Supplementary Online Content

Berwian IM, Wenzel JG, Collins AGE, et al. Computational mechanisms of effort and reward decisions in patients with depression and their association with relapse after antidepressant discontinuation. *JAMA Psychiatry.* Published online February 19, 2020. doi:10.1001/jamapsychiatry.2019.4971

This supplementary material has been provided by the authors to give readers additional information about their work.

eAppendix 1. Methods

S1.1 In- and Exclusion Criteria

Participants fulfilling the following inclusion criteria were eligible for participation in the study:

1. age 18-55 years
2. ability to consent and adhere to the study protocol
3. written informed consent
4. fluent in written and spoken German.

Patients had to additionally fulfil the following criteria:

1. currently under medical care with a psychiatrist or general practitioner for remitted Major Depressive Disorder and willing to remain in care for the duration of the study (approx. 9 months)
2. informed choice to discontinue medication (including willingness to taper the medication over at most 12 weeks) that was independent of study participation
3. clinical remission (Hamilton Depression Score of less than 7) had been achieved under therapy with antidepressants without having undergone manualized psychotherapy; with no other concurrent psychotropic medication and had been maintained for a minimum of 30 days
4. consent to information exchange between treating physician and study team members regarding inclusion/exclusion criteria and past medical history.

Any of the following exclusion criteria led to exclusion of participants. This included the following general criteria:

1. any disease of type and severity sufficient to influence the planned measurement or to interfere with the parameters of interest (this includes neurological, endocrinological, oncological comorbidities, a history of traumatic or other brain injury, neurosurgery or longer loss of consciousness.)
2. premenstrual syndrome (ICD-10 N94.3).

and criteria related to magnetic resonance imaging (MRI):

1. MRI-incompatible metal parts in the body
2. inability to sit or lie still for a longer period
3. possibility of presence of any metal fragments in the body
4. pregnancy
5. pacemaker, neurostimulator or any other head or heart implants
6. claustrophobia
7. dependence on hearing aid.

For patients the following additional criteria would led to exclusion:

1. current psychotropic medication other than antidepressants
2. questionable history of major depressive episodes without complicating factors
3. current acute suicidality
4. lifetime or current axis II diagnosis of borderline or antisocial personality disorder
5. lifetime or current psychotic disorder of any kind, bipolar disorder
6. current posttraumatic stress disorder, obsessive compulsive disorder, or eating disorder
7. current drug use disorder (with the exception of nicotine) or within the past 5 years.

Healthy controls were excluded if there was a lifetime history of DSM_IVTR axis I or axis II disorder with the exception of nicotine dependence.

S1.2 Questionnaires and Clinical Assessments

Clinical in- and exclusion criteria were assessed with the Structured Clinical Interview for DSM-IV (SCID) I and II to diagnose axis 1 disorders (major mental disorders) and axis II disorders (personality disorders), respectively[1]. The Structured Interview Guide for Hamilton Depression Rating Scale (SIGH-D)[2] consisting of 17 items was used to assess inclusion and the Inventory of Depressive Symptomatology Clinician Rated (IDS-C)[3] with 30 items to quantify residual depression. Neuropsychological assessments included the Mehrfachwahl Wortschatz Test[4], the Trial Making Test (TMT) A/B[5] and Digit Span Backwards from the Wechsler Adult Intelligence Scale[6] for verbal intelligence, cognitive processing speed, executive functions and working memory, respectively. Additionally, we applied the German version of the Response Style Questionnaire (RSQ-10D)[7] measuring brooding and reflection as components of rumination with 5 items each and the Temporal Experience of Pleasure Scale (TEPS)[8] measuring individual trait dispositions in both anticipatory and consummatory experiences of pleasure with 10 and 8 items each.

S1.3 Effort task

Unlike in the original version of the task[9] rewards were always delivered deterministically if sufficient button presses for the chosen option were emitted within 40s. Participants were seated in front of the laptop and read the instructions. They were instructed to attempt to earn as many points as possible by pumping up a balloon through button presses. The high-effort option yielded between 3 and 7 points on each trial. The order in which the high rewards were presented was pseudo-randomized, such that each was presented 12 times. If participants did not respond in time, the trial was aborted and the next trial was started. There was a fixed 1.5s inter-trial interval.

S1.4 Data Analysis

All analyses were performed using Matlab version 2016b.

S1.4.1 Standard Statistical Analysis

Trials with a response time under 0.7 secs were removed to exclude outliers and accidental fast responses (1.5% of trials in main sample) and allow computational modelling of the non-decision time (see section S1.5). For each participant, the fraction of high-effort choices, the average effort execution time, i.e. the average time between button presses, and the average decision time, i.e. the time participants take to decide between the high- and low-effort option, were calculated for all high-effort options. Participants with data more than three standard deviations from the mean for any high reward level were excluded.

To investigate group differences, mixed-design ANOVAs were used. Between-subjects independent variables were group (patients vs. HC or relapsers vs. non-relapsers) and the within-subjects independent variable was reward level of the high choice. We examined three dependent variables in separate ANOVAs, namely fraction of high-effort choices, mean effort execution time and mean decision time for each reward level. If the main or interaction term in the ANOVA was significant, we a) assessed group difference for these variables for each reward level post-hoc correcting for multiple comparison using false-discovery rate (FDR) and b) repeated the analysis for trials with high- and low-effort choices only for effort execution time and decision time.

To investigate the effect of discontinuation, we used mixed-design ANOVAs with group (discontinuation vs. no discontinuation between MA1 and MA2) as between-subjects independent variable and time of assessment (MA1 and MA2) as within-subjects independent variable. Fraction of high-effort choices, mean effort execution time and mean decision time over all reward levels were examined as dependent variables in three separate ANOVAs.

To examine whether the discontinuation effects related to relapse, we repeated the above analysis within the group of patients who discontinued and used relapsers vs. non-relapsers as between-subjects independent variable. Hence, all these analyses included between- and within-subjects factors, but no random effects.

On an exploratory level, we re-examined significant effects with linear mixed effects models and generalised linear mixed effects models including subjects as a random effects. With these analyses, we investigated if we can confirm the results of the ANOVAs.

S1.4.2 Clinical measures

Two-sample t-tests were used to examine group differences, where appropriate, in anticipatory and consummatory anhedonia, residual depression, disease severity, number of prior episodes and medication load, brooding, cognitive speed processing and executive functions, intelligence and working memory. We computed an overall measure of disease severity as the first principal component of number of past depressive episodes, age at illness onset, time in remission, time since depression onset, severity of last episode, time sick in total and time sick in the last five years as variables.

64% of our main sample took a selective serotonin reuptake inhibitor, 31% a serotonin-norepinephrine reuptake inhibtor and 5% an antidepressant from a different class. To compare medications between groups, we computed the medication load for each participant. Medication load was based on the dose prior to discontinuation divided by the maximal allowed dose according to the Swiss compendium (www.compendium.ch) and by the weight of the participant.

S1.4.3 Missing Data

We employed two approaches to examine the effect of dropouts. Firstly, we repeated all comparisons done between patients and controls between all patients who finished the study and were assessed during MA1 to those who dropped out after MA1. These included t-tests for all clinical and questionnaire measurements and parameter comparisons and mixed-design ANOVAs with finished vs. dropouts as between-subjects independent variable and reward level of the high choice as within-subjects independent variable. In separate ANOVAs, we examined three dependent variables, namely fraction of high-effort choices, mean effort execution time and mean decision time for each reward level. Furthermore, for each variable we computed two Cox proportional hazards models to assess its impact on time to relapse. In the first Cox regression we included all subjects who relapsed or dropped out after initiation of discontinuation and all subjects who finished the study. In the second Cox regression, we included only subjects who relapsed or dropped out after completion of discontinuation and subjects who finished the study.

S1.4.4 Split-Half and Test-Retest Reliability

To compute split-half reliability, we split the data from all subjects during MA1 in half and correlated the fraction of high effort choices, effort execution time and decision time from the first and second half. We also computed correlation coefficients after Spearman-Brown correction. Participants in arm 12W performed the task twice without discontinuing their medication. This allowed us to examine test-retest reliability of the behavioural outcomes. We computed the intraclass correlation (ICC) for the fraction of high effort choices, effort execution time and decision time using the data of all subjects in group 12W of the main sample who participated at MA1 and MA2.

