Supplementary information

Explicit knowledge of task structure is a primary determinant of human model-based action

Pedro Castro-Rodrigues[1-4,#], Thomas Akam[2,5,#], Ivar Snorasson[6], Marta Camacho, M[1,a], Vitor Paixão[2], Ana Maia[1-3,7], J. Bernardo Barahona-Corrêa[1-3], Peter Dayan[8,9], H. Blair Simpson[6,10], Rui M. Costa[2,3,11], Albino J. Oliveira-Maia[1-3,*]

[1]Champalimaud Clinical Centre, Champalimaud Centre for the Unknown, Lisbon, Portugal;

[2]Champalimaud Research, Champalimaud Centre for the Unknown, Lisbon, Portugal;

[3]NOVA Medical School, Lisbon, Portugal;

[4]Centro Hospitalar Psiquiátrico de Lisboa, Lisbon, Portugal;

[5]Department of Experimental Psychology, University of Oxford, Oxford, UK;

[6]Center for Obsessive-Compulsive & Related Disorders, New York State Psychiatric Institute, New York, USA;

[7]Department of Psychiatry and Mental Health, Centro Hospitalar de Lisboa Ocidental, Lisbon, Portugal;

[8]Max Planck Institute for Biological Cybernetics, Tübingen, Germany;

[9]The University of Tübingen, Tübingen, Germany;

[10]Department of Psychiatry, Columbia University, New York, USA;

[11]Zuckerman Mind Brain Behavior Institute, Columbia University, New York, USA

[#] Equally contributing authors.

*Corresponding author: albino.maia@neuro.fchampalimaud.org

[a]Current address: John Van Geest Center for Brain Repair, University of Cambridge, UK.

1) Supplementary methods: Information provided to study participants

A) Information before task

"You will now play a game in order to gain of as many rewards as possible.

Rewards will be represented in the screen as coins. Every time you get a coin, it will show up in the screen and it will be added to your total number of rewards. The number of coins you get will determine the value of the gift-card that you will receive at the end of your participation.

You will perform 1200 trials and in each trial you can get either one coin or no coin. At the end of those 1200 trials, 400 will be randomly chosen to count the final number of coins.

The minimum amount of money in your gift card will be 10 euros. For each coin that you get above 150 coins, you will get an increase of 20 cents in your gift card. Therefore, if you get 175 coins the amount will be 15 euros, 200 coins correspond to 20 euros and 225 coins correspond to the maximum amount that the gift-card can have, which is 25 euros. Amounts will be distributed rounded to the closer multiple of 5 euros.

At the top left corner of the screen, there will be a coin counter which shows how many coins you got in each session. That number may not have direct correspondence with the final amount, since that amount will be calculated using a random sample of trials.

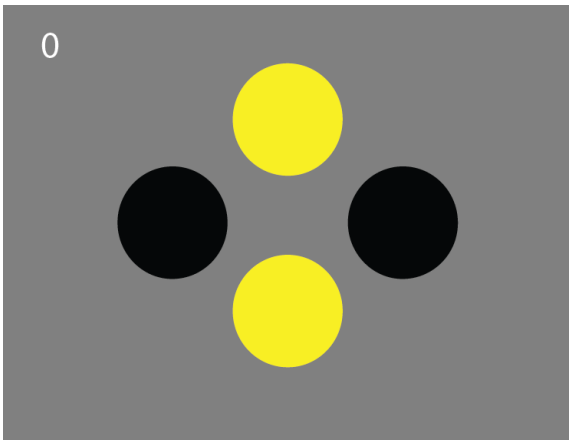
You will play the game using the arrow keys after stimuli show up in the screen.

Each session of the game will last for approximately 15 minutes. Once the session is completed, a sentence thanking you for your participation will show up in the screen. When that screen shows up you should leave the room.

B) Debriefing – Fixed transition probabilities version

We will now explain the structure of the game.

First the two central circles (upper and lower) are yellow, indicating that you can choose one of them.



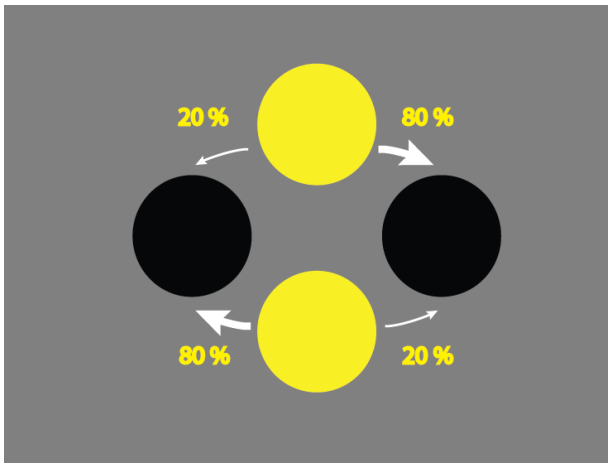
If you press the upper arrow key, you will choose the upper circle. If you press the lower arrow key, you will choose the lower circle.

After you choose the upper or the lower circle, one of the two side circles will light up, i. e., will turn yellow (left or right). After you press the arrow key that corresponds to the lateral circle that lit up (left or right), a coin may or may not appear.

The probability according to which the central circles give access to either one of the lateral circle also follows some rules.

If you choose the upper circle, one of two different things can happen. Most of the times (actually 80% of the times) the right side circle will light up. Rarely, the left side circle will light up.

If you choose the lower circle, most of the times (actually 80% of the times) the left side circle will light up. On the remaining occasions, the right side circle will light up.



The left and right circles give access to the rewards, which are symbolized as coins. However, the probability of winning a coin is not equal on the left or on the right: it is always higher on one of the sides.

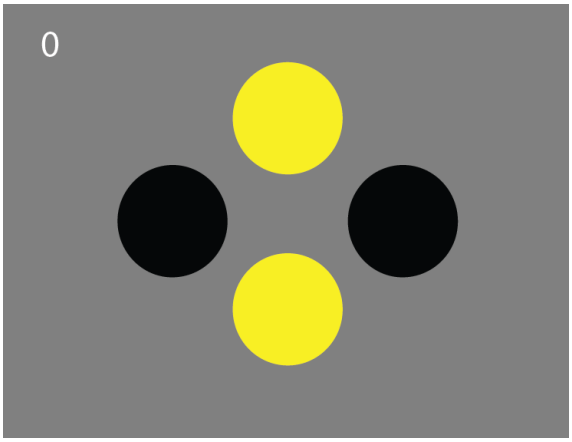
Sometimes it is higher on the left and sometimes it is higher on the right. The side in which that probability is higher changes after 20 or more trials.

You will now play a last session, with the same rules. Good luck!

C) Debriefing – Changing transition probabilities version

We will now explain the structure of the game.

First the two central circles (upper and lower) are yellow, indicating that you can choose one of them.



If you press the upper arrow key, you will choose the upper circle. If you press the lower arrow key, you will choose the lower circle.

After you choose the upper or the lower circle, one of the two side circles will light up, i. e., will turn yellow (left or right). After you press the arrow key that corresponds to the lateral circle that lit up (left or right), a coin may or may not appear.

The probability according to which the central circles give access to either one of the lateral circle also follows some rules. The game is divided in two types of blocks.

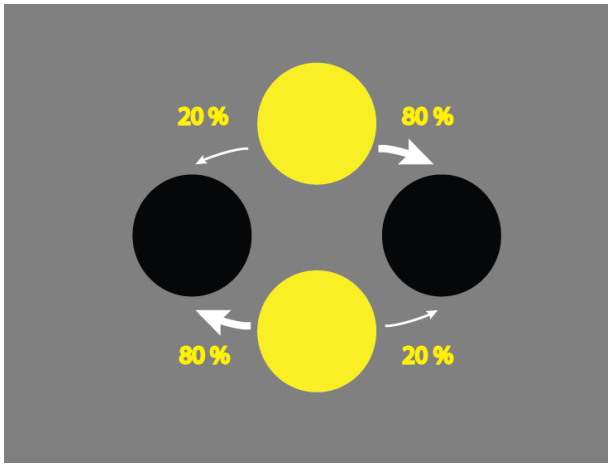
In “A” blocks, choosing the upper circle leads more frequently (80% of the times) to the lighting up of the right side circle. On the other hand, in these blocks, choosing the lower circle, leads more frequently (80% of the times) to the lighting up of the left side circle.

