforcement Learning Context: A Web-Based Study

Pål Ovanger Stensland

Faculty of Health Sciences – Department of Psychology

University of Tromsø – The Arctic University of Norway

Main thesis, cand. Psychol.

Main thesis, cand.psychol.

PSY-2901

Supervisor: Matthias Mittner

February 2022

UiT / NORGES ARKTISKE UNIVERSITET

Preface

This is submitted as the main thesis for the cand.psychol. at the University of Tromsø – The Arctic University of Norway, Faculty of Health Sciences - Department of Psychology.
My supervisor came up with the original idea for the experiment and introduced me to the possibility of conducting it over the internet, which I thought sounded very interesting, especially given the potential usefulness of such studies given the circumstances of being in the midst of a global pandemic.

My supervisor has given me methodological and continuous feedback throughout the writing process and input on which statistical analysis to use and also helped with the more advanced analyzes and was also responsible for the coding for building the online experiment. I have received extensive feedback on relevant literature and input on reinforcement learning, eye-movements and in-lab versus online differences in this process. I would like to thank my supervisor Matthias Mittner for his excellent guidance and contagious commitment to the field of reinforcement learning.

Abstract

The present thesis experimentally set out to try to answer if there was a correlation between reinforcement learning and eye-movements and what the implications of such a correlation might be. An important experimental factor here was the decision to do this online, to see if it was possible to get valid and reliable results, and furthermore perhaps reach out to a more diverse group of people than a typical in-lab study would, making the results more generalizable.

38 people were recruited via the website prolific.co. The participants then performed a learning test, where they were shown two symbols on the screen, and the objective was to find the symbol with the highest value out of the pair, with three different symbol pairs. After each presentation of a pair, the participant had to choose one symbol, which were then followed by a rewarding or non-rewarding feedback. Each symbol pair had a different ratio of positive relative to negative feedback. The participants eye-movements were tracked via their web-camera, to see if they fixated more on the most rewarding symbol. The results showed that the participants reliably learned to choose the higher value symbol, comparable in validity to that of in-lab studies. We also found a statistically significant correlation between learning and fixating on the most rewarding symbol, although the quality of the eye-tracking was of too poor quality to draw any conclusions about this correlation.

The present experiment reached a diverse group of people from all over the world and proved that it is possible to perform a reinforcement learning experiment online, although the technology of eye-tracking cannot match an in-lab study. Further online research is needed in many areas to determine what type of experiments can produce valid and reliable data, which is especially relevant to the generalizability of research, and the present situation of a global pandemic which limits the in-lab approach.

Keywords: Reinforcement learning, reward, eye-tracking, eye-movements, web-study, online, generalizability, validity, reliability, Q-learning

Introduction

Humans have evolved to adapt to our environment by interpreting and learn from experience and our immediate surroundings what actions will maximize reward. These basic and fundamental behaviors and mechanisms are crucial to survival, and therefore have an evolutionary basis rooted in the brain (Schultz, 2015). To understand and predict human behavior has been a topic of study since the beginning of psychological research (Leahey, 1991) with increasingly complex theories and algorithms that seek to accomplish this.

In this context, reinforcement learning has been an area of keen interest to many psychological researchers, especially in neuropsychology (and many other fields, such as AI, statistical economics etc.) as it has become a more and more sophisticated area of which to understand how the brain works in decision making (Shteingart & Loewenstein, 2014). Reinforcement learning is a theory about how humans make decisions, and has also become a branch of machine learning, in which to understand and predict how an "agent" make options which leads to the most valued outcome. In a human context this can be understood as being relevant in most decision-making contexts, considering the paradigm that the goal in learning is to maximize reward and avoid negative outcomes, and that this learning process takes place all the time, continuously updating the brain´s representation of value, learning from mistakes (i.e., penalty) and reward (Lee, Seo & Jung, 2012).

Reinforcement learning can be tested in experimental settings, where a specific choice or behavior that leads to a more favorable outcome will be integrated and thus influence which action is taken next based on learned reinforcements. The prediction error, which is the hypothetical representation of the discrepancies between expected and actual rewards or punishments is continuously updated and is a basis for understanding the underlying mechanism regarding motivated behaviour (Chase, Kumar, Eickhoff & Dombrovski, 2015).

Reinforcement learning is divided into two different methods for learning in decision making (Dayan & Berridge, 2014). One, called the model-free strategy is based on the notion that the agent makes decisions based on the previously learned outcomes of their actions or observations, where these values are progressively updated depending on the outcome of the observation/action (reward or punishment) and therefore the person becomes more and more efficient in making the more favorable choice based on cumulative knowledge. The other method, called model-based learning, is based on the person´s orientation of the environment, expectations and predictions and calculations in order to make cognitive judgement about what might be the best choice to maximize reward.

These preference in choice and learning from experience have a neural basis in the brain, such as in the ventral striatum, which is composed of the nucleus accumbens and the olfactory tubercle, where the nucleus accumbens plays a crucial part regarding reward and reinforcement. It is a part of the basal ganglia, with dopaminergic neurons being released from the mesolimbic pathway to the ventral striatum. Dopamine is projected from the ventral tegmental area, to the mesolimbic pathway, and this is essential in the processes of mediating reinforcement learning, motivation and reward.

Furthermore, the orbitofrontal cortex in the ventromedial prefrontal cortex, which also projects to the nucleus accumbens, is involved in decision making relating to the evaluation of value, learning and inhibition (Schultz, Tremblay & Hollerman, 2000). Further relevant to our study is also the fact that is has been showed that visual brain regions, the lateral occipital cortex, is active in the widespread networks that predicts forthcoming reward (Apitz & Bunzeck, 2012).

Eye-movements and reward

Ongoing cognitive processes related to reward is not only visible in brain imaging studies but can also be shown in our eyes. Both pupil size as measured by pupillometry and eye movements are strongly linked to our attention, where reward plays a crucial role. For example, Pietrock et al., (2019) found a significantly stronger increase in pupil diameter, longer gaze fixation time and shorter blinks was linked to reward-predicting cues in contrast to a control cue.

In relation to eye-movements it has for example previously been documented that irrelevant stimuli that is learned to be associated with reward captures the gaze more than just the physical salience of an object (Krajbich, Armel, & Rangel, 2010). In a different study, Liao & Anderson (2020) found that participants that had to fixate on a peripheral target, before fixating on one of four disks than then appeared in each cardinal position. This was then followed by a reward feedback depending on the direction the participant chose to look (not the actual target position). One specific direction gave a higher reward consistently, the participants learned to choose the target with the highest reward.

After learning the direction with the highest reward, a different visual trial was performed in extinction (i.e. the subsequent test was a visual task where the goal was to find the correct item on different locations on the screen, which then turned green to indicate the correct target was found. No rewarding stimuli was given in this extinction task. This was in order to remove the previously conditioned reward response of the first task). They observed

that the eye movements of the participants where reliably biased in the direction previously linked to the high-reward in the former task, suggesting that eye-movements and attention is influenced by a previously learned reward, even after extinction.

Therefore, the study of how human (and animal) eyes operate in experimental settings has thus played an important part in neuropsychology in regards to understanding how reward is reflected in, and can be observed and measured in how our eyes behave. With such studies as the one mentioned above, arises important questions about how these processes are mediated, and if they are under voluntary control, or is more of an automatic response phenomena.

For example, Theeuwes & Belopolsky, (2012) argue that eye-movements is automatic and not top-down driven in regards to reward. They found that a task-irrelevant stimuli previously associated with high monetary reward captures the eyes much more so than the same stimulus when previously associated with low monetary reward. The irrelevant stimuli captured the eyes and disrupted goal-directed behavior, and therefore, they argue, that this response is beyond voluntary top-down control, but rather is an automatic response effect mediated by reward.

Hypotheses

In the context of representation of value in a reinforcement learning context, we expected that the participants would learn which visual stimuli was most rewarding. Based on this we ask the question if there is a correlation between fixation time and learning what stimuli is most rewarding? And if so, what does that tell us about the role of eye-movements in reinforcement learning? Does a fixation time on the more valuable symbol reveal if there is an unconscious process of conditioning the gaze to the better symbol, or is it simply a top-down attentional phenomenon of looking at the symbol the participant learn to choose? Although similar experiments involving eye-movements and reward have been conducted in the past, the new factor here is how we chose to conduct it, which brings us to the next segment:

The Online Approach. The relevance of online studies during the current COVID-19 pandemic situation has made it an area of which it is important to explore the benefits and limitations of such experiments, as such information would be of great significance as there may be many potential future similar scenarios where in-lab experiments are limited or not possible.

It is maybe doubtful that web-based studies involving neurological parameters such as eye-movements will be as sophisticated as in a lab in the near future, due to the nature of necessary control required in a lab setting to get reliable and valid data. In-lab studies have the advantage of controlling every aspect of the experiment, for example being able to observe the test subject is an important aspect (Schmidt, 2009). Furthermore, the researchers presence makes the process precise in a way where any confounding factors may be removed, which would be factors that are very difficult to control for in an online based study. When the participant is not observed in a lab, the equipment each individual use can be of varying quality (Schmidt, 2009) i.e. computer speed, poor web camera and so on. Also, their behavior during testing, with outside distractions or other interferences might make the results less valid. That said, advances in technology regarding eye-tracking online has made considerable progress, for example with the use of a real-time, web-browser-based webcam eye-tracker such as WebGazer (Papoutsaki, Gokaslan, Tompkin, He & Huang, 2018).

A web-based approach could be a very useful tool in general, as it has the advantage of potentially reaching hundreds or thousands of participants in matter of minutes, performing studies that would take weeks or months in the lab, just over the course of minutes or hours (Riva, Teruzzi & Anolli, 2003). An online also study also has the potential advantage of reaching a much more diverse group of people from different cultures and socio-economic environments (Riva et al., 2003). As such, it can be an important area if interest to replicate and confirm data from an in-lab experiment, to have a larger degree of generalizability to the general population.

Regarding the generalizability of many small in-lab studies that often recruits people from western, educated, industrialized, rich, democratic societies, where a large proportion is university students, is then often generalized to other populations (Henrich, Heine & Norenzayan, 2010). As participants from the group just mentioned often are outliers compared to the rest of the population (Tiokhin, Hackman, Munira, Jesmin, & Hruschka, 2019), it can clearly be problematic to accept the notion that data acquired from these types of studies can be applied as completely relevant to the whole population. That said, people from less privileged backgrounds without the resources to access online experiments may be excluded from participation to a degree (Lourenco & Tasimi, 2020), so its generalizability comes with some limitations. Besides its potential pitfalls in interpreting the generalizability, studies involving less complex variables, such as questionnaires or the like would be less liable to be confounded by the limitations of more complex measurements such as in neurological testing.

As eye-tracking as mentioned is one area where neurological variables actually can be tested online, in this experiment we sought out to see if it was possible to conduct this type of experiment in a web-based manner, and if the experiment would provide valid and reliable data, and if so, if both the eye-tracking- and reinforcement learning data could be of value in a generalizable way.

Methods

Participants

40 participants of both genders (18 men and 22 women) were recruited via the website Prolific.co. The participants were between 18 to 50 years of age (M_{age} = 28.5) from all over the world; including 4 different countries in Africa, 9 different in Europe, 1 in Asia and 2 different in South-America. 11 were in full time jobs, 1 was unemployed, 13 were students (2 with full time- and 4 with part time jobs on the side, 7 unemployed 2 identified as "other"). The rest was either "other" or had not answered.

The requirements for participation where that they were between 18 to 50 years of age, where fluent in English, had no mental impairments, no mental illness and intact eye vision and had access to a computer equipped with a web camera.

We excluded participants that had 3 stops or more, i.e. where the participant did not choose a symbol before the trial time of 3.3 seconds run out (see figure 1 for visual illustration). We did not set an absolute exclusion criteria for times exiting the test window, but rather made the decision if they should be included based on number of times exiting in addition to being outside the test window for several seconds, which would be indicating they were not focusing on the experiment, making the data unreliable.

Two participants were excluded from the study due to stopping too many times, where the two whom were excluded stopped 3 and 6 times, respectively.

Study Setup

The website Prolific.co is a website dedicated to online research, where people from all over the world can register on the site, where their identity is confirmed via picture of their passport, e-mail and phone number. They then can answer different question about themselves, such as land of origin, age, gender, ethnicity etc. such that researchers can filter their participants on certain criteria in order to focus the group in a manner that best serves the experiment.

The participants gets paid a certain amount of money for participating in research projects of their own choosing, as long as they are within the eligible group chosen by the researcher. Besides the information provided, the participants identity remains completely anonymous to the researcher.

The researcher sets up a server site to their experiment where the participant is then sent to via prolific, and upon completing the study, gets a code which they then enter in prolific to confirm completion of the study. In this experiment, the data was sent to the server JATOS (https://www.jatos.org) where we then could immediately see the total time spent, time spent on reading the instructions, how many times the participant did not choose a symbol, when they entered full screen mode, how long they were in full screen mode, how many times they switched windows and for how long they had been in a different window after switching.

Experimental Design. The online task was adapted from (Turi et al., 2015), originally from Frank, Seeberger, and O`Reilly (2004), and was written in jsPsych (https://www.jspsych.org). As in Turi et al. (2015), the experiment had a training phase before the main test, in order to familiarize the participants with the structure of the test (see figure 1) borrowed from Turi et al. (2015), with some alteration regarding the timeframes. Three pairs of Japanese symbols were presented together, the same pairs as used in the experiment by Turi et al. (2015), one pair at a time, in a random order. Here labeled pair AB, CD and EF, where each pair contained one symbol connected to a higher reward than the other.

As shown in figure 1, the symbol A was the “better” choice with the reward of a happy face in 80 % of the times it appeared on the screen, and 20 % a sad face emoji. Vice versa, symbol B had the same reward 20 % of the time, and 80 % the sad emoji. As with the symbol pair AB, the other pairs also had a percentage of reward probability of one symbol over the other, albeit a different one, where CD had a balance of CD 70% and 30% respectively, and EF 60 % reward for E and 40 % for F.

As adopted from (Turi et al., 2015) the training phase had 1 block where the 3 symbol pairs were shown a total of 15 times, 5 trials for each symbol. With each presentation of a pair, one symbol was placed on the left and the other on the right of the screen. The “better” symbol was randomly presented on either the left or right side of the screen. The main test consisted of 3 blocks with 60 trials in each block (with 180 presentations in total).

Between each block the participant could take a short brake. Each block had a new set of symbol pairs. Each trial lasted for a total of 3.3 seconds, where it started with showing of a fixation cross placed in the middle of the screen with a timeframe of randomly chosen either

200, 400 or 800 milliseconds. The symbol pair then appeared on the screen for 1700 milliseconds where the participant had to choose one of the symbols using the key “F” on the keyboard for the left symbol or the key “J” for the right one. The chosen symbol was then highlighted for 200 milliseconds. If the participant had not made a choice within the given timeframe the trial was terminated, followed by a feedback symbol representing a “confused” face for 200 milliseconds. If a symbol was chosen, the feedback was either a “happy” or a “sad” face. The feedback faces were designed in a typical emoji fashion and shown for 200 milliseconds before moving on to the next trial (for reference, see figure 1).

