nature human behaviour

Article

https://doi.org/10.1038/s41562-023-01519-7

A highly replicable decline in mood during rest and simple tasks

In the format provided by the authors and unedited

Supplementary Materials

Contents

List of Figures

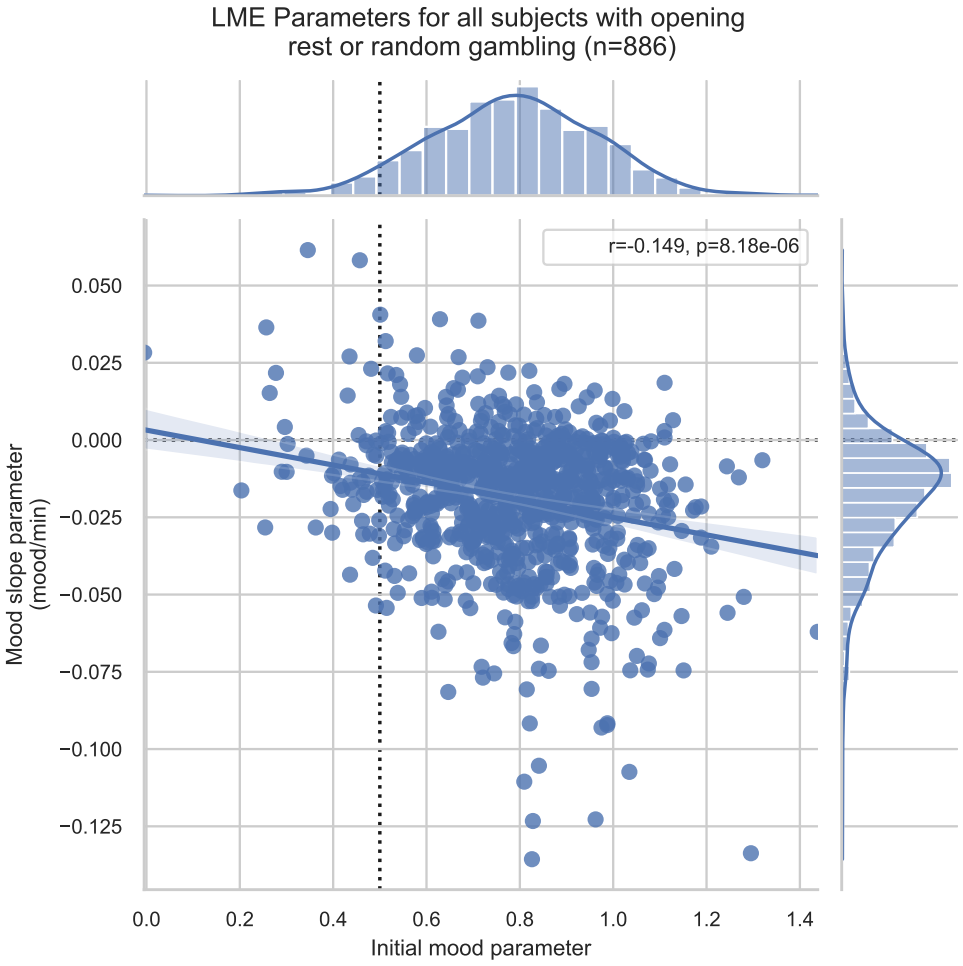
List of Tables

A. Cohorts

A list and summary of the cohorts used in this study can be found in Supplementary Table 1.

B. Linear Mixed Effects Model

A large-scale linear mixed effects (LME) model was used to quantify the Mood Drift Over Time ("mood drift" for short) observed in the online participants. The model is discussed in the Methods section, and many results are described in the Results section. Additional results are included below.



Supplementary Figure 1: Joint plot of LME slope and intercept parameters for all online participants receiving opening rest periods. The line represents the best linear fit, and the error patch represents a 95% confidence interval for that fit. The r and p in the legend refer to a Spearman correlation (2-sided, not corrected for multiple comparisons).

Mood Drift Over Time's Uncertain Relationship to Age

Our large-scale LME model reported that participants with ages 16-18 had a significantly lower initial mood ($-8.8 \pm 2.8\% mood$, $t_{879} = -3.1, p = 0.002$) and higher slope ($0.9 \pm 0.4\% mood/min$, $t_{898} = 2.31, p = 0.021$, both 2-sided) than those with ages 18-40. No other age group had significant differences in these parameters. The

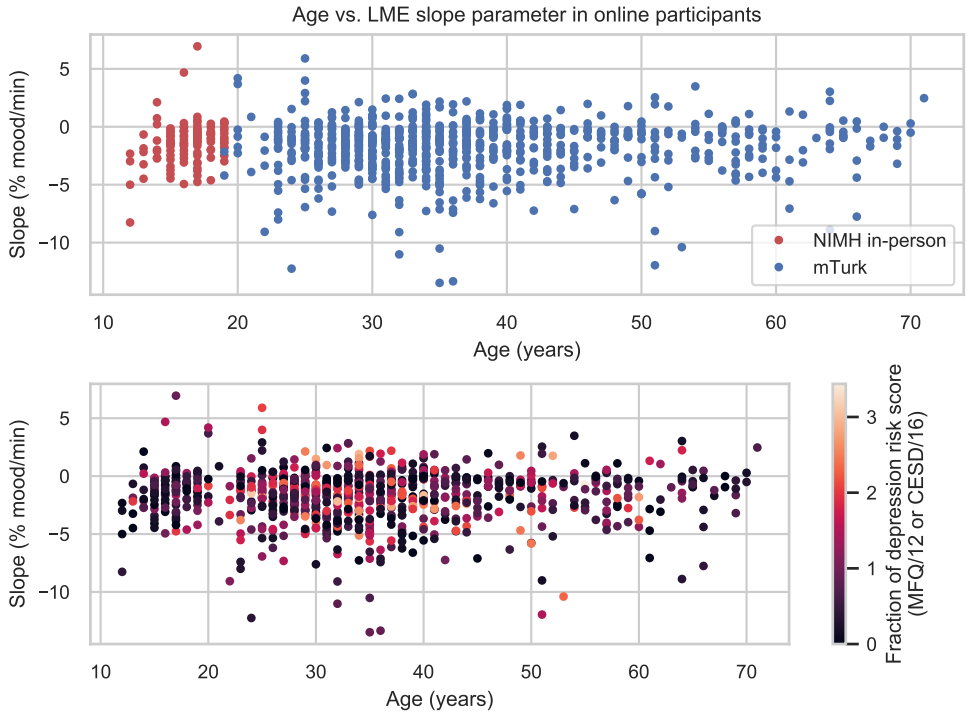
Opening Rest Cohort	nParticipants	Block 0	Block 1	Block 2	Block 3
15sRestBetween	40	rest15 * 30	closed+ * 54		
30sRestBetween	37	rest30 * 18	closed+ * 54		
7.5sRestBetween	38	rest7.5 * 45	closed+ * 54		
60sRestBetween	39	rest60 * 10	closed+ * 54		
AlternateRating	32	rest15 * 30	closed+ * 54		
Expectation-7mRest	64	rest15 * 18	random * 22	closed- * 22	closed+ * 22
Expectation-12mRest	67	rest15 * 18	random * 22	closed- * 22	closed+ * 22
RestDownUp	58	rest15 * 18	closed- * 33	closed+ * 33	
Daily-Rest-01	66	rest15 * 18	closed+ * 18	rest15 * 18	closed+ * 18
Daily-Rest-02	53	rest15 * 18	closed+ * 18	rest15 * 18	closed+ * 18
Weekly-Rest-01	196	rest15 * 18	closed+ * 22	closed- * 22	closed+ * 22
Weekly-Rest-02	164	rest15 * 18	open+ * 22	open- * 22	open+ * 22
Weekly-Rest-03	160	rest15 * 18	open+ * 22	open- * 22	open+ * 22
Adolescent-01	116	rest15 * 18	closed+ * 22	closed- * 22	closed+ * 22
Opening Task Cohort					
Visuomotor	37	task15 * 30	closed+ * 54		
Visuomotor-Feedback	30	task15 * 30	closed+ * 54		
Opening Gambling Cohort					
RestAfterWins	25	closed+ * 54	rest15 * 30		
Daily-Closed-01	68	closed+ * 32	closed- * 32	closed+ * 32	
Daily-Random-01	66	random * 32	random * 32	random * 32	
App-Exploratory	5000	random * 30			
App-Confirmatory	21896	random * 30			
Follow-Up Cohorts					
BoredomBeforeAndAfter	150	rest15 * 18	closed- * 33	closed+ * 33	
BoredomAfterOnly	150	rest15 * 18	closed- * 33	closed+ * 33	
MwBeforeAndAfter	150	rest15 * 18	closed- * 33	closed+ * 33	
MwAfterOnly	150	rest15 * 18	closed- * 33	closed+ * 33	
Activities	450	break420 * 1	closed- * 33	closed+ * 33	

Supplementary Table 1: A list and description of cohorts collected. nParticipants contains the number of participants who completed both the task and survey in this cohort. The columns beginning with "Block" denote the type, parameter, and number of trials used in that block of trials. "Rest" denotes looking at a fixation cross, and "task" denotes a simple visuomotor task in which a cross moves predictably across the screen and the subject is asked to press a button when it crosses the center line. The number that follows these labels is the time in seconds between mood ratings. "Break" denotes a free period where participants could leave to do anything they chose. "Closed" and "random" denote the closed-loop and random gambling task conditions described in the Methods section. ("open" denotes open-loop gambling not described in this paper; these blocks were not used in analyses). The + or - after the "closed" label indicates whether mood was being manipulated upwards (+) or downwards (-). The number after the * indicates how many trials of this type were included in the block. Certain cohort names also contain information. The AlternateRating cohort rated their mood with a single button press rather than moving a slider. The Expectation cohorts received opening instructions stating that the upcoming rest period would be up to 7 minutes or 12 minutes. Groups beginning with "Daily" or "Weekly" returned 1 day or 1 week apart to complete a similar task again (e.g., the Daily-Rest-02 cohort is the same participants as Daily-Rest-01, returning to complete the same task one day later). The Adolescent-01 cohort is a group of adolescents recruited in person rather than on Amazon Mechanical Turk.