S1.5 Computational Modelling

In order to identify the computational mechanisms underlying the behavior, we built computational models and then exposed them to series of thorough testing[10]. These were generative, i.e. they could be run on the very version of the task each individual was exposed to and generate decision and decision-time data as if a subject had performed the task. To ascertain the validity of a model, we asked whether the model itself was well-behaved by generating data from a particular set of parameters and ensuring that we could recover the true parameters through model fitting. We report the correlations between true and estimated parameters for the final model below. Next, we asked whether different models were identifiably by generating data from each model, and asking whether we could correctly identify which model had generated the data using Bayesian model comparison. Thereafter, we fitted each model to the data and generated data from the model with fitted parameters to see if it could generate patterns and variability that were similar to that observed. Finally, we performed Bayesian model comparison to penalise overly complex models that could fit any dataset simply because they have too many parameters. The goal was to identify a model with the optimal balance between parsimony and complexity. The code for the computational modelling is freely available at www.quentinhuys.com/pub/emfit.

S1.5.1 Models

We describe a series of models that capture increasing details of the behavioral data. The models are variations of drift-diffusion models (eFigure 1)[13] that capture both choice identity and the delay to the first button press. Drift-diffusion models standardly include parameters for the drift rate v, the starting point s_0, the boundary separation b and the non-decision time τ_{nd}.

We augmented an analytical approximation to these models[11] such that the difference in value between the high- and low-effort choices $V(h) - V(l)$ presented on each trial would determine the drift rate v in the drift-diffusion model. A large value for the high-effort option would result in a higher positive drift towards the high boundary, and a large value for the low-effort in a negative drift towards the low boundary. Choices in the model would be emitted when the process reached the boundary.

In the most basic 'constant' model, the reward value of options were disregarded, allowing only a constant bias θ to capture the effect of the differential effort across all trials, setting $V(l) = 0$ and $V(h) = \theta$, where l was the low effort option requiring 20 button presses, and h the high-effort option requiring 100 button presses.

In the 'scaling' model, the value of each choice a traded off the effort (the number of required button presses $e(a)$) and the reward $r(a)$ for that option, each scaled by individual sensitivity parameters:

$$V(a) = \beta_{rew} * r(a) - \beta_{eff} * e(a) \qquad 1$$

The 'deviation' model took some probability mass from the high-effort choice decision-time distribution $p(h,\tau)$ and added it to the low-effort choice decision-time distribution $p(l,\tau)$. Specifically, on trials where $r(h) < 5$, we let:

$$p(h,\tau) = p(h,\tau) * (1 - p_{switch}) \qquad 2$$

$$p(l,\tau) = p(l,\tau) + p(h,\tau) * p_{switch} \qquad 3$$

where τ is the decision time.

In order to capture progressively faster decision times over the course of the experiment, we also examined the inclusion of a 'boundary scaling' parameter that linearly reduced the boundary on each trial $b(t) = b\text{-}\beta_{scale} *t$. Finally, we examined fixing the starting point half-way between the boundaries, i.e. letting $s_0 = b/2$.

We compared parameter estimates between groups (patients vs. healthy controls and relapsers vs. non-relapsers). As the size of the actual boundary was determined by both the starting boundary b and the boundary scaling β_{scale} and the starting boundary b interacted with the starting point s_0 to determine the size of the low and high boundary, we computed the size of the average boundary and the low boundary using parameter estimates of the starting boundary s_0, the boundary scaling β_{scale} and the starting point s_0. Hence, we conducted a re-parameterisation such that the average boundary was computed as $b_{av} = (b - \beta_{scale} * (b/60) * 30))$ and the average low boundary as $b_{low} = b_{av} * (1 - s_0)$. Of note, the code was setup such that the boundary to the high choice was at the bottom and the boundary to the low choice at the top. Parameters were transformed using exponential or sigmoid transformations to respect natural boundaries in models.

S1.5.2 Model fitting

Models were formulated using an analytical approximation to the first-passage-time probability distribution of the drift-diffusion model (DDM)[**11**]. Parameters were estimated using an empirical hierarchical Bayesian procedure described in detail in the section "Model fitting procedure" by Huys and colleagues[**12**]. Priors are assumed to have a Gaussian distribution which serves to regularise the inference and prevent parameters from taking on extreme values in case they are not well-constrained. The parameters of the prior distribution θ are set to the maximum likelihood of all trials of all N subjects

$$\hat{\theta} = argmax(\prod_{i=1}^{N} \int d^N \, h_i \, p(A_i|h_i)p(h_i|\theta) \,) \qquad 4$$

where h is the parameter vector specified by each model for each subject and A_i are all actions for the i^{th} subject, which are assumed to be independent. In this analysis, all main sample data from main assessment 1 (MA1) were jointly fitted, i.e. with the prior assumption of no differences between any groups.

The maximisation of θ involves an iterative update procedure using an expectation-maximisation algorithm. Using a Laplacian approximation the E-step at the k^{th} iteration is computed as:

$$p(h|A_i) \approx \mathcal{N}(m_i^{(k)}, \Sigma_i^{(k)}) \qquad 5$$

$$m_i^{(k)} = argmax_h \, p(A_i|h)p(h|\theta^{(k-1)}) \qquad 6$$

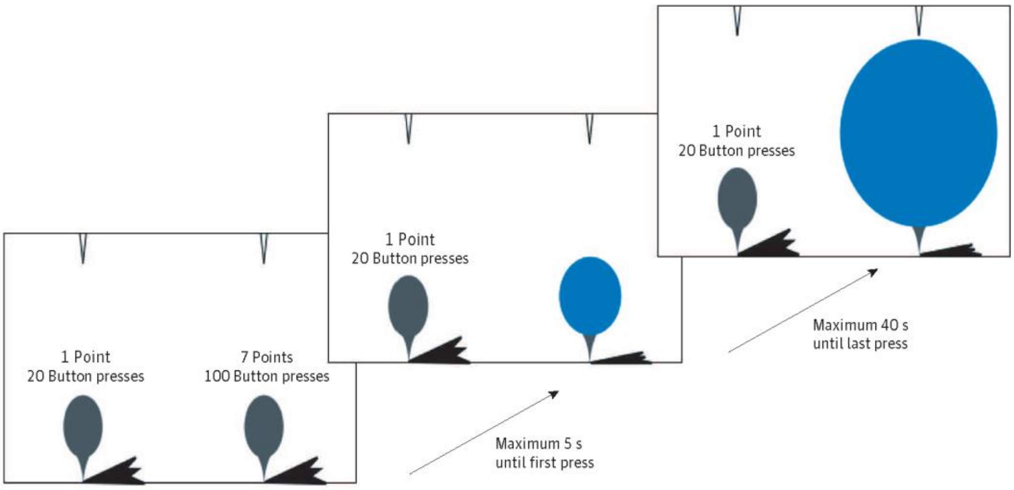
where $\mathcal{N}(\cdot)$ denotes a normal distribution over h with mean $m_i^{(k)}$ and variance $\Sigma_i^{(k)}$. The hyperparameters θ are estimated computing the mean μ and the variance ν^2 in the following way:

$$\mu^{(k)} = \frac{1}{N}\sum_i m_i^{(k)} \qquad 7$$

$$\left(\nu^{(k)}\right)^2 = \frac{1}{N}\sum_i \left[\left(m_i^{(k)}\right)^2 + \Sigma_i^{(k)}\right] - \left(\mu^{(k)}\right)^2 \qquad 8$$