In “B” blocks, choosing the upper circle leads more frequently (80% of the times) to the lighting up of the left side circle. On the other hand, in these blocks, choosing the lower circle, leads more frequently (80% of the times) to the lighting up of the right side circle.

Therefore, in “A” blocks, if you choose the upper circle, one of two things can happen. Most of the times (actually 80% of the times), the right side circle will light up. Rarely (20% of the time), the left side circle will light up.

In these same “A” blocks, if you choose the lower circle, one of two things can happen. Most of the times (actually 80% of the times), the left side circle will light up. Rarely (20% of the time), the right side circle will light up.

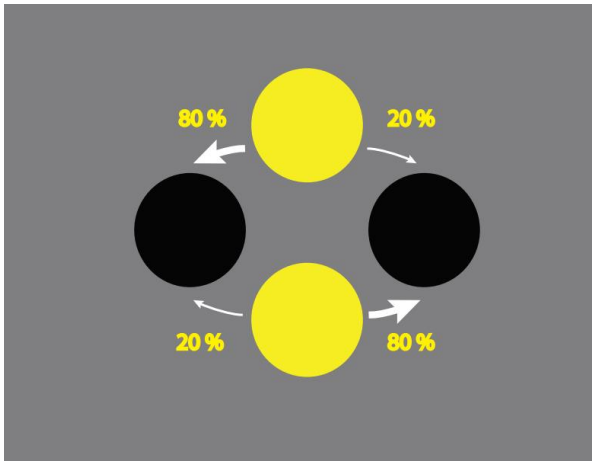
Schematic representation of the structure of “A” blocks:



In “B” blocks, if you choose the upper circle, one of two things can happen. Most of the times (actually 80% of the times), the left side circle will light up. Rarely (20% of the time), the right side circle will light up.

In these same “B” blocks, if you choose the lower circle, one of two things can happen. Most of the times (actually 80% of the times), the right side circle will light up. Rarely (20% of the time), the left side circle will light up.

Schematic representation of the structure of “B” blocks:



“A” blocks and “B” blocks alternate between them after 20 or more trials.

The left and right circles give access to the rewards, which are symbolized as coins. However, the probability of winning a coin is not the equal on the left or on the right: it is always higher on one of the sides. Sometimes it is higher on the left and sometimes it is higher on the right. The side in which that probability is higher changes after 20 or more trials.

You will now play a last session, with the same rules. Good luck!

2) Supplementary results

Changing transition probabilities inhibit model-based control

In 42 healthy volunteers recruited in Lisbon, we tested a version of the task in which transition probabilities linking the first step actions and second-step states underwent periodic reversals. In this Changing version of the task (Supplementary Figure 9), subjects were still able to successfully track which first step action was correct, choosing the correct option at the end of block well above chance level, (session 1: 0.77, session 3: 0.76). The proportion of subjects for whom a likelihood ratio test indicated model-based RL was being used in session 3 (6/42) was similar to that observed in the Fixed task (10/67). However, the stay-probability analysis did not show statistically significant evidence for a change in the influence of either the transition (null 95% CI [-0.21,0.22], coefficient change=0.14, *P*=0.21) or transition-outcome interaction (null 95% CI [-0.21,0.21], coefficient change=0.13, *P*=0.21) in session 3 relative to session 1 (Supplementary Figure 9a,b). On the contrary, there was an increase in the influence of trial outcome (null 95% CI [-0.37,0.36], coefficient change=0.46, *P*=0.013), associated with a model-free direct reinforcement strategy[1,2]. Similarly to the Fixed task, there was a significant correlation between loading on the transition–outcome interaction parameter and the number of rewards obtained (rho=0.4, *P*<0.01). Model comparison indicated that the mixture and model-free-only RL models fitted the data much better than the model-based-only model, and that the difference in BIC scores between the mixture model and model-free-only model was negligible (ΔiBIC = 3; Supplementary Figure 9d left panel). Again, similarly to the Fixed task, the model including both "bias" and "perseveration" parameters fits the data better than a model lacking these parameters (Supplementary Figure 9d, right panel). For consistency with analysis of the Fixed task, we used the mixture model to look for differences in behaviour between sessions 1 and 3 but found no statistically significant change in model parameters (Supplementary Figure 9e). These data suggest that, while model-free RL was dominant for the non-instructed sessions of both tasks, the dynamically changing action-state transition probabilities in the Changing task further reduced the ability to learn a model-based strategy. This is consistent with uncertainty based arbitration between model-based and model-free control[33,4], as the changing transition probabilities would be expected to increase uncertainty in the model-based system.

In this version of the task, we did not find statistically significant evidence for changes in the use of model-based RL after the debriefing, among subjects for whom a likelihood ratio test indicated model-based RL was not being used significantly in session 3 (n=36; Supplementary Figure 10). The fraction of subjects identified as using a model-based strategy at session 4 was the same in the debriefing and no-debriefing groups (debriefing group 2/12, no-debriefing group 4/24; z = 0, *P*=1, z-test for difference of proportions; Supplementary Figure 10a, f). The logistic regression analysis showed an increased influence of the trial outcome on subsequent choice in session 4 compared to 3 in the debriefing group (null 95% CI [-0.83,0.84], coefficient change=0.95, *P*=0.024, Supplementary Figure 10d), but the session by group interaction did not reach significance (null 95% CI [-0.92,0.88], group difference in coefficient change=0.87, *P*=0.050). The influence of the transition (null 95% CI [-0.48,0.47], coefficient change=0.01, *P*=0.95) and transition-outcome interaction (null 95% CI [-0.28,0.28], coefficient change=-0.14, *P*=0.40) on subsequent choice were unaffected by debriefing (Supplementary Figure 10d) and no parameters of the RL model differed significantly pre and post-debriefing (P>0.14, Supplementary Figure 10e). As expected, no significant differences were observed in any analyses between sessions 3 and 4 in the no-debriefing group (Supplementary Figure 10g-j). These results suggest that in the more complex Changing task, most subjects either did not understand the debriefing or decided the effort of trying to use information about the task structure was not worthwhile.

Explicit reports about task structure are dissociated from uninstructed behaviour

Before the debriefing, and after session 3, participants were given a pen-and-paper questionnaire to explore their understanding of the task structure. Participants were required to answer four questions. For the first two questions, participants were required to provide open answers. Afterwards, the same questions were asked but participants had to choose one from four multiple-choice options.

The questions were asked after showing the participants the following image in a sheet of paper:

Specifically, the first question was: “If you press the upper arrow key when this screen is being shown, what is most likely to happen next?”. The second question was: “If you press the lower arrow key when this screen is being shown, what is most likely to happen next?”

After the participants wrote their answers, the researcher collected the first sheet of paper and provided them a second sheet of paper, which also contained the same image at the top and the same two questions. However, instead of writing an open answer, participants were asked to choose one of four different options for each question:

a) The left side circle will turn yellow

b) The right side circle will turn yellow

c) A coin will appear on the left side

d) A coin will appear on the right side

To classify the open answers that participants gave as correct or incorrect, we created a set of criteria that the answers had to fulfil in order to be considered correct. According to these criteria, each answer had to contain at least one of the following elements: a) “the right circle would turn yellow / light up / be highlighted”; b) “the yellow circle moves [from the top circle] to the right circle”); c) “most of the times, the right circle would turn yellow, although in a minority of times the left circle turns yellow”. Two independent raters (PCR & AM) assessed open answers independently. Then, when in disagreement, they discussed and reached a consensus if those answers should be considered correct or incorrect. The rate of concordance between independent raters was 92%.

Considering all participants together, we found no statistically significant association between participants providing correct or incorrect answers and pre-debriefing behavioural measures, neither in open answers ($-1.7<t's<0.2$; $P's>0.09$; independent sample t-test’s) nor in multiple-choice questions ($-1.2<t's<0.1$; $P's>0.2$). However, we found that subjects who gave correct multiple-choice answers had a higher influence of model-based action values on choice after the debriefing ($t = -2.66$; $P=0.009$; independent sample t-test).

Concerning open answers that did not fill the criteria for being considered correct, we identified a frequent type of wrong belief/model, specifically that the second-step circle is random (~25% of open answers). Regarding the remaining incorrect open answers, they consisted of different types of answers each occurring with a small frequency (<10%), such as ignorance of the second step (“immediately after the first-step choice, a coin may appear or not”) or an incorrect transition model (common transitions identified as rare & rare transitions identified as common). As the first type of wrong model (random second-step) was particularly frequent, we divided subjects in three groups (correct model, random second-step, other wrong models) and tested if their pre-debriefing behaviour was different. We did not find statistically significant

differences between groups in terms of behavioural strategy, either in logistic regression or model fitting analyses (F's<2.0; P's>0.14; one-way ANOVA).