Factor	Estimate	2.5_ci	97.5_ci	SE	DF	T-stat	P-val	Sig
(Intercept)	0.784	0.756	0.812	0.0141	875	55.6	$< 10^{-6}$	*
Time	-0.0189	-0.0226	-0.0153	0.00185	864	-10.3	$< 10^{-6}$	*
isMale	-0.0144	-0.0395	0.0107	0.0128	877	-1.12	0.262	
meanIRIOver20	0.000698	-0.000585	0.00198	0.000655	901	1.07	0.287	
totalWinnings	-0.000332	-0.00435	0.00369	0.00205	898	-0.162	0.872	
meanRPE	0.158	-0.0104	0.326	0.0859	898	1.84	0.0662	
fracRiskScore	-0.186	-0.202	-0.169	0.00828	877	-22.4	$< 10^{-6}$	*
isAge0to16	-0.0456	-0.108	0.0168	0.0318	879	-1.43	0.152	
isAge16to18	-0.0883	-0.144	-0.0325	0.0285	879	-3.1	0.002	*
isAge40to100	-0.00712	-0.0351	0.0208	0.0143	877	-0.5	0.617	
Time:isMale	0.00159	-0.00171	0.00488	0.00168	869	0.944	0.345	
Time:meanIRIOver20	-0.000103	-0.000267	$6.1 * 10^{-5}$	$8.4 * 10^{-5}$	810	-1.23	0.219	
Time:totalWinnings	$-1.9 * 10^{-5}$	-0.000566	0.000529	0.00028	$1.04 * 10^{3}$	-0.0664	0.947	
Time:meanRPE	-0.00743	-0.0304	0.0155	0.0117	$1.05 * 10^{3}$	-0.634	0.526	
Time:fracRiskScore	0.00515	0.00303	0.00728	0.00109	869	4.75	$2 * 10^{-6}$	*
Time:isAge0to16	-0.00144	-0.00967	0.00678	0.0042	895	-0.344	0.731	
Time:isAge16to18	0.00869	0.00131	0.0161	0.00376	898	2.31	0.0212	*
Time:isAge40to100	0.00302	-0.000638	0.00668	0.00187	865	1.62	0.106	

Supplementary Table 2: Results of the LME model trained on all naïve online adult and adolescent participants who received opening rest, visuomotor task, or random gambling periods; as produced by the pymer software package. The first column lists each factor in the model as described in the Methods section. Factors beginning with "is" are binary (0 or 1). "Time" is the mood slope parameter we use to quantify mood drift. Mood ratings ranged from 0-1, and time was in minutes. totalWinnings and meanRPE were in points, whose monetary value is unknown to naïve subjects. fracRiskScore was the score on a clinical depression questionnaire divided by a clinical cutoff. Age was in years. Factors preceded by "Time:" indicate the interaction of that parameter and the elapsed time. The next four columns describe the effect size: "Estimate" is the estimated coefficient of each factor in the model, 2.5 and 97.5 ci are the 95 percent confidence interval of the estimate, and SE is its standard error. DF is the degrees of freedom, T-stat is the t statistic, and P-val is the 2-sided p value. All values are rounded to 3 decimal places. The Sig (significance) column contains * if $p < 0.05$.

slope parameters produced by an LME without age factors included are plotted against age in Supplementary Figure 2. The relationship between age and mood slope was not clear from these plots; more research will be required to clarify the relationship between mood drift and age.



Supplementary Figure 2: Mood slopes (produced by an LME model with age-related terms removed) plotted against participant age.

C. Eliminating Methodological Confounds

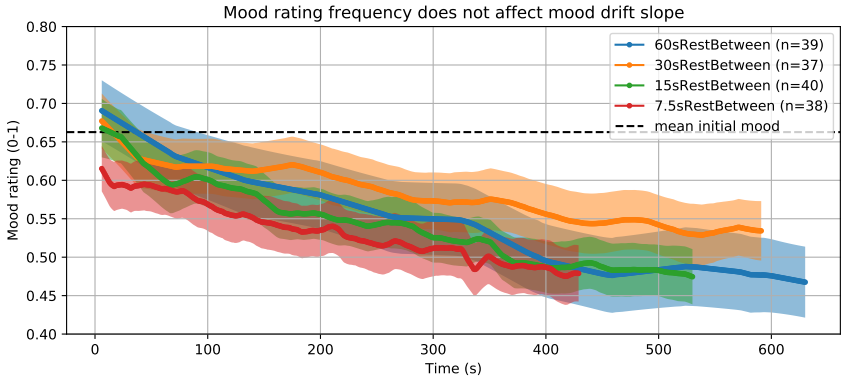
Because this finding is new, we wanted to examine the impact of possible methodological confounds. We therefore created slightly modified versions of the task to see whether the observed decline in mood ratings might be due to:

1. The aversive nature of rating one's mood
2. The method of rating mood and its susceptibility to fatigue
3. The expected duration of the rest period
4. Multitasking or task switching

Mood Drift Is Not a Product of Aversive Mood Ratings

To investigate whether the decline in mood might be driven by the ratings themselves, we varied the frequency of mood ratings. We reasoned that, if mood ratings were decreasing mood, more frequent ratings would cause mood to decline more quickly. We observed that participants with 60 s, 30 s, 15 s, and 7.5 s of rest between ratings (cohorts 60sRestBetween, 30sRestBetween, 15sRestBetween, and 7.5sRestBetween, in Supplementary Table 1) all had mood ratings that declined at roughly the same rate (Supplementary Figure 3). This finding was later confirmed by our multi-cohort LME model, in which a participant's mean inter-rating interval did not have a significant relationship with their slope parameter (inter-rating-interval x time interaction =

-0.0103 %mood, 95%CI = (-0.0267, 0.0061), $t_{810} = -1.23, p = 0.219$, 2-sided, Supplementary Table 2). From this, we conclude that mood ratings were not aversive enough that an increase in mood rating frequency led to an increase in mood drift.



Supplementary Figure 3: Mean ± STE mood rating at each time in the 4 cohorts with 60 s, 30 s, 15 s, and 7.5 s of rest between mood ratings (cohorts 60sRestBetween, 30sRestBetween, 15sRestBetween, and 7.5sRestBetween, respectively). The magnitude of mood drift did not vary with the frequency of mood ratings.

Mood Drift Over Time Is Not an Artefact of the Rating Method

Participants had thus far rated their mood with a slider that started in the middle of the scale (0.5). We therefore wondered whether participants' mood ratings were converging on 0.5 because they were becoming more fatigued and ratings near the middle of the slider required the least effort. In another modified version of the task, we asked participants (cohort AlternateRating in Supplementary Table 1) to press a single number key (1-9) to indicate their happiness during the mood ratings, where 1 was "unhappy" and 9 was "happy". In this way, we made each mood require roughly equal time and effort. We did not find evidence of a difference between the LME slope parameters collected from this task and those of the original cohort (-2.22 vs. -2.45 %mood/min, 95%CI = (-0.772, 1.23), $t_{70} = 0.427, p = 0.671$, 2-sided).

Mood Drift Over Time Is Not Driven by Expectations

We examined whether the mood ratings might be affected by the expected duration of the rest period. This would suggest that the mood drift observed during rest was a product of rumination about the amount of rest time remaining. To test this, we gave identical tasks to two groups, preceded by slightly different instructions: one was told that the initial rest period would be up to 7 minutes (cohort Expectation-7mRest, $n = 64$), and the other was told it would be up to 12 minutes (cohort Expectation-12mRest, $n = 67$). After these instructions, both groups actually received rest periods of approximately 6.4 minutes. Participants were randomised to a group at the time of participation. We did not find evidence of a difference in LME slope parameters between these two groups (Expectation-7mRest vs. Expectation-12mRest (-1.47 vs. -1.53%mood/min, 95%CI = (-0.613, 0.743), $t_{104} = 0.185, p = 0.854$, 2-sided).

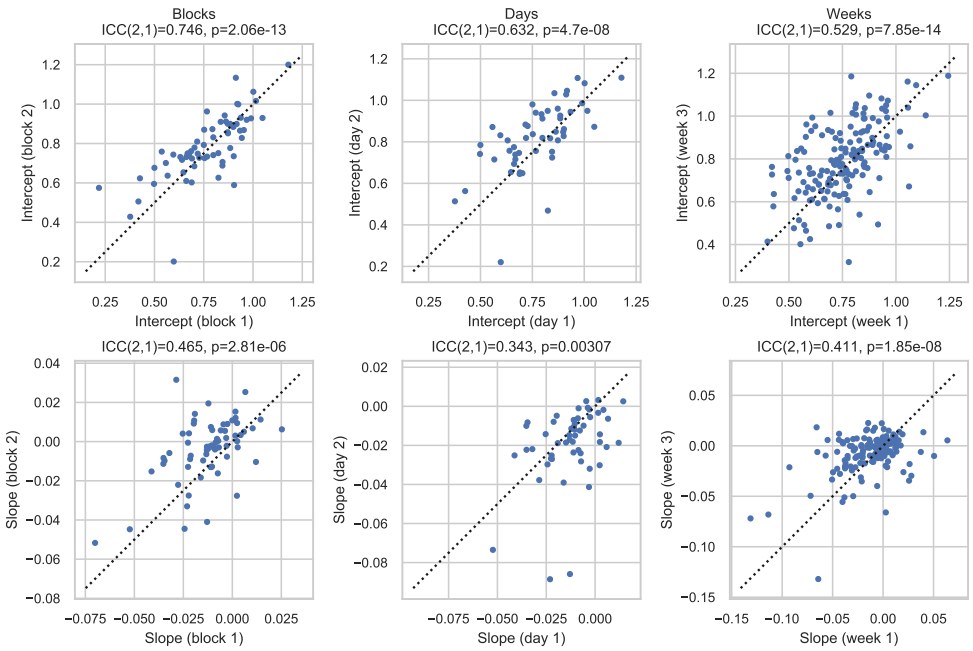
Mood Drift Over Time Is Not Driven by MultiTasking

Mood drift's generalizability across task conditions speaks to the concern that online participants were multitasking on their computers or phones during rest periods. Online participants included in the large-scale LME moved or locked in their mood rating slider on 97.7% of rest trials, suggesting that any multitasking was not so engaging as to stop them from noticing the next mood rating. Cohorts with short rest periods between mood ratings likely had to make responses too frequently to multitask, but the time between ratings

did not change participants' level of mood drift (see section titled "Mood Drift Over Time Is Not a Product of Aversive Mood Ratings" above). This evidence does not rule out that people were multitasking, but it suggests that any multitasking taking place did not reliably change the observed levels of mood drift.

D. Stability Over Time

We examined the stability of the LME intercept and slope parameters within an individual. One cohort (Daily-Rest-01 in Supplementary Table 1) repeated a task with a rest block lasting 6.8 minutes on average, a closed-loop positive gambling block lasting 3.5 minutes on average, another 6.8-minute rest block, and another 3.5-minute closed-loop positive gambling block. This cohort was invited to return the following day to complete the same task again (Daily-Rest-02). This allowed us to assess stability both (a) across blocks within a run, and (b) across days. A second cohort (Weekly-Rest-01) completed an initial rest block lasting 6.8 minutes on average, followed by three 4.3-minute closed-loop gambling blocks (1 positive, 1 negative, 1 positive). They were invited back one and two weeks later to complete the same task again (Weekly-Rest-02/03). This allowed us to assess stability across weeks.



Supplementary Figure 4: Stability of LME coefficients estimating the initial mood (top) and mood drift over time (bottom) for each participant across rest periods one block apart (left), 1 day apart (middle), and 2 weeks apart (right). ICC denotes the intra-class correlation coefficient for each comparison.P values shown are one-sided (since ICC values are expected to be positive) with no correction for multiple comparisons.