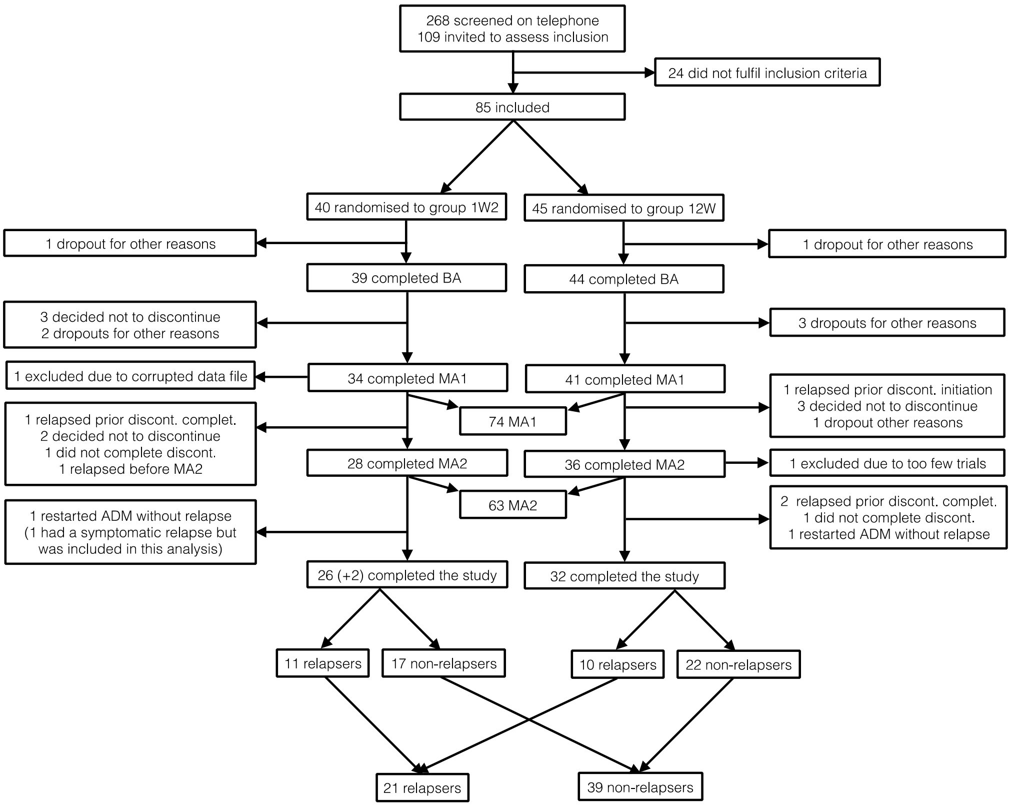
To implement this method, we needed the analytical derivatives of the likelihood with respect to all the parameters. To achieve this, we had to introduce a cut-off, exclude trials with a decision time faster than the cut-off, and restrict the τ_{nd} not to exceed the cut-off. This was set at 700ms based on visual inspection of the data, length of non- decision times presented in the literature and to avoid excluding too many trials.

eFigure 1. Physical Effort Task



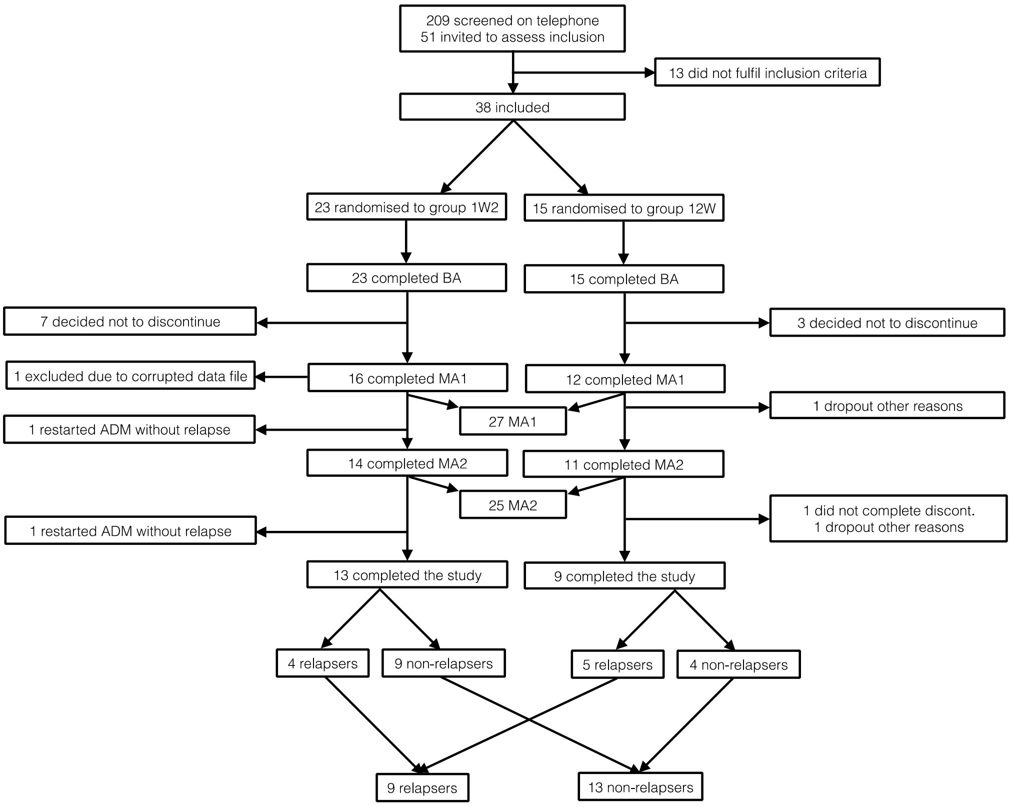
Depicted is a trial in the physical effort task. On each of 60 trials, participants had 5 seconds to decide between 20 button presses for 1 point or 100 button presses for a higher amount of points (3-7 points). Index fingers of the left and right hand were used on keys “j” and “k” for the option presented on the left and right side, respectively. Participants had 40 seconds to pump up the respective balloon, until it exploded when the required number of button presses was reached. Switching after the first button press was not possible. This task was adapted from with permission from Gold et al [13].

eFigure 2. CONSORT Diagrams of the Main Sample


268 screened on telephone
109 invited to assess inclusion
24 did not fulfil inclusion criteria
85 included
40 randomised to group 1W2
45 randomised to group 12W
1 dropout for other reasons
1 dropout for other reasons
39 completed BA
44 completed BA
3 decided not to discontinue
2 dropouts for other reasons
3 dropouts for other reasons
1 excluded due to corrupted data file
34 completed MA1
41 completed MA1
74 MA1
1 relapsed prior discont. initiation
3 decided not to discontinue
1 dropout other reasons
1 relapsed prior discont. complet.
2 decided not to discontinue
1 did not complete discont.
1 relapsed before MA2
28 completed MA2
36 completed MA2
1 excluded due to too few trials
63 MA2
2 relapsed prior discont. complet.
1 did not complete discont.
1 restarted ADM without relapse
1 restarted ADM without relapse
(1 had a symptomatic relapse but
was included in this analysis)
26 (+2) completed the study
32 completed the study
11 relapsers
17 non-relapsers
10 relapsers
22 non-relapsers
21 relapsers
39 non-relapsers

Depicted are reasons for dropouts and exclusion for the main sample. Group 1W2 withdraw their antidepressants between MA1 and MA2 and group 12W underwent both MAs before withdrawal. BA = baseline assessment, MA1= main assessment 1, MA2 = main assessment 2, ADM = antidepressant medication

eFigure 3. CONSORT Diagrams of the Replication Sample


209 screened on telephone
51 invited to assess inclusion
13 did not fulfil inclusion criteria
38 included
23 randomised to group 1W2
15 randomised to group 12W
23 completed BA
15 completed BA
7 decided not to discontinue
3 decided not to discontinue
1 excluded due to corrupted data file
16 completed MA1
12 completed MA1
27 MA1
1 dropout other reasons
1 restarted ADM without relapse
14 completed MA2
11 completed MA2
25 MA2
1 did not complete discont.
1 dropout other reasons
1 restarted ADM without relapse
13 completed the study
9 completed the study
4 relapsers
9 non-relapsers
5 relapsers
4 non-relapsers
9 relapsers
13 non-relapsers

Depicted are reasons for dropouts and exclusion in the replication sample. Group 1W2 withdraw their antidepressants between MA1 and MA2 and group 12W underwent both MAs before withdrawal. BA = baseline assessment, MA1= main assessment 1, MA2 = main assessment 2, ADM = antidepressant medication

eAppendix 2. Results

S2.1 Participants

The CONSORT diagram in eFigures 2 and 3 show reasons for exclusion and dropouts for patients in the course of the study for the main sample and the replication sample, respectively. Of the 74 patients in the main sample, 59 finished the study according to protocol, of which 20 patients had a relapse. We categorized one additional patient as relapser, who fulfilled DSM-IV criteria for a relapse for 10 days, but quickly improved after antidepressant medication re-initiation after 10 days. Of the 27 patients in the replication sample, 9 finished the study with a relapse and 13 without a relapse.