When comparing different clinical groups, we found that individuals with OCD had a smaller proportion of correct open answers than healthy volunteers, but this difference did not reach statistical significance ($\chi 2$ = 2.1, P=0.14, chi-squared test). We also did not observe differences between OCD and healthy volunteers in their proportion of correct multiple-choice answers ($\chi 2$ =0.45, P=0.51). Individuals with mood and anxiety disorders had the same proportion of correct/incorrect answers that healthy volunteers, both in open answers ($\chi 2$ =0.2, P=0.7) and in multiple-choice ($\chi 2$ =0.03, P=0.87).

Our findings complement previous data provided by other groups using an operant conditioning outcome devaluation paradigm in which OCD subjects did not differ from healthy volunteers in their capacity to provide explicit assessment of the contingencies governing action and outcome, although their behaviour did not reflect these (Gillan et al., 2014; Gillan et al., 2015). However, in those studies outcomes were aversive and the experimental induction of habits in humans using operant conditioning paradigms has been questioned (De Wit et al., 2018; Robbins et al., 2019).

3) Supplementary tables

Supplementary table 1 – Permutation test results for differences in learning and debriefing effects between healthy volunteers and individuals with OCD.

Parameter	Learning effect			Debriefing effect		
	Group diff. in parameter change	null 95% CI	P value	Group diff. in parameter change	null 95% CI	P value
Logistic regression						
Outcome	0.277	-0.464, 0.477	0.2512	-0.355	-0.528, 0.537	0.1812
Transition	-0.073	-0.299, 0.295	0.6484	-0.198	-0.489, 0.485	0.4404
Trans. outcome	-0.434	-0.350, 0.359	**0.0152**	-0.202	-0.584, 0.597	0.5356
RL model						
Model-free strength (MF)	1.523	-1.681, 1.684	0.0732	-0.043	-1.615, 1.568	0.9440
Model-based strength (MB)	-0.266	-0.579, 0.574	0.3668	0.041	-0.831, 0.834	0.9368
Value learning rate (αQ)	0.012	-0.282, 0.265	0.9092	-0.079	-0.253, 0.262	0.5360
Eligibility trace (λ)	0.048	-0.204, 0.229	0.6668	-0.008	-0.198, 0.203	0.9340
Transition learning rate (αT)	0.045	-0.411, 0.374	0.7536	0.016	0.315, 0.308	0.9068
Choice bias	-0.076	-0.299, 0.299	0.6200	0.237	-0.355, 0.354	0.1804
Choice perseveration	0.220	-0.577, 0.572	0.4776	-0.864	-0.842, 0.800	0.0424

Permutation tests (5000 permutations) were used to assess differences in the fitted model parameter loadings between healthy volunteers (n=67) and individual with OCD (n=46) in the effect of learning (defined as change between session 1 and 3) and debriefing (defined as change between session 3 and 4, taking only subjects who are MF at session 3).

Supplementary table 2 – Permutation test results for differences in learning and debriefing effects between healthy volunteers and individuals with mood and anxiety disorders

Parameter	Learning effect			Debriefing effect		
	Group diff. in parameter change	null 95% CI	P value	Group diff. in parameter change	null 95% CI	P value
Logistic regression						
Outcome	0.321	-0.441, 0.437	0.1592	-0.149	-0.487, 0.507	0.5708
Transition	0.060	-0.291, 0.291	0.6764	-0.102	-0.507, 0.518	0.6844
Trans. outcome	-0.188	-0.378, 0.387	0.3500	0.126	-0.594, 0.617	0.6848
RL model						
Model-free strength (MF)	0.265	-1.342, 1.262	0.6664	0.218	-1.229, 1.241	0.7548
Model-based strength (MB)	0.055	-0.550, 0.550	0.8728	0.489	-0.750, 0.732	0.2020
Value learning rate (αQ)	0.080	-0.261, 0.253	0.5288	-0.082	-0.254, 0.250	0.5304
Eligibility trace (λ)	-0.080	-0.170, 0.198	0.3660	0.078	-0.200, 0.191	0.4276
Transition learning rate (αT)	-0.155	-0.556, 0.516	0.6500	-0.124	-0.439, 0.443	0.6024
Choice bias	0.111	-0.273, 0.271	0.4204	0.114	-0.336, 0.328	0.5140
Choice perseveration	0.501	-0.598, 0.609	0.1140	-1.400	-0.821, 0.840	0.0012

Permutation tests (5000 permutations) were used to assess differences in the fitted model parameter loadings between healthy volunteers (n=67) and individual with mood and anxiety disorders (n=49) in the effect of learning (defined as change between session 1 and 3) and debriefing (defined as change between session 3 and 4, taking only subjects who are MF at session 3).

Supplementary table 3 - Differences in learning and debriefing effects in healthy volunteers between the Lisbon and New York samples.

Parameter	Learning effect			Debriefing effect		
	Group diff. in parameter change	null 95% CI	P value	Group diff. in parameter change	null 95% CI	P value
Logistic regression						
Outcome	-0.233	-0.635, 0.646	0.4844	0.216	-0.742, 0.736	0.5744
Transition	-0.088	-0.347, 0.346	0.6392	-0.255	-0.777, 0.762	0.5336
Trans. outcome	-0.128	-0.475, 0.511	0.6392	-0.197	-0.916, 0.868	0.6564
RL model						
Model-free strength (MF)	0.015	-2.180, 2.017	0.9492	0.103	-1.771, 2.141	0.9056
Model-based strength (MB)	-0.457	-0.712, 0.725	0.2096	-0.744	-1.200, 1.219	0.2180
Value learning rate (αQ)	-0.167	-0.347, 0.344	0.3320	0.139	-0.315, 0.330	0.4600
Eligibility trace (λ)	-0.166	-0.261, 0.272	0.2412	0.202	-0.284, 0.274	0.1616
Transition learning rate (αT)	0.283	-0.603, 0.573	0.3632	-0.130	-0.428, 0.463	0.6260
Choice bias	0.175	-0.379, 0.411	0.4168	0.097	-0.481, 0.497	0.7180
Choice perseveration	0.158	-0.763, 0.750	0.7124	0.642	-1.313, 1.198	0.3164

Permutation tests (5000 permutations) were used to assess differences in the fitted model parameter loadings between Fixed task Lisbon (n=40) and New York (n=27) samples in the effect of learning (defined as change between session 1 and 3) and debriefing (defined as change between session 3 and 4, taking only subjects who are MF at session 3).

Supplementary table 4 - Differences in learning and debriefing effects in individuals with OCD between the Lisbon and New York samples.

Parameter	Learning effect			Debriefing effect		
	Group diff. in parameter change	null 95% CI	P value	Group diff. in parameter change	null 95% CI	P value
Logistic regression						
Outcome	0.477	-0.677, 0.705	0.2052	-0.298	-0.830, 0.777	0.4824
Transition	0.393	-0.477, 0.525	0.1400	0.514	-0.618, 0.693	0.1412
Trans. outcome	0.005	-0.452, 0.481	0.9968	0.333	-0.858, 0.902	0.4764
RL model						
Model-free strength (MF)	1.081	-2.746, 2.756	0.3912	-3.128	-2.864, 2.511	0.0372
Model-based strength (MB)	-0.656	-0.745, 0.644	0.0852	0.286	-1.152, 1.271	0.6452
Value learning rate (αQ)	0.057	-0.393, 0.414	0.8376	0.349	-0.447, 0.420	0.1108
Eligibility trace (λ)	-0.079	-0.282, 0.339	0.6508	0.095	-0.331, 0.321	0.6164
Transition learning rate (αT)	-0.068	-0.503, 0.452	0.8972	0.241	-0.413, 0.395	0.2812
Choice bias	0.037	-0.537, 0.534	0.9088	-0.170	-0.610, 0.609	0.5576
Choice perseveration	0.733	-0.934, 1.005	0.1564	-0.129	-1.048, 1.052	0.8032