The LME intercept parameter (i.e., initial mood) showed high stability across blocks ($ICC(2,1) = 0.746, p < 0.001$), days ($ICC(2,1) = 0.632, p < 0.001$), and weeks ($ICC(2,1) = 0.529, p < 0.001$, each 1-sided since ICC values are expected to be positive), confirming the stability of subjective momentary mood ratings. The Slope parameter showed moderate stability that was statistically significant, across blocks ($ICC(2,1) = 0.465, p < 0.001$), days ($ICC(2,1) = 0.343, p < 0.001$), and weeks ($ICC(2,1) = 0.411, p < 0.001$, each 1-sided as before). Scatter plots are shown in Supplementary Figure 4. This level of stability suggests that inter-individual differences in initial mood and slope are driven by stable traits rather than random fluctuations.

E. Mood Drift Over Time Is Inversely Related to Depression Risk

In the main text, we found that the relationship between a participant's mood drift and their depression risk was statistically significant, but that its impact on model fit was very small. In this section, we expand upon these depression-related findings from the main text.

First, we investigated whether participants' mood drift correlated with trait-level depressive characteristics. In our online participant LME model, higher depression risk score was significantly associated with lower initial mood ($Mean \pm SE = -18.6 \pm 0.8\% mood$, $t_{877} = -22.4, p < 0.001$) and less negative mood drift (depression-risk * time interaction, $Mean \pm SE = 0.515 \pm 0.109\% mood/min$, $t_{869} = 4.75, p < 0.001$). This relationship is visually characterised in several ways in Figure 5. Each analysis supports the relationship between mood slope and trait-level depression.

Including the interaction between time and depression-risk in the LME model improved model fit ($\chi^2(1, N = 14) = 21.5, p < 0.001$). But the effect of its inclusion was very small: the within-individual variance explained $(R_1^2)^{1,2}$ increased from $R_1^2 = 0.291$ (without this new term in the model) to $R_1^2 = 0.293$ (with it). The inclusion of time's interaction with depression-risk in our model produced a very small effect ($f^2 = 0.00289$[3,4]). Similarly, between-individual variance explained $(R_2^2)^{1,2}$ increased from $R_2^2 = 0.1127$ to $R_2^2 = 0.1134$ ($f^2 = 0.000886$).

The inverse relationship between depression risk and mood slope was later replicated in our follow-up cohorts (i.e., cohorts MwBeforeAndAfter, MwAfterOnly, BoredomBeforeAndAfter, and BoredomAfterOnly, n=600). As before, a higher depression risk score was significantly associated with lower initial mood ($Mean \pm SE = -18.1 \pm 0.9\% mood$, $t_{593} = -20.3, p < 0.001$) and less negative mood drift (depression-risk * time interaction, $Mean \pm SE = 0.510 \pm 0.140\% mood/min$, $t_{594} = 3.64, p < 0.001$, 2-sided).

This relationship was also observed in the mobile app cohort. Using each participant's life happiness rating as a proxy for (lack of) depression risk, we found a significant negative correlation between life happiness and β_T ($r_s = -0.0658, p < 0.001$, 2-sided) (Figure 10, bottom right).

We took care to examine the possibility that regression to the mean or floor effects were driving these results. These possibilities are examined in Supplementary Notes F. and G..

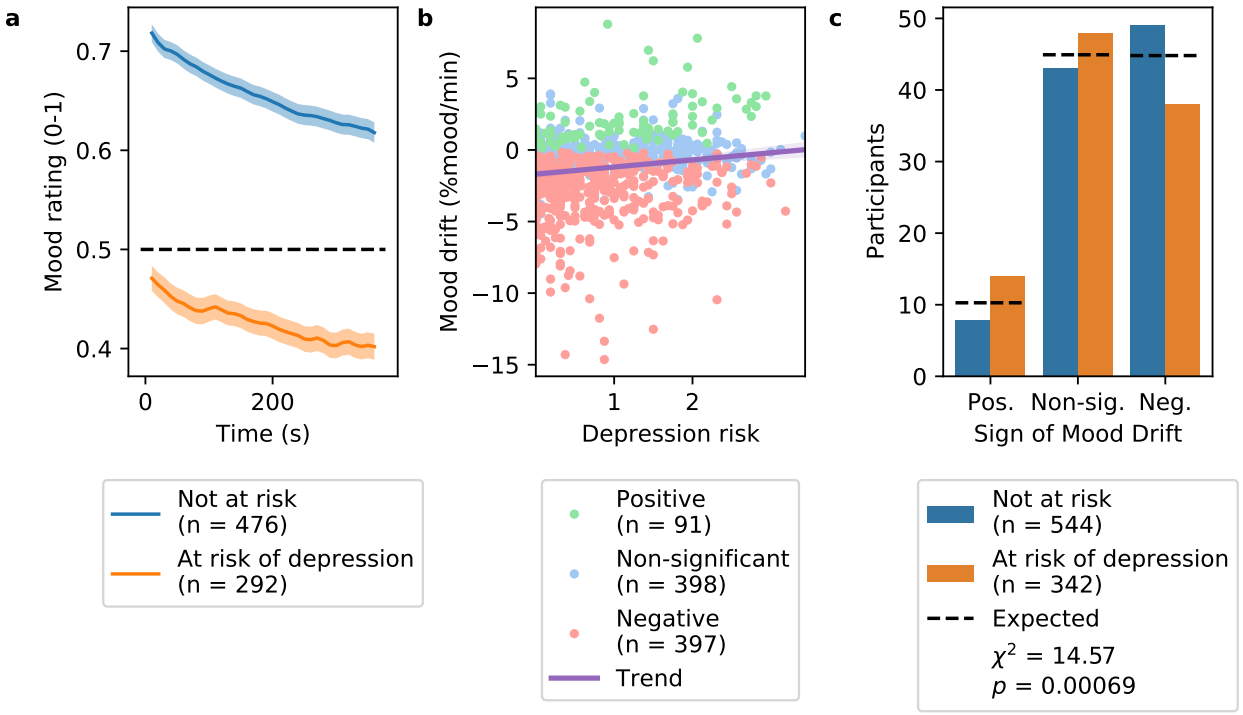
F. Examining Possible Regression to the Mean

We were concerned that our results concerning depression and mood drift might be an artefactual result of regression to the mean: for a purely random process, values starting high will tend to go down over time, and values starting low will tend to go up over time. Thus, slope parameters might be less negative for people with higher depression risk simply because their initial mood happened to be lower.

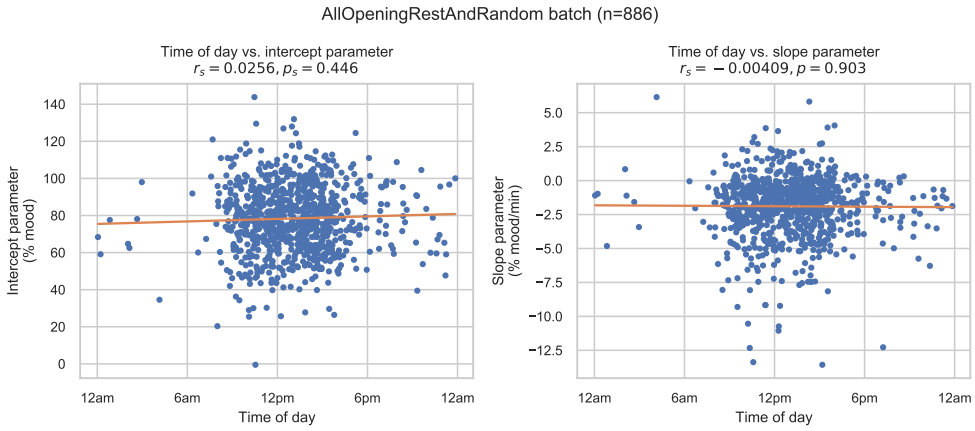
In addition to the stability analyses in D., we also examined the specific effect of time of day on mood. Past research has shown that affective ratings vary consistently with time of day, with reports of pleasantness being lowest in the morning and highest in the evening.[5] Time of day also impacts loss sensitivity during risky decision-making.[6] If time of day were related to initial mood or mood slope, our individual difference results could possibly be explained by depressed individuals participating at different times of day than non-depressed participants. In the dataset of online participants, however, we did not observe a significant relationship between the time of day when the task was completed and the intercept or slope parameter (Supplementary Figure 6). This suggests that inter-individual differences in initial mood and slope were not driven by periodic daily fluctuations in mood.

G. Examining Possible Floor Effects

Individuals reporting greater depressive symptoms on average reported lower initial mood at the onset of the task. If their mood declined further, they therefore had less of the mood scale available to them to express it.



Supplementary Figure 5: Relationship between mood drift and depression risk. (a) Mood ratings over time of online participants at risk of depression (defined as MFQ>12 or CES-D>16) vs. those not at risk for the 768 participants with at least 6 minutes of resting mood data (error bars are SEM). The dotted line represents the mean initial rating (mean of cohort means). (b) We fitted simple regressions of time versus mood within each individual and determined significance of the time term with Benjamini-Hochberg false-discovery rate correction ($\alpha = 0.5,\ p < 0.05$) to better understand the relationship between depression risk and the change in mood over time. Depression risk is operationalised as score on the CES-D or MFQ divided by the threshold for depression risk on each measure (16 and 12 respectively). (c) Proportion of individuals with or without risk of depression (i.e., depression risk >1 or <1) with positive (significantly greater than zero), non-significant (no evidence of a significant difference from zero), and negative (significantly less than 0) slopes of mood over time. 13 more individuals at risk of depression have a positive slope than the 35 expected based on the rates in individuals not at risk of depression, $\chi^2(1, N = 886) = 14.57, p < 0.001$ (2-sided Pearson's chi-squared statistic with no correction for multiple comparisons).



Supplementary Figure 6: Intercept and slope parameters learned by the LME model, plotted against time of day in the online cohorts.Lines show best linear fit. r_s denotes Spearman correlation coefficient. P values shown are 2-sided with no correction for multiple comparisons.

This could lead to "floor effects" where the mood of depressed individuals appears to decline more slowly with time simply because they have reached the bottom of the scale and are forced to level out.

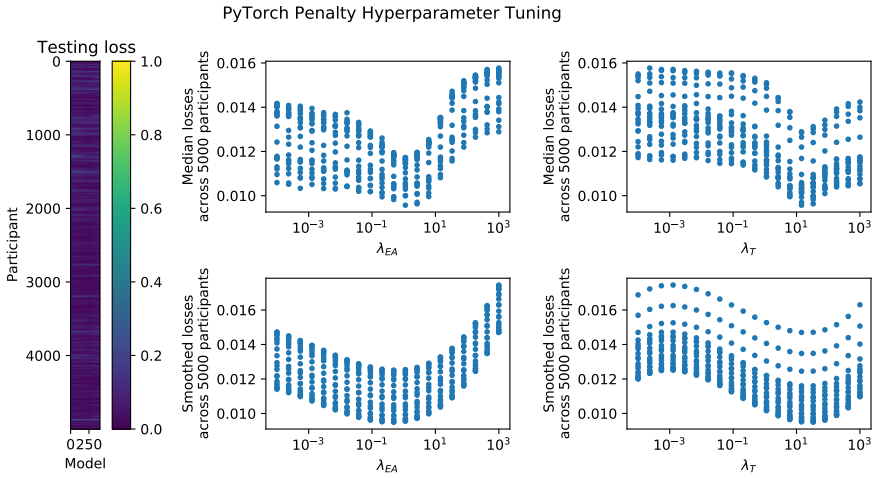
In a sensitivity analysis, we excluded the 27/600 participants in the follow-up cohorts (See Supplementary Table 1) who reached the floor of the mood scale (i.e., mood = 0) at any time during the rest period. We then re-fit the LME model of mood. The significant effect of the interaction between depression risk and time (i.e., the relationship between depression risk and mood drift) persisted in this analysis ($t_{566} = 4.06, p < 0.001$, 2-sided). Thus, the effect is not driven by depressed participants reaching the absolute minimum of the scale.