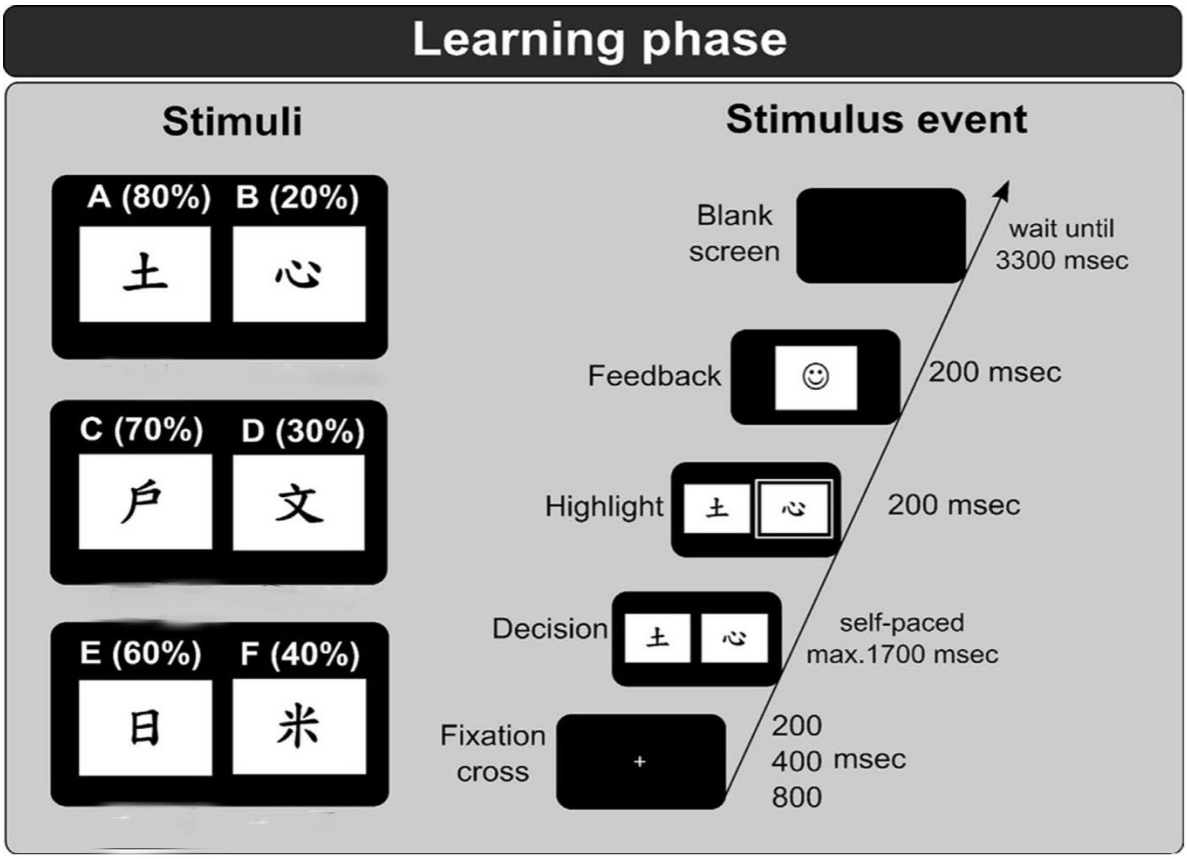


Figure 1 Visual presentation of one block of the main test

Before starting the experiment the participants were provided detailed instructions on how to perform the experiment. They were given 2 GPB (pound sterling) if they performed the task correctly. As an incentive to motivate the participants to focus and be motivated to do well on the test, they were informed that if they performed well, they would be given a bonus of 1 GBP in addition to the original payment. They were also notified that we measured the keypresses, eye-movements and switching between windows, and given a upper limit time frame of completion of the test of 30 minutes (M = 17.9, SD = 2.04) Furthermore, upon data

analysis we would make the decision whether they had followed the instructions correctly, and if not, they would not get the payment.

Eye-Tracking. After the instructions and information on the test where given, an eye calibration was performed using the software WebGazer (Papoutsaki et al., 2016) paired with the participants web camera. They were then given the task of holding their head still within a green box and then instructed to look at 5 little black dots on the screen and click on them using their mouse. After this, the calibration continued with the participant looking at the dots again, this time without clicking on them. This was performed to get the most reliable data from the eye tracking during the experiment.

Reinforcement Learning. As detailed in Turi et al. (2015), processes and behaviour can be described, modeled and in retrospect fitted to the behavoiural data from the experiment using specific reinforcement learning algorithms, in this case, Q-learning. We used here the same modified version of the Rescorla-Wagner algorithm as in Turi et al. (2015): $Q_{t+1}(i) = Q_t(i) + \alpha(r_t - Q_t(i))$ for $i \in \{A,B,C, D; E, F\}$, where *t* represents the trial number. The so-called action values Q for each single item where initialized to zero and then updated progressively. Furthermore, the prediction error defined as $r_t - Q_t(i)$, as the difference between the given and expected feedback, and r_t meaning the reward given on trial *t*.

As further described in Turi et al. (2015), retrieved from Frank, Moustafa, Haughey, Curran & Hutchison (2007) & Jocham, Klein & Ullsperger (2011). Q-learning involves a learning rate parameter; α, which represents the "*the difference between the previous outcome estimate and the actual estimate after a certain action"* and where a higher α-value indicates a more irregular shifting pattern between the choices made, and a lower α-value indicates a more "*gradual value integration and more stable value estimation*" (Turi et al., 2015), with reference to Frank et al. (2007). Furthermore, in the Q-learning algorithm used here (Frank et al., 2007; Jocham, et al., 2011) the β-value parameter points to the test-subjects leaning towards either choosing a strategy either towards exploitation or exploration, where higher β-values indicates choosing more randomly, whereas lower ones indicate exploitation, where the better option is chosen.

To sum up; as detailed in Turi et al. (2015), the relationship between α and β represents a model thought to show the α-value as to what degree the test-subject learn from previous choices and therefore predicts and chooses the most optimal choice. We then used

the soft-max rule to calculate the probability of choosing one symbol in the pair presented, as in for example choosing A when the pair AB was shown: $P_t(A) = \exp(Q_t(A)/\beta)/[\exp(Q_t(A)/\beta) + \exp(Q_t(B)/\beta)]$.

Results

Participants used an average of 17 minutes and 9 seconds on the test (*SD* = 2.04) including eye calibration, pauses between blocks and reading the instructions. Number of stops (where the participant did not choose a symbol before the trial time of 3.3 seconds run out) was overall few (*M* = 0.18, *SD* = 0.6). We also recorded how often participants switched windows on their computers. The number of times switching into a different window than the test had an average of 2.2 switching times (*SD* = 1.58), but the time being spent opening/switching windows was short with the longest time averagely spent outside the test window was 0.28 seconds (*SD* =0.58).

We calculated the accuracy of choosing the right symbol for each participant. To see if there was a learning effect, we split each block in two parts of equal size, to compare if there was a difference in accuracy in the first and second 30 trials of each block. If participants learned the correct stimulus values over time, they should choose the higher-valued symbol more often in the second compared to the first part.

We used a 2 (parts: first half vs. second half) x 3 (symbol pairs: AB, CD and EF) repeated measures ANOVA, with accuracy as the dependent variable (see table 1 for descriptive statistics). Accuracy is here defined as when the symbol with a higher percentage of reward feedback was picked, therefore it was also defined as correct also when the feedback was negative, and vica versa when the wrong symbol was chosen but positive feedback was given, it was labeled as the wrong choice. Assumption of sphericity was tested using Mauchly´s test of sphericity, which were violated, therefore the Greenhouse-Geiser correction was used. Significance level was set a level of $p < .05$.

Table 1

Accuracy for choosing the right symbol, comparing the first and second half of each of the 3 blocks, where each block was 60 trials. 30 trials in the first half and 30 in the second half of one block was compared, with 3 blocks with a total of 180 trials.

	First		Second	
Variable	*M*	*SD*	*M*	*SD*
AB	.59	.17	.67	.15
CD	.53	.17	.66	.13
EF	.50	.17	.53	.19

Note. M and *SD* represent mean and standard deviation, respectively. (N = 38)

There was a learning effect observed, where the participants learned which symbol had the highest probability of reward (see table 1). The repeated measures ANOVA revealed a main significant learning effect in general between the first and second half of the blocks ($F(1.37) = 21.8$ $p < .001$, $\eta_p^2 = .37$). Furthermore, there was a statistical significant difference between the symbol pairs ($F(1.9,69.2) = 9.2$ $p < .001$, $\eta_p^2 = .2$). There was also a statistical significant interaction, indicating that the difference in learning effect when comparing the different symbol pairs between the first and second half of the three blocks was not uniform ($F(1.9,73.9) = 3.78$ $p < .028$, $\eta_p^2 = .09$), (see figure 2).

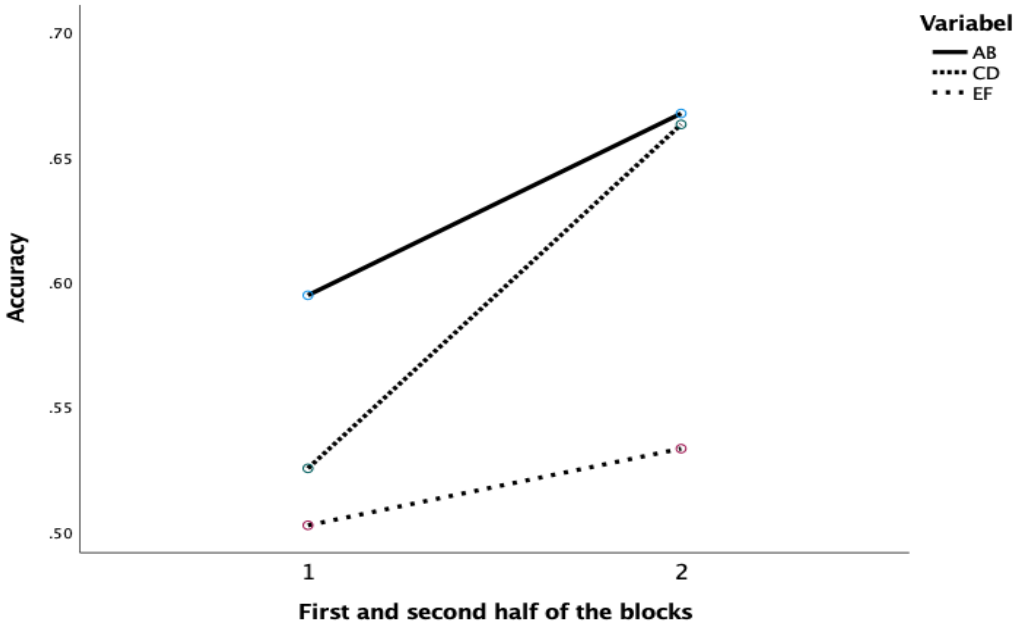


Figure 2. Shows the three symbol pairs, with the mean of each participant´s accuracy between first and second half of the three blocks

Reinforcement Learning

As detailed in (Turi et al., 2015), processes and behaviour can be described, modeled and in retrospect fitted to the behavioral data from the experiment using specific reinforcement learning algorithms, in this case, Q-learning. We used here the same modified version of the Rescorla-Wagner algorithm as in (Turi et al., 2015): $Q_{t+1}(i) = Q_t(i) + \alpha(r_t - Q_t(i))$ for $i \in \{A,B,C, D; E, F\}$, where *t* represents the trial number. The so-called action values Q for each single item were initialized to zero and then updated progressively. Furthermore, the prediction error defined as $r_t - Q_t(i)$, as the difference between the given and expected feedback, and r_t meaning the reward given on trial *t*.

As further described in (Turi et al., 2015), retrieved from Frank, Moustafa, Haughey, Curran & Hutchison (2007); Jocham, Klein & Ullsperger (2011), Q-learning involves a learning rate parameter; α, which represents the "*the difference between the previous outcome estimate and the actual estimate after a certain action"* and where a higher α-value indicates a more irregular shifting pattern between the choices made, and a lower α-value indicates a more "*gradual value integration and more stable value estimation*" (Turi et al., 2015), with reference to Frank et al. (2007).

Furthermore, in the Q-learning algorithm used here, (Frank et al., 2007; Jocham, et al., 2011) the β-value parameter points to the test-subjects leaning towards either choosing a strategy either towards exploitation or exploration, where higher β-values indicates choosing more randomly, whereas lower ones indicates exploitation, where the better option is chosen. To sum up; as detailed in Turi et al. (2015), the relationship between α and β represents a model thought to show to what degree the test-subject learn from previous choices and therefore predicts and chooses the most optimal choice.

We then used the soft-max rule to calculate the probability of choosing one symbol in the pair presented, as in for example choosing A when the pair AB was shown: $P_t(A) = \exp(Q_t(A)/\beta)/[\exp(Q_t(A)/\beta) + \exp(Q_t(B)/\beta)]$.

Table 2

Average α and β values for the participants, N=38

Variable	*M*	*SD*
α value	.15	.11
β value	.32	.05

Note. M=Mean, *SD*=Standard deviation.

The average α and β values for all the participants is presented in table 2. An example of the Q-values for one individual participant across the blocks (see figure 3) shows that the participant learned to identify the most rewarding symbols, as represented in the Q-values (and the mean α and β values) representing the internal, accumulated value associated with each symbol. The Q-value for the most rewarding symbol went up, and Q-values for the less rewarding symbols went down. This demonstrates a reinforcement learning effect, where the cumulative knowledge given from reward feedback gradually updated the individual's representation of value, and therefore gradually choosing the higher valued symbol more often (for Q-values for all the participants, see Appendix).

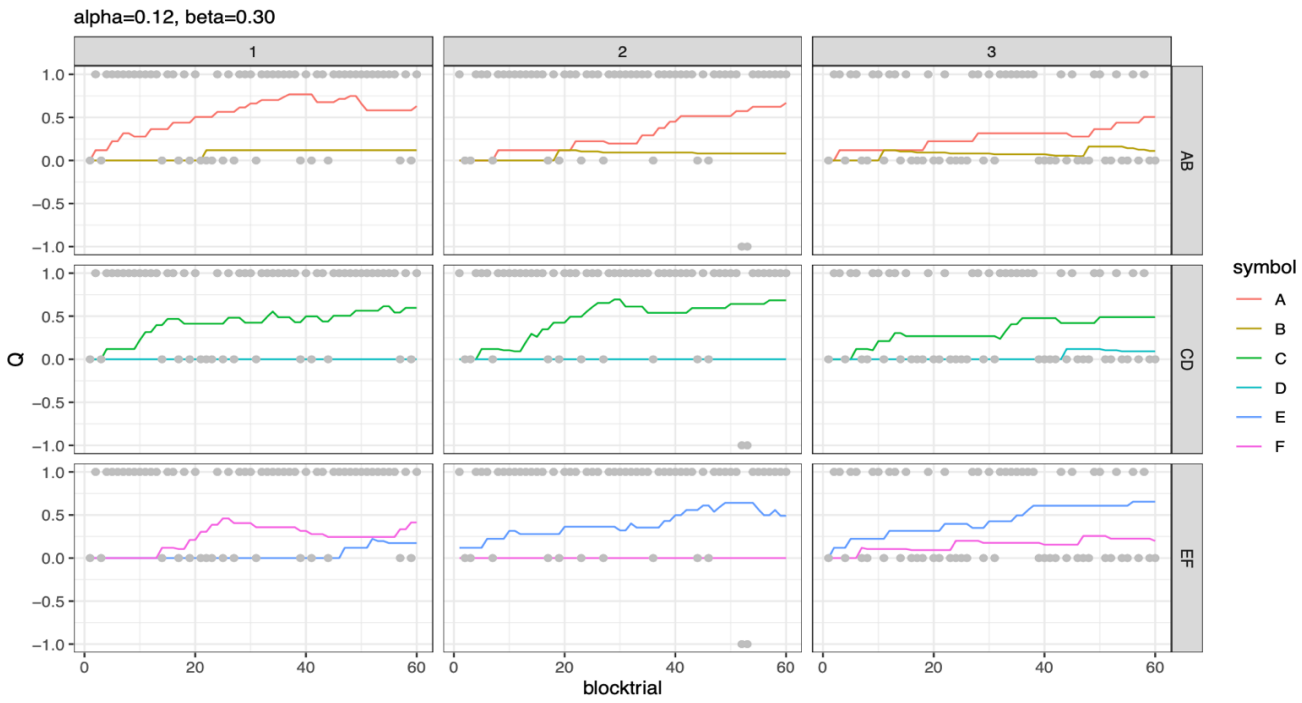


Figure 3. Q-values for one participant, showing each symbol pair across the three blocks. Average α and β value for this individual displayed at the top. Grey dots indicated individual decisions, with correct decisions at y=1 and incorrect decisions at y=0.

Eye-Tracking

Regarding the eye-tracking calibration, the results varied quite a bit in accuracy where around 21 participants had an acceptable calibration (evaluated subjectively by the analyst, see figure 4 left for an example), 7 were of poorer quality but nevertheless mostly acceptable to determine in what direction the participant had fixated, and 10 were of poor quality. Our main analysis quantified the percentage of time our participants spent looking at either the left or the right part of their screen. Hence, it was most important that they looked in the direction of the dots presented on the screen during calibration (or the symbols during trials) while the absolute accuracy of fixation (especially in the vertical direction) was less important. 10 participants had poor quality eye tracking, as shown in the figure 4 (right), where one can infer that something was not optimal to track the eyes, thus making the eye-movement data of this participant less valid in the analysis compared to the one to the left with much more precise data. The main analysis was nevertheless run on the full dataset including all subjects' data.