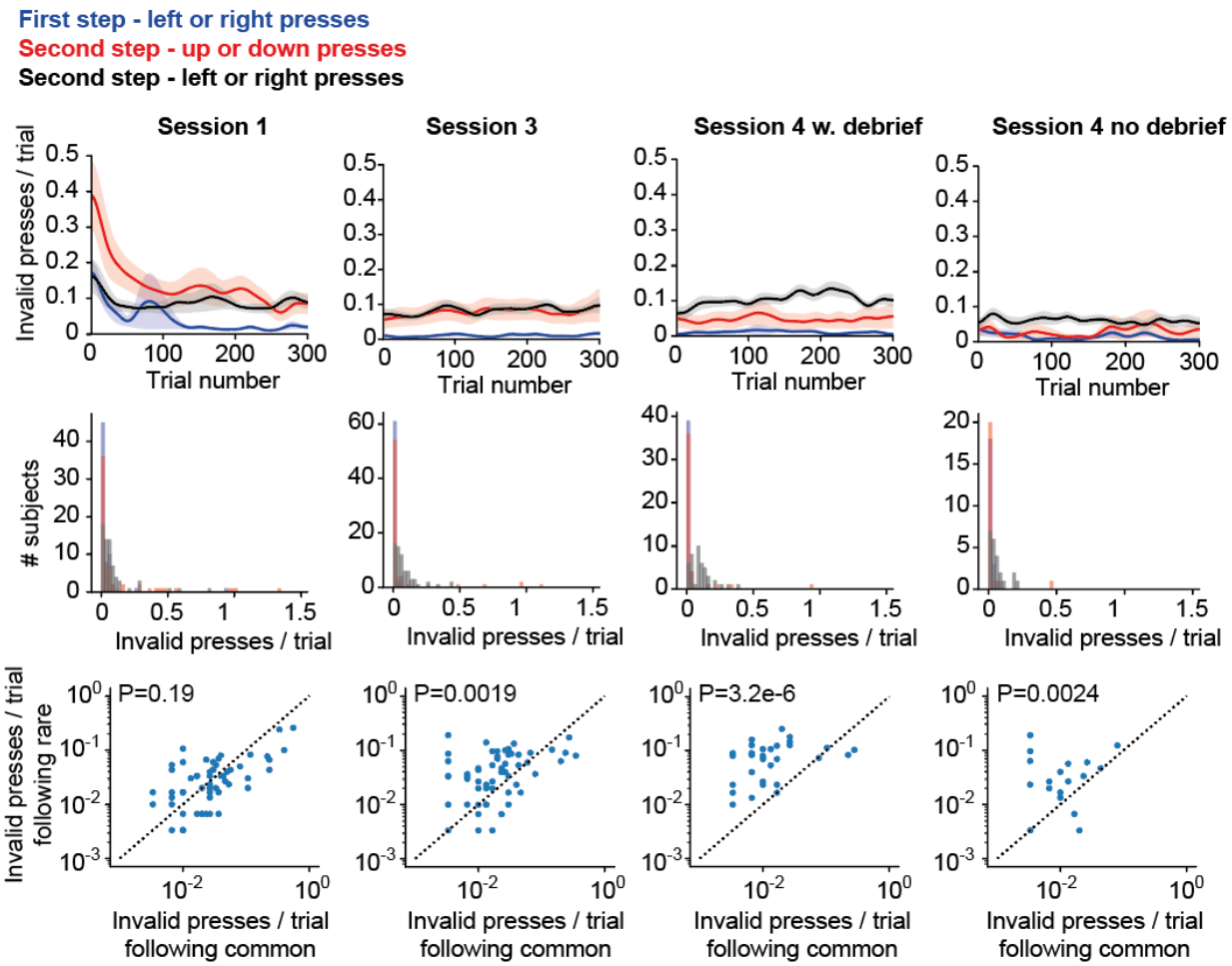
Permutation tests (5000 permutations) were used to assess differences in the fitted model parameter loadings between Fixed task Lisbon (n=16) and New York (n=30) samples in the effect of learning (defined as change between session 1 and 3) and debriefing (defined as change between session 3 and 4, taking only subjects who are MF at session 3).

Supplementary table 5 - Differences in learning and debriefing effects in individuals with other mood and anxiety disorders between the Lisbon and New York samples.

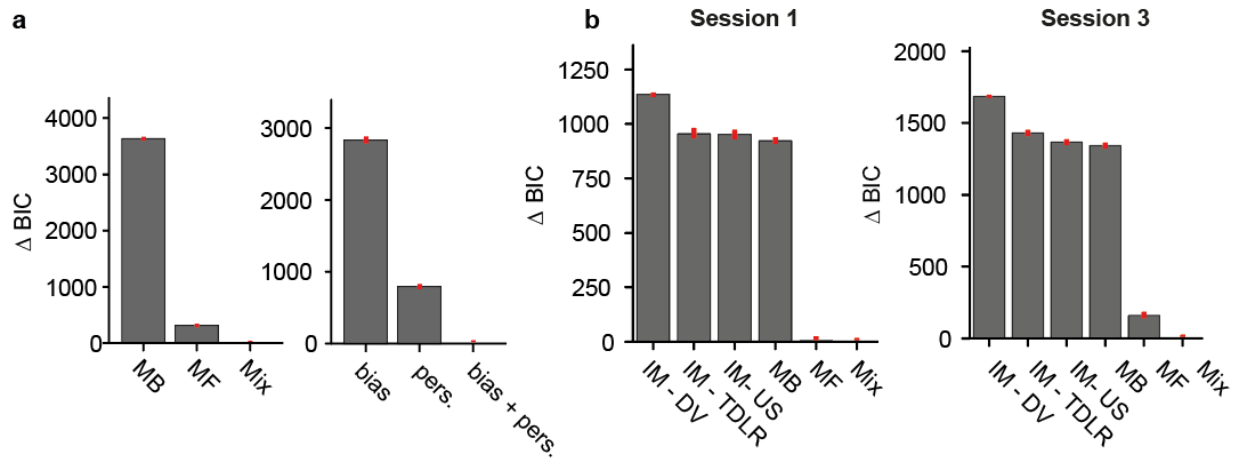
Parameter	**Learning effect**			**Debriefing effect**		
	Group diff. in parameter change	null 95% CI	P value	Group diff. in parameter change	null 95% CI	P value
Logistic regression						
Outcome	-0.432	-0.614, 0.660	0.1796	0.150	-0.788, 0.798	0.7076
Transition	0.229	-0.526, 0.533	0.3976	-0.055	-0.777, 0.773	0.9784
Trans. outcome	0.521	-0.617, 0.649	0.1160	-0.367	-0.940, 0.981	0.5092
RL model						
Model-free strength (MF)	0.341	-1.611, 1.910	0.6388	-0.544	-1.764, 1.517	0.5136
Model-based strength (MB)	0.114	-0.915, 1.011	0.8672	-0.102	-1.246, 1.123	0.9592
Value learning rate (αQ)	-0.088	-0.403, 0.360	0.7328	0.132	-0.374, 0.431	0.5828
Eligibility trace (λ)	-0.119	-0.229, 0.265	0.2912	0.029	-0.263, 0.307	0.8732
Transition learning rate (αT)	-0.020	-0.730, 0.720	0.9616	-0.195	-0.601, 0.688	0.5728
Choice bias	0.086	-0.410, 0.420	0.7124	0.102	-0.541, 0.549	0.6796
Choice perseveration	0.595	-0.996, 1.002	0.2532	-0.458	-0.969, 1.061	0.3928

Permutation tests (5000 permutations) were used to assess differences in the fitted model parameter loadings between Fixed task Lisbon (n=16) and New York (n=33) samples in the effect of learning (defined as change between session 1 and 3) and debriefing (defined as change between session 3 and 4, taking only subjects who are MF at session 3).

