Supplementary Material for **A double-blind study assessing the impact of orbitofrontal theta burst stimulation on goal-directed behavior**

**Computational modeling of habit override task performance**

Results reported in the main manuscript use summary statistics of behavior to describe performance on the habit override task. However, learning tasks can also be analyzed with generative models, which describe trial-level learning patterns and can arbitrate among different influences on behavior with increased precision. Development, estimation, and results of a computational model of habit override are presented below; results are generally consistent with the summary statistics presented in the manuscript but are described below for the interested reader.

Model development: To our knowledge, computational learning models have not been developed for habit override tasks, so we developed a model and possible variants to account for differences in override behavior. Specifically, we developed models to test whether variations in override behavior are due to differences in model-based revaluation of stimuli, learning from experienced outcomes, or random responding.

Reinforcement learning models assumed participants learned the value of responding (pressing a button) for each stimulus based on outcomes (shock or no shock) and instructions. The values of responding for each stimulus (CS+A, CS+B, and CS-) were assumed to be independent of one another. Therefore, the probability of responding on a trial ($P(a_t)$) when presented with a stimulus s was based on arbitrating the expected value of responding ($Q(a_t)$) versus not responding ($1\text{-}Q(a_t)$) for that stimulus (s) based on a softmax function, with an inverse temperature parameter β representing choice stochasticity or randomness:

$$P(a_t|s_t) = \frac{\exp\left[\beta * Q(a_t)\right]}{\exp[\beta * Q(a_t)] + \exp\left[\beta * [1 - Q(a_t)]\right]}$$

For each trial, the experienced outcome ($O_t$) was coded as the value of making a response to avoid a shock: 1 if the participant either: responded and did not receive a shock, or did not respond and received a shock, and 0 if the participant did not respond and did not receive a shock. For the sake of consistency, the value of responding was 0 if the participant responded and received a shock as well, which would occur if participants pressed the wrong button (such errors were rare: 1.8% of responses). This design allowed the model to capture habitual avoidance responses, which are common (Solomon & Wynne, 1954), by coding a non-shocked outcome as 1 regardless if the button press prevented the shock or if a shock would not have been delivered regardless. Therefore, the expected value of responding for a stimulus was updated with the following equation after an outcome, with a learning rate parameter α representing how quickly expected values were updated based on the experienced difference between actual and expected outcomes:

$$Q(a_{t+1}) = Q(a_t) + \alpha * [O_t - Q(a_t)]$$

At the beginning of the override block, one of the shock electrodes attached to the participant's wrist was disconnected and the participant was instructed that one of the stimuli (CS+B) would no longer lead to shock. For the sake of modeling, any change in the value of responding for this devalued stimulus at the beginning of the override block is termed 'instructed devaluation' (though 'observed devaluation' would also be appropriate, as participants both observed and were instructed about the change in

contingencies, and the model is agnostic as to whether instructions, observed disconnection of shock electrodes, or both, caused changes in behavior). This instructed devaluation was represented in the model by a change in the expected value of the devalued stimulus to the parameter τ at the beginning of each block of new stimuli (when t=1, similar to (Atlas, Doll, Li, Daw, & Phelps, 2016)):

$$Q[a_1(CS-)] = 0$$

$$Q[a_1(CS+A)] = 1$$

$$Q[a_1(CS+B)] = \tau$$

Given the near-ceiling performance for CS- and CS+A during both acquisition and override, the expected value of responding for CS- was set at 0 and for CS+A was set at 1 at the beginning of each block.

Model comparison: Additional variants of the model separated learning rates for learning from responses vs. nonresponses ($\alpha_{act}$ and $\alpha_{no\_act}$) and represented the value of responding as a free parameter ω (instead of 1) to account for individual differences in avoidance. However, model simulation revealed that posterior values of the α (learning rate) parameter remained close to the prior, suggesting that participants' behavior was uninformative about the value of α. This effect was similar for models using one parameter (α) versus two ($\alpha_{act}$ and $\alpha_{no\_act}$) and suggests that, regardless of whether an action was performed or not, little learning was occurring during the override portion. Specifically, the large number of trials with stable, deterministic outcomes resulted in small discrepancies between actual and expected values (prediction errors), leading to small value updates that could not arbitrate among different values of α. Therefore, another model variant was tested with α fixed at its mean (0.5). Simulation also showed that parameters β (inverse temperature) and ω (value of responding) were collinear, since both affect the relationship between expected value and choices (Huys, Pizzagalli, Bogdan, & Dayan, 2013). Only the β parameter was retained in most models, and differences in this parameter can therefore be interpreted as differences in choice stochasticity or the value of responding.