In the main sample, 40 healthy controls (HC) were included in the study. 4 HC dropped out of the study due to a lack of time or interested to continue. Data from one HC was not available and one HC was excluded due to non-adherence to study instructions, leading to final sample of 34 HC. In the replication sample, 26 HC were included, 4 dropped out of the study and the data file of one HC was not available. Hence, data from 21 HC was used for the analyses.

S2.1.1 Replication sample participant characteristics

eTable 1. Participant Characteristics for the Replication Sample from Berlin

	Patients vs. HC			**Relapsers vs. Non-relapsers**		
Variable[a]	Patients (n = 27)	HC (n = 21)	P value	Relapsers (n = 9)	Non-relapsers (n = 13)	P Value
Demographics						
Age	36.8 (12.1)	35.6 (11.5)	0.73	41.1 (13.5)	34.5 (11.7)	0.24
Male sex, No. (%)	5 (19)	4 (19)	0.18	2 (22)	2 (15)	0.95
Neuropsychology						
Intelligence[b]	29.7 (3.7)	28.0 (3.7)	0.12	31.4 (3.3)	28.8 (3.6)	0.09
Working memory[b]	6.9 (1.7)	8.6 (2.4)	0.008	7.0 (1.4)	7.3 (1.7)	0.65
Cognitive processing speed[b]	24.6 (10.5)	22.0 (6.0)	0.31	26.3 (8.6)	21.0 (7.4)	0.14
Executive functions[b]	51.5 (17.4)	53.6 (14.4)	0.66	59.8 (22.2)	45.3 (14.7)	0.08
Clinical measures						
Number of prior episodes	NA	NA	NA	2.7 (1.4)	1.4 (0.5)	0.01
Residual depression[b]	4.96 (5.4)	0.76 (1.3)	<0.001	7.9 (7.5)	2.3 (2.3)	0.02
Disease severity[c]	NA	NA	NA	0.04 (0.55)	-0.12 (0.19)	0.34
Medication load [c]	NA	NA	NA	0.007 (0.005)	0.007 (0.003)	0.72
Variables of interest						
Anticipatory pleasure[b]	3.9 (0.4)	4.3 (0.5)	0.004	3.7 (0.6)	4.5 (0.7)	0.18
Consummatory pleasure[b]	4.7 (0.6)	4.9 (0.5)	0.15	4.5 (0.7)	4.8 (0.7)	0.45
Brooding[b]	9.5 (2.8)	7.8 (2.6)	0.04	8.9 (2.3)	8.8 (2.3)	0.91

a) Unless stated otherwise, mean (SD) are shown; b) Determined as follows: intelligence: Mehrfachwahl Wortschatz Test[4]; working memory: digit span backwards test from the Wechsler Adult Intelligence Scale[6]; cognitive processing speed: Trail Making Test A[5]; executive processing speed: Trail Making Test B[5]; residual depression: Inventory of Depressive Symptomatology-Clinician Rated[3]; anticipatory pleasure: subscale of anticipatory pleasure of the Temporal Experience of Pleasure Scale[8]; consummatory pleasure: subscale of consummatory pleasure of the Temporal Experience of Pleasure Scale[8]; brooding: brooding subscale of the German version of the Response Style Questionnaire[7]; c) Computation of the variables is described in the supplement S1.5.2; HC = healthy controls; NA = not applicable

Demographic, neuropsychology, clinical measures and covariates of interest for the replication sample are shown in eTable 1. HC had a better working memory capacity, reported enhanced capacity to anticipate pleasure and reported less brooding rumination. As in the main sample brooding did not correlate with decision time (p=0.50).

Relapsers had more prior episodes and more residual symptoms than non-relapsers. These findings are as anticipated and shown for relapse risk in general. This was not seen in the larger main sample (Table 1) where instead relapsers had less residual symptoms than non-relapsers.

S2.1.2 Dropout comparisons and intention-to-treat analyses

As listed in eTable 2, patients who dropped out after MA1 compared to patients who completed the study as either relapsers or non-relapsers did not differ on any measure included in the present analyses. Furthermore, Cox proportional hazard models showed that time to relapse after antidepressant reduction and after completion of discontinuation was significantly predicted by the same variables that differed between relapsers and non-relapsers, namely decision time and lower boundary.

eTable 2. Dropout Comparisons and Cox Regressions

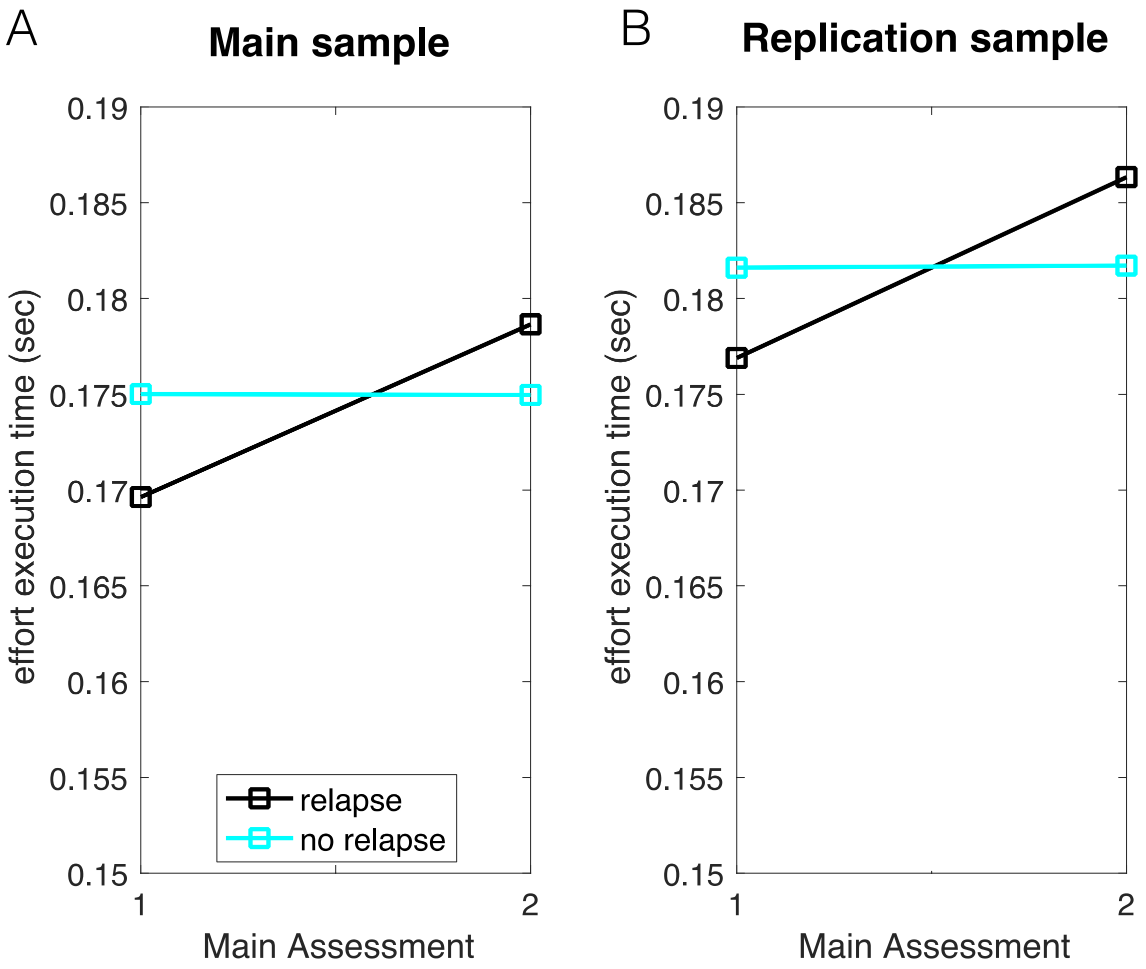
	Study completers (n=60) vs. Dropouts (n=14) at MA1	**CR from start of ADM reduction**	**CR from end of ADM reduction**
Variable	P value	P value	P value
Demographics			
Age	0.52	0.58	0.48
Sex	0.88	0.07	0.82
Neuropsychology			
Intelligence[a]	0.92	0.69	0.35
Working memory[a]	0.85	0.39	0.44
Cognitive processing speed[a]	0.05	0.09	0.92
Executive functions[a]	0.98	0.33	0.35
Clinical measures			
Number of prior episodes	0.49	0.83	0.72
Residual depression[a]	0.13	0.06	0.06
Disease severity[b]	0.63	0.58	0.49
Medication load[b]	0.30	0.91	0.41
Covariates of interest			
Anticipatory pleasure[a]	0.16	0.58	0.97
Consummatory pleasure[a]	0.37	0.30	0.38
Brooding[a]	0.60	0.09	0.03
Behavioural measures			
Probability of high choice	0.40/0.99[c]	0.19	0.58
Decision time	1.00/0.82[c]	0.02	0.008
Effort execution time	0.41/0.05[c]	0.68	0.86
Model parameter comparisons			
Reward sensitivity	0.68	0.76	0.96
Effort sensitivity	0.85	0.52	0.71
Non-decision time	0.23	0.74	0.47
Probability of switching	0.32	0.22	0.25
Starting boundary	0.40	0.10	0.08
Average boundary	0.27	0.07	0.04
Lower boundary	0.23	0.01	0.01