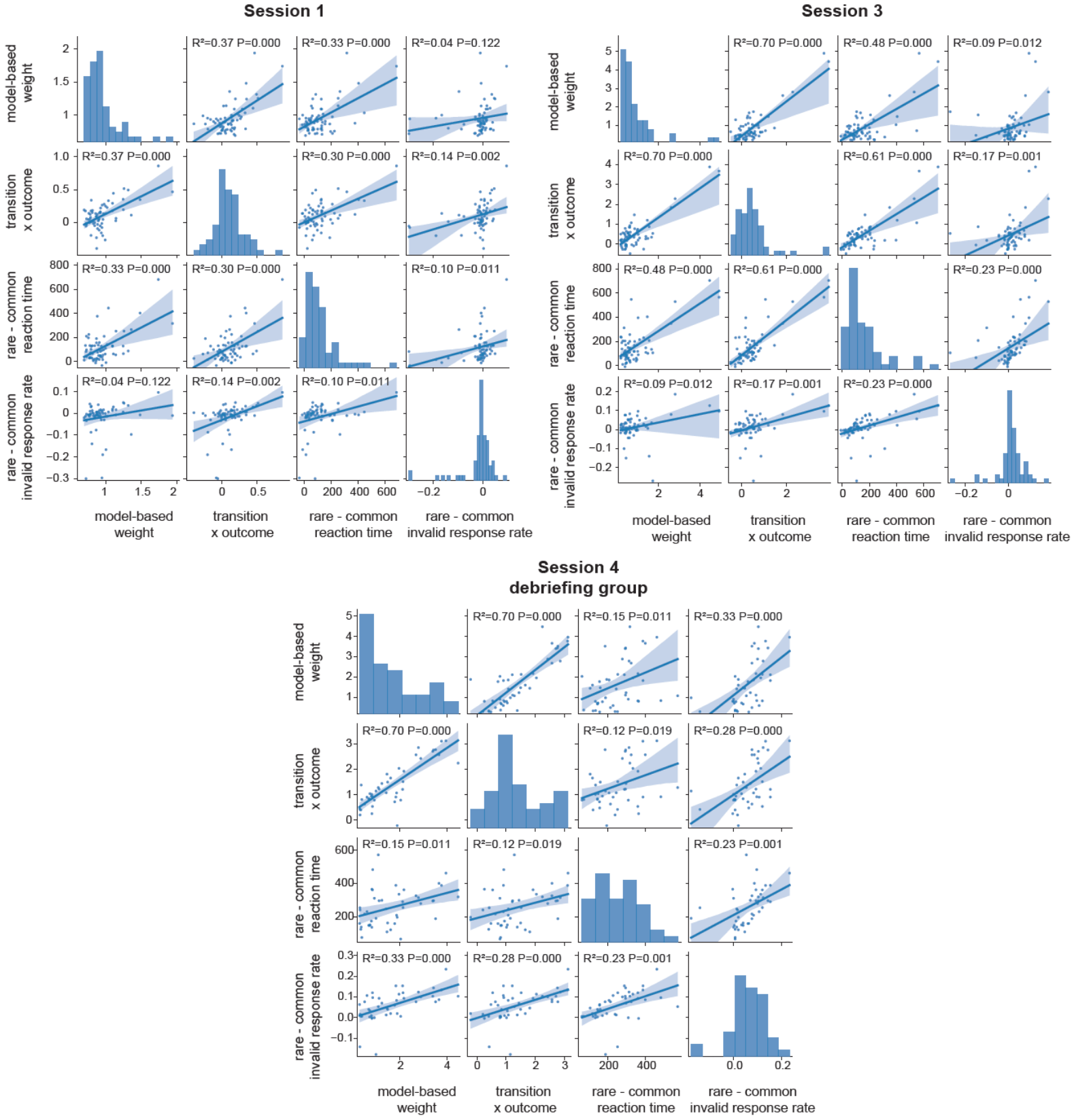
4) Supplementary figures



Supplementary figure 1. Invalid key presses. Top panels show the mean number of invalid key presses per trial as a function of trial number, Gaussian smoothed with an SD of 10 trials, shaded area shows SEM across subjects. Middle panels show histogram of the mean number of invalid presses per trial across the entire session for each subject. Bottom panels show the rate of invalid left/right presses at the second-step for each subject following common (x axis) and rare (y axis) transitions. P values are for Wilcoxon signed rank test for difference in rate of incorrect presses following common vs rare transitions.

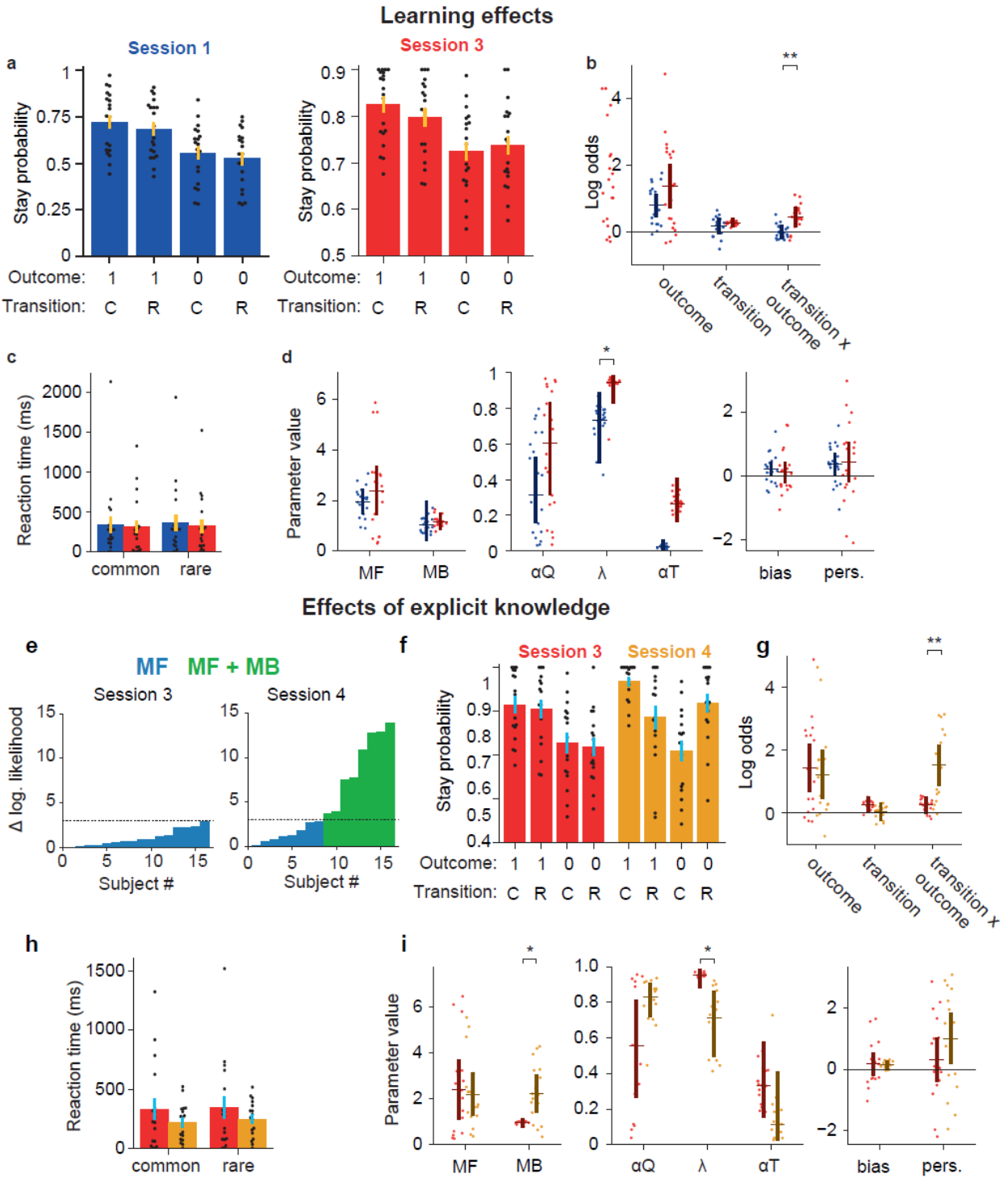


Supplementary figure 2. Model comparison. **a**) Model comparison for sessions 1-3 of the Fixed task. Panels show the difference in BIC score relative to the best fitting model. Left panel, comparison of model-based (MB), model-free (MF) and MB-MF mixture (Mix) models. Right panel, comparison of mixture model with bias parameter, perseveration parameter, and bias + perseveration parameters. **b**) Model comparisons including additional model-based agents with incorrect models of the task structure; one which believed state transitions were deterministic but volatile (IM-DV), one with transition dependent learning rates at the second-step (IM-TDLR) and one which believed that one first step option was unlucky and reduced reward probability at the second-step (IM-US).