We also considered whether participants might be reluctant to reach the floor of the scale but could still reach a sort of "individual" mood floor, a point under which they would be reluctant to rate themselves. In our follow-up cohorts, rest periods were followed a period of negative mood induction (via increasing the probability of monetary losses in a block of trials). We have demonstrated before[7] that this form of mood induction produces potent changes in mood with effect sizes of Cohen's $d = -1.75$. We took the lowest point during this mood induction to represent a (conservative) individual mood floor. This allowed us to check whether participants reached an individual mood floor during the preceding rest period. In a sensitivity analysis, we excluded the 101/600 participants who reached such an "individual mood floor" (i.e., we excluded all those participants who during resting state reached the minimum mood that they had reached during the negative mood induction). This sensitivity analysis also had minimal effect on our results, in which the interaction effect of depression risk and time remained significant. ($t_{493} = 3.43, p < 0.001$, 2-sided).

H. Computational Model

Our computational model was based on the one described and validated in,[7] which accurately modelled subjective mood ratings in a very similar gambling game. The computational model fit the data well for most of our mobile app participants. In the tuning step, the hyperparameters minimizing testing loss were determined to be $\lambda_{EA} = 0.483, \lambda_T = 33.6$. The relationship between these hyperparameters and the smoothed testing loss is shown in Figure 7.

When using these hyperparameters, the median testing loss (defined as the mean squared errors for the 2 testing trials) across the 5,000 exploratory/tuning participants used to tune parameters was 0.00486. When those hyperparameters were used on the 21,896 confirmatory app participants, the median loss on testing trials was 0.00325. The mean (across participants) Spearman correlation coefficient between each participant's



Supplementary Figure 7: Tuning of penalty term hyperparameters. The two penalty parameters λ_{EA} and λ_T were varied systematically, and the computational model was fit to all but the final two ratings for each participant. Top graphs show the median testing loss (i.e., the sum of squared errors on the final two ratings) across participants. Bottom graphs show these same losses after smoothing with a polynomial fit. The parameters with the lowest smoothed loss on this exploratory mobile app cohort were used in the final model fit to the confirmatory mobile app cohort.

model fits and actual mood ratings was $r_s = 0.715$, 95% CI = (0.754, 0.759).

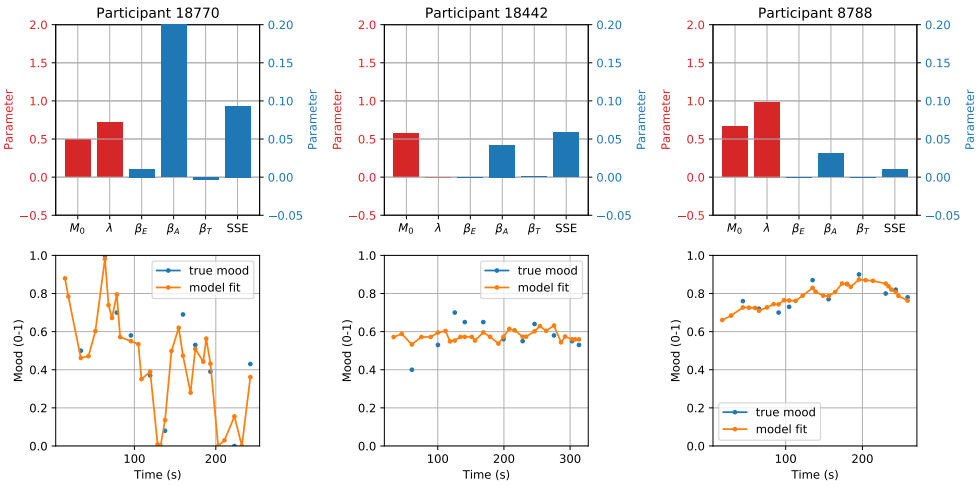
Sample fits are shown in Supplementary Figure 8. Histograms of the learned parameters are shown in Supplementary Figure 9. Relationships between β_T and the other model parameters are shown in Supplementary Figures 10 and 11.

I. Linking Subjective Momentary Mood Ratings to Life Happiness

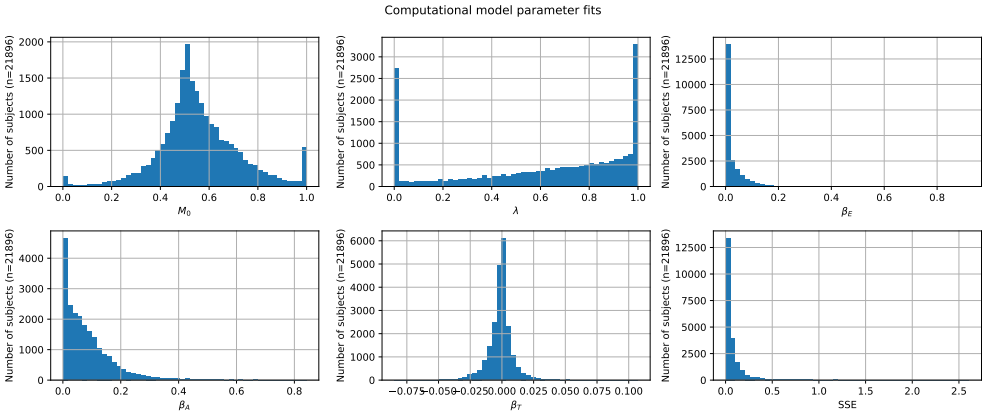
To measure the psychometric validity of the subjective momentary mood ratings, we correlated the initial mood (or "Intercept") parameter of the online cohort's LME model (left) and the mobile app cohort's computational model (right) with the life happiness ratings. Results showed that both estimates of initial mood correlated significantly with ratings of life happiness (Supplementary Figure 12)

J. Impact of Methodology Choices on Mobile App Slope Estimates

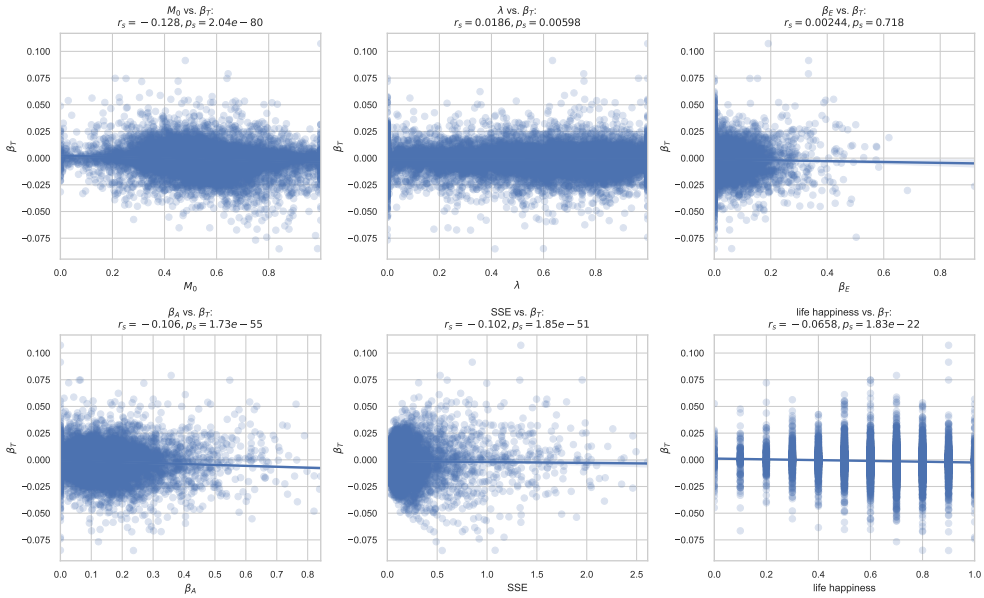
Results showed that mobile app participants experienced significantly less mood drift than online participants. This difference is larger if we use the computational model's time sensitivity parameter rather than the LME analysis' slope parameter. This is likely related to the regularization hyperparameter used in the computational model but not the LME analysis. If an LME analysis is used on both cohorts instead of the computational model, the difference between the two groups' medians is considerably smaller, shrinking from 1.49%*mood*/*min* to 0.774%*mood*/*min* (Supplementary Figure 14). It is also possible that participants experiencing greater mood drift "self-selected" out of the mobile app game: frustrated mobile app participants could exit at any time without penalty, whereas online participants would lose compensation if they dropped out. However, no relationship was observed between the time sensitivity parameter of our computational model and the number of times a participant played the game (Supplementary Figure 13). Finally, since no participants are known to have participated in both experiments, we cannot rule out more general cohort effects: the participants choosing to play the mobile app game could simply have different sensitivity to time on task than those participating in the online experiment.



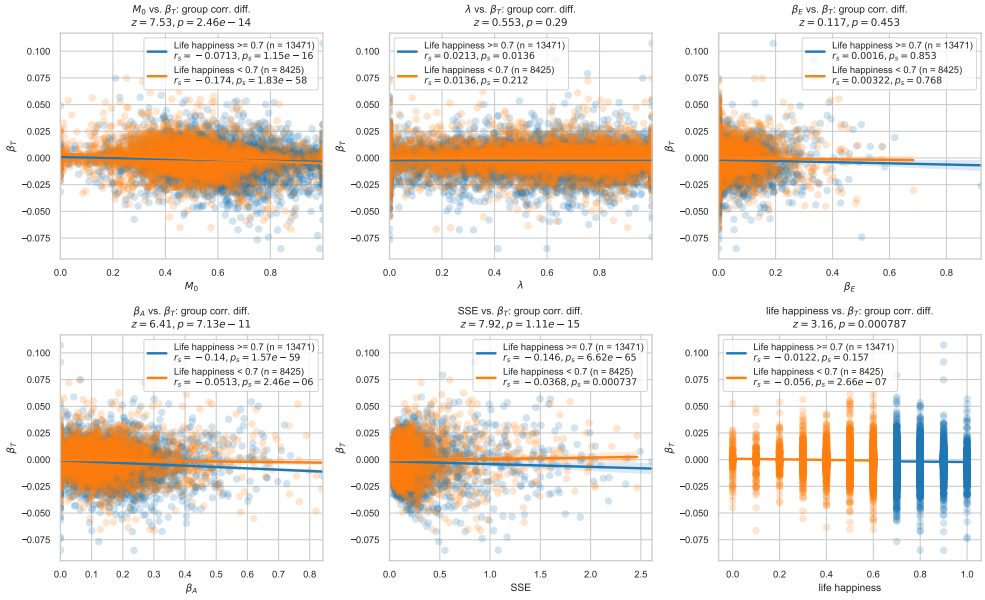
Supplementary Figure 8: Sample fits of the computational model for three random subjects in the confirmatory mobile app cohort. SSE = sum squared error, a measure of goodness of fit to the training data. In the top plots, the red bars are in units of the left-hand y axis, and the blue bars are in units of the right-hand y axis.