a) Determined as follows: intelligence: Mehrfachwahl Wortschatz Test[4]; working memory: digit span backwards test from the Wechsler Adult Intelligence Scale[6]; cognitive processing speed: Trail Making Test A[5]; executive processing speed: Trail Making Test B[5]; residual depression: Inventory of Depressive Symptomatology-Clinician Rated[3]; anticipatory pleasure: subscale of anticipatory pleasure of the Temporal Experience of Pleasure Scale[8]; consummatory pleasure: subscale of consummatory pleasure of the Temporal Experience of Pleasure Scale[8]; brooding: brooding subscale of the German version of the Response Style Questionnaire[7]; b) Computation of the variables is described in the supplement S1.4.2; c) main effect/interaction effect; MA1 = main assessment one; CR = Cox regression; ADM = antidepressant medication;

S2.2 Vigor

eFigure 4A shows how effort execution time changes due to discontinuation in patients who go on to relapse and patients who remain well. As mentioned in the main text, the interaction between relapse and discontinuation does not reach significance according to our preset significance level of α=0.017 (F(1,25)=3.914, p=0.06)). Relapsers and non-relapsers did not differ before discontinuation (t(25)=-0.521, p=0.607). On an exploratory level, paired t-tests indicated that patients who would go on to relapse had longer effort execution times after discontinuation (MA1: 0.170(0.019), MA2: 1.79(0.019), t(9)=-2.753, CI:-0.016- -0.002, p=0.02, FDR-corrected p=0.09), i.e. less vigor during effort execution, whereas no change occurred in patients who discontinued and would remain well (MA1: 0.1750(0.029), MA2: 1.750(0.020), t(16)=0.014, CI:-0.006-0.006, p=1.00, FDR-corrected p=1.00). After discontinuation, the group comparison, however, did not reach significance (t(25)=0.471, p = 0.64). The direction of the effects can be replicated in our second sample numerically (relapser: MA1: 0.177(0.009) MA2: 0.186(0.020); non-relapser: MA1: 0.182(0.016), MA2: 0.182(0.019)), but none of the effects reaches significance (eFigure 4B).

eFigure 4. Relapse Discontinuation Interaction Effect for Effort Execution Time



Depicted is the change in effort execution time between main assessment 1 and main assessment 2 for patients who discontinued separated into patients who relapsed and patients who did not relapse during the follow-up period for the main sample (A) and the replication sample (B).

S2.3 Analysis Comparisons

Applying linear mixed effect models to re-examine significant result from the mixed-design ANOVAs, showed that the pattern of results is similar with both analyses approaches (eTable 3), although the exact values, naturally, differ. Of note, linear mixed effect models were not more sensitive but for the interaction effect of relapse and discontinuation on effort execution time which reaches our pre-set significance level of α=0.017 according to the linear mixed effect models analysis.

eTable 3. ANOVA and LMEM Comparisons

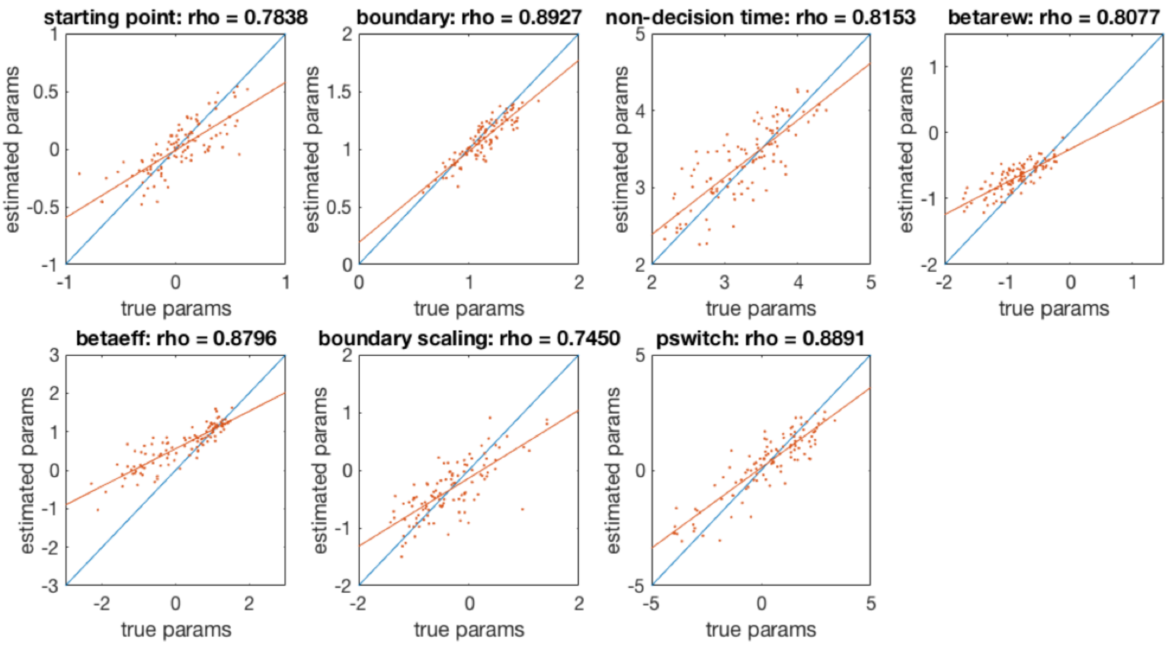
	ANOVAs	LMEM
Effect	P value	P value
Interaction of depression and reward on probability of high choice	<0.001	0.04
Main effect of depression on decision time	0.02	0.03
Main effect of relapse on decision time	<0.001	0.003
Interaction of relapse and discontinuation on effort execution time	0.06	0.002

For all effects found to be significant in an analysis of variance (ANOVA), we re-examined the effect using (generalised) linear mixed effect models (LMEM).