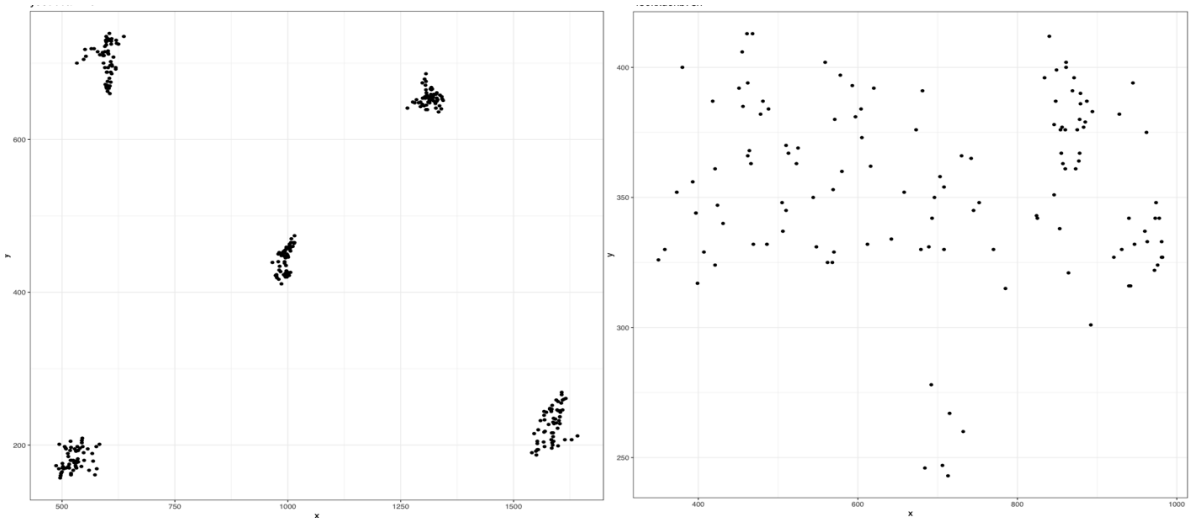


Figure 4. The calibration on the left has clear and defined fixations on the dots that appeared on the screen, whereas the one to the right has an almost random pattern.

Q-Values and Eye-Movements. We calculated the percentage of time our participants focused on the left vs. the right side of the stimulus display during the time the stimulus was displayed on the screen and correlated it to the Q-value derived from fitting the computational model to the response data.

Figure 5 shows the average correlation (correlations were transformed using Fisher's Z-transformation) between these quantities for intervals of increasing duration (i.e., the percentage of time participants focused on the left side was calculated based on data collection during the interval from stimulus onset to 100 ms, 200 ms, 300 ms and so on). The rationale behind this analysis was to identify the time window during which the higher-valued stimulus was openly attended to (viewed). The confidence interval for the average, transformed correlation coefficient excludes zero for intervals up to [0, 800] ms. From 900 ms onwards, the correlation is significantly different from zero, Z=0.09, 95% CI=[0.01, 0.16], indicating that the higher-valued symbol (that was also more likely to be chosen) received a higher percentage of viewing time.

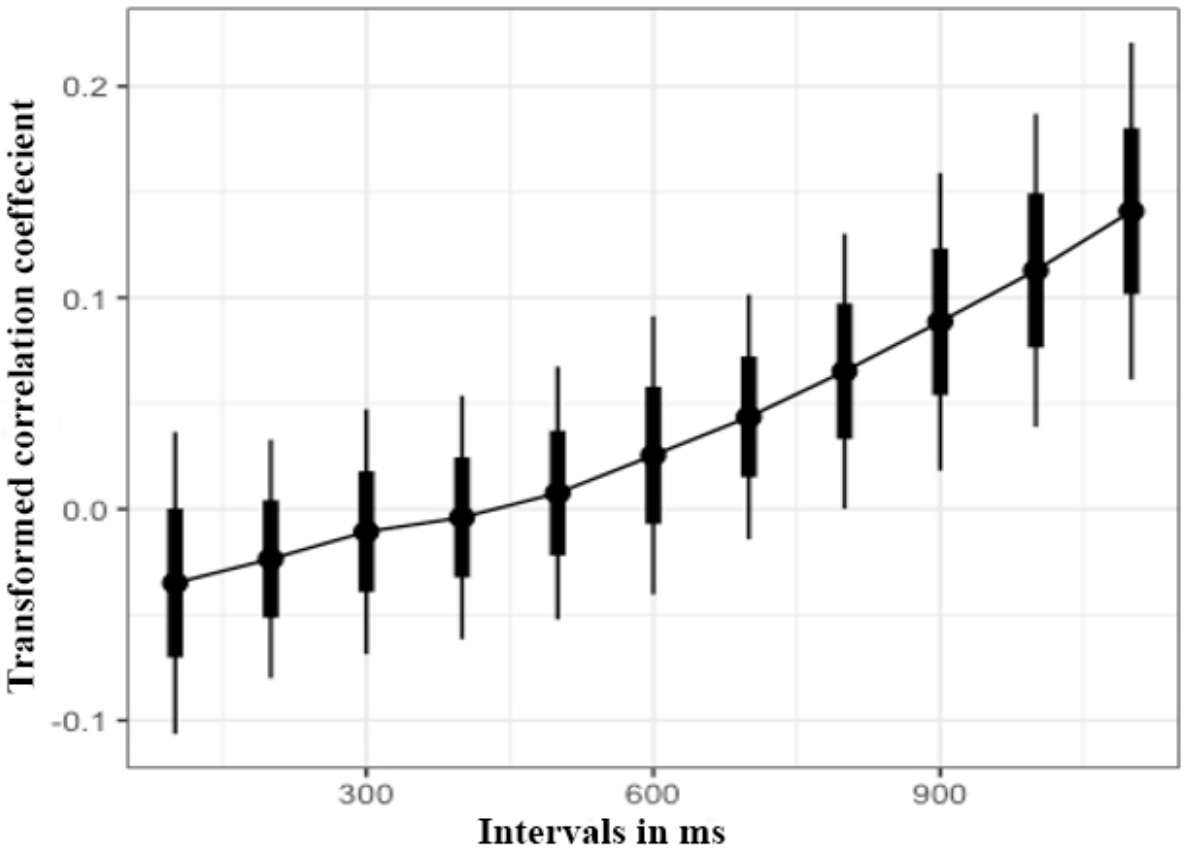


Figure 5. Each point represents percentage of gaze of all participants (N=38) with the intervals from 100 to 1100 millisecond and the Fischer´s Z-transformation on Y-axis. The lines in bold are the standard error and the thinner lines are the confidence interval.

Discussion

In this study we found that participants reliably learned to pick the higher value symbol, and that the higher the percentage of reward/accurate feedback in a symbol pair, the higher the percentage of choosing the most valued symbol when comparing the first and second half of the three learning blocks. The AB-pair had a higher percentage accuracy than the CD-pair, and the CD- higher than the EF-pair. We further found that these results were statistical significant, both in the main learning effect between the first and second half of the three blocks and the difference between the symbol pairs throughout the study. Furthermore, we also found that the learning effect between the pairs from the first half to the second half of the three blocks were statistically significant. The highest valued symbol pair AB received a higher mean total number of correct answers than the higher valued symbol pairs CD and EF in the three pairs, respectively, between the first and second half of the blocks. These effects are well-known and established in similar studies using the same task in a lab-based setting (Turi et al., 2015).

Furthermore, the Q-learning algorithms fitted to the data set confirmed the general learning across the group, where the participants reliably seemed to update (learn) the representation of value of the symbols, as reflected in the Q-values and the mean of the α- and β values, confirming a reinforcement learning effect. The estimated parameters are very similar in magnitude to those estimated in previous studies (Turi et al., 2015) establishing the general ability of our web-based study to capture the same phenomenon measured in the lab.

The Q-value derived from fitting the computational model to the response data in regards to the correlation between the reinforcement learning and eye-movements; more specifically, the time to fixation in milliseconds and its correlation to the actual key press of choosing the right symbol: We found a significant correlation between the time to fixation at the higher value symbol and the reinforcement learning of choosing the right symbol being at the same mean millisecond interval for which the participants made a choice.

The eye-calibration and general eye-tracking data was of varying quality, where we found that the quality of the eye-tracking was good enough to track where the participant had fixated in the majority of participants, but there was however a significant number of participants where the eye-tracking data was in the range of poor to very poor quality.

As expected, the participants learned to pick the higher value symbol progressively, and we expected the participants ending up in the end of each block with a percentage of correct answers close to the percentage of the positive feedback percentage for each pair, which was also the case (i.e. for example in the AB pair ending closer and closer to 80 % accuracy in the end of the block, because that was the percentage positive feedback for A in the pair). Surprisingly, the CD-pair had a higher-than-expected percentage of right answers relative to its feedback, compared to the AB-pair, given the higher accuracy in feedback in AB. The reason for this remains unclear. Even the EF pair had a learning effect, although as expected, with a lower mean learning accuracy compared to the two other pairs.

The reinforcement learning parameters reliably confirmed the hypotheses that the participants continuously updated the representation of value and making inferences based on the previously learned values in which symbol to choose next. The correlation between reinforcement learning and eye-movements found at 900 ms and later could hypothetically point to a top-down driven process and being just a natural response to shifting the attention to the symbol chosen. Theeuwes & Belopolsky (2012) points to the fact that attentional capture and holding of attention are bottom-up and top-down driven respectively, and is two different processes, where one is salience-driven and the other following exogenous capture of attention.

Their study was based on the notion that stimuli previously associated with reward captured the eyes in favor of a top-down and goal-oriented task of looking for a target. In our study the time of eye-fixation and answering was coincident, and in addition, the target to look for was also the stimuli with the highest reward. So, in order to answer the question if there was an unconscious and more automatic response to the more rewarding symbol would maybe have to be based on whether the eye-fixation preceded the choice, which our results show that it didn´t. Another way would be to have more accurate reading of fast saccades, as fast saccades are more predictive of an automatic response than just a fixation (Theeuwes & Belopolsky, 2012). The time to fixation co-occuring with choice may just mean these fixations are just part of the top-down decision process.

That said, since the eye-data was of much less valid quality than the reinforcement learning data, where there was a significant amount of participants where the eye tracking was either completely scrambled or of very poor quality, one could hypothesize that if the

experiment only included the participants with optimal eye-tracking, this result would maybe be different. Though these hypotheses are pure speculation, it could also be that the if the eye-movements were measured in a lab and analyzed with the reinforcement learning data from our experiment, we would get a completely different result in correlation.

The reasons for the shortcomings in eye-movement data could be due to many factors. For example, in the participants where the eye-tracking/calibration was poor, one could also see the lower sample rate of the camera, simply by the difference in how much fewer fixation points were recorded (as seen in the difference between the two calibrations in figure 4). This could be due to an old or slow computer, poor quality web camera etc. Although interesting that we did find a correlation, it nevertheless cannot compare to the results found in a lab with more advanced eye-tracking equipment where the sampling rate is so much higher that saccades can be measured accurately. So as of now, this type of online eye-tracking experiment is not something that can replace any in-lab experiment or give valid and reliable data. Here one could not compensate for poor measurements with a larger sample size either, as the technology just is not there yet to get the data results needed for such complex measurements. As mentioned before, measuring eye-movements requires a large degree of control of the experiment situation, which is not possible as of now in a web-situation, as we experienced in this study.

All that said, eye-tracking can still be used as measurements in studies as a way of ensuring participant compliance. For example, if we conducted this experiment only in regards to reinforcement learning, the eye-tracking data could help in determining whether the person is actually paying attention or not, as this does not require the same sample rate needed in a correlation study such as this, to just monitor if someone has an eye-movement pattern compatible with the experiment.

Regarding the demographic data, the participants were a highly diverse group in terms of ethnicity, country of origin and occupation. Although a small sample size, this suggest that this type of online testing can be of high value in terms of a higher degree of generalizability, especially if used in conjunction with in-lab studies, when attributing test results to the general population, although as pointed out earlier, with some limitations (Lourenco & Tasimi, 2020), regarding for example people from third-world countries with poor living conditions without access to the internet. But still, the fact that we were able of being able to

reach such a diverse group in this small study, where many did not belong to a typical group of people from Western, Educated, Industrialized, Rich, and Democratic (WEIRD) societies (Henrich et al., 2010), may have important implications regarding the results of future studies, where more and more are conducted online.

Furthermore, the learning effect and high compliance in this experiment was seen in most of the participants. As pointed out, that there was a learning effect even in the EF-pair points to the fact that the participants were clearly focusing during the test, as one would expect it to be difficult to learn the higher value symbol with a ratio of 60/40 % reward. This compliance is proven by the fact that most of the participants followed the instructions; the data results shows that most of them answered in every trial, and that the times they exited the test window were for such a brief time that it most likely was due to maybe clicking away incoming messages, by mistake, or other smaller interferences.

The reasons for this compliance may be due to the fact that the design of the experiment was pretty simple and straight forward in its task principles, combined with the salient and simple layout of the test together with the visual feedback of emojis which have been shown to elicit affective arousal (Fischer & Herbert, 2021), making it user friendly and maybe even entertaining. We hypothesize that these factors contributed to the learning effect and the motivational focus seen in most of the individuals who participated in the study.

Furthermore, the motivation for doing well on these tests is most likely also mediated by the fact that they get paid for participation, and maybe more importantly, gets extra paid for doing well. The implication of this is that giving monetary rewards for participation and extra rewards for doing well modulates the degree of involvement and focus and may be a necessary prerequisite to ensure that the data is reliable and valid, as people probably would not just use 20 minutes on an online experiment without these incentives.

Further Research/Ethical Considerations

Considering ethical issues that might arise from web-based studies, it doesn´t require much imagination to think of a scenario were personal information in a study done online is misused or even susceptible to hacking attempts, so the privacy of the individual is a consideration which is of utmost importance if the website ensures anonymity, especially if the study involves information about the individual that is of a sensitive character or

something that can be taken advantage of, so the importance of using serious websites cannot be overstated enough. That said, the data from the video during eye-tracking is sent to our server are just meaningless coordinates, not the actual images of people, since the video is stored only on the individual's own computer, hacking would have to take place on that individual's computer (which the website obviously would not be responsible for).

New research technology develops continuously, building upon experience, just as the reinforcement learning in this experiment. The limitations of some of the data collection we encountered might not necessarily be one of the areas which will develop enough in the near future to discard of the lab just yet (if that for some reason was someone's goal). And although it may not be a revolutionary discovery that rewarding stimuli is reinforcing (Pavlov, 1928), the value of knowing that such experiments can be done online sets the stage for further exploration of how web-based research can be further developed. Taken together, although the limitations of the eye-tracking makes the hypotheses regarding eye-movements in this experiment less valid and thus less informative than the learning part, that is precisely also the reason for doing such experiments, as it gives very valuable information about what kinds of experiments are possible to conduct at this point in time.

Conclusion

The online eye-tracking part of the experiment is not valid due to the limitations of technology as of now. In conclusion, to do an experiment such as ours, trying to find a correlation between reinforcement learning and eye-movements must be performed in a lab. That said, regarding reinforcement learning, the total learning effect of the participants suggests that a web-based experiment measuring these parameters gives valid and reliable data if the experiment is set up in a user-friendly way and is conducted through serious and secure websites. All in all, with such a diverse demographic group of people, experiments such as this have a generalizability (with some limitations) that exceeds that of most in-lab studies performed in western, educated, industrialized, rich democratic societies. And even in a relatively small sample such as ours (disregarding the eye-movement part) the behavioral and modeling results were similar to that which is observed in lab studies

References

Apitz, T., & Bunzeck, N. (2012). Reward modulates the neural dynamics of early visual category processing. *NeuroImage*, *63*(3), 1614–1622. https://doi.org/10.1016/j.neuroimage.2012.08.046

Chase, H. W., Kumar, P., Eickhoff, S. B., & Dombrovski, A. Y. (2015). Reinforcement learning models and their neural correlates: An activation likelihood estimation meta-analysis. *Cognitive, affective & behavioral neuroscience*, *15*(2), 435–459. https://doi.org/10.3758/s13415-015-0338-7

Dayan, P., & Berridge, K. C. (2014). Model-based and model-free Pavlovian reward learning: revaluation, revision, and revelation. *Cognitive, affective & behavioral neuroscience*, *14*(2), 473–492. https://doi.org/10.3758/s13415-014-0277-8

Fischer, B., & Herbert, C. (2021). Emoji as Affective Symbols: Affective Judgments of Emoji, Emoticons, and Human Faces Varying in Emotional Content. *Frontiers in psychology*, *12*, 645173. https://doi.org/10.3389/fpsyg.2021.645173

Frank, M.J., Moustafa, A.A., Haughey, H.M., Curran, T., & Hutchison, K.E. (2007). Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. *Proceedings of the National Academy of Sciences, 104*, 16311 - 16316.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world?. *The Behavioral and brain sciences*, *33*(2-3), 61–135. https://doi.org/10.1017/S0140525X0999152X

Jocham, G., Klein, T. A., & Ullsperger, M. (2011). Dopamine- mediated reinforcement learning signals in the striatum and ventromedial prefrontal cortex underlie value-based choices. The Journal of Neuroscience, 31(5), 1606e1613.