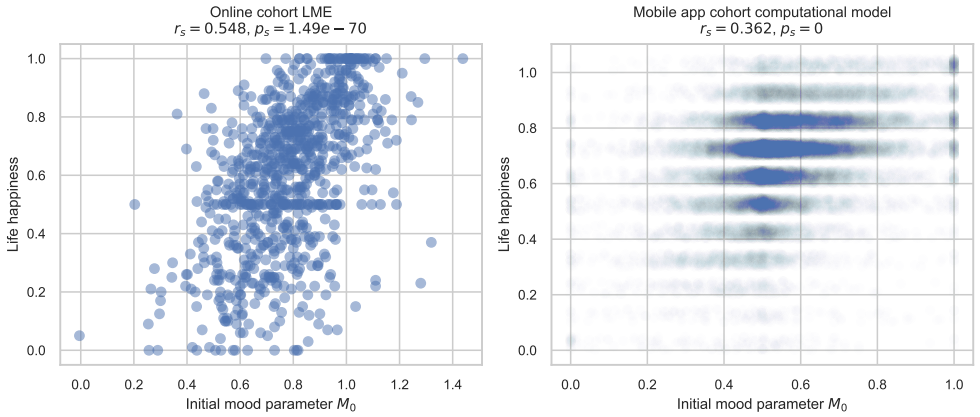
Supplementary Figure 9: Histogram of computational model parameters across the 21,896 confirmatory mobile app subjects.



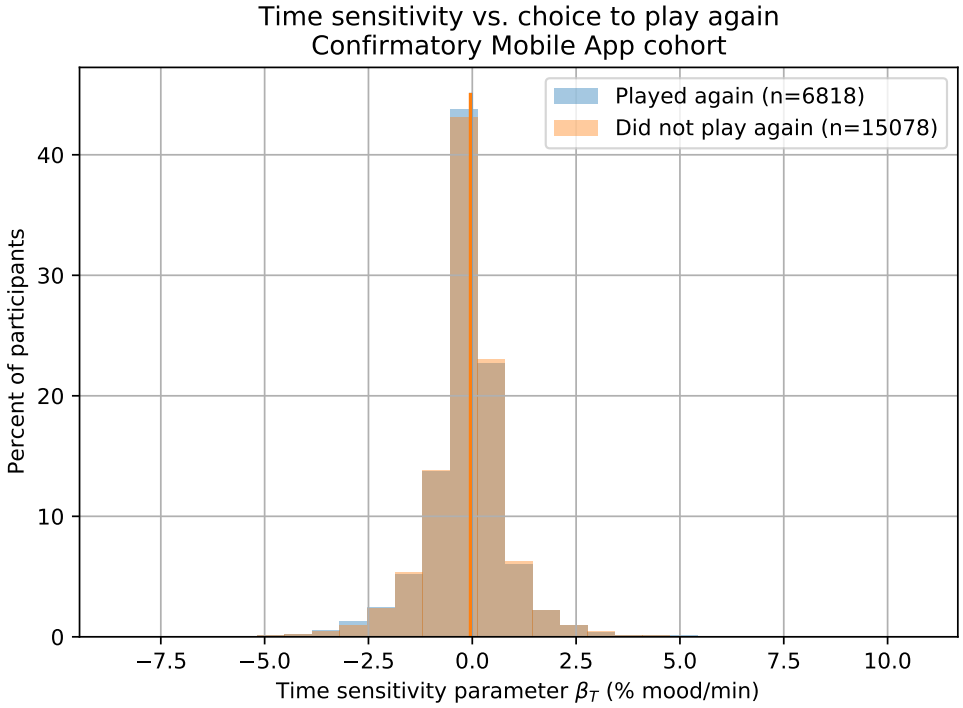
Supplementary Figure 10: Time sensitivity parameter β_T vs. other parameters in the confirmatory mobile app cohort. Each dot is a participant (n=21,896). Each line is a linear best fit, and patches show the 95% confidence interval of this fit. r_s denotes Spearman correlation coefficient. P values shown are 2-sided with no correction for multiple comparisons



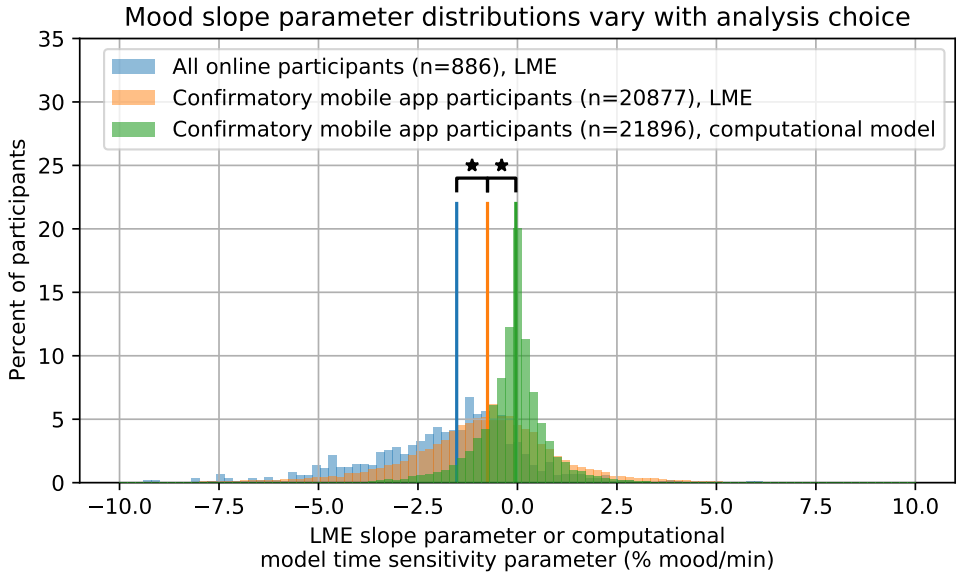
Supplementary Figure 11: Time sensitivity parameter β_T vs. other parameters in the confirmatory mobile app cohort, in 2 groups separated by high (blue) or low (orange) life happiness. Each dot is a participant (n=21,896). Each line is a linear best fit, and patches show the 95% confidence interval of this fit. r_s denotes Spearman correlation coefficient. Group differences in Spearman correlations were statistically assessed using a z statistic. P values shown are 2-sided with no correction for multiple comparisons.



Supplementary Figure 12: Initial mood parameter vs. life happiness rating in the online cohort (left) and the confirmatory mobile app cohort (right). Life happiness ratings were always multiples of 0.1; small positive random values were added during plotting to reduce overlap between data points. Each dot is a participant (left: n=886, right: n=21,896). r_s denotes Spearman correlation coefficient. P values shown are 2-sided with no correction for multiple comparisons.



Supplementary Figure 13: Histogram of the computational model time sensitivity parameter for subsets of the confirmatory mobile app cohort that chose to play again later (blue) and those that did not (orange). No significant difference in the distributions was observed (median = -0.0392 vs. -0.0449, IQR = 0.766 vs. 0.758 %mood/min, 2-sided Wilcoxon rank-sum test, $W_{21894} = 0.804, p = 0.421$). Vertical lines represent group medians.

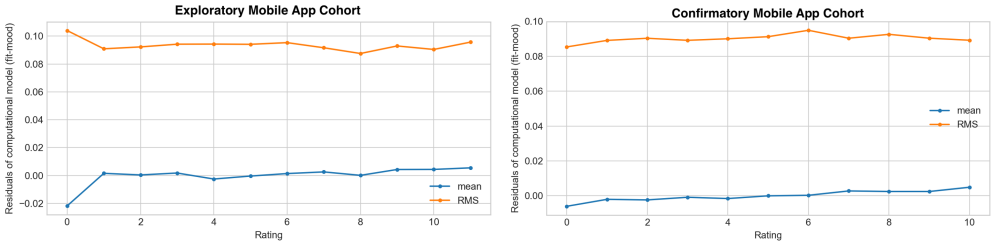


Supplementary Figure 14: Histogram of the LME mood slope parameters for the online cohort (blue) and the confirmatory mobile app cohort (orange), along with the computational model time sensitivity parameter for the confirmatory mobile app cohort (green). Mobile app participants with outlier task completion times were excluded from the LME analysis (see Methods). Note that the use of LME modeling to analyze the mobile app data significantly lowered the distribution of slopes compared to when the computational model was used (median= -0.752 vs. -0.0408, IQR= 2.10 vs. 0.764 %mood/min, 2-sided Wilcoxon rank-sum test, $W_{42771} = -54.2, p < 0.001$), but the LME slopes from the mobile app were still significantly greater than those of the online cohort (median = -1.53 vs. -0.752, IQR= 2.34 vs. 2.1 %mood/min, 2-sided Wilcoxon rank-sum test, $W_{21761} = 14.5, p < 0.001$). Vertical lines represent group medians. Stars indicate $p < 0.05$. P values were not corrected for multiple comparisons.

K. Sensitivity analysis: Excluding First Rating

We chose to include the first mood rating in our linear trend estimation, despite the fact that this rating appeared to be an outlier in our exploratory cohort's computational model fits (Supplementary Figure 15, left). To check the sensitivity of our conclusions to this choice, we performed the same analyses while excluding this first mood rating from our model fitting procedure.

In our confirmatory cohort, this pattern (in which the first rating was an outlier) was not observed (Supplementary Figure 15, right). Nevertheless, we preregistered this sensitivity analysis, and we therefore report the results for the confirmatory cohort below.

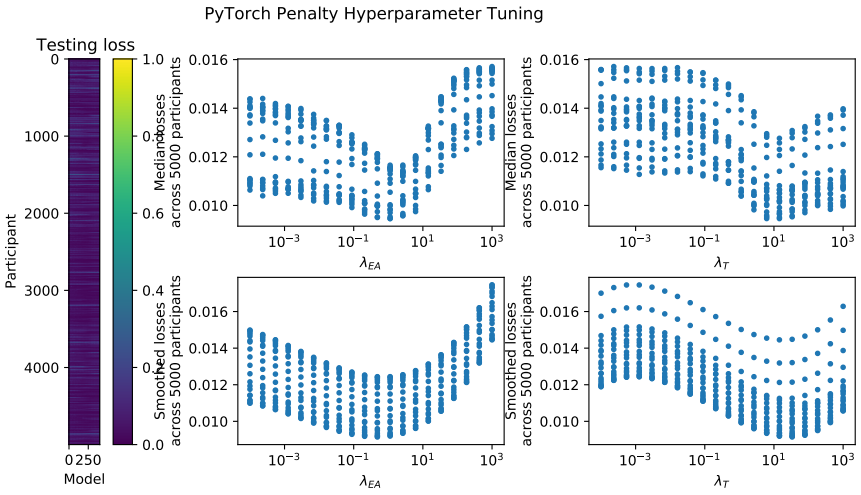


Supplementary Figure 15: Mean (blue) and root-mean-square (RMS, orange) residuals across the exploratory (left) and confirmatory (right) mobile app subjects of the computational model fit for each rating number. In the exploratory cohort, the first rating appeared to be an outlier, inspiring our preregistered sensitivity analysis. In the confirmatory cohort (right), this pattern was not observed. But we still report our preregistered sensitivity analysis on the confirmatory cohort.