S2.4 Model and Parameter Recovery

For all six models, the model that generated the data was also the most parsimonious model in formal model comparison, i.e. all models were identifiable (eTable 4). Parameters in the final model were also identifiable, as shown by the correlations between true and estimated parameters for the winning model (eFigure 5). As evident from visual inspection of eFigure 6 the most complex model including individually fitted parameters for reward sensitivity (β_{rew}) and effort sensitivity (β_{eff}), which scale the reward and effort options on a trial, hence termed scaling, as well as the switching probability (p_{switch}) of the deviation process, the starting point (s_0), the non-decision time (τ_{nd}), the starting boundary (*b*) and a linear boundary scaling over trials (β_{scale}) can replicate the behavior of all groups best. This model was confirmed to be the most parsimonious model in the main sample (eFigure 7) and the replication sample (eFigure 8). Furthermore, this model also had the best fit for 88 % of the subjects in the main sample (eTable 5).

eFigure 5. Parameter Recovery



For our most parsimonious model including all 7 parameters we used the parameters identified from our data to generate new data and estimated the parameters again. Orange dots show the estimated parameter for each true parameter, i.e. the parameter used to generate the data. The orange line represents a regression line fitted to the scatter values. The correlation for each parameter are depicted above the according panel. Blue identity lines are for visual orientation only.

eTable 4. Model Recovery

generating/recovering model	S+NDT+B+SP+BS+Pswitch	S+NDT+B+SP+BS	S+NDT+B+SP+Pswitch	S+NDT+B+SP	S+NDT+B	C+NDT+B+SP
S+NDT+B+SP+BS+Pswitch	0	318.6	277.3	611.9	641.7	3771.1
S+NDT+B+SP+BS	22.2	0	526.7	507.4	565.5	3659.4
S+NDT+B+SP+Pswitch	45.1	258.2	0	234.8	287.8	3449.4
S+NDT+B+SP	51	45.6	37.1	0	53	3392.4
S+NDT+B	59.9	41.5	22.4	6	0	3215.6
C+NDT+B+SP	223.3	198.9	192.4	179.7	284.2	0

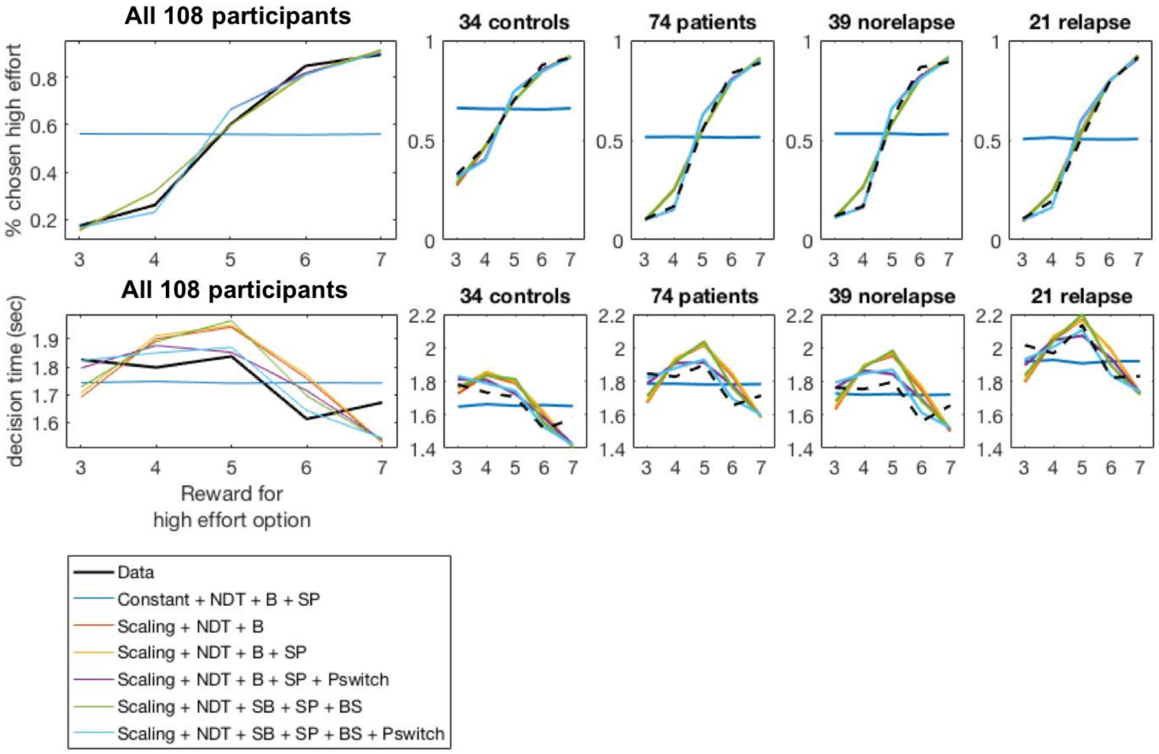
For each model listed on the left, we generated new data using the estimated parameters of our sample. Values in the corresponding row indicate the integrated Baysian Information Criterion (iBIC) for all models when fitted to the generated data. Zeros indicate the smallest iBIC and, therewith, the most parsimonious model. Constant and scaling indicate that the drift-rate was determined as in the according model described in supplement S1.5. S = scaling; C = constant; NDT = non-decision time; B = boundary; SP = starting point; BS = boundary scaling

eTable 5. Winning Models

model	S+NDT+B+SP+BS+Pswitch	S+NDT+B+SP+BS	S+NDT+B+SP+Pswitch	S+NDT+B+SP	S+NDT+B	C+NDT+B+SP
number of subjects	95	5	5	0	0	3

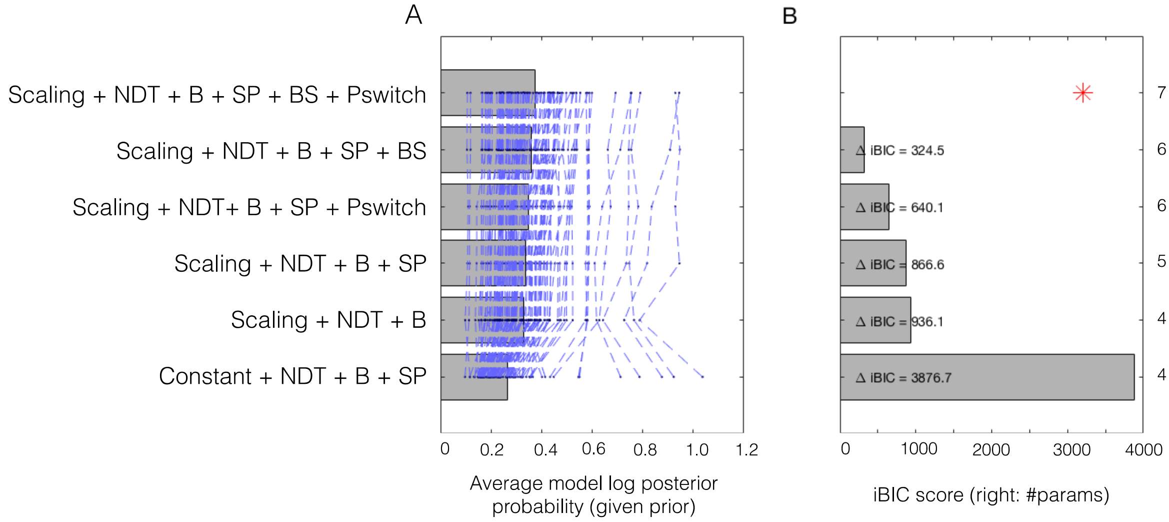
For each subject, we examined which model was the most parsimonious model. We report for each model how many subjects were best fitted by that model. Constant and scaling indicate that the drift-rate was determined as in the according model described in supplement S1.5. S = scaling; C = constant; NDT = non-decision time; B = boundary; SP = starting point; BS = boundary scaling