Krajbich, I., Armel, C. & Rangel, A. (2010) Visual fixations and the computation and comparison of value in simple choice. *Natural Neuroscience* 13

Leahey, T. H. (1991). *A history of modern psychology*. Englewood Cliffs, N.J: Prentice Hall.

Lee, D., Seo, H., Jung, M.W. (2012) Neural Basis of Reinforcement Learning and Decicion Making. *Annual Review of Neuroscience*, 35, 287-308

Liao, M. R., & Anderson, B. A. (2020). Reward learning biases the direction of saccades. *Cognition*, *196*, 104145. https://doi.org/10.1016/j.cognition.2019.104145

Lourenco, S. F., & Tasimi, A. (2020). No Participant Left Behind: Conducting Science During COVID-19. *Trends in cognitive sciences*, *24*(8), 583–584. https://doi.org/10.1016/j.tics.2020.05.003

Papoutsaki, A., Gokaslan, A., Tompkin, J., He, Y., & Huang, J. (2018). The eye of the typer. Proceedings of the 2018 ACM *Symposium on Eye Tracking Research & Applications*, 1–9. https://doi.org/10.1145/3204493.3204552

Papoutsaki, A., Sangkloy, P., Laskey, J., Daskalova, N., Huang, J., & Hays, J. (2016). *WebGazer: Scalable Webcam Eye Tracking Using User Interactions*. *IJCAI*.

Pavlov, I. P. (1928). *Lectures on conditioned reflexes*. (Translated by W.H. Gantt) London: Allen and Unwin.

Pietrock, C., Ebrahimi, C., Katthagen, T. M., Koch, S. P., Heinz, A., Rothkirch, M., & Schlagenhauf, F. (2019). Pupil dilation as an implicit measure of appetitive Pavlovian learning. *Psychophysiology*, *56*(12), e13463. https://doi.org/10.1111/psyp.13463

Riva, G., Teruzzi, T., & Anolli, L. (2003). The use of the internet in psychological research: comparison of online and offline questionnaires. *Cyberpsychology & behavior: the impact of the Internet, multimedia and virtual reality on behavior and society*, *6*(1), 73–80. https://doi.org/10.1089/109493103321167983

Schmidt, W. C. (2009). Technical considerations when implementing online research. In *Oxford Handbook of Internet Psychology* (1st ed.). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199561803.013.0029

Schultz, W., Tremblay, L., & Hollerman, J. R. (2000). Reward processing in primate orbitofrontal cortex and basal ganglia. *Cerebral cortex (New York, N.Y. : 1991)*, *10*(3), 272–284. https://doi.org/10.1093/cercor/10.3.272

Schultz W. (2015). Neuronal Reward and Decision Signals: From Theories to Data. *Physiological reviews*, *95*(3), 853–951.

Shteingart, H., & Loewenstein, Y. (2014). Reinforcement learning and human behavior. *Current opinion in neurobiology*, *25*, 93–98. https://doi.org/10.1016/j.conb.2013.12.004

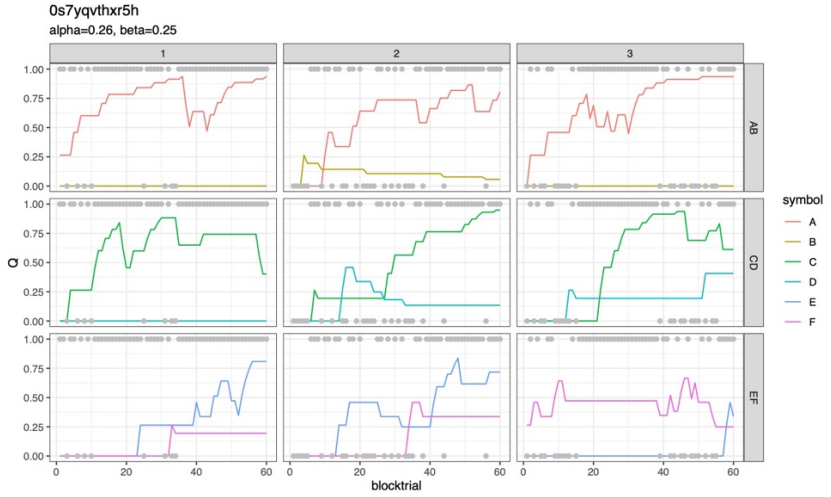
Theeuwes, J., & Belopolsky, A. V. (2012). Reward grabs the eye: oculomotor capture by rewarding stimuli. *Vision research*, *74*, 80–85. https://doi.org/10.1016/j.visres.2012.07.024

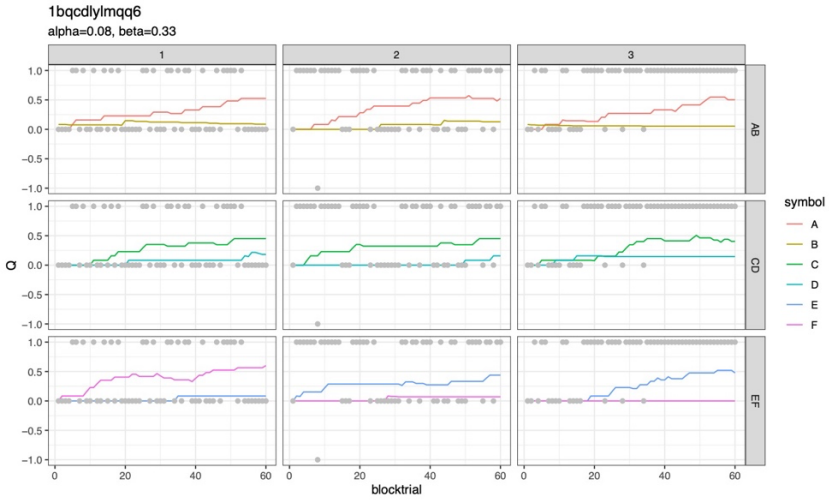
Tiokhin, L., Hackman, J., Munira, S., Jesmin, K., & Hruschka, D. (2019). Generalizability is not optional: Insights from a cross-cultural study of social discounting. *Royal Society open science*, *6*(2), 181386. https://doi.org/10.1098/rsos.181386

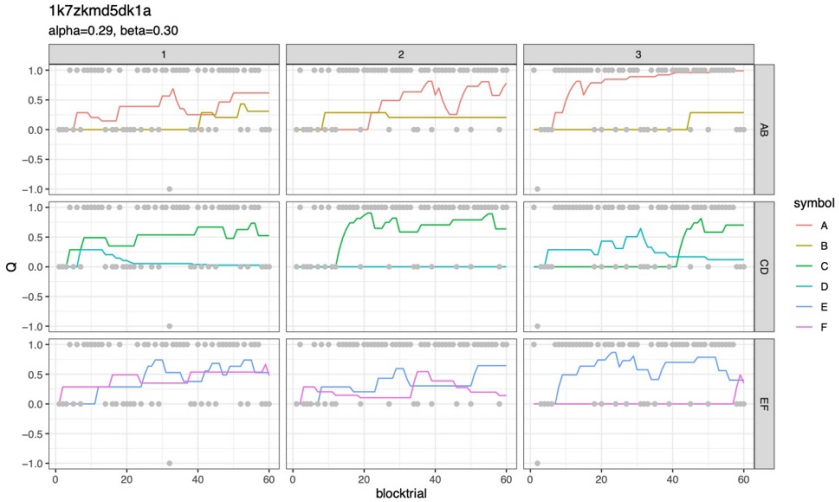
Turi, Z., Mittner, M., Opitz, A., Popkes, M., Paulus, W., & Antal, A. (2015). Transcranial direct current stimulation over the left prefrontal cortex increases randomness of choice in instrumental learning. *Cortex; a journal devoted to the study of the nervous system and behavior*, *63*, 145–154. https://doi.org/10.1016/j.cortex.2014.08.026

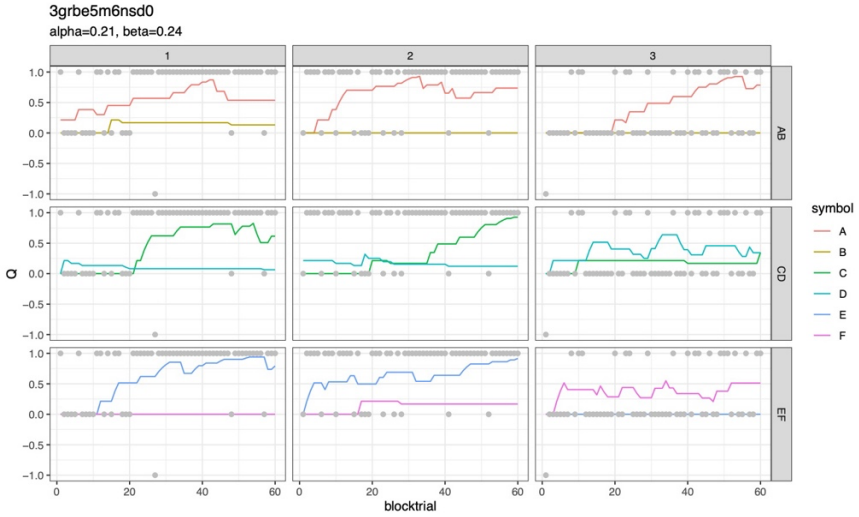
Appendix

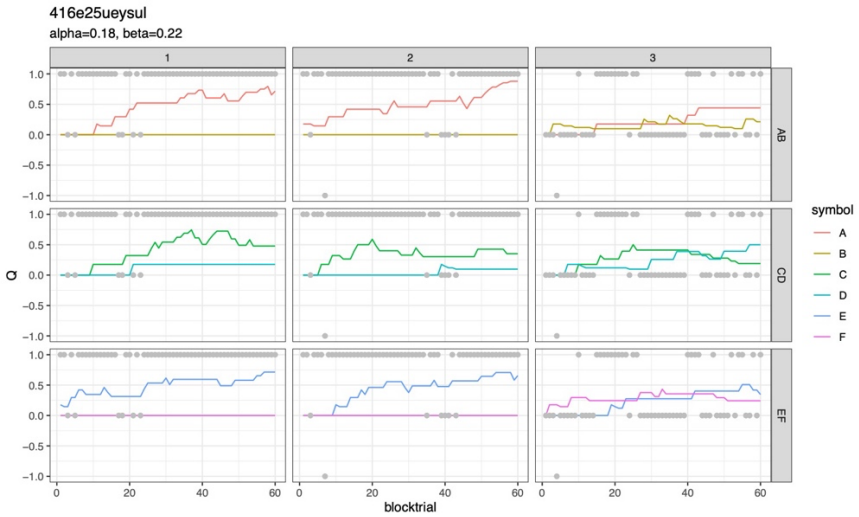
The Q-learning plots for all participants in this study.

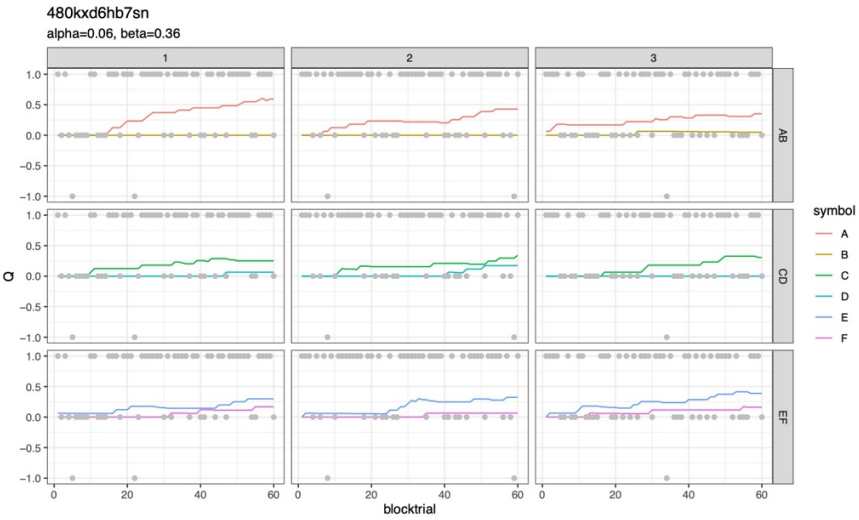


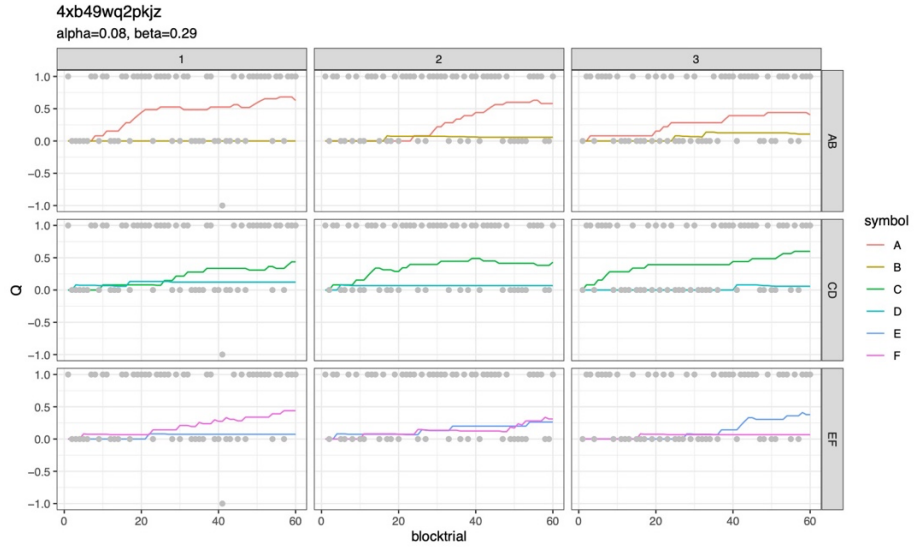


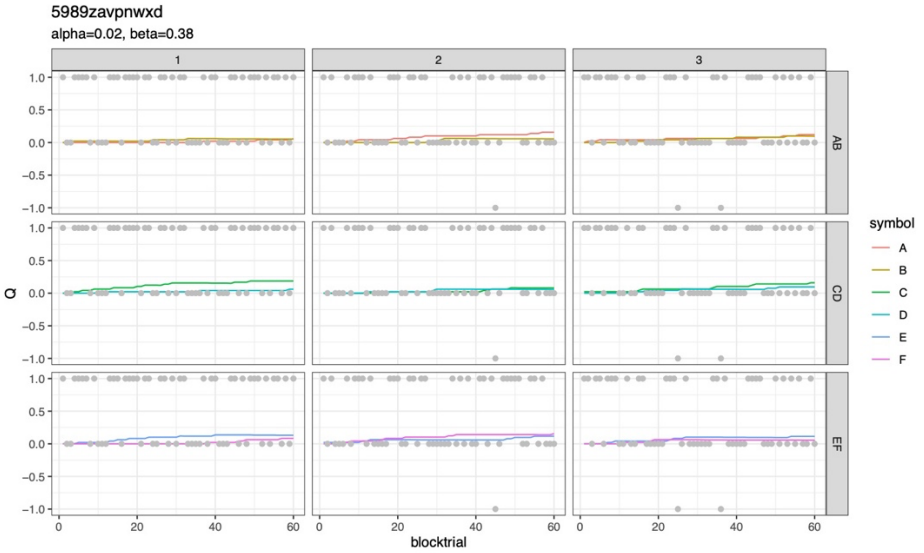


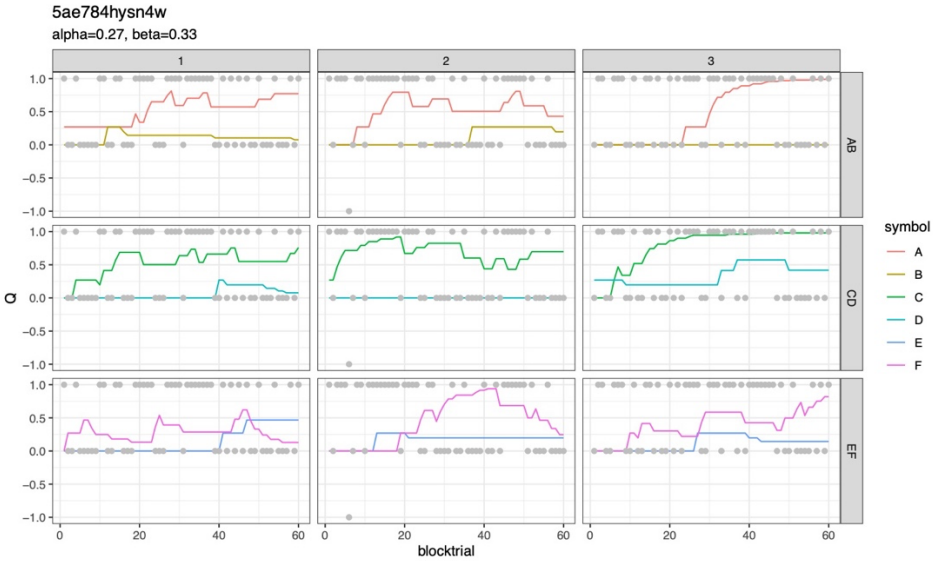

3grbe5m6nsd0
alpha=0.21, beta=0.24
1
2
3
AB
CD
EF
Q
blocktrial
symbol
A
B
C
D
E
F