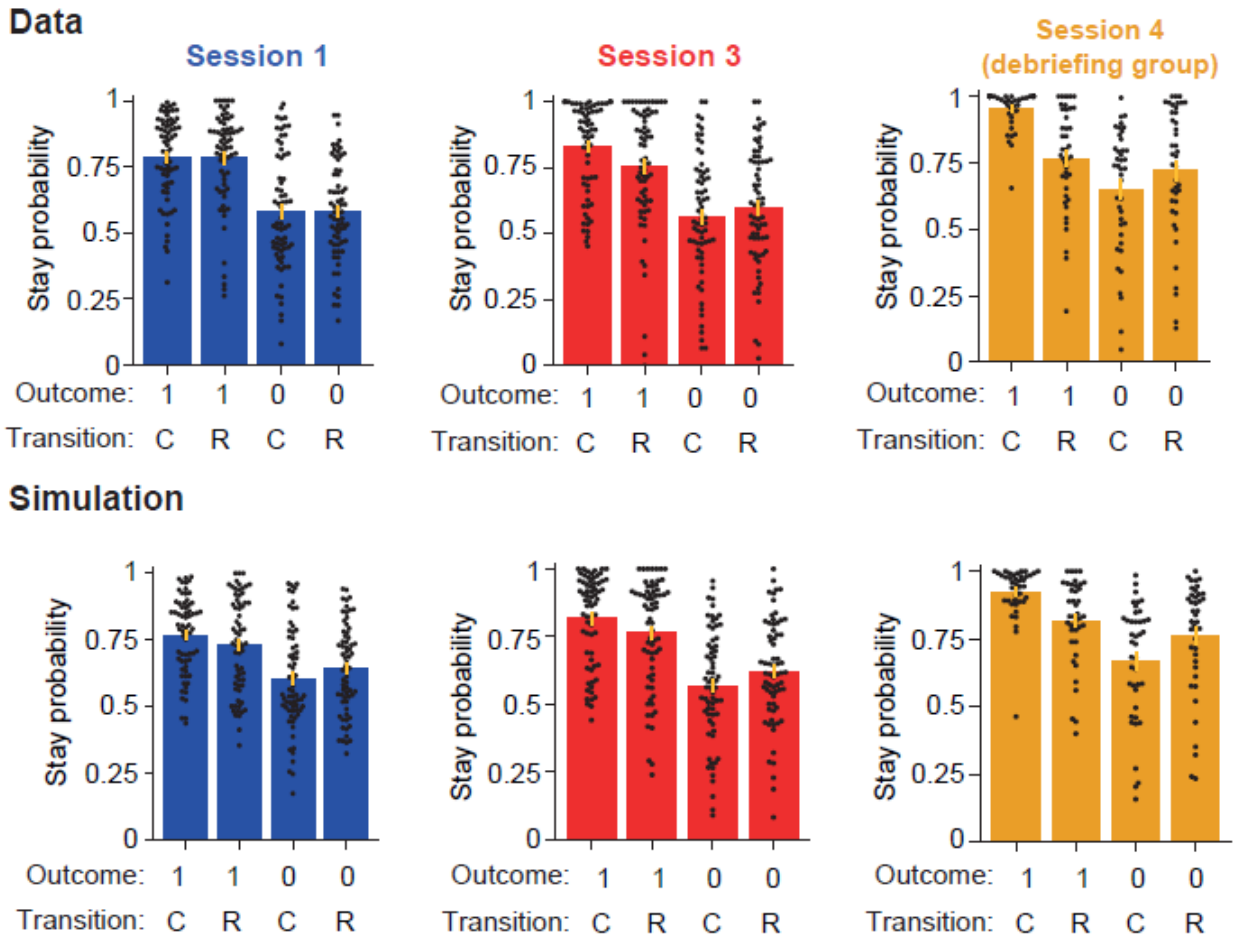
Supplementary figure 3. Correlation between choice and implicit measures of task structure learning. Correlation across subjects between different measures of task structure learning in the fixed task at sessions 1, 3 and 4. Two choice-based measures were used; the model-based weight parameter from the RL model fit and the transition x outcome predictor loading from the logistic regression, and two implicit measures; the difference in second-step reaction times following common vs rare transitions, and the difference in the rate of invalid

second-step responses (e.g. pressing left when the right state was active) following common vs rare transitions. Points show individual subjects. Lines show linear fit with 95% confidence interval on fit indicated by shaded region. In both session 1 and 3, when subjects are learning only from experience, both measures of model-based choice are correlated with the rare-common reaction time difference.

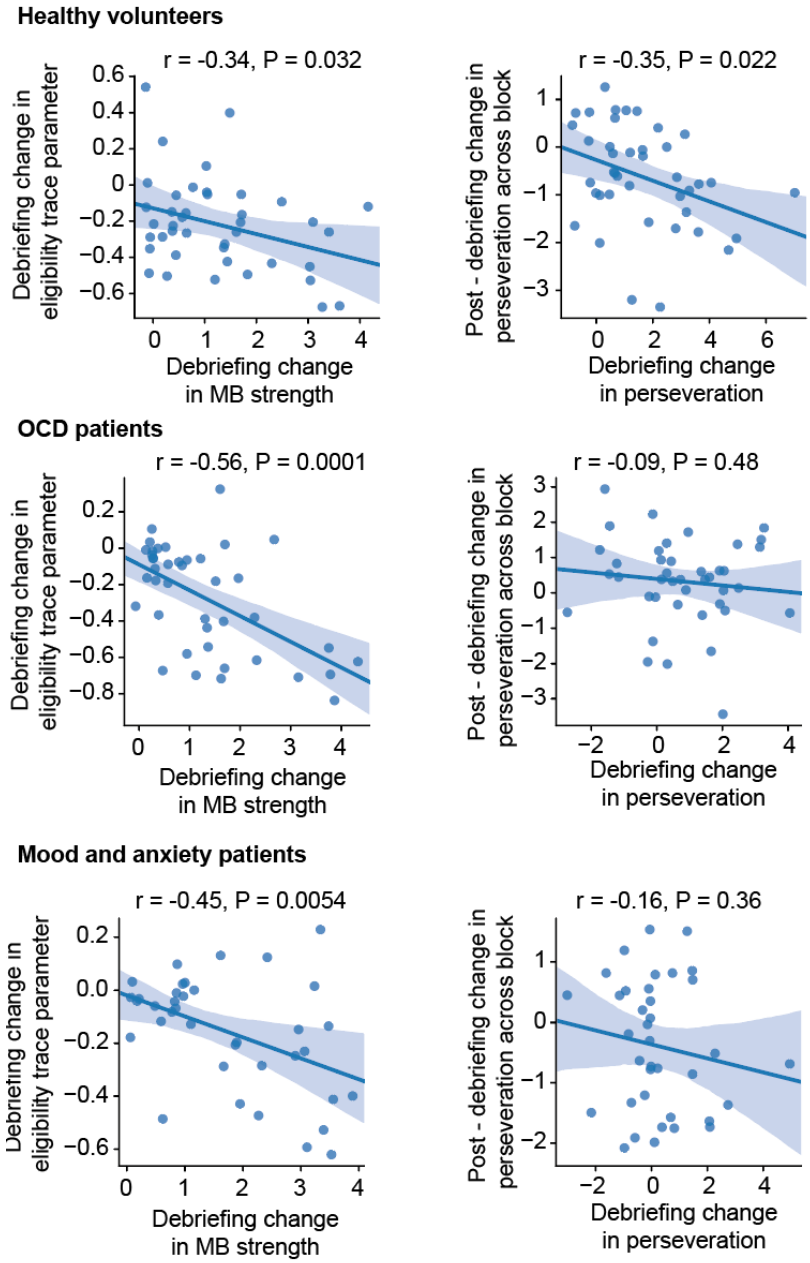


Supplementary figure 4. Slow-paced task. (a-d) Learning effects a) Stay probability analysis. **b)** Logistic regression analysis of stay probabilities. **c)** Reaction times after common and rare transitions in session 1 and 3. **d)** Comparison of mixture model fits between session 1 and session 3. **(e-i) Effects of explicit knowledge. e)** Per-subject likelihood ratio test for use of model-based strategy on session 3 (left panel) and session 4 (right panel). **f)** Stay probability analysis. **g)** Logistic regression analysis of stay probabilities. **h)** Reaction times after common and rare transitions in session 1 and 3. **I)** Comparison of mixture model fits between session 1 and session 3. RL model parameters: MF, Model-free strength; MB,