Four models in total were tested on participants' data from the baseline visit: α+β+τ (3 free parameters); $\alpha_{act}$+$\alpha_{no\_act}$+β+τ (4 free parameters); β+τ+ω (3 parameters); and β+τ (2 parameters). See model estimation below for estimation details. Integrated BIC (iBIC; (Huys et al., 2013)) was used to compare model fits by averaging each model's likelihood over the number of samples (marginal posterior likelihood) and correcting for the number of parameters. A lower iBIC indicates a better fit. The β+τ+ω model did not converge ($\hat{R}$>1.05; likely due to the collinearity noted in simulations) and so is not included in comparisons. The other model fits are noted in the table below.

| **Model** | **Negative Log Likelihood** | **iBIC** |
|---|---|---|
| α+β+τ | -2386.881 | 4802.862 |
| $\alpha_{act}$+$\alpha_{no\_act}$+β+τ | -2386.707 | 4812.218 |
| β+τ | -2382.877 | 4785.155 |

**Supplementary Table 1. Model fits for variants of habit override model.**

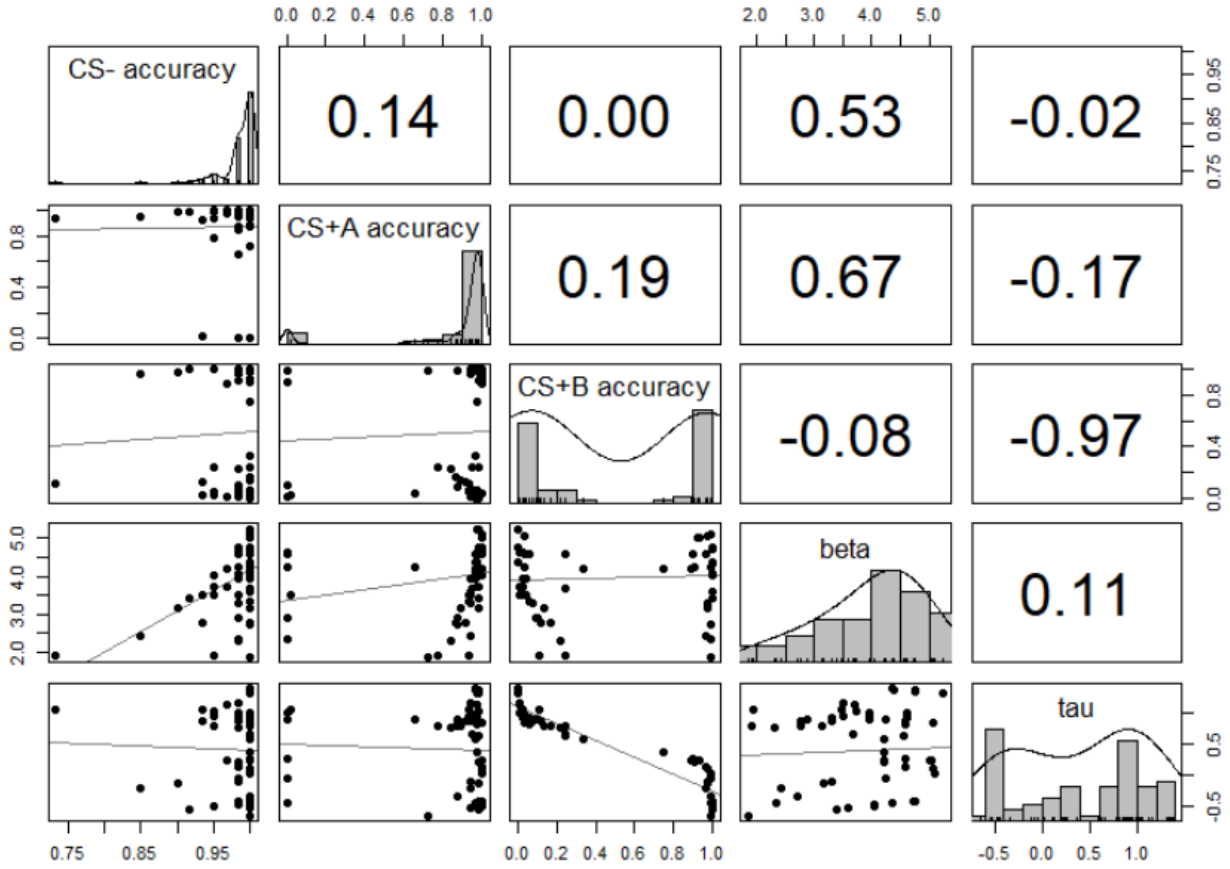
The model with a fixed, single learning rate (β+τ) fit the best (lowest iBIC), suggesting that differences in behavior were due to either differences in instructed devaluation (represented by differences in τ) or in choice stochasticity or relative valuation of responding (represented by differences in β), and not by differences in learning. To understand the relative effects of choice stochasticity and instructed devaluation, the best-fitting model was fit to participants' choices and parameters were analyzed.