eFigure 6. Model Fits



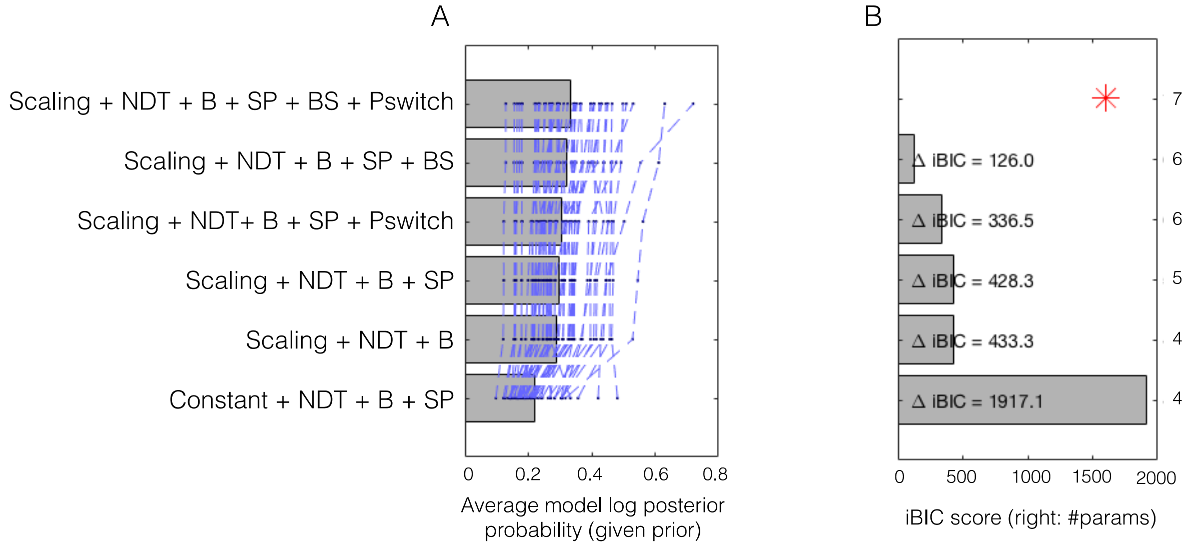
The actual data is plotted alongside the data generated from all our models for all participants and separately for the four participant groups. At the top, the probability of choosing the high-effort option depending on the magnitude of the high-reward option is plotted and at the bottom, the decision time, again depending on the magnitude of the high-reward option. Constant and scaling indicate that the drift-rate was determined as in the according model described in supplement S1.5. In the legend, all additional parameters included in each model are listed. NDT = non-decision time; B = boundary; SP = starting point; BS = boundary scaling

eFigure 7. Model Comparison in the Main Sample



Better models have lower iBIC scores. A) shows the average model log posterior probability given the prior. Dashed blue lines show these values for all individual subjects. B) The red star indicates the most parsimonious model with the smallest integrated Bayesian Information Criterion (iBIC). Constant and scaling indicate that the drift-rate was determined as in the according model described in supplement S1.5. All additional parameters included in each model are listed on the left of the model. NDT = non-decision time; B = boundary; SP = starting point; BS = boundary scaling

eFigure 8. Model Comparison in the Replication Sample



eFigure 8: Model comparison in the replication sample. A) shows the average model log posterior probability given the prior. Dashed blue lines show these values for all individual subjects. B) The red star indicates the most parsimonious model with the smallest integrated Bayesian Information Criterion (iBIC).Constant and scaling indicate that the drift-rate was determined as in the according model described in supplement S1.5. All additional parameters included in each model are listed on the left of each model. NDT = non-decision time; B = boundary; SP = starting point; BS = boundary scaling

S2.5 Split-Half and Test-Retest Reliability

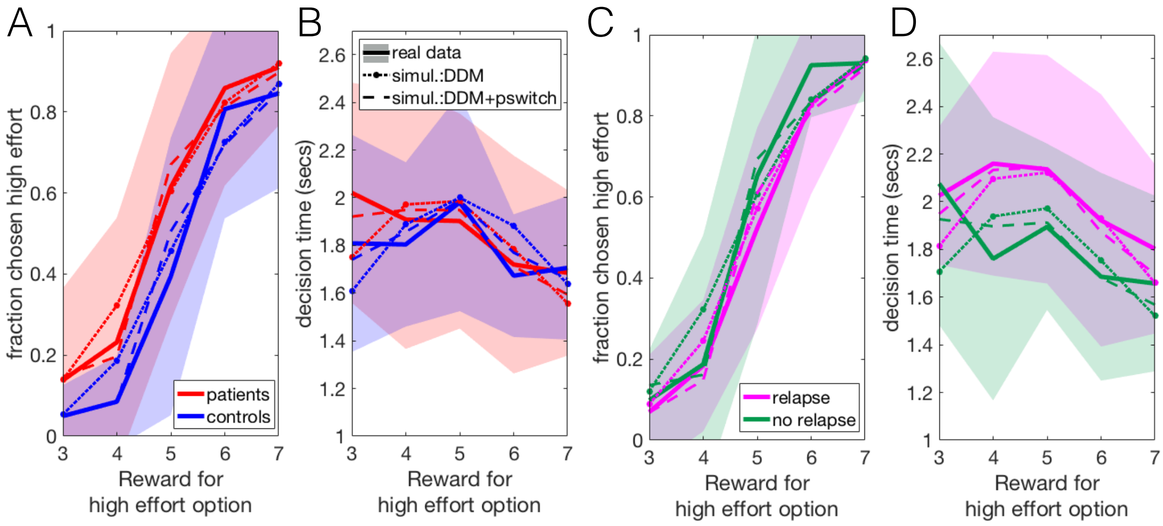
Correlation coefficients and significance levels for all behavioural variables are listed in eTable 6. As can be inferred from the coefficients listed in the table, split-half reliability is high and test-retest reliability is moderate in our sample. One potential reason why test-retest reliability is not higher is that we did not include practice trials. Hence, participants changed their behavior due to learning the options of the tasks over trials when they did the task for the first time, and only did less so, when they did the task for the second time. Thus, we would recommend practice trials in future versions of this task.

eTable 6. Split-Half and Test-Retest Reliability

	Split-half reliability			Test-retest reliability	
Measure	r	P value	r[a]	r	P value
Probability of high choice	0.74	<0.001	0.85	0.41	0.008
Effort execution time	0.89	<0.001	0.94	0.61	<0.001
Decision time	0.68	<0.001	0.81	0.49	0.01

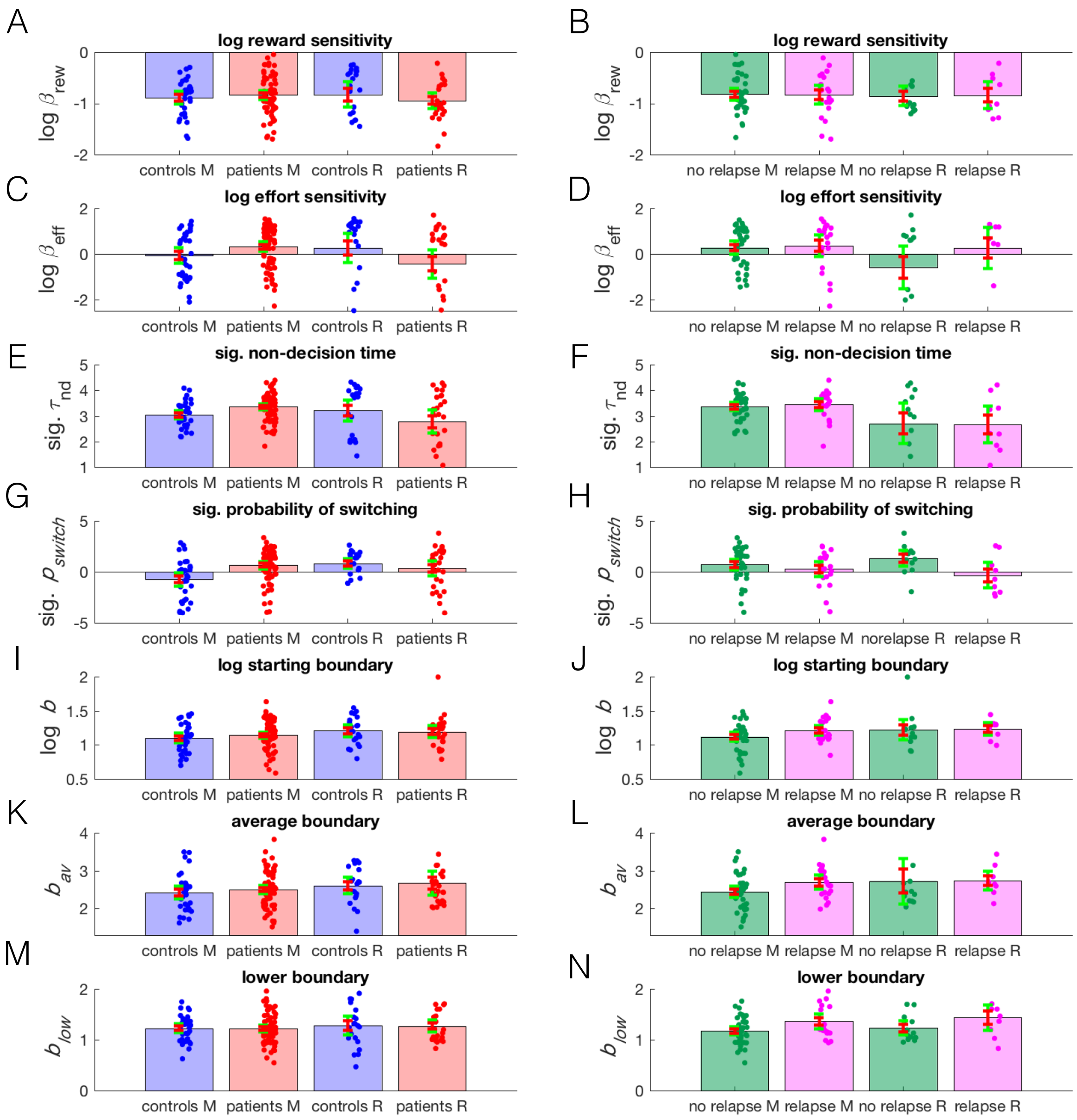
a) listed are correlation coefficients after Spearman-Brown correction