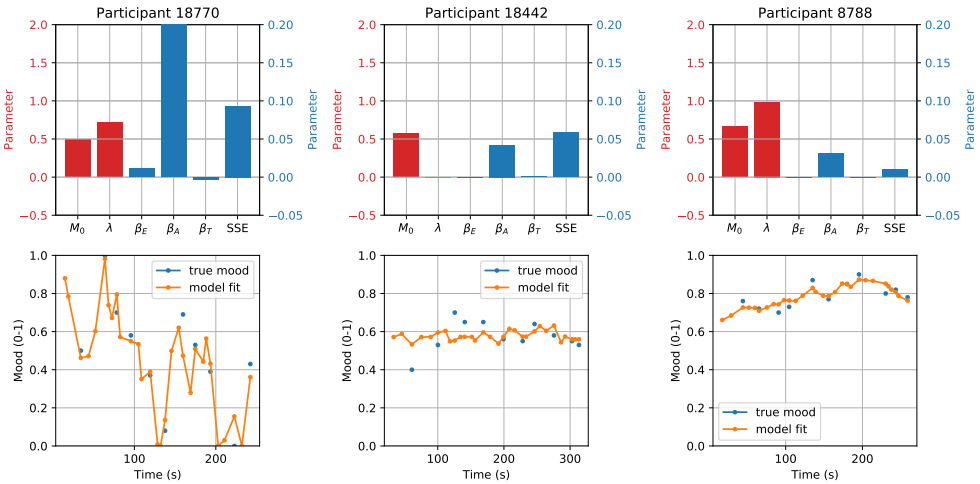
- Model tuning:
  - best fitting penalty hyperparameters (model WITH β_T): $[\lambda_{EA} = 0.483,\ \lambda_T = 33.6]$
  - best fitting penalty coefficients (model WITHOUT β_T): $\lambda_{EA} = 0.207$
  - median MSE (model WITH vs. WITHOUT β_T): 0.0032388 vs. 0.0033644
  - IQR of MSE difference (model WITH vs. WITHOUT β_T): 0.0000214
  - 2-sided Wilcoxon sign-rank test on the difference between models with and without β_T: $W_{499} = 0.0, p < 0.001$
- Distribution of β_T:
  - Mean $\pm$ standard error β_T: -0.129% mood/min $\pm$ 0.00667
  - 2-sided Wilcoxon sign-rank test on β_T vs. 0: $W_{21895} = 1.00 * 10^8, p < 0.001$
  - 2-sided Wilcoxon rank-sum test of LME time coefficients vs. Computational Model β_T: $W_{42771} = -18.4, p < 0.001$
- Individual differences:
  - life happiness vs. β_T: $r_s = -0.0654$, $p < 0.001$, 2-sided
  - β_A vs. β_T: $r_s = -0.106$, $p < 0.001$, 2-sided
  - β_A vs. β_T (life happiness>=0.7): $r_s = -0.140$, $p < 0.001$, 2-sided
  - β_A vs. β_T (life happiness<0.7): $r_s = -0.0510$, $p < 0.001$, 2-sided
  - β_A vs. β_T correlation difference between high and low life happiness groups: $z = 6.43$, $p < 0.001$, 2-sided

L. Results of Boredom, MW, and Free Activities Preregistration

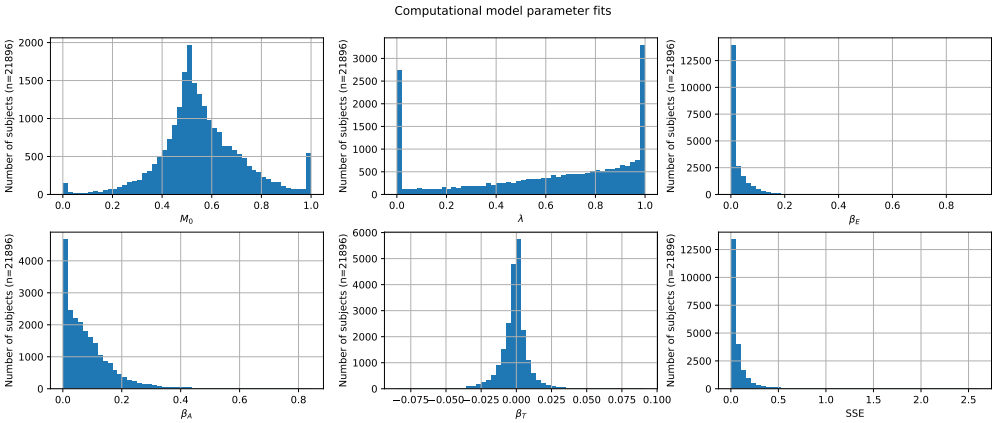
We performed a follow-up set of preregistered tasks and analyses on boredom, mind-wandering, and freely chosen activities (https://osf.io/gt7a8). The purpose of the boredom and MW analyses was to quantify the ability of these factors to explain individual subjects' mood drift. After carrying out the preregistered analyses, we reexamined our analysis method and adopted a different approach to address this question more specifically. We will use this section to motivate and present the results as originally preregistered.



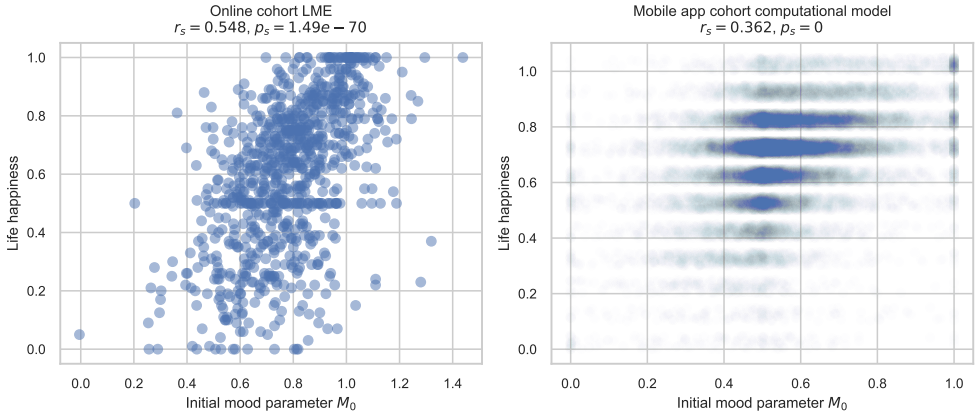
Supplementary Figure 16: Sensitivity analysis with first rating excluded from model fit: Tuning of penalty term hyperparameters. The two penalty parameters λ_{EA} and λ_T were varied systematically, and the computational model was fit to all but the final two ratings for each participant. Top graphs show the median testing loss (i.e., the sum of squared errors on the final two ratings) across participants. Bottom graphs show these same losses after smoothing with a polynomial fit. The parameters with the lowest smoothed loss on this exploratory mobile app cohort were used in the final model fit to the confirmatory mobile app cohort.



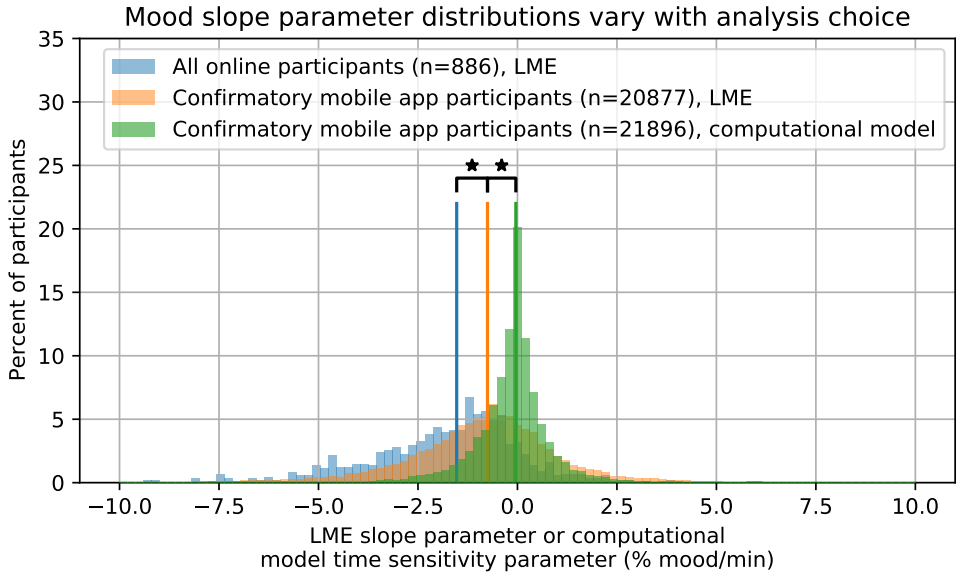
Supplementary Figure 17: Sensitivity analysis with first rating excluded from model fit: Sample fits of the computational model for three random subjects. SSE = sum squared error, a measure of goodness of fit to the training data. In the top plots, the red bars are in units of the left-hand y axis, and the blue bars are in units of the right-hand y axis.



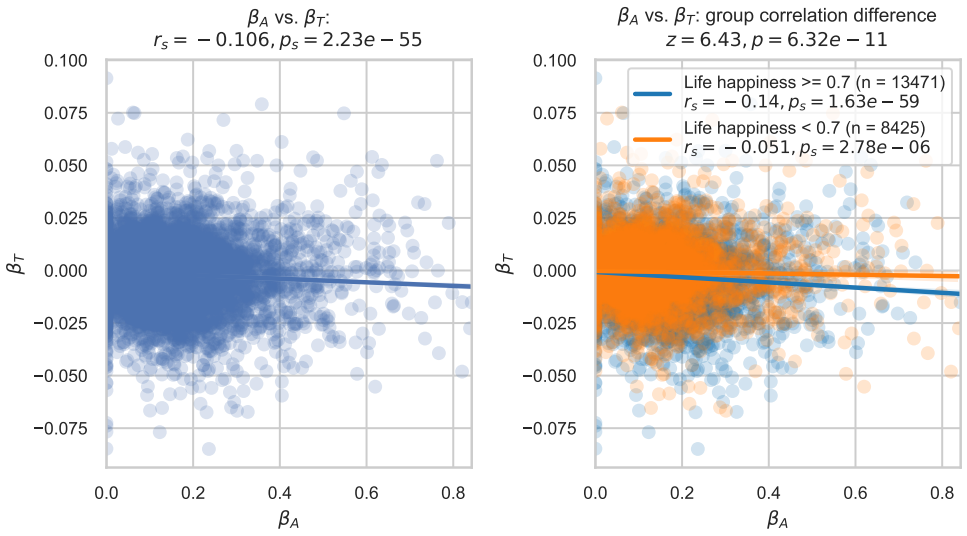
Supplementary Figure 18: Sensitivity analysis with first rating excluded from model fit: Histogram of computational model parameters across the confirmatory mobile app subjects.