Model estimation: The best-fitting habit override model was fit using hierarchical Bayesian estimation in Stan (Carpenter et al., 2016). Hyperparameters on mean values of parameters were given a normal distribution centered at 0 and standard deviations of 1 (τ) or 3 (β). The mean value of β was constrained to be greater than 0. Hyperparameters on standard deviations of parameters were given a Student's t distribution with three degrees of freedom and standard deviations of 1 (τ) or 2 (β). Parameter values per participant and visit (as applicable) were estimated using non-centered parameterization (Betancourt & Girolami, 2015) to improve estimation.

Comparison to behavioral summaries: Free parameters β and τ were estimated per participant for the baseline visit and compared to summaries of accuracy for each parameter (see Supplementary Figure 1 below). The model-estimated instructed devaluation parameter τ was strongly correlated with override performance (overall CS+B accuracy, Spearman correlation = -0.970), meaning that a higher model-calculated value of responding at the beginning of each override block was negatively related to the average number of trials where responding was (correctly) withheld.

The instructed devaluation parameter was not correlated with inverse temperature or either CS- or CS+A accuracy ($rs < |0.2|$), while inverse temperature was moderately correlated with both CS- ($r = 0.53$) and CS+A ($r = 0.67$) accuracy. Therefore, while the instructed devaluation parameter was essentially equivalent to devaluation success and unrelated to performance with other stimuli, the inverse temperature parameter was related to performance accuracy on non-devalued stimuli but was unrelated to devaluation success.

In summary, the modeling results show that variations in performance on the habit override task are due to a combination of differences in instructed devaluation and random responding (or choice stochasticity). The proportion of correctly devalued responses, reported in the main text, was strongly related to the model-based measure of instructed devaluation and not to random responding, confirming that this statistic measures the ability to re-value stimuli in a model-based manner, and not other behavioral effects. Meanwhile, the proportion of correct responses for non-devalued stimuli was related to the model-based parameter measuring random responding but not the parameter measuring instructed devaluation, supporting that random responding was related to accuracy of responses to other stimuli besides the devalued one.




**Supplementary Figure 1**: Relationships between model-estimated parameters and summaries of behavioral performance on the override task at baseline. Diagonal plots are histogram and density plots for each variable, values above the diagonal are Spearman correlations between variables, and plots below the diagonal are scatterplots and regression lines for individual data points.

**Relationships between symptoms (OCI Total Score) or devaluation success and all measures on two-step task.**

The main text reports relationships between symptoms or devaluation success and the main behavioral measure of interest on the two-step task, model-based planning (reward x transition interaction). Supplementary Table 2 below reports all results for these analyses. Model-free learning is represented by the main effect of reward and the intercept represents the overall likelihood of staying with the previous action (perseveration).

| | **Intercept** | | **Reward** | | **Transition Type** | | **Reward x Transition Type** | |
|---|---|---|---|---|---|---|---|---|
| | z | p | z | p | z | p | z | p |
| **OCI Total** | | | | | | | | |
| **Main effect** | 4.134 | <.001 | 2.225 | 0.0261 | 0.143 | 0.8862 | 1.37 | 0.172 |
| **Interaction with OCI** | -1.672 | 0.095 | -0.264 | 0.792 | -0.773 | 0.440 | -0.244 | 0.807 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Devaluation task success** | | | | | | | | |
| **Main effect** | 2.78 | 0.005 | 2.16 | 0.031 | -0.780 | 0.436 | 0.060 | 0.952 |
| **Interaction with devaluation success** | 2.00 | 0.046 | 1.63 | 0.104 | -0.301 | 0.763 | 2.85 | 0.004 |