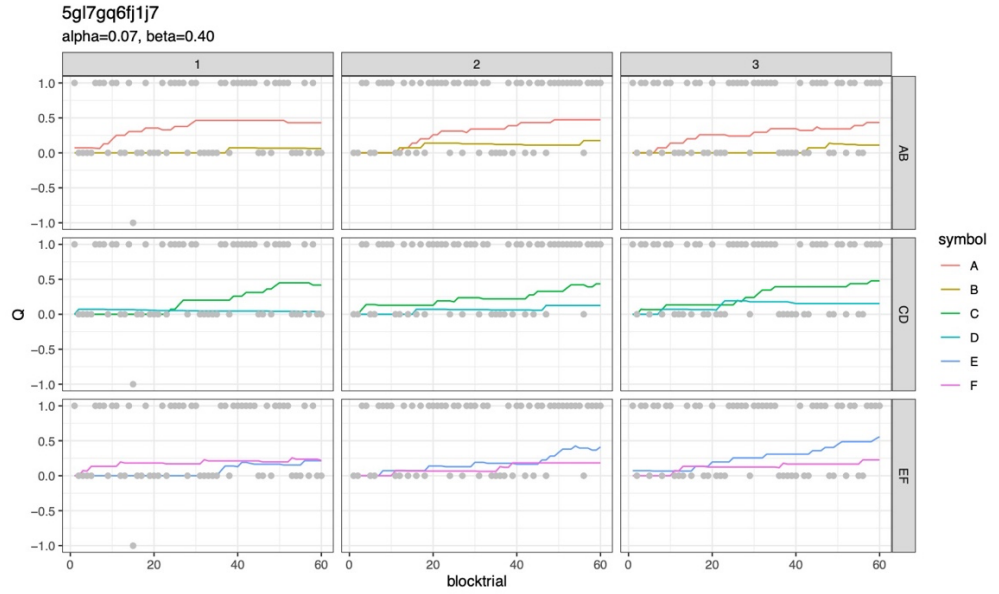

416e25ueysul
alpha=0.18, beta=0.22
1
2
3
AB
CD
EF
Q
blocktrial
symbol
A
B
C
D
E
F

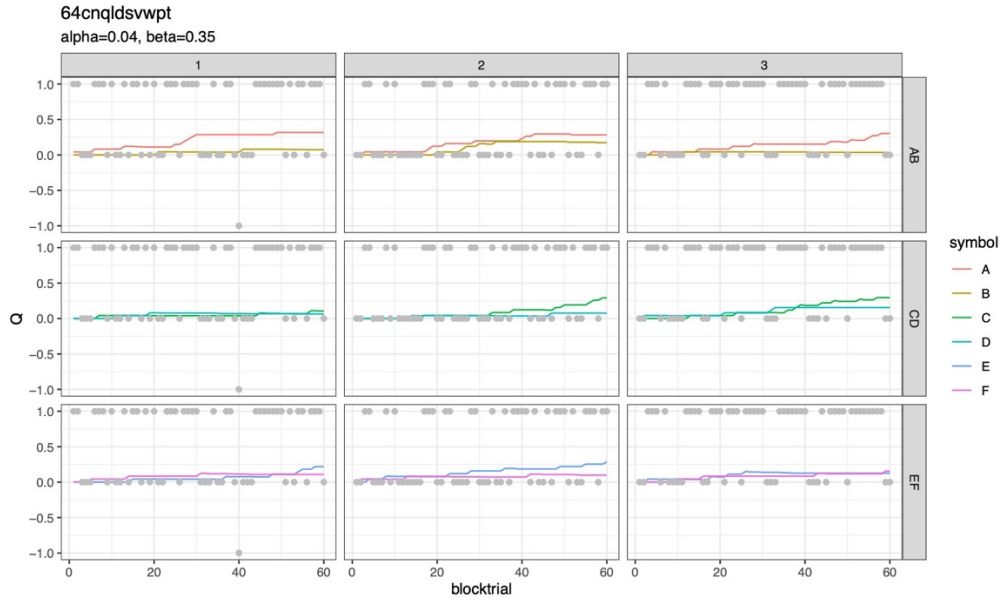

480kxd6hb7sn
alpha=0.06, beta=0.36
1
2
3
AB
CD
EF
Q
blocktrial
symbol
A
B
C
D
E
F

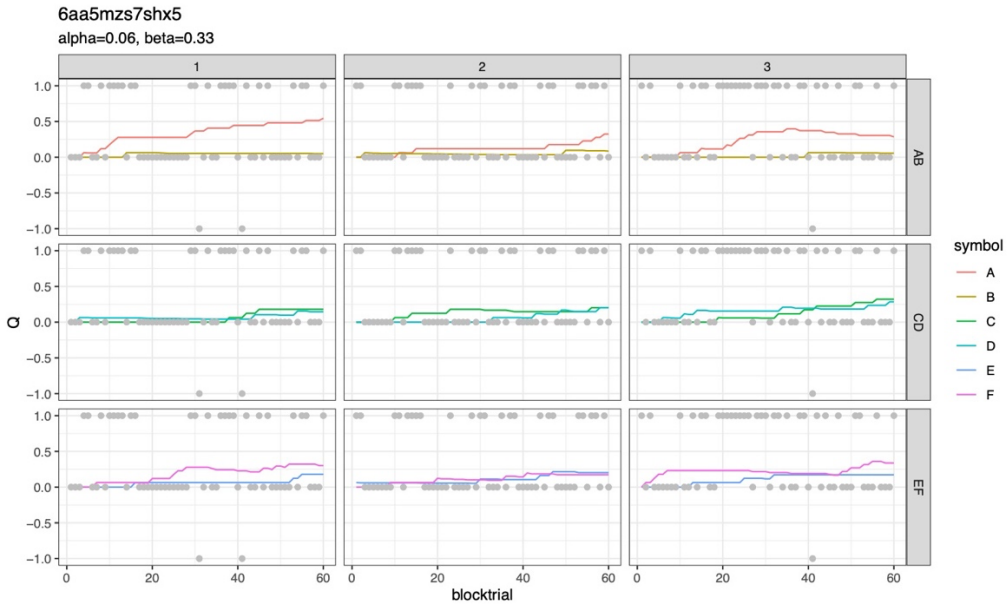

4xb49wq2pkjz
alpha=0.08, beta=0.29
1
2
3
AB
CD
EF
Q
blocktrial
symbol
A
B
C
D
E
F

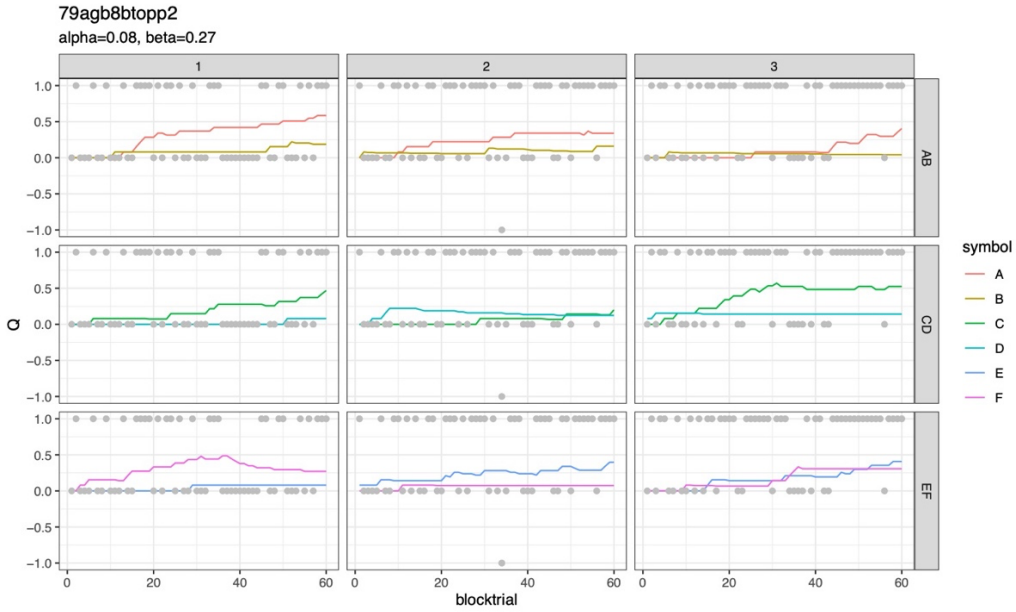

5989zavpnwxd
alpha=0.02, beta=0.38
1
2
3
AB
CD
EF
Q
blocktrial
symbol
A
B
C
D
E
F

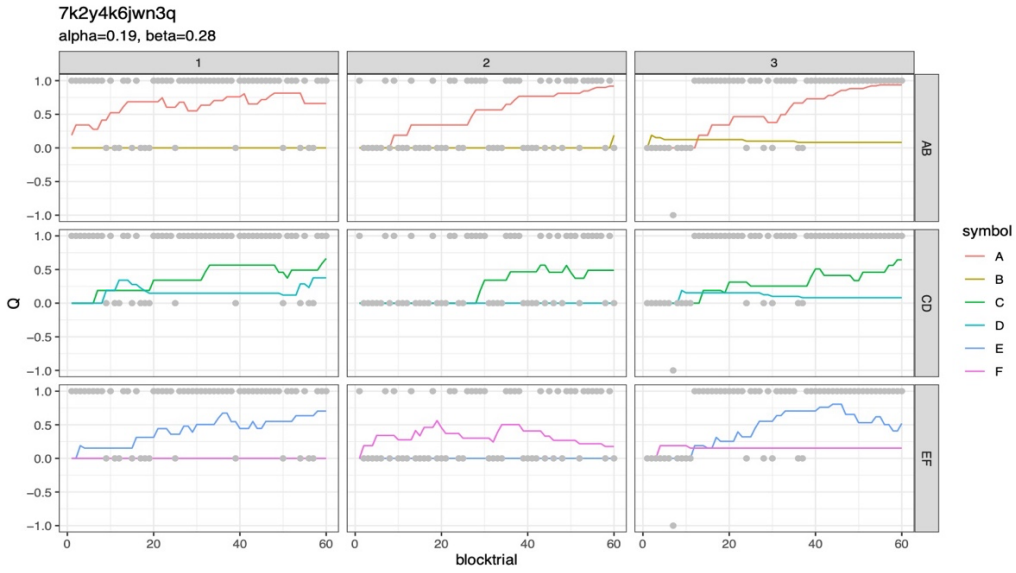

5ae784hysn4w
alpha=0.27, beta=0.33
1
2
3
AB
CD
EF
Q
blocktrial
symbol
A
B
C
D
E
F

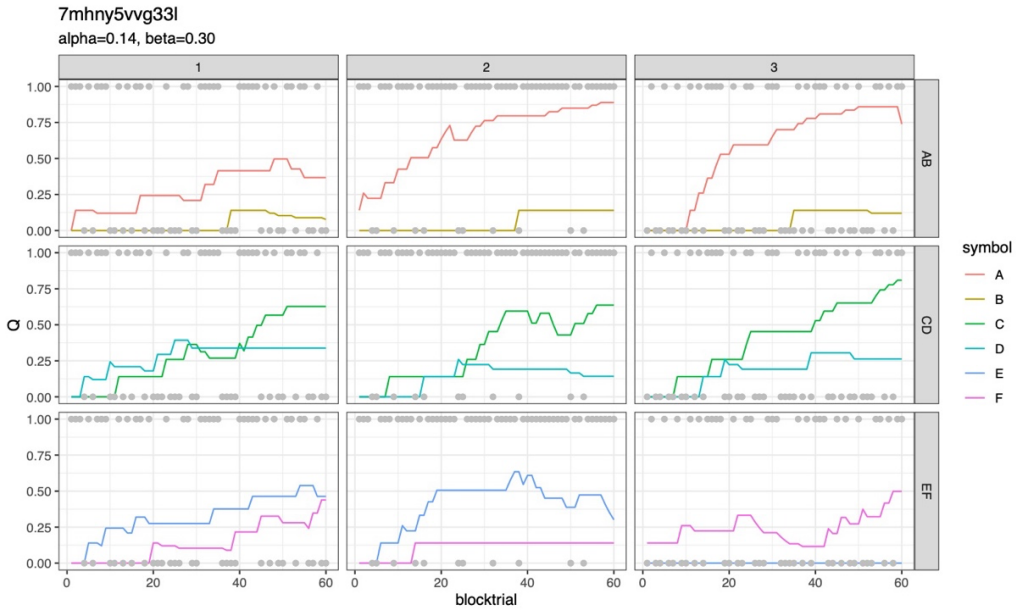

5gl7gq6fj1j7
alpha=0.07, beta=0.40
1
2
3
AB
CD
EF
Q
blocktrial
symbol
A
B
C
D
E
F

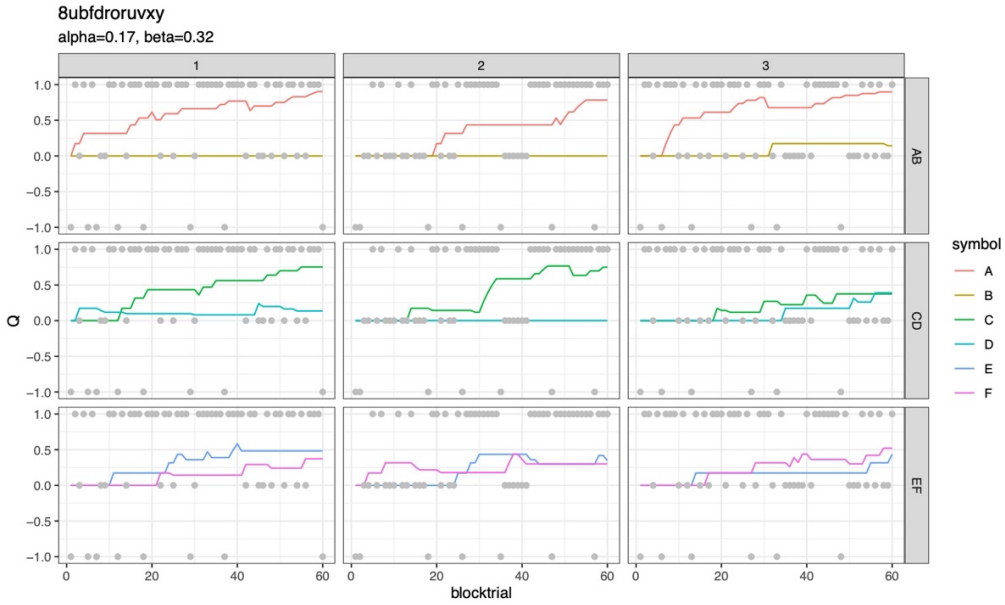

64cnqldsvwpt
alpha=0.04, beta=0.35
1
2
3
AB
CD
EF
Q
blocktrial
symbol
A
B
C
D
E
F