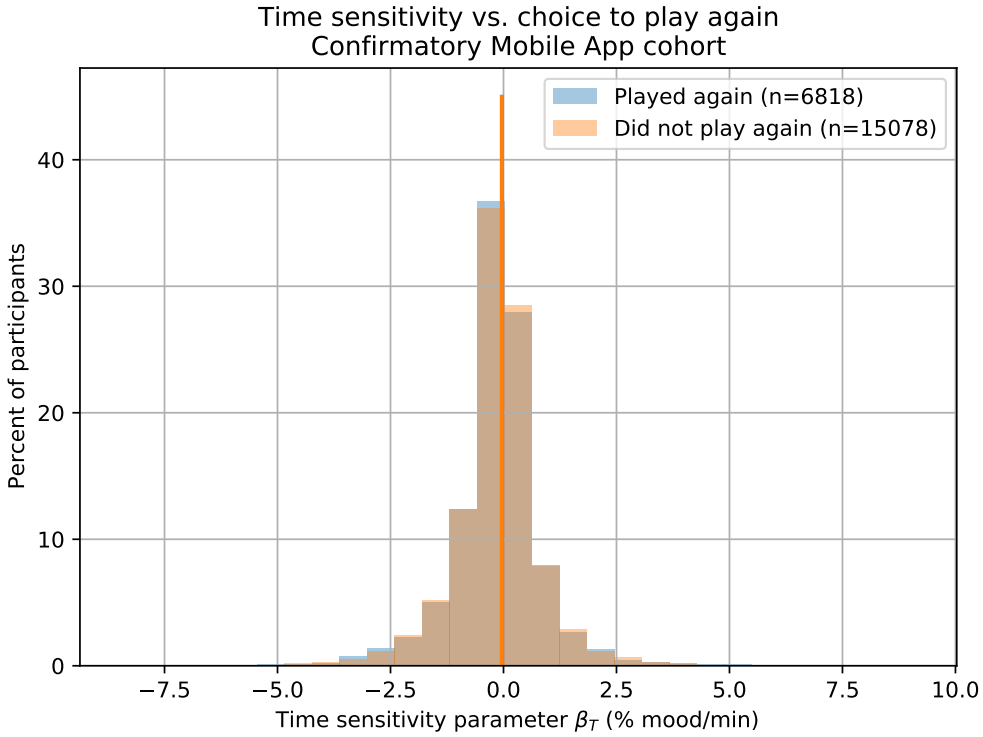
Supplementary Figure 19: Sensitivity analysis with first rating excluded from model fit: Initial mood parameter vs. life happiness rating in the online cohort (left) and the confirmatory mobile app cohort (right). Life happiness ratings were always multiples of 0.1; small positive random values were added during plotting to reduce overlap between data points. Each dot is a participant (left: n=886, right: n=21,896). r_s denotes Spearman correlation coefficient. P values shown are 2-sided with no correction for multiple comparisons.



Supplementary Figure 20: Sensitivity analysis with first rating excluded from model fit: Histogram of the LME mood slope parameters for the online cohort (blue) and the confirmatory mobile app cohort (orange), along with the computational model time sensitivity parameter for the confirmatory mobile app cohort (green). Mobile app participants with an inter-rating interval (IRI) > 38 seconds were excluded from analysis. Note that the use of LME modeling to analyze the mobile app data significantly lowered the distribution of slopes compared to when the computational model was used (median= -0.331 vs. -0.0404, IQR= 2.2 vs. 0.764 %mood/min, 2-sided Wilcoxon rank-sum test, $W_{42771} = -18.4, p < 0.001$), but the LME slopes from the mobile app were still significantly greater than those of the online cohort (median= -0.331 vs. -1.43, IQR= 2.2 vs. 2.12 %mood/min, 2-sided Wilcoxon rank-sum test, $W_{21761} = 18.9, p < 0.001$). Vertical lines represent group medians. Stars indicate $p < 0.05$. P values were not corrected for multiple comparisons.



Supplementary Figure 21: Sensitivity analysis with first rating excluded from model fit: Individual differences in sensitivity to the passage of time relate to other individual differences. The computational model's time sensitivity parameter β_T for each participant in the confirmatory mobile app cohort is plotted against that participant's reward sensitivity parameter β_A in the whole group (left) and separated by life happiness (right). Each dot is a participant (n=21,896). Each line is a linear best fit, and patches show the 95% confidence interval of this fit. r_s denotes Spearman correlation coefficient. The group difference in Spearman correlations was statistically assessed using a z statistic. P values shown are 2-sided with no correction for multiple comparisons.



Supplementary Figure 22: Sensitivity analysis with first rating excluded from model fit: Histogram of the computational model time sensitivity parameter for subsets of the confirmatory mobile app cohort that chose to play again later (blue) and those that did not (orange). No significant difference in the distributions was observed (median= -0.045 vs. -0.0393, IQR= 0.758 vs. 0.767 %mood/min, 2-sided Wilcoxon rank-sum test, $W_{21894} = 0.838, p = 0.402$). Vertical lines represent group medians.

Order	Activity	Frequency
1.	I thought.	50.2%
2.	I consumed the news.	28.2%
3.	I looked at photos.	20.2%
4.	I listened to music, podcasts, or radio.	23.5%
5.	I did some work for my (non-MTurk) job.	16.3%
6.	I looked for a (non-MTurk) job.	10.4%
7.	I paid bills, banked, or invested.	10.2%
8.	I did something else on my computer or phone.	44.7%
9.	I read texts or emails.	22.5%
10.	I wrote something.	12.2%
11.	I watched videos.	18.5%
12.	I went on social media.	20.3%
13.	I shopped.	9.44%
14.	I did something on MTurk.	15.4%
15.	I called/videochatted with someone.	8.22%
16.	I played a computer/phone game.	13.6%
17.	I did something on my computer/phone not listed here.	15.6%
18.	I read something NOT on a computer/phone.	11.8%
19.	I wrote something NOT on a computer/phone.	8.5%
20.	I watched TV.	12.8%
21.	I ate or drank something.	21.6%
22.	I spoke with someone in person.	13.5%
23.	I did a craft.	8.17%
24.	I stood up.	26.2%
25.	I did something physically active.	15.5%
26.	I went to the restroom.	14.1%
27.	I did something OFF a computer/phone not listed here.	17.6%

Supplementary Table 3: Activities reported during the rest period by the (n=450) participants in the Activities cohort (in the order in which the activities were rated).

Supplementary Note M. will then present the improved approach referenced in the main text.

In a new "Activities" cohort ($n = 450$), participants were allowed to choose their own activities during a 7-minute rest period, as described in the main text. Afterwards, participants could indicate how much time they spent on each activity using a slider ranging from "Not at all" (scored at 0%) to "The whole time" (scored at 100%). Their rating of each activity (in the order in which they were rated) is shown in Supplementary Table 3. The most frequent activities reported were thinking (mean 50.2%), reading the news (28.2%), and standing up (26.2%). The rest were performed for less than a quarter of the average break period. Those who reported thinking also reported other activities; most participants apparently used this response to indicate not exclusively sitting and thinking, but rather thinking about the things they were doing.

Two new cohorts were collected to quantify the degree to which mood drift could be explained by boredom. Each received a rest period with mood ratings 20 seconds apart, followed by the Multidimensional State Boredom Scale's short form (MSBS-SF).[8] The first (cohort BoredomBeforeAndAfter, $n = 150$) completed the MSBS-SF both before and after this rest period. The second (cohort BoredomAfterOnly, $n = 150$) completed the MSBS-SF only after this rest period. Both cohorts completed a survey that included the short boredom proneness scale (SBPS) to assess trait boredom.[9] Using a one-sided t-test to remain conservative, we determined that repeated administration of the MSBS-SF did affect later responses: that is, participants who were asked about boredom before the rest period reported lower boredom after the rest period than

those who were not asked about boredom before the rest period (Cohen's $d = -0.411$). Because we could not rule out the possibility of a large effect (H0: Cohen's $d < -0.5,\ t_{298} = 0.987,\ p = 0.163$), we did not combine across the two cohorts in subsequent analyses.

Past research has found that it is not mind-wandering in the general or "traditional" sense (i.e., any task-unrelated thought) that decreases mood, it is mind-wandering with negative affective content.[10] This notion is supported by current theories of mind-wandering not as a monolith, but as a collection of thoughts whose content shapes brain activity and behaviour.[11] Research has linked thought probe responses about the affective content of this ongoing thought to brain activity patterns in the mOFC.[12] The method described in[13] provides a way to quantify the negative affective content of this ongoing thought that more robustly separates affective tone from the mere presence of task-unrelated thought (see Methods).

Two new cohorts were collected to quantify the degree to which mood drift could be explained by the content of ongoing thought (including MW with negative emotional content). Each received a rest period with mood ratings 20 seconds apart, followed by a 13-item Multidimensional Experience Sampling (MDES) as described by Turnbull et al.[13] The first (cohort MwBeforeAndAfter, $n = 150$) completed the MDES only after this rest period. The second (cohort BoredomAfterOnly, $n = 150$) completed the MDES only after this rest period. As described by Ho et al.,[14] we applied principal components analysis (PCA) on participants' MDES responses to find a component whose primary loading was on the "emotion" item (in which they reported their thoughts as being negative or positive). The "emotion dimension" of each MDES response was then quantified as the amplitude of this component. The sign of PCA components is not meaningful, so we arbitrarily chose that increased emotion dimension would represent more negative thoughts. Both cohorts completed a survey that included the 5-item mind-wandering questionnaire (MWQ), which quantifies a person's proneness to mind-wandering without regard to the valence of those spontaneous thoughts.[15] Using two one-sided t-tests to remain conservative, we determined that repeated administration of the MDES did not affect later responses in the emotion dimension: that is, participants did not report different emotional valences after the rest period if they were also asked about their thoughts before the rest period (Cohen's $d = 0.0739$; $H0 : d < -0.5 :\ t_{298} = 7.52,\ p < 0.001$; $H0 : d > 0.5 :\ t_{298} = 5.58,\ p < 0.001$).

Our preregistration contained ten specific hypotheses. Below, we reproduce them and follow each with a concise summary of whether the hypothesis was supported by the data.

1.1) In the validation of short interval state boredom scale repeat administration, we hypothesize that the effect of including an initial administration will have an absolute effect size (cohen's d) less than 0.5. We will test this with two, one-sided t-tests (TOST). With an alpha of 0.01 and sample size of 150 participants per arm, TOST has 99.22% power to reject the null hypothesis of an absolute effect greater than 0.5 and 83.04% power for an absolute effect greater than 0.35.