**Supplementary Table 2**: All relationships on the two-step task with symptoms and devaluation success.

**Relationships between other symptom measures and behavior.**

The main text results report on total OCI score as a measure of compulsive symptom severity. As an alternate measure of severity that encompassed different presentations of compulsive behavior, the total number of self-report measures of compulsivity (OCI subscales, MGH Hairpulling Scale, Skin Picking Scale, and Threat-Related Reassurance Seeking scale) that each participant scored more than 1 SD above the mean on. Using this measure showed similar patterns as the total OCI score: significantly higher scores for those with unsuccessful devaluation ($t_{52}$ = 2.35, p=0.02) and no relationship with model-based planning (z = -0.799, p > 0.1).

To examine if specific compulsive symptoms may relate more to behavioral alterations than others, we assessed relationships between OCI subscores and other measures of compulsive behaviors (OCI Washing, Ordering, Checking, Thought Neutralizing; MGH Hairpulling Scale, Skin Picking Scale, and Threat Related Reassurance Scale) and devaluation success. The strongest effect was with the Washing OCI subscale ($t_{52}$ = 2.70, p=.009). This relationship did not hold when covarying for other OCI subscales (significance of beta value predicting devaluation success from logistic regression: z = -1.47), suggesting that behavioral impairments were more strongly related, but not specific, to this subscale.

**Relationship between model-based planning and devaluation success by visit and group.**

As reported in the main text, participants with lower model-based planning showed better devaluation success after active cTBS (reward*transition type*devaluation success*active TBS type: z = 2.68, p = 0.007). To understand the direction of results in this interaction, the relationship between model-based planning and devaluation success (reward*transition type*devaluation success) was investigated separately by visit and group, and then the change in the relationship between model-based planning and devaluation success with active vs. sham TBS (reward*transition type*devaluation success*active TBS) was assessed separately by group (see **Supplementary Table 2** for results for each model). Briefly, these results showed that the relationship between model-based planning and devaluation success decreased with active TBS for the cTBS group and, to a lesser extent, increased with active TBS for the iTBS group. The model specification used here (where model-based planning is measured via trial-level interactions of reward and transition type on stay behavior) does not allow for running separate models by level of model-based planning, but further investigation of individual random effect coefficients, as plotted in Figure 4C, confirmed that the reduction in the relationship between model-based planning and devaluation success was driven by lower model-based planning in participants showing successful devaluation after active cTBS.

| *Reward*transition type*devaluation success* | | | | |
|---|---|---|---|---|
| **Group** | **Visit** | **Coefficient** | **SE** | **z** |
| cTBS | Baseline | 0.240 | 0.143 | 1.68 |
| cTBS | Active | -0.039 | 0.180 | -0.217 |
| iTBS | Baseline | 0.136 | 0.131 | 1.04 |
| iTBS | Active | 0.423 | 0.161 | 2.62 |
| *Reward*transition type*devaluation success*active TBS* | | | | |
| **Group** | | **Coefficient** | **SE** | **z** |
| cTBS | | -0.360 | 0.179 | -2.02 |
| iTBS | | 0.288 | 0.185 | 1.56 |

**Supplementary Table 3. Relationship between model-based planning and devaluation success by visit and group.**

**Supplemental References**


Atlas, L. Y., Doll, B. B., Li, J., Daw, N. D., & Phelps, E. A. (2016). Instructed knowledge shapes feedback-driven aversive learning in striatum and orbitofrontal cortex, but not the amygdala. *ELife*, *5*(e15192). https://doi.org/10.7554/eLife.15192

Betancourt, M., & Girolami, M. (2015). Hamiltonian Monte Carlo for Hierarchical Models. In *Current Trends in Bayesian Methodology with Applications* (pp. 79–101).

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., … Riddell, A. (2016). Stan: A probabilistic programming language. *Journal of Statistical Software*, *20*.

Huys, Q. J., Pizzagalli, D. A., Bogdan, R., & Dayan, P. (2013). Mapping anhedonia onto reinforcement learning: A behavioural meta-analysis. *Biology of Mood & Anxiety Disorders*, *20*(1), 1–29. https://doi.org/10.1186/2045-5380-3-12

Solomon, R. L., & Wynne, L. C. (1954). Traumatic avoidance learning: The principles of anxiety conservation and partial irreversibility. *Psychological Review*, *61*(6).