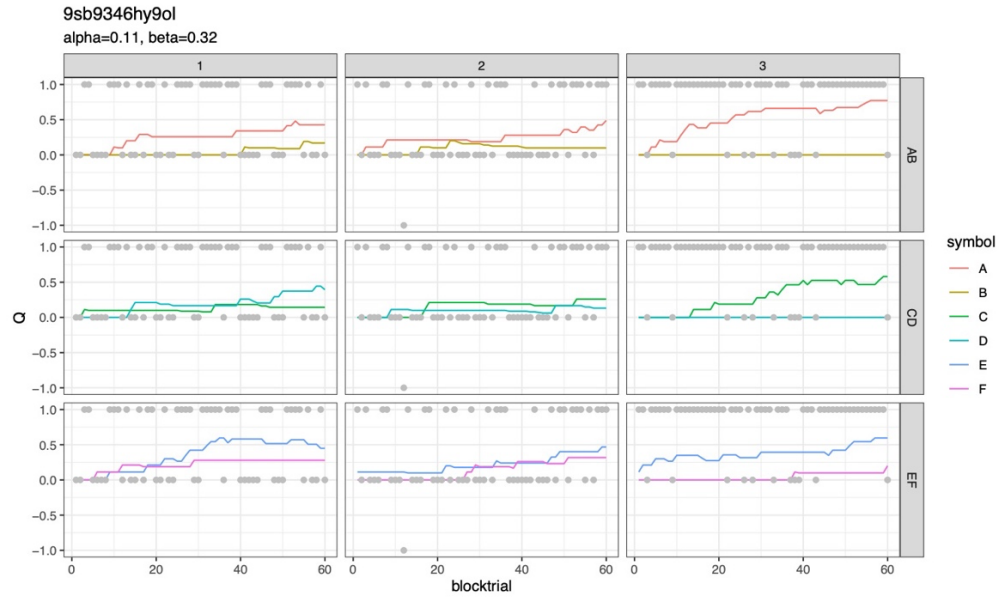

6aa5mzs7shx5
alpha=0.06, beta=0.33
1
2
3
AB
CD
EF
Q
blocktrial
symbol
A
B
C
D
E
F

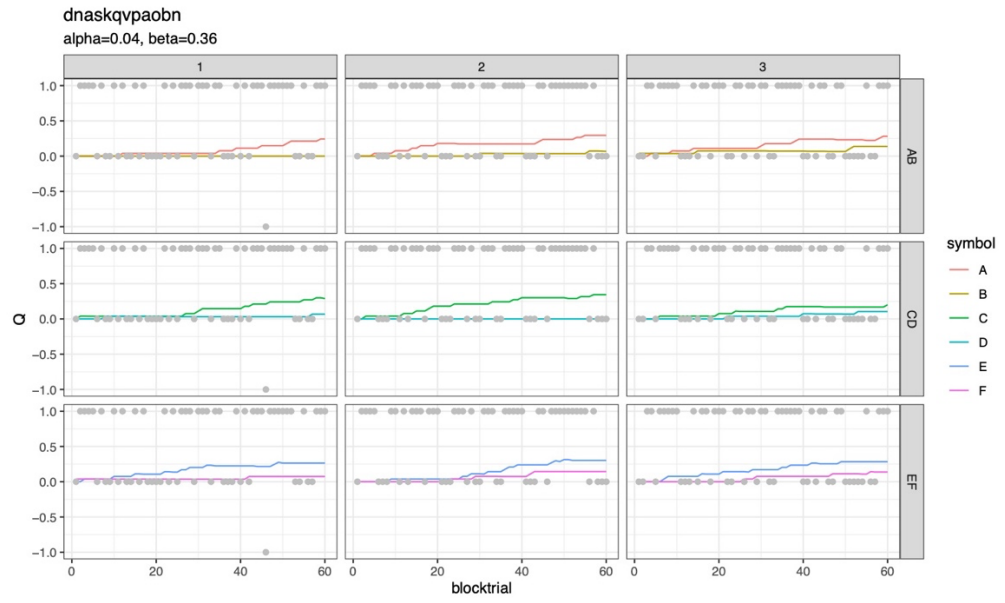

79agb8btopp2
alpha=0.08, beta=0.27
1
2
3
AB
CD
EF
Q
blocktrial
symbol
A
B
C
D
E
F

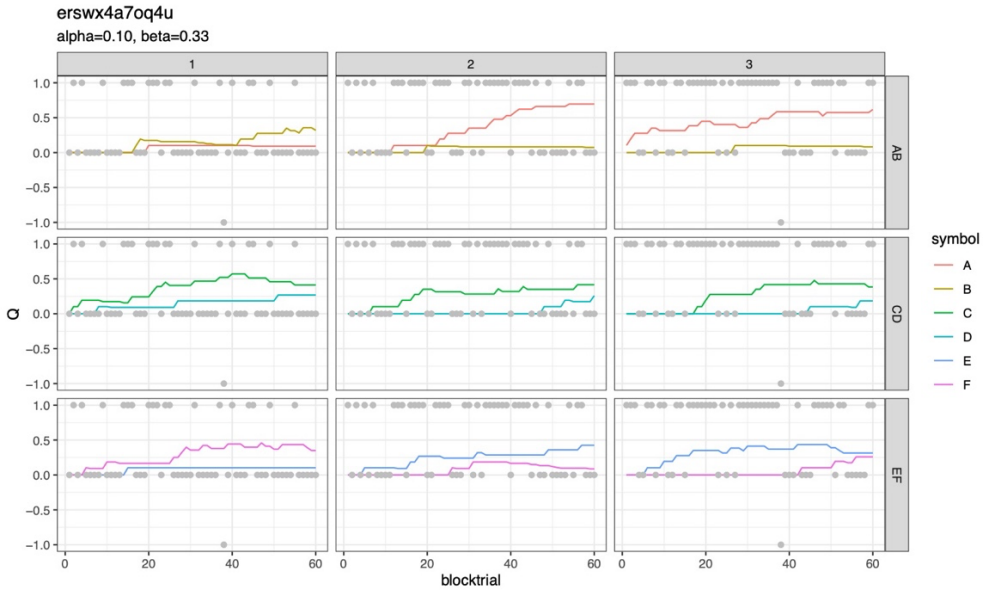

7k2y4k6jwn3q
alpha=0.19, beta=0.28
1
2
3
AB
CD
EF
Q
blocktrial
symbol
A
B
C
D
E
F

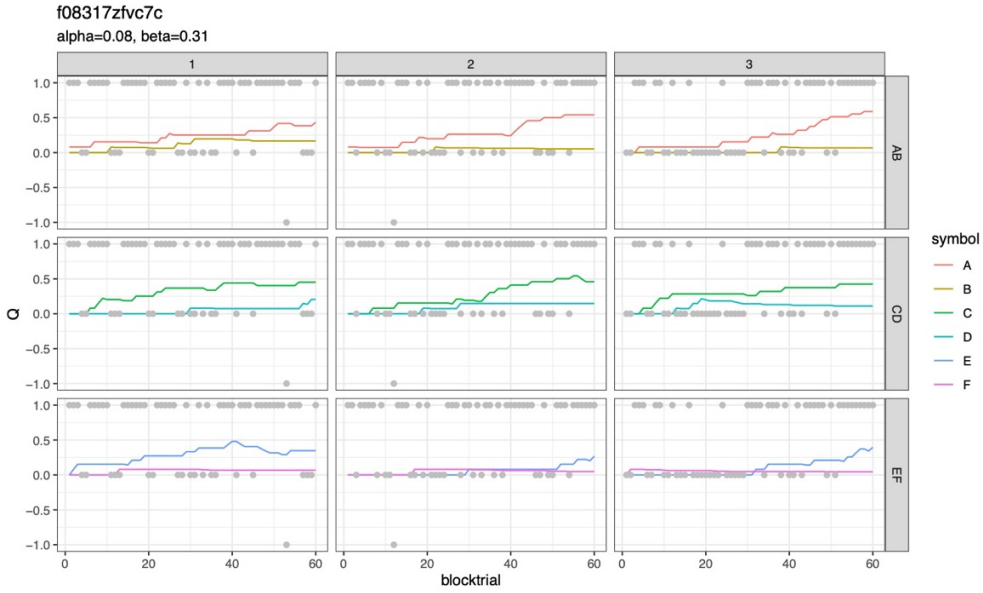

7mhny5vvg33l
alpha=0.14, beta=0.30
1
2
3
AB
CD
EF
Q
blocktrial
symbol
A
B
C
D
E
F

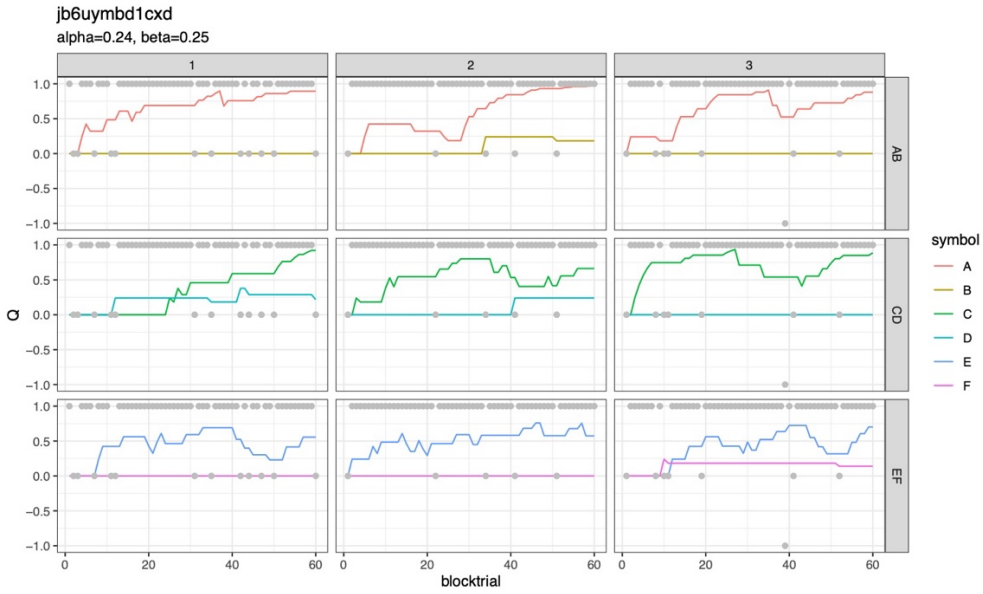

8ubfdroruvxy
alpha=0.17, beta=0.32
1
2
3
AB
CD
EF
Q
blocktrial
symbol
A
B
C
D
E
F

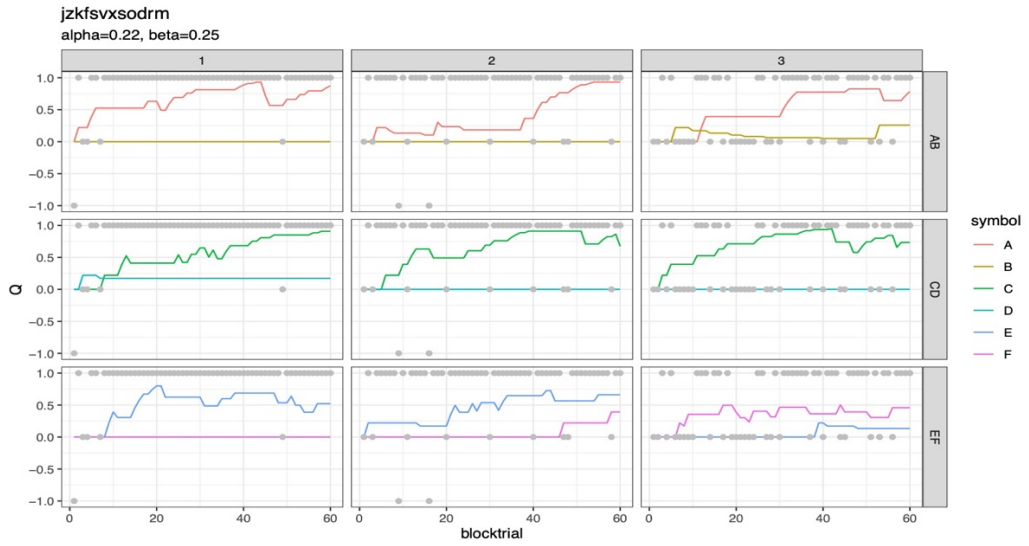

9sb9346hy9ol
alpha=0.11, beta=0.32
1
2
3
AB
CD
EF
Q
blocktrial
symbol
A
B
C
D
E
F

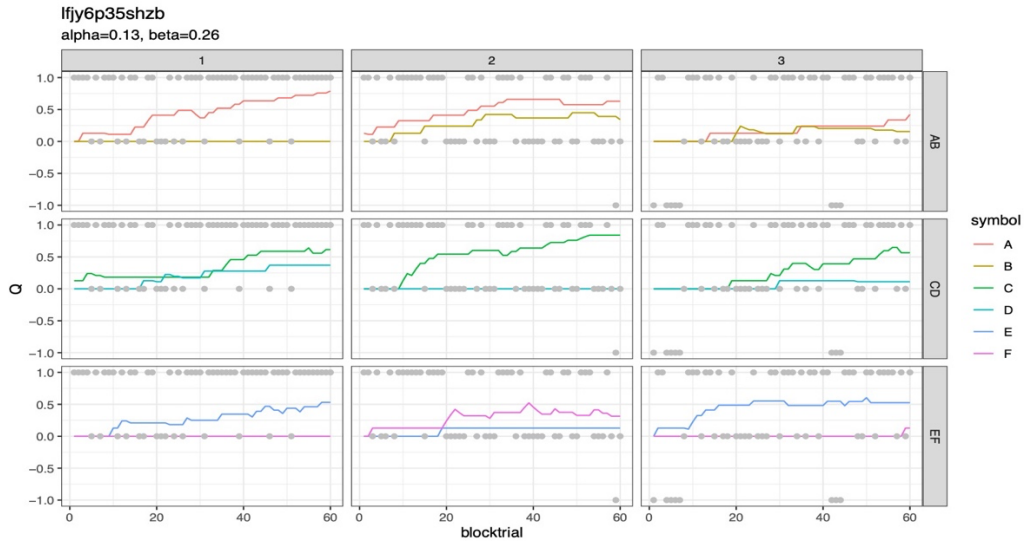

dnaskqvpaobn
alpha=0.04, beta=0.36
1
2
3
AB
CD
EF
Q
blocktrial
symbol
A
B
C
D
E
F

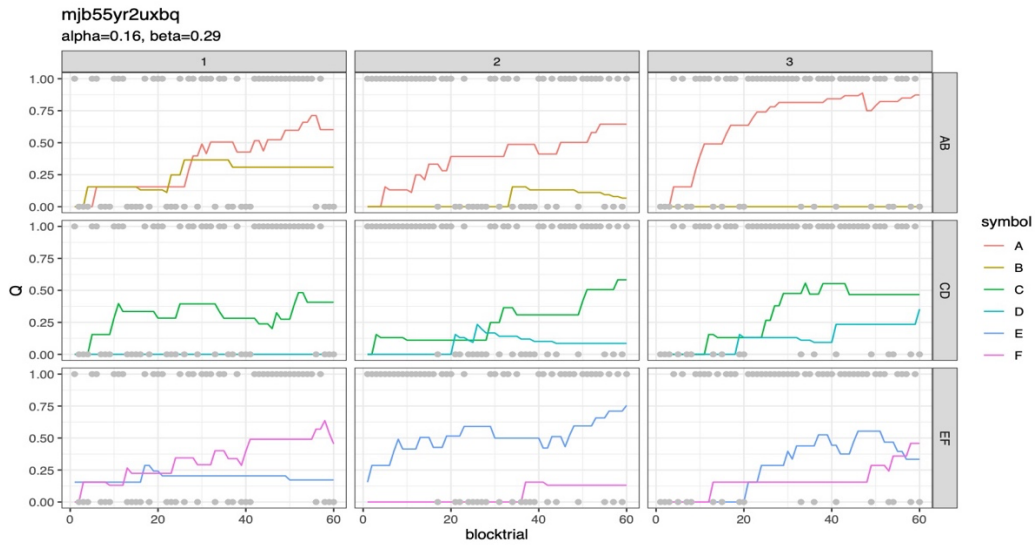

erswx4a7oq4u
alpha=0.10, beta=0.33
1
2
3
AB
CD
EF
Q
blocktrial
symbol
A
B
C
D
E
F