This hypothesis was NOT confirmed.

- BoredomBeforeAndAfter vs. BoredomAfterOnly: Cohens D=-0.411
- Is BoredomBeforeAndAfter < BoredomAfterOnly with Cohens $d > -0.5 :\ T_{298} = 0.987,\ p = 0.163$
- Is BoredomBeforeAndAfter > BoredomAfterOnly with Cohens $d < 0.5 :\ T_{298} = -10.1,\ p < 0.001$
- Presenting boredom questions before start of task leads to DECREASED responses after block0. because we cannot exclude H0:|D|>=0.5, we will use only the BoredomAfterOnly cohort in subsequent analyses.

1.2) We hypothesize that final state boredom will explain variance in subject-level POTD slope. This is a one-sided hypothesis.

This hypothesis was confirmed ($\chi^2(2, N = 16) = 8.77,\ p = 0.0125$).

1.3) We hypothesize that the change in boredom will explain variance in subject-level POTD slope. This is a one-sided hypothesis.

This hypothesis was confirmed ($\chi^2(2, N = 16) = 18.6,\ p < 0.001$).

1.4) We hypothesize that trait boredom will explain variance in subject-level POTD slope. This is a one-sided hypothesis.

This hypothesis was NOT confirmed ($\chi^2(2, N = 16) = 2.375,\ p = 0.305$).

2.1) In the validation of short interval state MDES repeat administration, we hypothesize that the effect of including an initial administration will have an absolute effect size (cohen's d) less than 0.5. We will test this with two, one-sided t-tests (TOST). With an alpha of 0.01 and sample size of 150 participants per arm, TOST has 99.22% power to reject the null hypothesis of an absolute effect greater than 0.5 and 83.04% power for an absolute effect greater than 0.35.

This hypothesis was confirmed.

- MwBeforeAndAfter vs. MwAfterOnly: Cohens D=0.0739
- Is MwBeforeAndAfter < MwAfterOnly with Cohens $d > -0.5:\ T_2 98 = 7.52,\ p < 0.001$
- Is MwBeforeAndAfter > MwAfterOnly with Cohens $d < 0.5:\ T_2 98 = -5.58,\ p < 0.001$
- Presenting MW questions before start of task DOES NOT change responses after block0. Because we can exclude H0:|D|>0.5, we will use both MW cohorts in subsequent analyses.

2.2) We hypothesize that the final emotion dimension score will explain variance in subject-level POTD slope. This is a one-sided hypothesis.

This hypothesis was confirmed ($\chi^2(2, N = 16) = 44.0,\ p < 0.001$).

2.3) We hypothesize that the change in emotion dimension score will explain variance in subject-level POTD slope. This is a one-sided hypothesis.

This hypothesis was confirmed ($\chi^2(2, N = 16) = 7.30,\ p = 0.0260$).

2.4) We hypothesize that trait mind-wandering will explain variance in subject-level POTD slope. This is a one-sided hypothesis.

This hypothesis was NOT confirmed ($\chi^2(2, N = 16) = 1.20,\ p = 0.548$).

3.1) We hypothesize that final mood ratings will be lower on average than the initial mood ratings in our real-world task. This is a one-sided hypothesis.

This hypothesis was NOT confirmed.

- Mean pre-break mood: 65.7%, post-break mood: 66.6%, change in mood: 0.909% (0.13%/min)
- happinessBeforeActivities < happinessAfterActivities (PAIRED): $T = -1.33,\ p = 0.0918$
- happinessBeforeActivities > happinessAfterActivities (PAIRED): $T = -1.33,\ p = 0.908$
- Free time break DOES NOT change mood ratings in block 0.

3.2) We hypothesize that the decrease in mood ratings will be smaller than that observed in the boredom task. This is a one-sided hypothesis.

This hypothesis was confirmed.

- activities < boredom: $T = 6.28,\ p = 1$
- activities > boredom: $T = 6.28,\ p < 0.001$
- Free time break happiness change is GREATER than boredom happiness change in block 0.

M. Amended Analyses on Boredom and Mind-Wandering

After completing the boredom and MW analyses described in the previous section, we realised that boredom and MW factors explained significant variance in initial mood (i.e., model intercept terms) in addition to mood slope. For example, $finalBoredom$ and $Time : finalBoredom$ interaction each explained separate amounts of variance. Because our research question was specifically about these factors' ability to explain changes in mood over time, we decided that our research questions would be better answered by comparing models with and without these additional factors' interactions with time. Both expanded and reduced models included the additional factor (e.g., $finalBoredom$), and the expanded model also included the factor's interaction with time (e.g., $Time : finalBoredom$).

We have also switched from a general residual sum-of-squares R^2 to the more specific $R_1^{2\,1,\,2}$ to capture the ability of the new factor's interaction with time to explain *within-participant* variance. We use the difference in R_1^2 values between the expanded model (with the new factor's interaction with time) and the reduced model (without it) to calculate a Cohen's f^2 value to describe the effect size. This approach more specifically addresses the question of how well the new factor can capture each participant's mood drift.

In response to reviewer comments, we considered not only the emotion dimension of the MDES scores, but all 13 principal components, thus more comprehensively investigating whether any aspect of the content of ongoing thought could explain mood drift.

We have included the results of the analyses exactly as they were preregistered in Supplementary Note L.. In the Results section of the main text, we have reported the amended results described below. The Results section focused primarily on within-individual variance explained R_1^2 and its associated f^2 values. For completeness, below we also report the between-individual variance explained R_2^2 and its associated f^2 values.

The interaction between time and final state boredom (i.e., at the end of the rest block) improved model fit (Likelihood ratio test: $\chi^2(1, N = 16) = 6.47,\ p = 0.0110$). But the effect on model fit was very small: the within-individual variance explained increased from $R_1^2 = 0.370$ (without this new term in the model) to $R_1^2 = 0.374$ (with it) ($f^2 = 0.00578$). Similarly, the between-individual variance explained increased from $R_2^2 = 0.125$ (without this new term in the model) to $R_2^2 = 0.126$ (with it) ($f^2 = 0.00144$).

The change in state boredom across the rest block produced similar results. A model including time's interaction with change-in-state-boredom improved model fit ($\chi^2(1, N = 16) = 12.3,\ p < 0.001$). The effect on model fit was again very small: the within-individual variance explained increased from $R_1^2 = 0.413$ to $R_1^2 = 0.410$ ($f^2 = 0.0111$). Similarly, the between-individual variance explained increased from $R_2^2 = 0.156$ to $R_2^2 = 0.159$ ($f^2 = 0.00300$).

An LME model including time's interaction with all final (i.e., after the rest period) MDES components improved model fit ($\chi^2(13, N = 40) = 34.2,\ p = 0.00113$), however the effect on within-individual variance explained was small $R_1^2 = 0.596$ to $R_1^2 = 0.604$ ($f^2 = 0.0227$). The effect on between-individual variance explained was very small $R_2^2 = 0.198$ to $R_2^2 = 0.201$ ($f^2 = 0.00372$).

The change in MDES components across the rest block produced similar results. A model including time's interaction with change-in-all-MDES-components improved model fit ($\chi^2(13, N = 40) = 36.4,\ p < 0.001$), however, the effect on within-individual variance explained was small $R_1^2 = 0.408$ to $R_1^2 = 0.430$ ($f^2 = 0.0380$). The effect on between-individual variance explained was very small $R_2^2 = 0.156$ to $R_2^2 = 0.164$ ($f^2 = 0.00987$).

References

[1] Snijders, T. A. B. & Bosker, R. J. Modeled variance in two-level models. *Sociological methods & research* **22** (3), 342–363 (1994) .

[2] Nakagawa, S. & Schielzeth, H. A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in ecology and evolution* **4** (2), 133–142 (2013) .

[3] Cohen, J. *Statistical Power Analysis for the Behavioral Sciences* (Routledge, 2013). https://www.taylorfrancis.com/books/9781134742707.

[4] Selya, A. S., Rose, J. S., Dierker, L. C., Hedeker, D. & Mermelstein, R. J. A Practical Guide to Calculating Cohen's f2, a Measure of Local Effect Size, from PROC MIXED. *Frontiers in Psychology* **3**, 111 (2012). https://doi.org/10.3389/fpsyg.2012.00111 .

[5] Egloff, B., Tausch, A., Kohlmann, C. W. & Krohne, H. W. Relationships between time of day, day of the week, and positive mood: Exploring the role of the mood measure. *Motivation and Emotion* **19** (2), 99–110 (1995). https://doi.org/10.1007/BF02250565 .

[6] Bedder, R. L., Vaghi, M. M., Dolan, R. J. & Rutledge, R. B. Risk taking for potential losses but not gains increases with time of day. *psyarxiv* (2020). https://doi.org/10.31234/osf.io/3qdnx

[7] Keren, H. *et al.* The temporal representation of experience in subjective mood. *eLife* **10**, 1–24 (2021). https://doi.org/10.7554/elife.62051 .

[8] Hunter, J. A., Dyer, K. J., Cribbie, R. A. & Eastwood, J. D. Exploring the utility of the Multidimensional State Boredom Scale. *European Journal of Psychological Assessment* **32** (3), 241–250 (2016). https://doi.org/10.1027/1015-5759/a000251 .

[9] Struk, A. A., Carriere, J. S. A., Cheyne, J. A. & Danckert, J. A short boredom proneness scale: Development and psychometric properties. *Assessment* **24** (3), 346–359 (2017) .

[10] Poerio, G. L., Totterdell, P. & Miles, E. Mind-wandering and negative mood: Does one thing really lead to another? *Consciousness and Cognition* **22** (4), 1412–1421 (2013). https://doi.org/10.1016/j.concog.2013.09.012 .

[11] Smallwood, J. *et al.* The neural correlates of ongoing conscious thought. *iScience* **24** (3), 102132 (2021). https://doi.org/10.1016/J.ISCI.2021.102132 .

[12] Tusche, A., Smallwood, J., Bernhardt, B. C. & Singer, T. Classifying the wandering mind: Revealing the affective content of thoughts during task-free rest periods. *NeuroImage* **97**, 107–116 (2014). https://doi.org/10.1016/j.neuroimage.2014.03.076 .

[13] Turnbull, A. *et al.* The ebb and flow of attention: Between-subject variation in intrinsic connectivity and cognition associated with the dynamics of ongoing experience. *NeuroImage* **185** (September 2018), 286–299 (2019). https://doi.org/10.1016/j.neuroimage.2018.09.069 .

[14] Ho, N. S. P. *et al.* Facing up to why the wandering mind: Patterns of off-task laboratory thought are associated with stronger neural recruitment of right fusiform cortex while processing facial stimuli. *NeuroImage* **214** (March), 116765 (2020). https://doi.org/10.1016/j.neuroimage.2020.116765 .

[15] Mrazek, M. D., Phillips, D. T., Franklin, M. S., Broadway, J. M. & Schooler, J. W. Young and restless: validation of the Mind-Wandering Questionnaire (MWQ) reveals disruptive impact of mind-wandering for youth. *Frontiers in psychology* **4**, 560 (2013) .