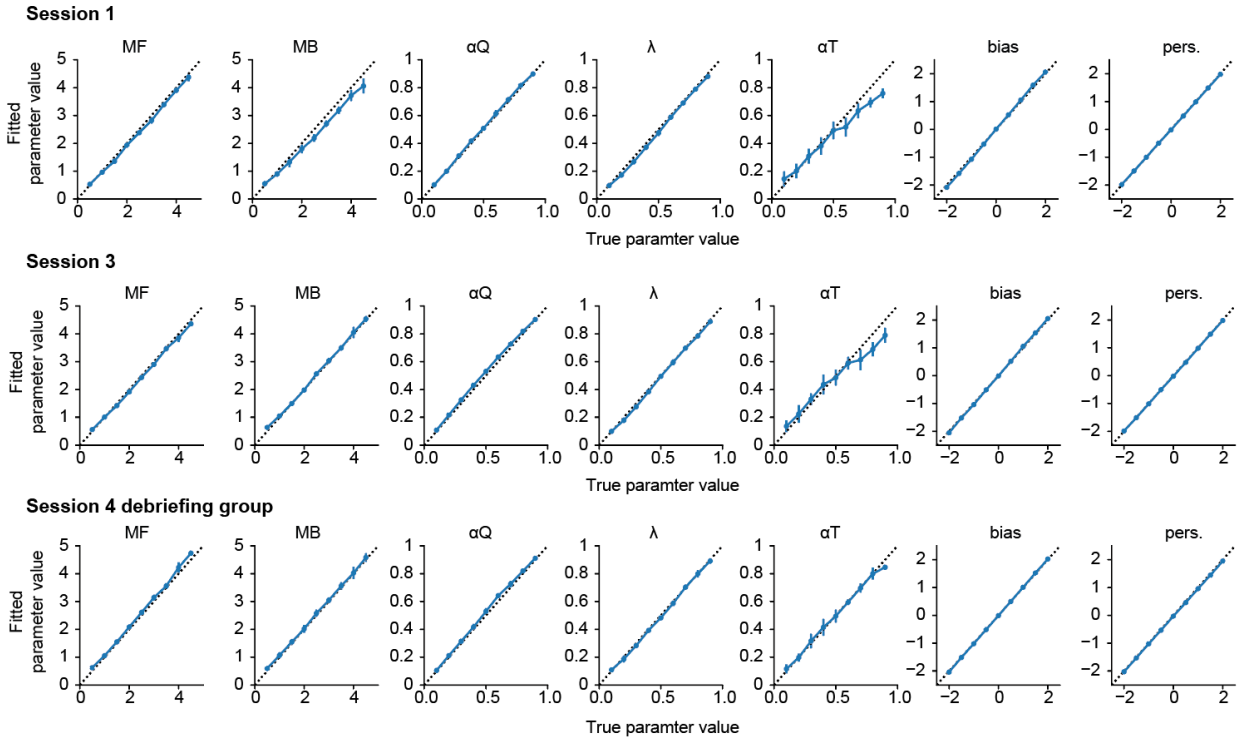
Model-based strength; αQ, Value learning rate; λ, Eligibility trace; αT, Transition probability learning rate; bias, Choice bias; pers., Choice perseveration.



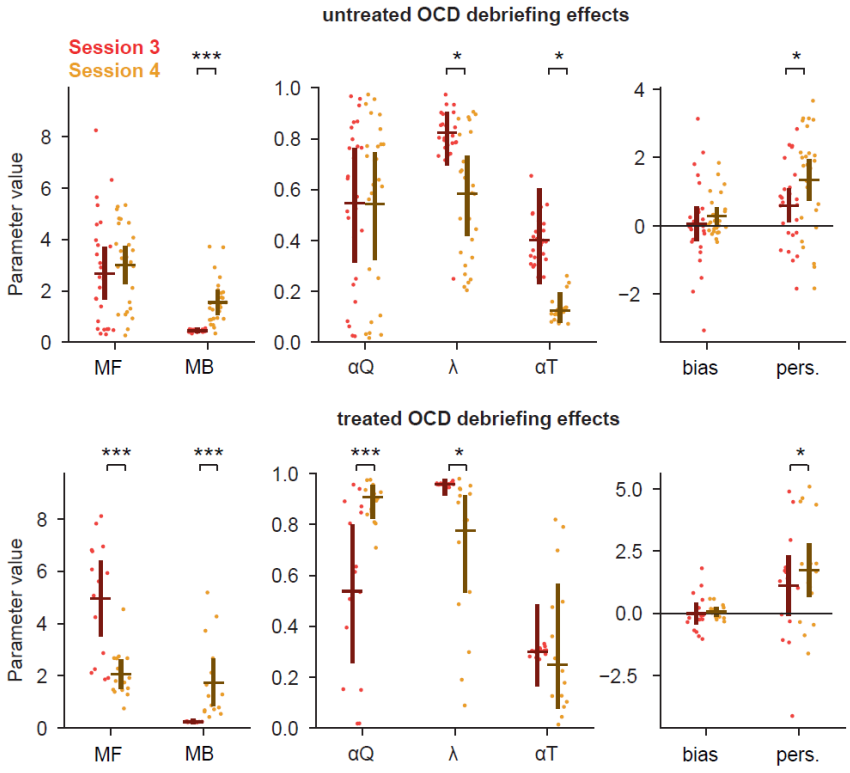
Supplementary figure 5. Stay probabilities for best fitting RL model Stay probability analysis showing the probability of repeating the first step choice on the next trial as a function of trial outcome (rewarded or not rewarded) and state transition (common or rare). Top panels show experimental data, bottom panels show behaviour simulated from the best fitting RL model, which used a mixture of model-based and model-free. Error bars indicated the cross subject standard error of the mean (SEM). In each group data was analysed separately for session 1 (blue graph), session 3 (red graph) and session 4 (gold graph).



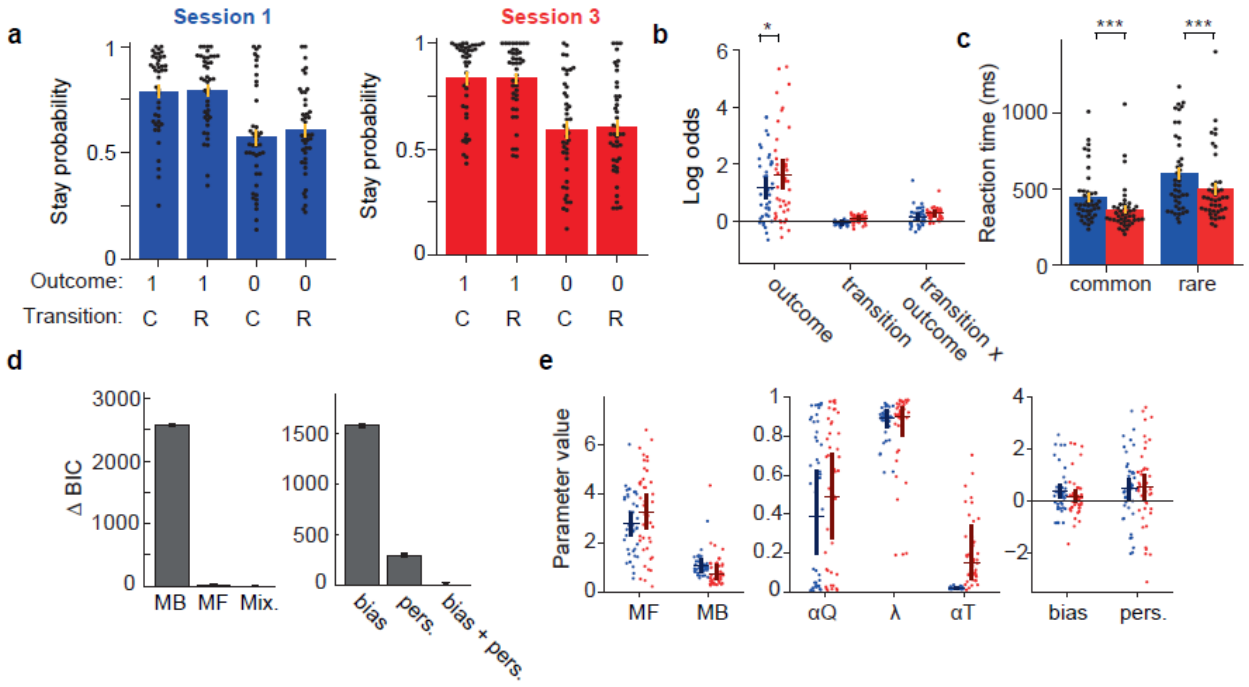
Supplementary figure 6. Debriefing effect correlations. Left panels – Correlation across subjects between the effect of debriefing on subjects use of model-based RL (as assessed by the RL model's model-based weight parameter) and that on the RL model's eligibility trace parameter. Right panels – Correlation between the effect of debriefing on subjects overall perseveration (as assessed by the RL models perseveration parameter), and post debriefing, their change in perseveration from early to late in blocks, assessed using logistic regression analysis of data from early (10-20 trials post block transition) and late (30-40 trials) in each block.