eFigure 9. Raw Behavioral Data and Model Fits in the Replication Sample



Fraction of high-effort choices (A,C) and times to first button press (B,D) as a function of reward offered for the high-effort choice comparing patients vs. controls (A,B) and relapsers vs. non-relapsers (C,D). Solid lines show group means in the raw data, surrounding shades the according standard deviations of the raw data. Black stars indicate significant post-hoc tests corrected for false-discovery rate for the individual reward levels. Dotted and dashed lines show means of the surrogate data generated from models in all panels. The standard drift diffusion model (dotted lines) forces fast decisions to accompany deterministic behavior, and hence a prominent inverted U-shape dependence of decision-times on reward levels (B,D). Inclusion of the deviation process allows the deterministic decisions to be accompanied by longer decision-times (dashed lines).

eFigure 10. Parameter Comparisons



Depicted are all parameter comparisons for the main sample (M) and the replication sample (R) next to each other. A, C ,E ,G ,I ,K ,M) show comparisons of data from healthy controls and patients and B, D, F, H, J, L, N) show comparisons of patients who remained well and patients who went on to relapse. Bars indicate the group mean, dots indicate the data points of the individual participants, red error bars show standard errors and green error bars 95% confidence intervals. Meaning and computation of the parameters is explained in detail in Supplementary Section S1.5. Titles of subpanels spell out the parameter names which label the y-axis. "Log" indicates that a parameter underwent an exponential transformation and "sig." indicates that a parameter underwent a sigmoidal transformation to ensure natural boundaries.

eTable 7. Behavioral Effects for Fraction of Effortful Choices and Decision Time in the Replication Sample

	Patients vs. HC			**Relapsers vs. Non-relapsers**		
Variable (mean (SD))	Patients (n = 27)	HC (n = 21)	P Value[a]	Relapsers (n = 9)	Non-relapsers (n = 13)	P Value[a]
Probability of high choice	.553 (.403)	.436 (.409)	.009[b]/.12[c]	.509 (.387)	.559 (.424)	.43[b]/.69[c]
separate for each reward level						
3	.109 (.149)	.050 (.076)	.14	.069 (.133)	.099 (.120)	.98
4	.207 (.270)	.085 (.109)	.14	.184 (.153)	.187 (.306)	.98
5	.627 (.307)	.394 (.334)	.10	.523 (.234)	.652 (.353)	.93
6	.887 (.170)	.807 (.263)	.23	.832 (.212)	.925 (.123)	.93
7	.934 (.077)	.846 (.228)	.14	.937 (.065)	.930 (.090)	.98
Decision time	1.774 (.357)	1.795 (.380)	.82[b]/.04[c]	2.009 (.435)	1.815 (.490)	.26[b]/.14[c]
separate for each reward level						
3	1.935 (.306)	1.809 (.445)	.68	2.025 (.273)	2.076 (.569)	.82
4	1.789 (.333)	1.804 (.336)	.88	2.160 (.443)	1.761 (.570)	.44
5	1.826 (.352)	1.981 (.445)	.68	1.136 (.452)	1.895 (.335)	.44
6	1.655 (.361)	1.673 (.251)	.88	1.922 (.499)	1.686 (.419)	.44
7	1.666 (.347)	1.251 (.294)	.88	1.802 (.334)	1.657 (.354)	.46

a) Unless stated otherwise, P Values are FDR-corrected posthoc tests; b) P Values of main effect; c) P Values of interaction effect. HC = healthy controls

eReferences

[1] Wittchen HU, Fydrich T. *Strukturiertes klinisches Interview für DSM-IV. Manual zum SKID-I und SKID-II*. Göttingen, DE: Hofgrefe; 1997.

[2] Williams JB. A structured interview guide for the hamilton depression rating scale. *Arch Gen Psychiatry*. 1988; 45(8):742–7.

[3] Rush AJ, Gullion CM, Basco MR, Jarrett RB, Trivedi MH. The inventory of depressive symptomatology (IDS): psychometric properties. *Psychol Med*. 1996;26(3):477–86.

[4] Lehr S. *Mehrfachwahl-Wortschatz-Intelligenztest MWT-B*. Balingen, DE: Spitta; 2005.

[5] Reitan RM. Validity of the trial making test as an indicator of organic brain damage. *Percept Mot Ski*. 1958;8:271–276. doi:10.2466/pms.1958.8.3.271.

[6] Wechsler D. *Wechsler Adult Intelligence Scale - Fourth eEdition (WAIS-IV)*. San Antonio, Texas: Psychological Corporation; 2014.

[7] Huffziger S, Kühner C. Die Ruminationsfacetten Brooding und Reflection: Eine psychometrische Evaluation der deutschsprachigen Version RSQ-10D. *Z Klin Psychol und Psychother*. 2012;41(1):38–46. doi:10.1026/1616-3443/ a000118.

[8] Gard DE, Germans Gard M, Kring AM, John OP. Anticipatory and consummatory components of the experience of pleasure: A scale development study. *J Res Pers*. 2006;40:1086–1102. doi:10.1016/j.jrp.2005.11.001.

[9] Gold JM, Strauss GP, Waltz JA, Robinson BM, Brown JK, Frank MJ. Negative symptoms of schizophrenia are associated with abnormal effort-cost computations. *Biol Psychiatry*. 2013;74(2):130–6. doi:10.1016/j.biopsych.2012.12. 022.

[10] Huys QJM. Bayesian approaches to learning and decision-making. In A Anticevic, J Murray, eds., *Computational Psychiatry: Mathematical Modelling of Mental Illness*. Cambridge, MA: Academic Press, 2017; 247–271.

[11] Navarro DJ, Fuss IG. Fast and accurate calculations for first-passage times in wiener diffusion models. *J Math Psychol*. 2009;53:222–230.

[12] Huys QJM, Cools R, Gölzer M, et al. Disentangling the roles of approach, activation and valence in instrumental and pavlovian responding. *PLoS Comput Biol*. 2011;7(4):e1002028. doi:10.1371/journal.pcbi.1002028.

[13] Gold JM, Strauss GP, Waltz JA, Robinson BM, Brown JK, Frank MJ. Negative symptoms of schizophrenia are associated with abnormal effort-cost computations. *Biol Psychiatry*. 2013;74(2):130-136. doi:10.1016/j.biopsych.2012.12.022