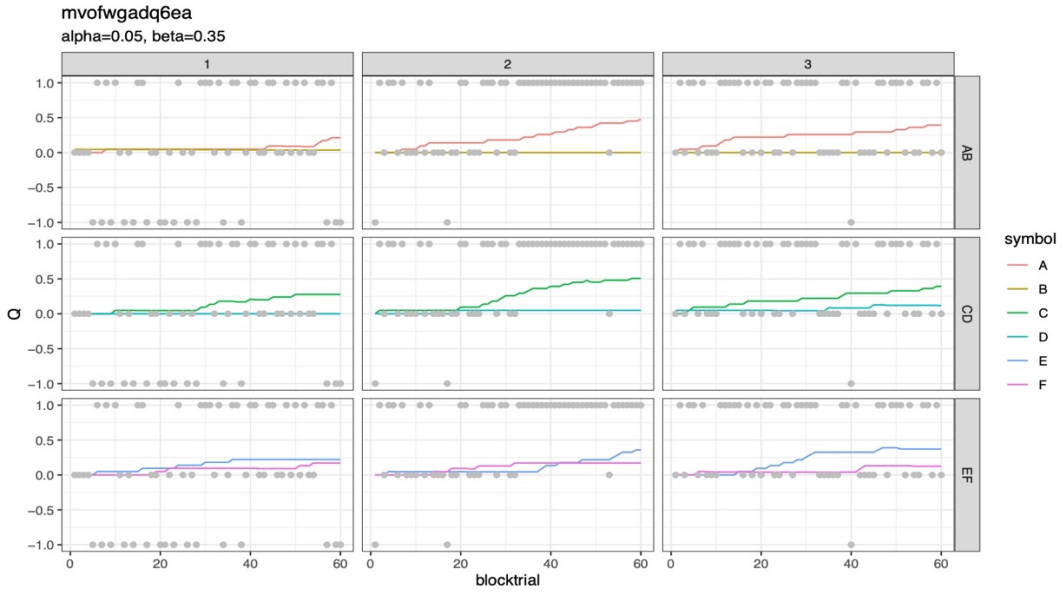

f08317zfvc7c
alpha=0.08, beta=0.31
1
2
3
AB
CD
EF
Q
blocktrial
symbol
A
B
C
D
E
F

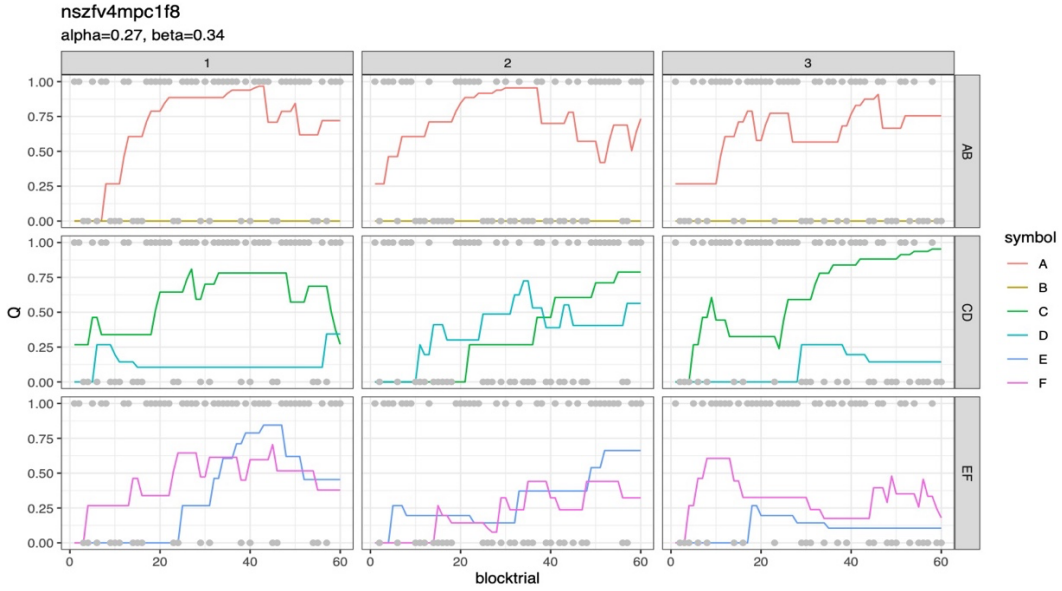

jb6uymbd1cxd
alpha=0.24, beta=0.25
1
2
3
AB
CD
EF
Q
blocktrial
symbol
A
B
C
D
E
F

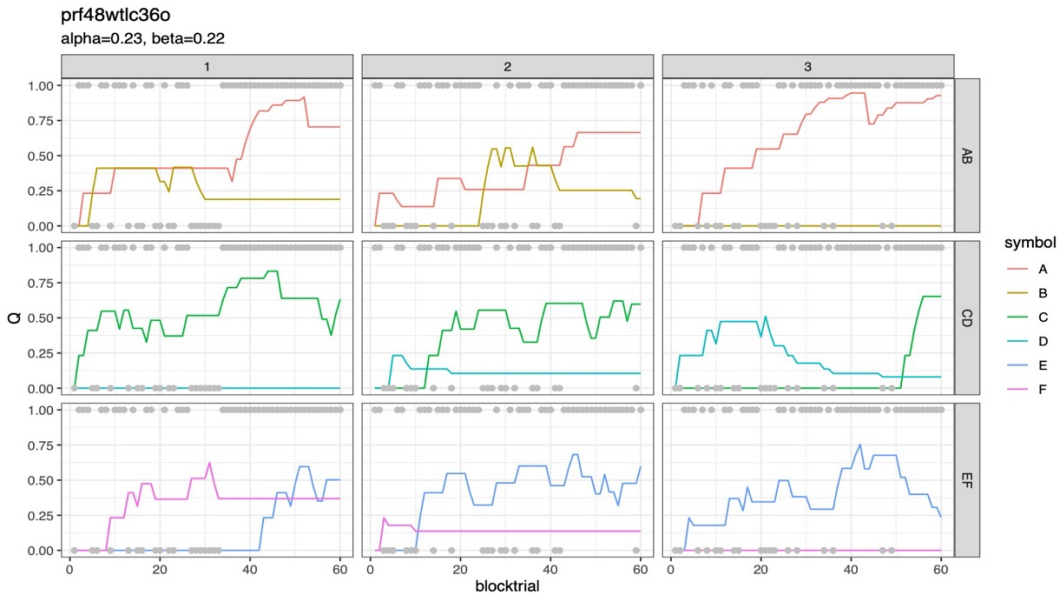

jzkfsvxsodrm
alpha=0.22, beta=0.25
1
2
3
AB
CD
EF
Q
blocktrial
symbol
A
B
C
D
E
F

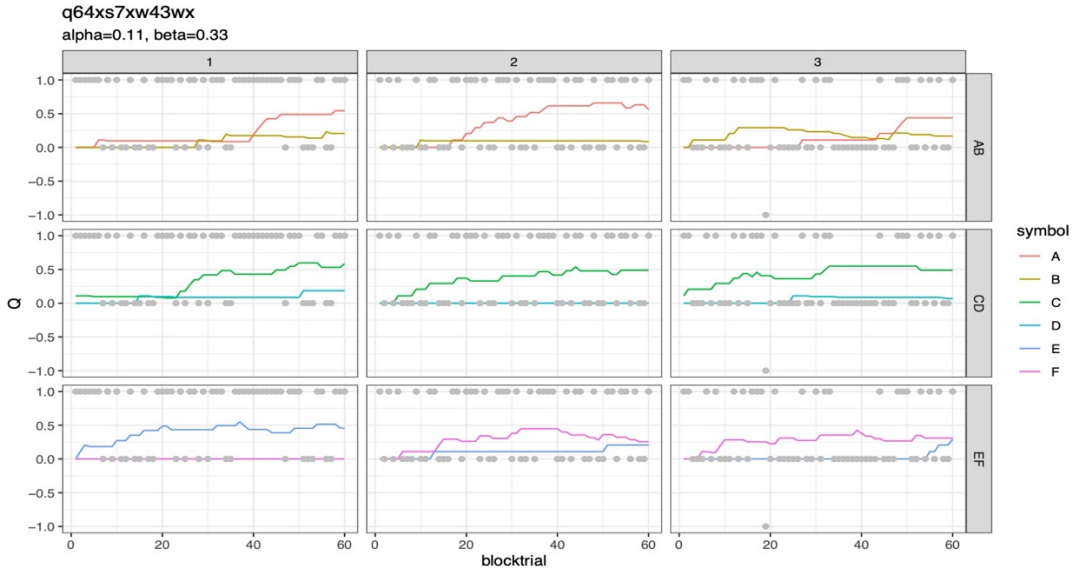

lfjy6p35shzb
alpha=0.13, beta=0.26
1
2
3
AB
CD
EF
Q
blocktrial
symbol
A
B
C
D
E
F

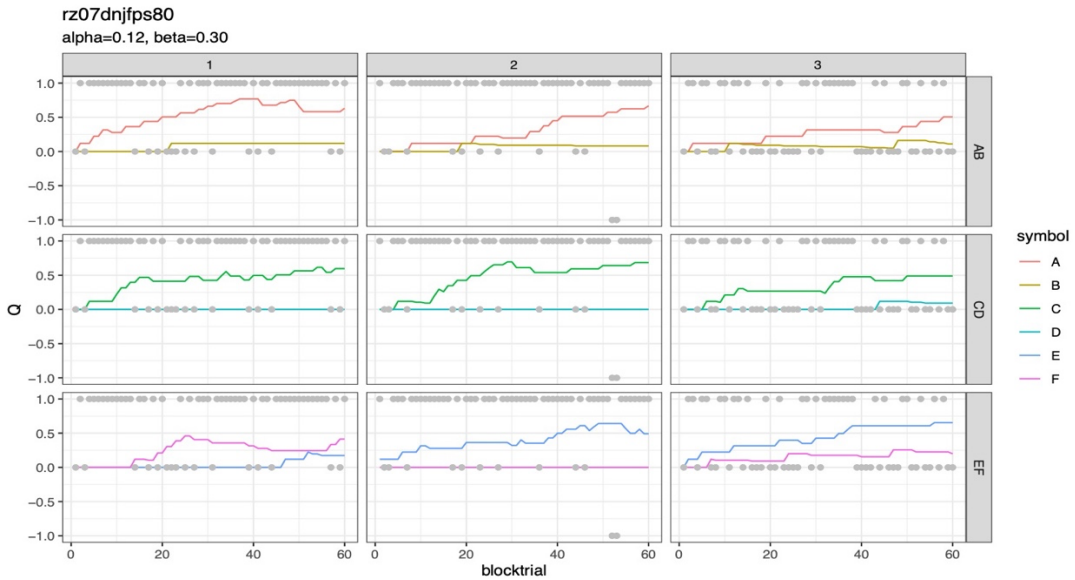

mjb55yr2uxbq
alpha=0.16, beta=0.29
1
2
3
AB
CD
EF
Q
blocktrial
symbol
A
B
C
D
E
F

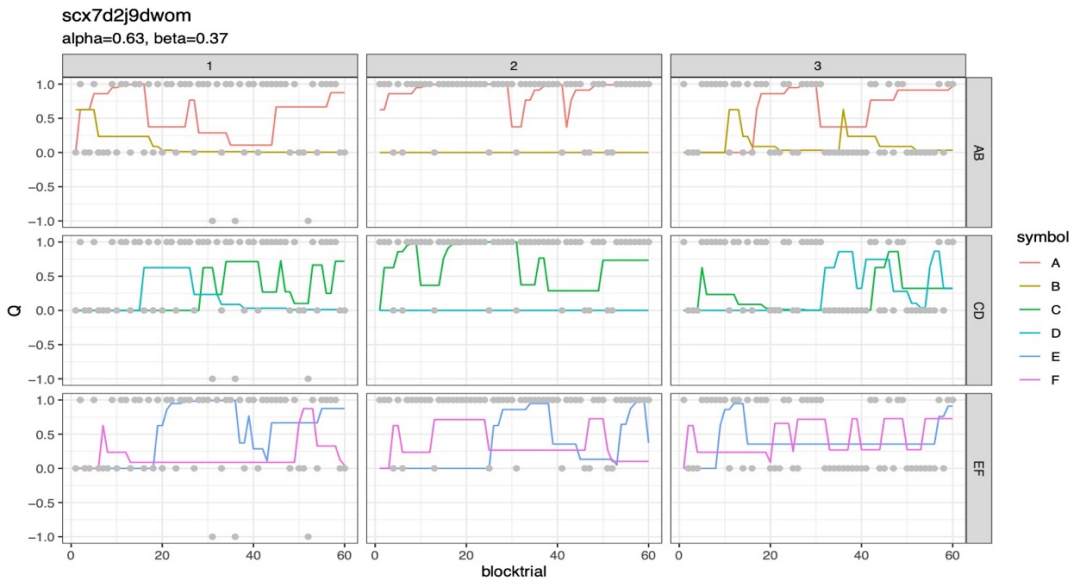

mvofwgadq6ea
alpha=0.05, beta=0.35
1
2
3
AB
CD
EF
Q
blocktrial
symbol
A
B
C
D
E
F

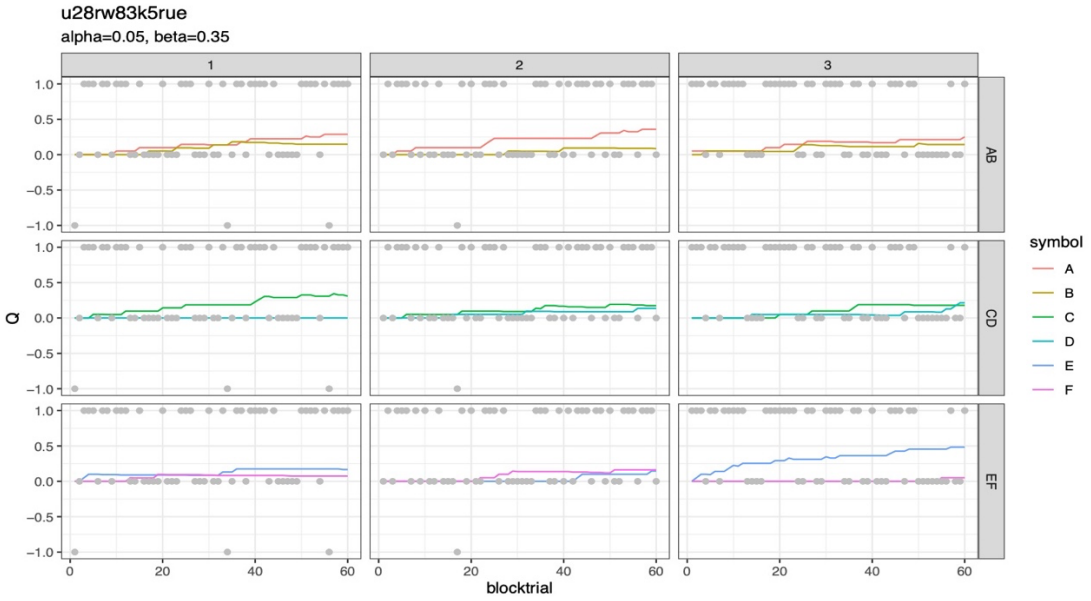

nszfv4mpc1f8
alpha=0.27, beta=0.34
1
2
3
AB
CD
EF
Q
blocktrial
symbol
A
B
C
D
E
F