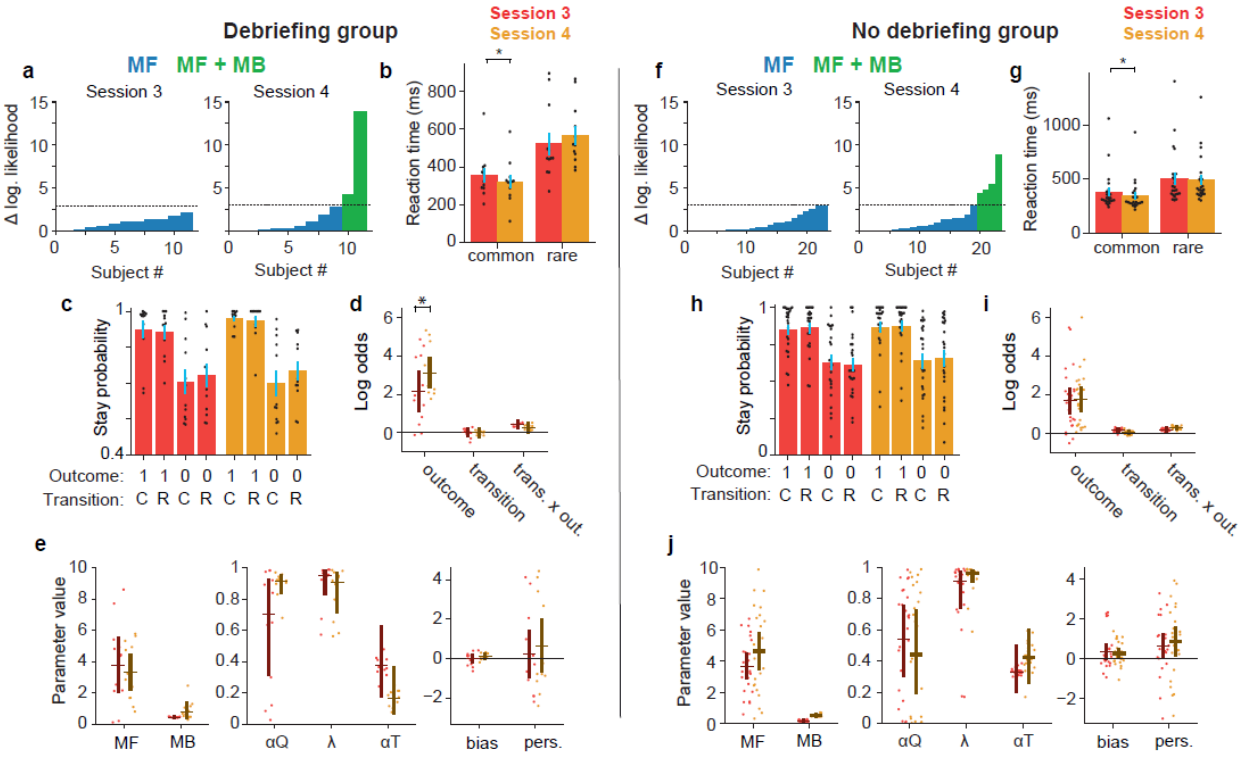
Supplementary figure 7. RL model parameter recovery. Test of the accuracy with which RL model parameters could be recovered from simulated data. Panels show mean and standard deviation of recovered parameters across 10 repeated simulation runs when the parameter under investigation was fixed at the specified 'true parameter value' and the other parameters were drawn randomly for each subject from the population level distributions fit to the specified dataset (top row- fixed task session 1, middle row – fixed task session 3, bottom row – fixed task debriefing group session 4). Overall the accuracy of parameter recovery was very good, with a slightly reduced accuracy for the transition learning rate (parameter α_T) in sessions 1 and 3 where the influence of model-based RL was small.



Supplementary figure 8 – Comparison of debriefing effects in individuals with OCD between Lisbon and New York Samples. For each RL model parameter, dots indicate maximum a posteriori model parameter loading for individual subjects, bars indicate the population mean and 95% confidence interval on the mean. For each model parameter, permutation tests were used to assess effect of debriefing on the fitted loadings, separately for individuals with OCD recruited in New York (n=30), who were tested in the absence of pharmacological treatment (top panel), and in Lisbon (n=16), the majority of whom were under pharmacological treatment (bottom panel). These analyses were not performed in the remaining groups, because significant group-debriefing interactions were found only for the OCD sample (see Supplementary tables 1-3). Regarding the two variables for which such interactions were significant (*P*=0.04 for both; Supplementary table 2), debriefing was found to reduce the strength of model-free RL (MF) and increase the value learning rate (αQ) in treated, but not untreated, individuals with OCD. RL model parameters: MF, Model-free strength; MB, Model-based strength; αQ, Value learning rate; λ, Eligibility trace; αT, Transition probability learning rate; bias, Choice bias; pers., Choice perseveration.



Supplementary figure 9. Learning effects in the Changing transition probabilities task. a) Stay probability analysis showing the probability of repeating the first step choice on the next trial as a function of trial outcome (rewarded or not rewarded) and state transition (common or rare). Error bars indicate the cross subject standard error (SEM). The left panel shows data from the first session, the right panel shows data from session 3. **b)** Logistic regression analysis of how the outcome (rewarded or not), transition (common or rare) and their interaction, predict the probability of repeating the same choice on the subsequent trial. Positive loading on the 'outcome' predictor indicates a tendency to repeat rewarded choices. Positive loading on the 'transition' predictor reflects a tendency to repeat choices followed by common transitions. Positive loading on the 'transition x outcome' interaction predictor indicates a tendency to repeat choices that were rewarded following a common transition, or that were not rewarded following a rare transition. Dots indicate maximum a posteriori loadings for individual subjects, bars indicate the population mean and 95% confidence interval on the mean. Statistical significance of differences in factor loadings for each predictor between session 1 (blue) and 3 (red) were evaluated using permutation tests. **c)** Reaction times after common and rare transitions in session 1 and 3. **d)** Bayesian Information Criteria (BIC) model comparison for sessions 1-3. Left panel, comparison of model-based (MB), model-free (MF) and mixture (MF+MB) models. Right panel, comparison of mixture model with bias parameter, perseveration parameter, and bias + perseveration parameters. **e)** Comparison of mixture model fits between session 1 and session 3. RL model parameters: MF, Model-free strength; MB, Model-based strength; αQ, Value learning rate; λ, Eligibility trace; αT, Transition probability learning rate; bias, Choice bias; pers., Choice perseveration.



Supplementary figure 10. Effects of explicit knowledge in the Changing transition probabilities task. **(a, f)** Per-subject likelihood ratio test for use of model-based strategy on session 3 (left panel) and session 4 (right panel). Data was analysed separately for groups with (A) and without (F) debriefing. Y-axis shows difference in log likelihood between mixture (model-free + model-based) RL model and model-free only RL model. Blue bars indicate subjects for which likelihood ratio test favours model-free only model, green bars indicate subjects for which test favours mixture model, using a $p<0.05$ threshold for rejecting the simpler model. We compared sessions 3 and 4 only in the subjects for whom a likelihood ratio test indicated that model-based RL was not used in session 3. **(b, g)** Reaction times after common and rare transitions in session 3 and 4. **(c, h)** Stay probability analysis showing the probability of repeating the first step choice on the next trial as a function of trial outcome (rewarded or not rewarded) and state transition (common or rare). Error bars indicated the cross subject standard error of the mean (SEM). In each group data was analysed separately for session 3 (red graph) and session 4 (gold graph). **(d, i)** Logistic regression analysis of how the outcome (rewarded or not), transition (common or rare) and their interaction, predict the probability of repeating the same choice on the subsequent trial. **(e, j)** Comparison of mixture model fits between session 3 (red) and session 4 (gold) in the group without instruction (left panels) and the group with instruction (right panels). RL model parameters: MF, Model-free strength; MB, Model-based strength; αQ, Value learning rate; λ, Eligibility trace; αT, Transition probability learning rate; bias, Choice bias; pers., Choice perseveration.