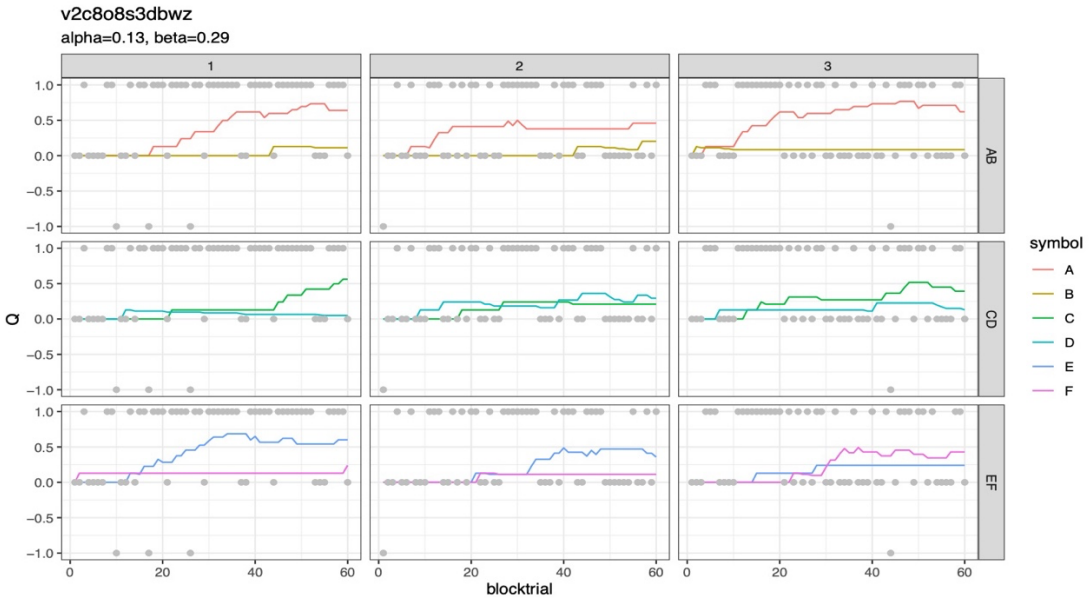

prf48wtlc36o
alpha=0.23, beta=0.22
1
2
3
AB
CD
EF
Q
blocktrial
symbol
A
B
C
D
E
F

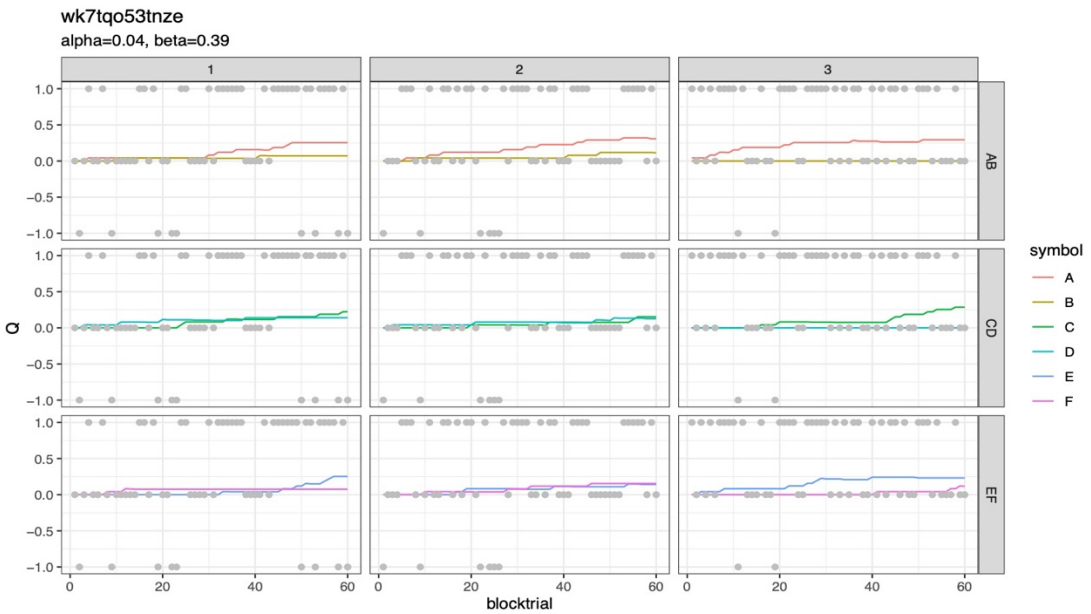

q64xs7xw43wx
alpha=0.11, beta=0.33
1
2
3
AB
CD
EF
Q
blocktrial
symbol
A
B
C
D
E
F

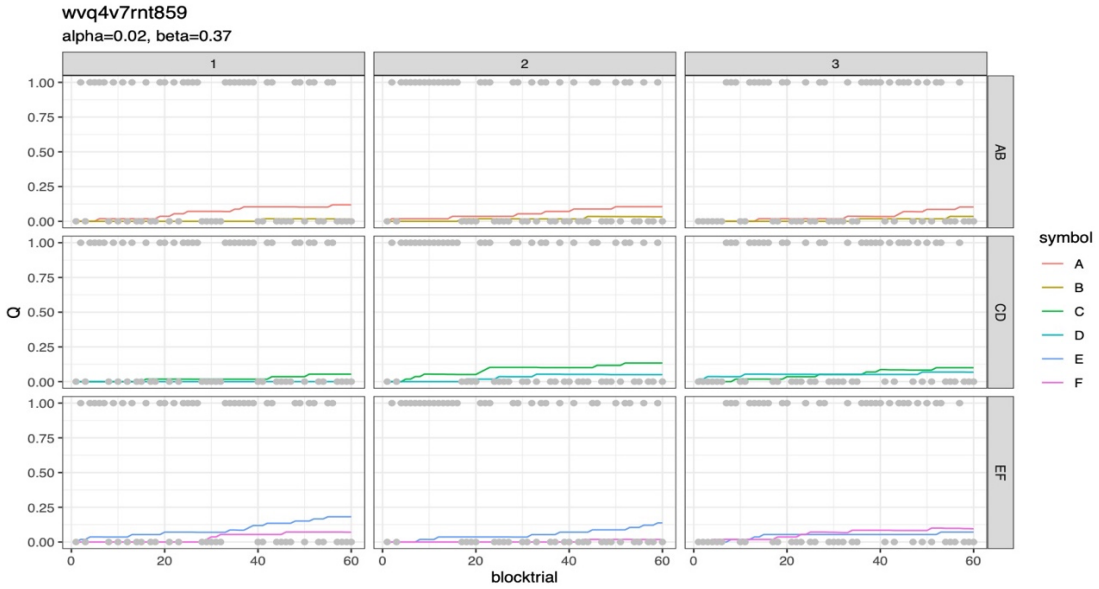

rz07dnjfps80
alpha=0.12, beta=0.30
1
2
3
AB
CD
EF
Q
blocktrial
symbol
A
B
C
D
E
F

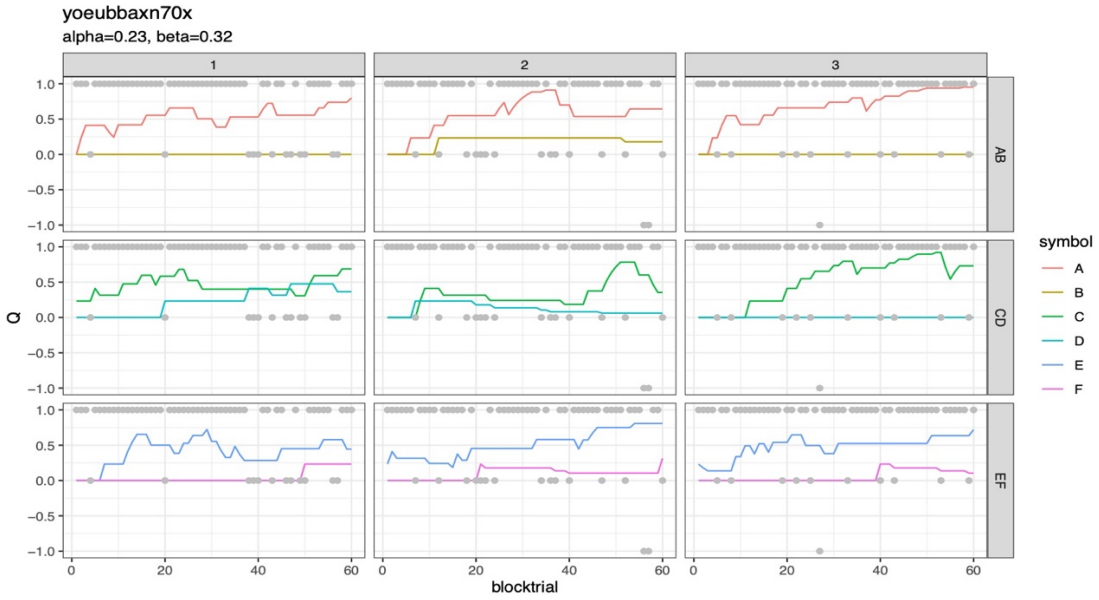

scx7d2j9dwom
alpha=0.63, beta=0.37
1
2
3
AB
CD
EF
Q
blocktrial
symbol
A
B
C
D
E
F

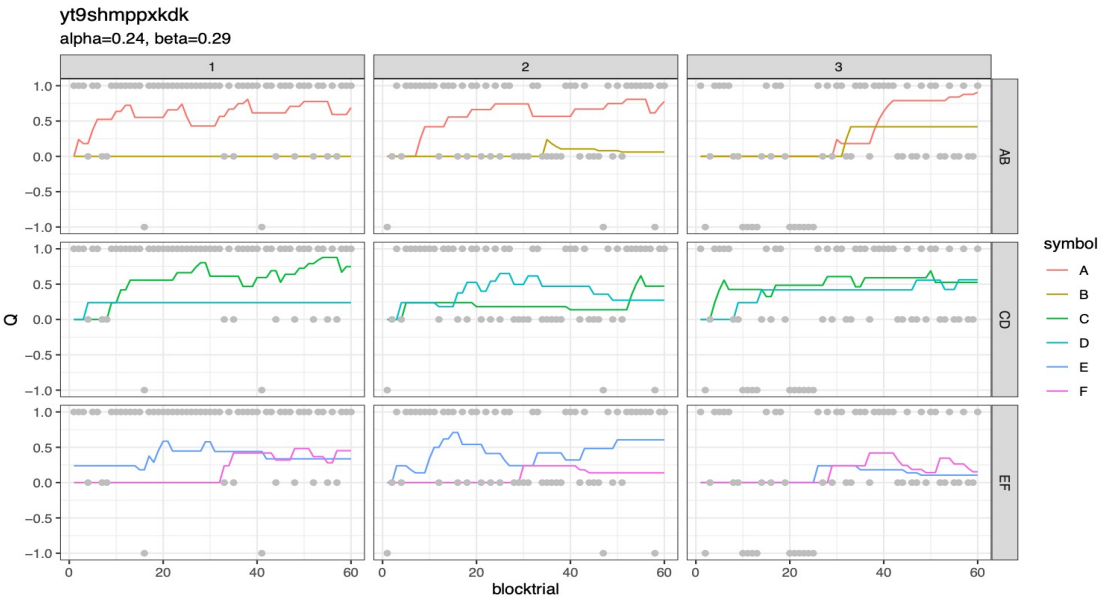

u28rw83k5rue
alpha=0.05, beta=0.35
1
2
3
AB
CD
EF
Q
blocktrial
symbol
A
B
C
D
E
F

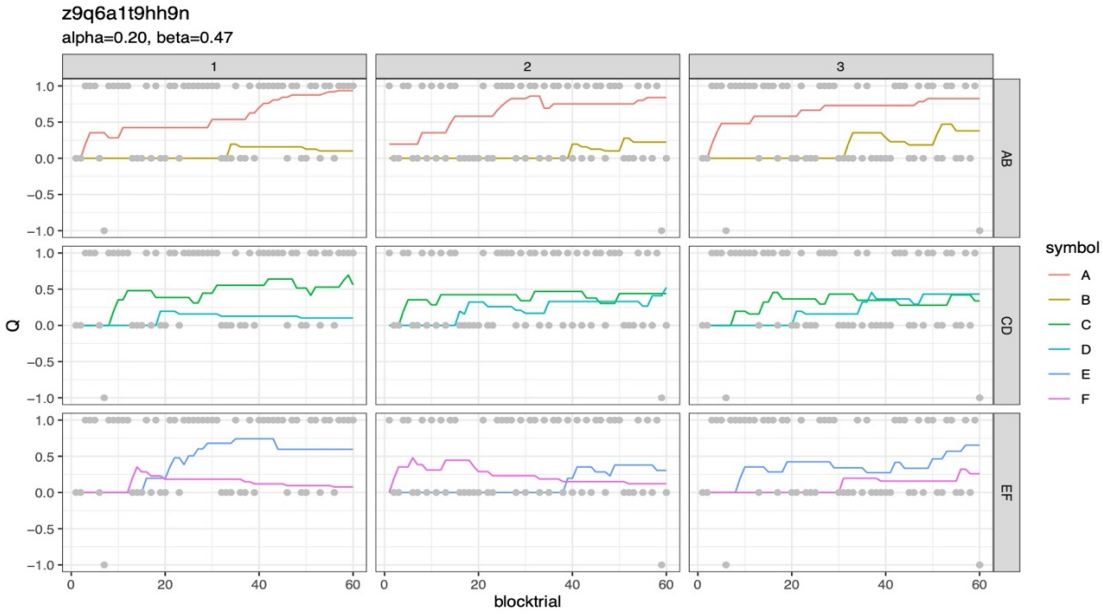

v2c8o8s3dbwz
alpha=0.13, beta=0.29
1
2
3
AB
CD
EF
Q
blocktrial
symbol
A
B
C
D
E
F

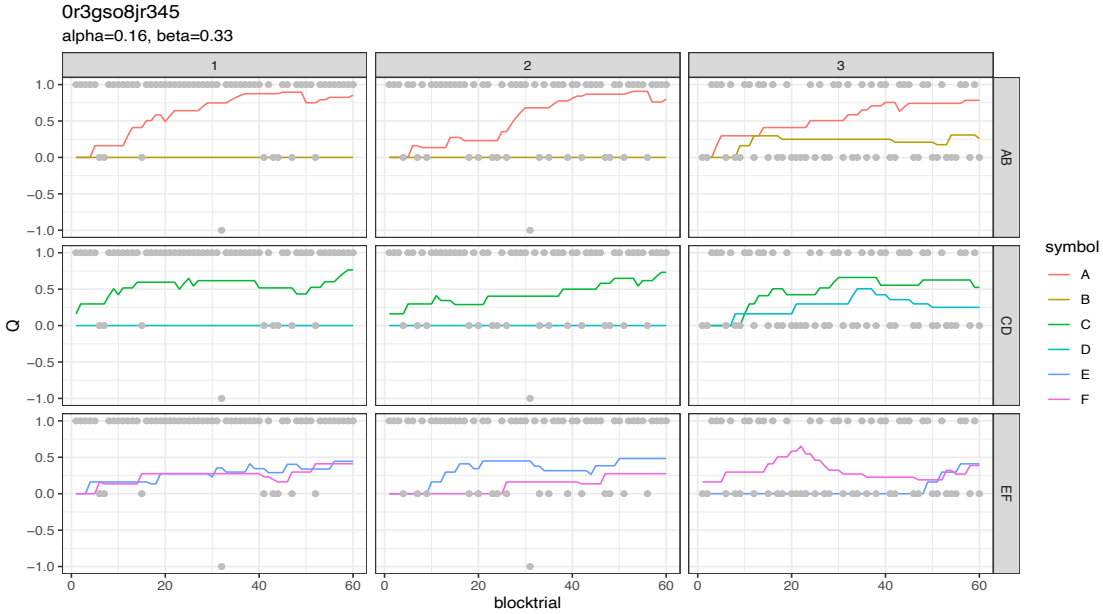

wk7tqo53tnze
alpha=0.04, beta=0.39
1
2
3
AB
CD
EF
Q
blocktrial
symbol
A
B
C
D
E
F


wvq4v7rnt859
alpha=0.02, beta=0.37
1
2
3
AB
CD
EF
Q
blocktrial
symbol
A
B
C
D
E
F


yoeubbaxn70x
alpha=0.23, beta=0.32
1
2
3
AB
CD
EF
Q
blocktrial
symbol
A
B
C
D
E
F


yt9shmppxkdk
alpha=0.24, beta=0.29
1
2
3
AB
CD
EF
Q
blocktrial
symbol
A
B
C
D
E
F


z9q6a1t9hh9n
alpha=0.20, beta=0.47
1
2
3
AB
CD
EF
Q
blocktrial
symbol
A
B
C
D
E
F


0r3gso8jr345
alpha=0.16, beta=0.33
1
2
3
AB
CD
EF
Q
blocktrial
symbol
A
B
C
D
E
F